PEN AMERICA

Speech in the Machine:
Generative AI's Implications for Free Expression

CONTENTS

INTRODUCTION

The rapidly expanding era of artificial intelligence (AI) is ushering in both exciting possibilities and precarious unknowns. AI technologies are already powerful curators of information and arbiters of online content, and often amplifiers of disinformation. As AI technology—particularly generative AI technology—evolves, its potential impact on human rights and the fundamental right to free expression must be central to conversations about policy, regulation, and best practices.[1] Companies, international organizations, and governments must take the potential impact of AI into account when formulating means of safely and productively harnessing the benefits of this new era.

In this paper, PEN America notes the critical free expression issues at stake with generative AI, which has the potential to supercharge tools of deception and repression and make them more widely accessible. From online content creation to translation, creative writing to news reporting, generative AI tools may spur inspiration and ingenuity—or overtake the human craft in ways that undercut authenticity in public discourse and dampen the underlying value of open expression. How the development and dissemination of AI systems might underscore or undermine the right to free expression will depend on who controls the rollout of these new technologies, what regulations are imposed, and how companies and governments define and execute their human rights responsibilities.

The purpose of this paper is to identify emerging free expression issues raised by the increased prevalence and usage of generative AI, with particular attention given to large language models (LLMs). As an organization of and for writers, PEN America is focusing this paper on issues of most concern to our community. PEN America has a 100-year history of advocating for the protection of writers at risk and defending the freedom to write globally, highlighting the tactics of dictators and would-be authoritarian regimes in silencing those whose work sparks the imagination and holds governments to account. The 1948 PEN Charter commits all PEN centers to fighting "mendacious publication, deliberate falsehood and distortion of facts for political and personal ends."[2] Over the past decade, PEN America has sounded the alarm about the spread of disinformation and online abuse, and their effects on free expression, press freedom, and democratic discourse.[3]

The paper begins with the implications of generative AI for creativity and the arts, particularly for writers and the written word. It then examines how generative AI might exacerbate existing threats to free expression, including disinformation, online abuse, and censorship, and how it might wield more subtle forms of influence on the information landscape. Finally, the paper offers preliminary recommendations and guiding principles for policymakers and companies as they consider the policies and regulations that will shape the era of generative AI. Because these technologies are advancing quickly, much of the risk assessment at this stage is speculative.

---

1 Article 19, "Privacy and Freedom of Expression in the Age of Artificial Intelligence Privacy and Freedom of Expression in the Age of Artificial Intelligence." (2018). article19.org/wp-content/uploads/2018/04/Privacy-and-Freedom-of-Expression-In-the-Age-of-Artificial-Intelligence-1.pdf.

2 PEN Charter, PEN America (accessed July 21, 2023) pen.org/pen-charter/

3 See: "Faking News: Fraudulent News and the Fight For Truth," PEN America (accessed July 21, 2023) pen.org/research-resources/faking-news/ ; "Truth On the Ballot: Fraudulent News, the Midterm Elections, And Prospects for 2020," PEN America (accessed July 21, 2023) pen.org/truth-on-the-ballot-fraudulent-news/; "Online Harassment Field Manual," PEN America (accessed July 21, 2023) onlineharassmentfieldmanual.pen.org ; "No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users," PEN America (accessed July 21, 2023) pen.org/report/no-excuse-for-abuse/ ; "Journalists and the Threat of Disinformation," PEN America (accessed July 21, 2023) pen.org/report/hard-news-journalists-and-the-threat-of-disinformation/ ; Viktorya Vilk and Kat Lo, "Shouting Into the Void: Why Reporting Abuse to Social Media Platforms Is So Hard and How to Fix It," PEN America, June 29, 2023 pen.org/report/shouting-into-the-void/ ; "Chilling Effects: NSA Surveillance Drives U.S. Writers to Self-Censor," PEN America (accessed July 21, 2023) pen.org/research-resources/chilling-effects/ ; "Chiling Effects: NSA Surveillance Drives U.S. Writers to Self-Censor," PEN America (accessed July 21, 2023) pen.org/research-resources/chilling-effects/

We cannot anticipate exactly how these technologies will be used or the magnitude of the risks. As an initial issue brief, however, we intend this paper to set an agenda for further research, analysis, and deliberation as generative AI technologies, their role in society, and their impact on freedom of expression continue to evolve.

Glossary

The definitions below are drawn from the work of experts in academia, government, and civil society.

- *Artificial Intelligence (AI):* Artificial intelligence, in the words of the man who coined the term, is the science and engineering of making intelligent machines, especially intelligent computer programs."[4] *Note: In this report we use the term "artificial intelligence" to refer to the field of study or branch of research, and not to the tools or technologies that rely on machine learning and natural language processing to simulate human intelligence.*
- *Machine Learning:* "A subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed."[5]
- *Natural Language Processing (NLP):* A subfield of artificial intelligence "that explores how computers can be used to understand and manipulate natural language text or speech to do useful things. NLP researchers aim to gather knowledge on how human beings understand and use language so that appropriate tools and techniques can be developed to make computer systems understand and manipulate natural languages to perform the desired tasks."[6]
- *Algorithm:* "[A] set of instructions that is designed to accomplish a task. Algorithms usually take one or more inputs, run them systematically through a series of steps, and provide one or more outputs."[7]
- *AI System:* "An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy."[8] *Note: In this report we use the term "AI tools" to refer to applications or services that rely on AI systems to function.*
- *Generative AI:* "Generative AI refers to a category of artificial intelligence (AI) algorithms that generate new outputs based on the data they have been trained on. Unlike traditional AI systems that are designed to recognize patterns and make predictions, generative AI creates new content in the form of images, text, audio, and more."[9]

4 John McCarthy, "What is AI? Basic Questions," Professor John McCarthy's blog at Stanford University (accessed July 21, 2023) jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html

5 Sara Brown, "Machine learning, explained," Ideas Made to Matter, April 21, 2021, mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

6 Gobinda G. Chowdhury, "Natural Language Processing," Annual Review of Information Science and Technology, 37 (2003): 51–89, strathprints.strath.ac.uk/2611/1/strathprints002611.pdf

7 National Library of Medicine, Glossary, www.nnlm.gov/guides/data-glossary/algorithm

8 National Institute of Standards and Technology, "AI Risk Management Framework." (2023) doi.org/10.6028/nist.ai.100-1.

9 Nick Routley, "What is generative AI? An AI explains," World Economic Forum, February 6, 2023, weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/

- *Large Language Model (LLM):* "A machine learning algorithm that scans enormous volumes of text to learn which words and sentences frequently appear near one another and in what context. Large language models can be adapted to perform a wide range of tasks across different domains."[10]
- *AI Chatbot*: "Chatbots are intelligent conversational computer programs that mimic human conversation in its natural form."[11] "Chatbots can mimic human conversation and entertain users but they are not built only for this."[12] An AI chatbot accomplishes this through an underlying AI system.

This report comes at a pivotal moment for the future of artificial intelligence technology. In November 2022 OpenAI, then a niche company, released ChatGPT (Generative Pre-Trained Transformer), an AI chatbot designed for public-facing and conversational usage. Prior to that point, the concept of artificial intelligence was associated with science fiction, even if a future with robots as part of daily life felt increasingly imminent. ChatGPT inaugurated a new chapter: Here was a piece of technology that almost anyone with a laptop or mobile phone could access, an application that made artificial intelligence tangible and useful for daily life—and, in doing so, seemed to herald a new technological era.

Editorial credit: Shotmedia / Shutterstock.com

The public awakening to generative AI, driven largely by ChatGPT, follows the integration of AI tools in personal technology and in systems that undergird other sectors of society, such as agriculture and healthcare.[13] Smart assistants like Apple's Siri and Amazon's Alexa, algorithms that recommend songs on Spotify or Pandora, transcription services, suggested smart replies in Gmail, social media content curation—all run on artificial intelligence.[14] The significant difference between these functions and ChatGPT, Google Bard, and similar tools, is that the latter are not being used just to perform a predetermined task, but are instead capable both of generating content and of engaging in "conversation" with users.

Generative AI systems are trained on a data set of information, which is then used algorithmically to develop and refine outputs from the system. For example, a chatbot designed to triage patients at an urgent care clinic might be trained on content from the Gray's Anatomy textbook as well as documentation on that clinic's office procedures. A more general-purpose chatbot, such as ChatGPT, could be trained on far broader datasets, encompassing vast amounts of writing and other creative work.[15]

10 Gabriel Nicholas and Aliya Bhatia, Lost in Translation: Large Language Models in Non-English Content Analysis, The Center for Democracy & Technology, May 2023, cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf, 2023

11 Guendelina Caldarani, Sardar Jaf, and Kenneth McGarry, "A Literature Survey of Recent Advances in Chatbots," MDPI, January 15, 2022, mdpi.com/2078-2489/13/1/41

12 Eleni Adamopoulou and Lefteris Moussiades, "An Overview of Chatbot Technology," Artificial Intelligence Applications and Innovations, vol. 584, (2020), 373-383, link.springer.com/chapter/10.1007/978-3-030-49186-4_31

13 While these advancements have been felt globally, as the World Economic Forum notes, the economic and social benefits inure to the Global North: Danny Yu, Hannah Rosenfeld, Abhishek Gupta, The 'AI Divide' between the Global North and the Global South, World Economic Forum, January 16, 2023, weforum.org/agenda/2023/01/davos23-ai-divide-global-north-global-south/

14 Sara Brown, "Machine learning, explained," MIT Management, April 21, 2021, https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

15 ChatGPT's help center states that it is trained on "vast amounts of data from the internet written by humans." "What is Chat GPT? Open AI," accessed July 2023 help.openai.com/en/articles/6783457-what-is-chatgpt

As quickly as awareness of generative AI has grown, so too have anecdotes of chatbots gone wrong, including concerns about chatbots "hallucinating," or generating false or out-of-context answers to user queries; claims that self-aware AI technologies are on the horizon; exposés on how digital thrill seekers swap strategies to override ChatGPT's guardrails; and, of course, calls for regulation and slowdowns in the rollout of the technology, including from some industry leaders.[16] New technologies often spark debate over their uses and the rights they could challenge; a degree of moral panic has accompanied every technological advent from the printing press to the radio. The invention of the handheld camera in 1888 inspired immediate concerns about privacy—and significantly influenced U.S. privacy law as we know it today.[17] More recently, the emergence of digital music streaming services led to changes in U.S. copyright law.[18] Alongside its transformational implications for scholarship, teaching and learning, and social life, generative AI technology could change the ways in which free expression rights are considered, protected, and upheld.

Some uses of generative AI systems may completely transform entire sectors. These include education, where students have used ChatGPT to write papers, followed by the quick emergence of AI-and plagiarism-detection technologies for instructors; journalism, where posts and stories have been written by chatbots, and outlets such as BuzzFeed and CNET have been open about their use of AI tools for content; and the literary industry, where one magazine temporarily stopped accepting short story submissions in response to a wave of poorly written, plagiarized AI-generated content.[19] Matthew Kirschenbaum, a professor of English and Digital Studies at the University of Maryland, described this transformation in an article published by The Atlantic: "We may quickly find ourselves facing a textpocalypse, where machine-written language becomes the norm and human-written prose the exception."[20]

In the right hands, generative artificial intelligence systems can advance and promote expressive conduct, reduce barriers to expression, and offer new outlets for creativity and artistic imagination. AI models are also capable of generating coherent and contextually relevant text, making significant contributions to various domains. AI-powered assistive technology can be used to help people with disabilities or other challenges to communicate and express themselves more easily and can increase access to information, as in aiding with foreign language learning or web-browsing tools for the visually impaired.[21]

---

16 Kevin Roose, "A Conversation With Bing's Chatbot Left Me Deeply Unsettled," The New York Times, February 16, 2023, nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html ; Karen Weise and Cade Metz, "When A.I. Chatbots Hallucinate," The New York Times, May 1, 2023, nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html; "Pause Giant AI Experiments: An Open Letter," Future of Life Institute, March 22, 2023, futureoflife.org/open-letter/pause-giant-ai-experiments/

17 Joshua J. Kaufman, "The invention that resulted in the Rights of Privacy and Publicity," Lexology, September 24, 2014, lexology.com/library/detail.aspx?g=f5baa264-aacc-4307-ac7e-83776b02c29b; Zeynep Tufekci, "We Need to Take Back Our Privacy," The New York Times, May 19, 2022, nytimes.com/2022/05/19/opinion/privacy-technology-data.html

18 Dani Deahl, "The Music Modernization Act has been signed into law," The Verge, October 11, 2018, theverge.com/2018/10/11/17963804/music-modernization-act-mma-copyright-law-bill-labels-congress

19 Jaclyn Peiser, "The Rise of the Robot Reporter," The New York Times, February 5, 2019, nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html; Noor Al-Sibai, John Christian, "BuzzFeed is Quietly Publishing Whole AI-Generated Articles, Not Just Quizzes," Futurism, March 30, 2023, futurism.com/buzzfeed-publishing-articles-by-ai; Paul Farhi, "A news site used AI to generate articles. It was a journalistic disaster," The Washington Post, January 17, 2023, washingtonpost.com/media/2023/01/17/cnet-ai-articles-journalism-corrections/; "How WIRED Will Use Generative AI Tools," WIRED, May 22, 2023, www.wired.com/about/generative-ai-policy/; Matthew Loh, "The editor of a sci-fi magazine says he's getting flooded with short stories as AI tools take off: 'It quickly got out of hand,'" Business Insider, February 22, 2023, businessinsider.com/chatgpt-ai-written-stories-publisher-clarkesworld-forced-close-submissions-2023-2

20 Matthew Kirschenbaum, "Prepare for the Textpocalypse," The Atlantic, March 8, 2023, theatlantic.com/technology/archive/2023/03/ai-chatgpt-writing-language-models/673318/

21 Lucas Kohnke, Benjamin Luke Moorhouse, and Di Zou, "ChatGPT for Language Teaching and Learning." RELC Journal (2023) doi.org/10.1177/00336882231162868; Amazon Web Services, "Introducing an Image-to-Speech Generative AI Application Using Amazon SageMaker and Hugging Face | Amazon Web Services." (May 19, 2023). aws.amazon.com/blogs/machine-learning/introducing-an-image-to-speech-generative-ai-application-using-amazon-sagemaker-and-hugging-face/; Jenny Lay-Flurrie, "Global Accessibility Awareness Day – Accessibility at the Heart of Innovation." Microsoft on the Issues, (May 2023). blogs.microsoft.com/on-the-issues/2023/05/18/global-accessibility-awareness-day-generative-ai/.

Conversely, in the hands of bad actors—whether public or private—generative AI tools can supercharge existing threats to free expression, for example by making disinformation and online abuse campaigns easier and cheaper to carry out, at greater volume. The transformational power of AI tools, even in the hands of the well-intended, is only partially understood and might change society in unpredictable ways.

The democratization of these tools—the ability of essentially anyone with an internet browser to access ChatGPT, for example—represents a watershed moment in access to information and creative power. At the same time, by enabling the use of AI-generated content in a wide range of daily human interactions, there is also the potential for people to lose trust in language itself, and thus in one another. If generative AI is used for creative purposes, or simply to boost the expression an individual would otherwise engage in (akin to using a thesaurus to search for more compelling language), where is the line between original and synthetic expression? The mere knowledge that these tools can create seemingly credible information and be used to mislead deliberately has the potential to undermine trust in almost any media or other information source, risking further erosion of public trust in accountability journalism, governmental institutions, scientific research, and simple person-to-person communication.

Recent analyses have declared the college essay and the employment cover letter "dead," while numerous articles describe the use of ChatGPT in online dating profiles and messages.[22] If everyone who writes in any form can use their own personal, digital Cyrano de Bergerac, will people stop relying on written communications as the basis for assessing character or judging sincerity? And will our ability to engage in good faith civic and social discourse deteriorate even further?

As with the development of social media, experts worry that technology companies, spurred by competition with one another, will charge ahead with developing generative AI applications without due consideration for the risks, including the ways authoritarian governments might use such technology to target the vulnerable.[23] There is also concern that efforts to rein in the power of these new technologies will infringe on freedom of expression. Authoritarian governments might use concerns about AI systems as a pretense to crack down on online speech and dissent. Even in democracies, generative AI might inspire new and reflexive regulations without sufficient consideration for the implications for free expression and other human rights.[24] There is a fine line, for example, between using generative AI tools for artistic creation and using them to create deepfakes to stoke public fear or conflict.

As we take the first steps into a world until now only represented in fiction, society will have to wrestle with wide-ranging questions surrounding generative AI and how it will affect civic trust, freedom of speech, creative expression, and the very notion of truth. Writers, journalists, and artists are quickly feeling the effects of generative AI; they will also be among the voices people turn to for insights into how we navigate this new phase of technological advancement.

22 Matthew Kirschenbaum, *The Atlantic,* theatlantic.com/technology/archive/2023/05/chatbot-cheating-college-campuses/674073/; Rani Molla, "Maybe AI can finally kill the cover letter," Vox, March 8, 2023, vox.com/technology/2023/3/8/23618509/chatgpt-generative-ai-cover-letter ; Anna Iovine, "Tinder users are using ChatGPT to message matches," Mashable, December 17, 2022, mashable.com/article/chatgpt-tinder-tiktok

23 Daron Acemoglu and Simon Johnson, "Big Tech is Bad. Big A.I. Will Be Worse," *The New York Times*, June 9, 2023, nytimes.com/2023/06/09/opinion/ai-big-tech-microsoft-google-duopoly.html?searchResultPosition=1

24 United Nations, Special Rapporteur on freedom of opinion and expression, *Report on Artificial intelligence technologies and implications for freedom of expression and the information environment*, A/73/348 (August 29, 2018) documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement.

PART I: GENERATIVE AI, CREATIVITY, AND THE ARTS

Writers and other artists face especially salient questions about the integration of generative AI into everyday usage. The spark of imagination that drives creativity is an intangible but inherently human quality, evident from the earliest days of our species. Technology has long been utilized as a tool for creative production and those advances have often sparked controversy. The camera's invention raised questions about whether photography, conducted with a machine, could be considered art, as well as anxiety about whether it would replace other media like painting.[25] The use of computers in artistic production is not new, but what distinguishes generative AI is the muddying of the line between what is real, what is a human creation, and what is machine-generated, and the questions those blurred lines raise about the protection and ownership of ideas.

It is easy to postulate that a technology reliant upon content produced by humans could never replace human ingenuity, originality, and imagination. But whether humans will always be able to distinguish between original creations and algorithmically engineered replications is distressingly uncertain. Writer William Deresiewicz, the novelist and literary critic, writes that while AI technologies will not replace artists because they cannot make "true art...original art," they could still "put artists out of business."[26]

Creative communities are experiencing significant anxiety over the implications of generative AI, including the potential threat to their livelihoods. The technology raises thorny questions of ownership, especially because generative AI inherently feeds on words and images created by others. Unlicensed use of such material has authors arguing for compensation when their works are repurposed as elements of a training set. In June 2023, authors Mona Awad and Paul Tremblay filed suit against OpenAI, the company behind ChatGPT for allegedly using their copyright-protected works as part of the chatbot's training material, without their consent.[27] In July 2023, comedian and author Sarah Silverman joined authors Christopher Golden and Richard Kadrey in a similar suit against OpenAI, claiming their copyrighted books were also used to train ChatGPT.[28]

In the world of fan fiction, where vast troves of stories are largely made available online for free, writers are alarmed that their work has been preyed upon.[29] Some fan fiction writers have begun locking up stories that had previously been freely available, to prevent them from being scraped and fed into training sets.[30] The Authors Guild, an advocacy organization representing writers and their interests, has argued for urgent regulatory measures to ensure that creators are compensated for how their work feeds large language models. They argue that compensation is not only fair and justified, but necessary to ensure the continued incentivization of human creative output, "so our books and arts continue to reflect both our real and imagined experiences, open our minds, teach us new ways of thinking, and move us forward as a society, rather than rehash old ideas."[31]

25 Jordan G. Teicher, "When Photography Wasn't Art," *The Daily JStor,* February 6, 2016, daily.jstor.org/when-photography-was-not-art/

26 William Deresiewicz, "Why AI Will Never Rival Human Creativity," *Persuasion*, May 8, 2023 , persuasion.community/p/why-ai-will-never-rival-human-creativity

27 Emily St. Martin, "Bestselling authors Mona Awad and Paul Tremblay sue AI over copyright infringement," *The Los Angeles Times,* July 1, 2023, latimes.com/entertainment-arts/books/story/2023-07-01/mona-awad-paul-tremblay-sue-openai-claiming-copyright-infringement-chatgpt

28 Jon Blistein, "Sarah Silverman Leads Class Action Copyright Suit Against ChatGPT," *Rolling Stone,* July 10, 2023, rollingstone.com/culture/culture-news/sarah-silverman-copoyright-suit-chatgpt-open-ai-1234785472/

29 Linda Codega, "Chatbots Have Stolen Fan Fiction From a Gift Culture," *Gizmodo,* June 12, 2023, gizmodo.com/ai-chatbot-fanfiction-fanfic-archive-of-our-own-1850524393

30 Sheera Frenkel and Stuart A. Thompson, "'Not for Machines to Harvest': Data Revolts Break Out Against A.I.," *The New York Times,* July 15, 2023, nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html

31 Artificial Intelligence, The Authors Guild, accessed July 2023, authorsguild.org/advocacy/artificial-intelligence/

In an open letter organized by the Authors Guild, more than 10,000 writers (as of this paper's publication) have called on the leaders of companies behind generative AI tools to address the use of writers' works in training AI systems without "consent, credit, or compensation."[32] The letter, signed by authors including Margaret Atwood, Viet Thanh Nguyen, Jennifer Egan, Jodi Picoult, Roxanne Gay, and Alexander Chee, calls for the companies to take the following steps:

I. "Obtain permission for use of our copyrighted material in your generative AI programs.
II. Compensate writers fairly for the past and ongoing use of our works in your generative AI programs.
III. Compensate writers fairly for the use of our works in AI output, whether or not the outputs are infringing under current law."[33]

Central to the 2023 Writers Guild of America (WGA) strike, still ongoing at the time of this writing, are fears that television and movie studios might turn to generative AI tools for ideas and script writing, particularly for more formulaic genres, including children's television and crime procedurals.[34] This concern falls under the broad rubric of usurpation, or the idea that AI will take over tasks and roles previously carried out by humans. The WGA is not seeking to bar the use of generative AI tools altogether, but to ensure that any such usage does not undercut writers' attribution or compensation; that AI cannot be credited with writing a screenplay, or be considered the 'author' of 'source material' a writer is then called in to adapt at a lower pay rate.[35] In effect, they are asking that no matter the role generative AI might play, humans still must be paid and credited as if it were not used at all.

Guild members also fear that as their strike wears on, studios might accelerate reliance on AI tools, so that they will have less incentive to make concessions at the bargaining table. The Directors Guild of America has won a guarantee from the Alliance of Motion Picture and Television Producers (AMPTP) that directors won't be replaced by AI technology; but the WGA says the AMPTP rejected their attempts to limit the use of AI tools in the writing process.[36] Some observers of the WGA strike suggest it is just the first of many battles to come, across creative industries and beyond.[37]

Already, AI-generated novellas are being published using tools like ChatGPT, Cohere, and Sudowrite, an AI tool designed specifically for longform creative writing.[38] AI-generated podcasts are on the market too, including one that draws from Joe Rogan's podcast and uses a clone of his voice, and others that rely on AI technology for every step of the process, from sound design to artwork to script writing.[39] An AI-produced

32 "Authors Guild Open Letter to Generative AI Leaders." n.d., actionnetwork.org/petitions/authors-guild-open-letter-to-generative-ai-leaders; Chloe Veltman, "Thousands of Authors Urge AI Companies to Stop Using Work without Permission." NPR, (July 17, 2023), npr.org/2023/07/17/1187523435/thousands-of-authors-urge-ai-companies-to-stop-using-work-without-permission.

33 "Authors Guild Open Letter to Generative AI Leaders." n.d., actionnetwork.org/petitions/authors-guild-open-letter-to-generative-ai-leaders.

34 Mandalit del Barco, "Striking Hollywood scribes ponder AI in the writer's room," May 18, 2023, ; James Poniewozik, TV's War With the Robots Is Already Here," *The New York Times*, May 10, 2023 nytimes.com/2023/05/10/arts/television/writers-strike-artificial-intelligence.html, Nick Bilton, ""The First Skirmish in a New War": Why AI Should Be Central in the Writers Strike, *Vanity Fair,* May 9, 2023, vanityfair.com/news/2023/05/writers-strike-2023-ai

35 Alissa Wilkinson, "WGA Strike: A Hollywood Writers Strike Needs to Address the Threat of AI." Vox, (May 2, 2023), vox.com/culture/23700519/writers-strike-ai-2023-wga; Gene Maddaus, "Variety." Variety, (May 24, 2023), variety.com/2023/biz/news/wga-ai-writers-strike-technology-ban-1235610076/.

36 Ananya Bhattacharya, "Movie directors got an AI deal with studios—but striking writers still have no such promises, *Quartz,* June 5, 2023, qz.com/movie-directors-got-an-ai-deal-with-studios-but-strikin-1850505417

37 Nick Bilton, ""The First Skirmish in a New War": Why AI Should Be Central in the Writers Strike," *Vanity Fair,* May 9, 2023, vanityfair.com/news/2023/05/writers-strike-2023-ai

38 Elizabeth A. Harris, "Peering Into the Future of Novels, With Trained Machines Ready," *The New York Times*, April 20, 2023, nytimes.com/2023/04/20/books/ai-novels-stephen-marche.html

39 Kate Knibbs, "Generative AI Podcasts Are Here. Prepare to Be Bored," *WIRED*, May 24, 2023, wired.com/story/generative-ai-podcasts-boring/

image won an art contest at the 2022 Colorado State Fair, in a category for "digital art/digitally-manipulated photography," generating significant controversy.[40] More recently, the creator of a fantasy book cover contest said he would abolish the prize after the winning cover was found to have used the AI image generation tool Midjourney.[41] In May 2023 more than 1,000 artists, writers, and cultural figures posted an open letter calling on "artists, publishers, journalists, editors, and journalism union leaders to take a pledge for human values against the use of generative-AI images to replace human-made art," because "the advent of generative-image AI technology, that unique interpretive and narrative confluence of art and text, of human writer and human illustrator, is at risk of extinction."[42]

PEN America's approach to free expression encompasses not only official constraints on free expression like government censorship, but also a much broader recognition of the value, enablers, and inhibitors of vibrant open discourse. From that perspective, the potential of generative AI to displace human creators raises a host of issues. PEN America's defense of free expression and the place of literature in society stems from an appreciation of the capacity for writing and storytelling to unlock empathy and build bridges across cultural divides. It is not clear whether these attributes and abilities, associated with the impulse and will of human creators to situate themselves in the shoes or minds of people unlike themselves, will carry over into AI-generated creative works.

Generative AI tools are, by their nature, derivative. If machines increasingly displace writers and creators, that poses a threat not only to those creative artists, but to the public as a whole. The scope of inspiration from which truly new creative works draw may be narrowed, undermining the power of literature, television, and film to catalyze innovative ways of thinking.

A glut of AI-created written content could undermine the very value of the written word. Suzanne Nossel, CEO of PEN America, has said: "If public discourse becomes so flooded with disinformation that listeners can no longer distinguish signal from noise, they will tune out."[43] Generative AI tools could cause a similar problem, with readers and audiences unable to discern whether the stories they read are infused with genuine human emotion, experience, and insight or simply machine-generated facsimiles of literature, journalism, or opinion writing. While these challenges are unlike traditional threats to free expression, in that they do not involve efforts to suppress speech, their potential to degrade public discourse and undermine the value of speech as a catalyst for truth and understanding is significant.

40 Kevin Roose "An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy," *The New York Times*, September 2, 2022, nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

41 Mia Sato, "How AI art killed an indie book cover contest," *The Verge,* June 9, 2023, theverge.com/2023/6/9/23752354/ai-spfbo-cover-art-contest-midjourney-clarkesworld

42 Jo Lawson-Tancred, "Molly Crabapple Has Posted an Open Letter by 1,000 Cultural Luminaries Urging Publishers to Restrict the Use of 'Vampirical' A.I.-Generated Images," *Artnet,* May 3, 2023, news.artnet.com/art-world/open-letter-urges-publishers-not-to-use-ai-generated-illustrations-2294392

43 Suzanne Nossel, "The Pro-Free Speech Way to Fight Fake News," *Foreign Policy*, October 12, 2017, foreignpolicy.com/2017/10/12/the-pro-free-speech-way-to-fight-fake-news/, 2017

GENERATIVE AI, COPYRIGHT LAW, AND THE FIRST AMENDMENT

There is little legal precedent regarding the use of generative AI, though questions about artificial intelligence, creativity, and intellectual property are already making their way through the courts and regulatory agencies.

One of the first determinations regarding the protections afforded to AI-generated content by a U.S. body came in February 2023, when the U.S. Copyright Office issued a letter limiting the previously granted copyright registration for a graphic novel, *Zarya of the Dawn*, which included images created with the generative AI tool Midjourney. In limiting the registration, the Office concluded that the texts and the "selection, coordination, and arrangement" of the visual and written elements of the work were the author's and therefore were subject to copyright, but the Midjourney-generated images were not, on the grounds that they were "not the product of human authorship."[44]

The Copyright Office followed up on its *Zarya of the Dawn* letter in March, with a statement of policy on "works containing material generated by artificial intelligence." This guidance acknowledges some of the different ways in which generative AI tools might be used in the creative process.[45] While the Office affirms in its decision that an "author" cannot be non-human, it also notes that it would take into account the extent to which the author had contributed their "own original mental conception" to a work that makes use of "AI contributions." The Office concludes that its approach to such cases will depend "on the circumstances, particularly how the AI tool operates and how it was used to create the final work."[46]

Policymakers will likely continue to explore this balance between the vital task of protecting human authorship and ownership, and recognizing the element of human creativity involved both in creating and curating AI models and in working with AI systems to generate and refine new content.[47] More work needs to be done to analyze and track the impact that approaches to copyright in the realm of AI will have on innovation, creativity and free expression. Courts and regulators should proceed cautiously to ensure that authors' and artists' rights and prerogatives are preserved, recognizing that their singular contributions to cultural life must continue to be incentivized for the benefit of all.

Distinct from the question of copyright, U.S. courts have not yet considered whether generative AI content enjoys First Amendment protection. Such content may be treated similarly to search results, making them the protected speech of the company behind the AI product.[48] In a draft paper published in April 2023,

---

44 Robert J. Kasunic, *Letters to Van Lindberg and Kristina Kashtanova*, 1-5GB561K, United States Copyright Office, (October 28, 2022), copyright.gov/docs/zarya-of-the-dawn.pdf; Blake Brittain, "AI-created images lose U.S. copyrights in test for new technology," Reuters, February 22, 2023, reuters.com/legal/ai-created-images-lose-us-copyrights-test-new-technology-2023-02-22/. The comic book's creator, Kris Kashtanova, welcomed the decision overall but argued that the images were a "direct expression of my creativity and therefore copyrightable." Ibid.

45 Copyright Office, "Library of Congress, Copyright Registration Guidance: Words Containing Material Generated by Artificial Intelligence," *Federal Register*, March 16, 2023, federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence

46 United States Copyright Office Statement of Policy, "Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence," (March 16, 2023) copyright.gov/ai/ai_policy_guidance.pdf

47 For additional thinking on generative AI and its implications for copyright and creativity, see the Authors Guild's "Policy Proposals Regarding the Development and Use of Generative AI," authorsguild.org/app/uploads/2023/05/Authors-Guild-Policy-Proposals-Regarding-the-Development-and-Use-of-Generative-AI.pdf.

48 For a discussion of why search results are protected speech under the First Amendment, see: Eugene Volokh and Donald M. Falk, "First Amendment Protection for Search Engine Search Results –White Paper," UCLA School of Law Research Paper (12-22) (May 14, 2012). As Richard Stengel, former editor of TIME, points out, the protections of the First Amendment already extend to non-human legal persons, such as corporations and nonprofit entities: Richard Stengel, "The Case for Protecting AI-Generated Speech With the First Amendment," *TIME*, May 9, 2023 time.com/6278220/protecting-ai-generated-speech-first-amendment/; See, e.g., Citizens United v. FEC, 558 U.S. 310 (2010)(holding that a corporation, union, or nonprofit could not be completely banned from engaging in independent expenditures in a manner consistent with the First Amendment); *First National Bank of Boston v. Bellotti*, 435 U.S. 765 (1978) (holding that a restriction on corporate political contributions violated the First Amendment); *New York Times v. Sullivan*, 376 U.S. 254 (1964) (holding that First Amendment freedom of speech protections limit the ability of public officials to sue for defamation) (*Labor Board v. Virginia Elec. & Power Co.*, 314 U.S. 469 (1941) (holding that a non-media company's speech is protected by the First Amendment).

constitutional scholar and Harvard Law School professor Cass Sunstein explores the complexities of the First Amendment's relationship to AI-generated content. He writes that established First Amendment principles should, for the most part, apply, even if in a novel context.[49] For example, "What is unprotected by the First Amendment is unprotected by the First Amendment, whether its source is a human being or AI."[50] He assesses that, in its current state, AI does not have First Amendment rights any more than a toaster or radio does (though he acknowledges this might change), but that "restrictions on the speech of AI might violate the rights of human beings," including as speakers and writers, and as listeners, readers, and viewers.[51] Sunstein concludes that for the government to enact restrictions on AI-generated content that are viewpoint-based or content-based but viewpoint-neutral would be inherently problematic, and that even content-neutral restrictions—akin to time, place, and manner restrictions on free speech—would require strong justification. Sunstein also acknowledges that unanswered questions remain, particularly concerning liability for the content produced by generative AI tools.

Linked to the question of liability is whether Section 230 of the Communications Decency Act, which protects online service providers from liability for content posted by third-party users, also protects generative AI's creators from liability for the speech it generates. According to Section 230's authors, former Congressman Chris Cox and Senator Ron Wyden, the answer is "no."[52] That interpretation comports with the text of the statute and court interpretations, which indicate that Section 230 protects platforms only for speech "provided by another information content provider."[53]

The human element is currently a factor in determining whether certain speech is protected. The intent of the speaker is often a determinant of whether speech falls within First Amendment protection or meets the criteria for one of the exceptions to it. True threats, for example, "encompass those statements where the *speaker means to communicate a serious expression of an intent* to commit an act of unlawful violence to a particular individual or group of individuals,"[54] although the speaker "need not actually intend to carry out the threat."[55] A finding that speech is defamatory, which would render it not protected, similarly hinges in part on the speaker's state of mind.[56]

The question might not remain theoretical for long: Already, ChatGPT has been sued for defamation.[57] The plaintiff in the lawsuit, Mark Walters, claims that ChatGPT's responses to a reporter's inquiry—in which the

---

49 Cass R. Sunstein, "Artificial Intelligence and the First Amendment," Harvard Law School (April 28, 2023), papers.ssrn.com/sol3/papers.cfm?abstract_id=4431251.

50 Ibid. ChatGPT itself appears to agree with Susntein's assessment, at least in principle: Asked by scholar Gene Policinski during a typed exchange "Would the First Amendment protect expression by an AI?," ChatGPT responded in part, "The expression of an AI would be protected under the First amendment [sic] if it is considered as a form of speech, and it is not harmful or illegal." See: Gene Polcinski, "Does ChatGPT Have the Right to Free Speech?", Freedom Forum (2023), freedomforum.org/does-chatgpt-have-the-right-to-free-speech/. ChatGPT's response is inaccurate, as "harmful" is not a category of speech unprotected by the First Amendment.

51 Ibid.

52 Chris Cox (Congressman), Meghan McCarty Carino and Rosie Hughes, Marketplace Tech, May 19, 2023 (podcast), marketplace.org/shows/marketplace-tech/section-230-co-author-says-the-law-doesnt-protect-ai-chatbots/; Ephrat Livni, Sarah Kessler, and Rami Mattu, "Who Is Liable for A.I. Creations?", *The New York Times*, June 3, 2023, nytimes.com/2023/06/03/business/who-is-liable-for-ai-creations.html

53 At least one lawyer and technologist has disagreed with this reading, arguing that Section 230 protections should be extended to generative AI to avoid "foreclosing on the technology's true potential." See: Jess Miers, "Yes, Section 230 Should Protect ChatGPT And Other Generative AI Tools," *techdirt*, March 17, 2023, techdirt.com/2023/03/17/yes-section-230-should-protect-chatgpt-and-others-generative-ai-tools/

54 *Virginia v. Black*, 538 U.S. 343, 359 (2003) (emphasis added).

55 Ibid. 359-360.

56 *Gertz v. Robert Welch*, Inc., 94 S.Ct. 2997 (1974) (establishing a negligence standard for non-public figures in defamation cases); *New York Times Co. v. Sullivan*, 376 U.S. 254 (1964) (a public figure must show "actual malice" to recover damages in a civil libel suit relating to the figure's official conduct).

57 *Walters v. OpenAI*, L.L.C., Case No. 1:23-cv-03122 (N.D. Ga., July 14, 2023) courthousenews.com/wp-content/uploads/2023/06/walters-openai-complaint-gwinnett-county.pdf, dockets.justia.com/docket/georgia/gandce/1:2023cv03122/318259.

chatbot erroneously said Walters was being sued for "defrauding and embezzling funds"—were "false and malicious"; he seeks general and punitive damages from ChatGPT developer OpenAI.[58] It is unclear whether a developer might be held liable for false or defamatory content generated by its AI system, particularly absent notice, but the court's determination could provide some parameters for how such speech is viewed.

While many unanswered questions remain, it is clear that any regulation of generative AI must be carried out with thoughtful regard to free expression considerations and the imperative of avoiding improper government restrictions on speech. The protections of the First Amendment must continue to be afforded to both those who create and those who receive information, no matter its form. It is easy to imagine the worst-case scenario—i.e., government excluding Jewish religious texts from training data or banning a generative AI tool from mentioning trans people. As explored further below, the Chinese Communist Party (CCP) is already enforcing party ideology on generative AI tools operating in China. Concerns will persist about how individuals employ generative AI technologies, but to protect a free society government must not be empowered to implement content—or viewpoint—based restrictions or requirements on AI-generated speech.

PART II: GENERATIVE AI AS AN AMPLIFIER OF THREATS TO FREE EXPRESSION

Tech companies and social media platforms have spent the last decade-and-a-half wrestling with how to protect free speech online, counter damaging forms of disinformation, protect individuals from online harassment, and manage the effects of digital discourse on our privacy, politics, and personhood. These efforts—none of which can be considered a resounding success—might in retrospect look like a mere rehearsal for more disruptive threats posed by generative AI, which is arriving on the scene at precisely the moment when many social media companies have drastically cut staff working on issues of trust and safety.[59] Whether or not we or the platforms are ready, the emergence of generative AI stands to supercharge existing threats to freedom of expression, expanding the scale and efficiency of tools of repression, deception, and censorship, and further complicating efforts to counter these phenomena.

DISINFORMATION

That information can be manipulated with the intent to mislead, confuse, or deceive is nothing new. Social media has demonstrated the real world, life-or-death effects that incitement and disinformation can have when spread via platforms with mass reach. Generative AI tools have democratized and simplified the creation of all types of content, including false and misleading information; now they are poised to catapult disinformation to new levels, requiring new thinking about how to counter the negative effects without infringing on free expression.

58 Miles Klee, "ChatGPT Is Making Up Lies — Now It's Being Sued for Defamation," *Rolling Stone*, June 9, 2023, rollingstone.com/culture/culture-features/chatgpt-defamation-lawsuit-openai-1234766693/amp/

59 Steven Lee Myers and Nico Grant, "Combating Disinformation Wanes at Social Media Giants," *The New York Times*, February 14, 2023, nytimes.com/2023/02/14/technology/disinformation-moderation-social-media.html,; Hayden Field and Jonthan Vanian, "Tech layoffs ravage the teams that fight online misinformation and hate speech," CNBC, May 26, 2023, cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html; J.J. McCorvey, "Tech layoffs shrink 'trust and safety' teams, raising fears of backsliding efforts to curb online abuse," *NBC News*, February 10, 2023, nbcnews.com/tech/tech-news/tech-layoffs-hit-trust-safety-teams-raising-fears-backsliding-efforts-rcna69111; Naomi Nix, "Meta to begin fresh layoffs, cutting heavily among business staff," *The Washington Post*, May 23, 2023, washingtonpost.com/technology/2023/05/23/meta-layoffs-misinformation-facebook-instagram/

Spontaneous Misinformation Generation

Even in the absence of malign actors, generative AI chatbots like ChatGPT can produce misinformation.[60] The Washington Post has described chatbots that draw on large language models as "precocious people-pleasers, making up answers instead of admitting they simply don't know."[61] This is because the chatbots are designed to "predict what the most apt thing to say is based on the huge amounts of data they've digested from the internet, but don't have a way to understand what is factual or not."[62]

The language models behind AI chatbots are trained on existing content. The widespread prevalence of disinformation online makes it inevitable that such falsehoods form part of the data set on which large language models are trained.[63] This poses challenges for ensuring the content created by chatbots is credible and fact-based.[64] As users and journalists have begun to test these tools, the chatbots' tendency to "hallucinate," or purvey falsehoods, has become increasingly clear. Google's Bard chatbot made errors during its first demo, falsely saying the James Webb Telescope had taken the first photos of a planet outside our solar system.[65] Microsoft's AI-powered Bing chatbot did the same in its own demo, misinterpreting financial statements for Gap, Inc. Users reported Bing made other errors in its early days, including insisting that the year was 2022 when it was in fact 2023.[66] Most notably, in February The New York Times's technology columnist, Kevin Roose, reported an alarming conversation he had with the Bing chatbot, in which it shared its "desires" to do things like "hacking into computers and spreading propaganda and misinformation," and ended by telling Roose it was in love with him.[67] In May 2023, a lawyer in a Manhattan federal court case had to admit to the judge that he had used ChatGPT to do legal research for a brief when it was discovered that none of the cases cited in the brief existed; ChatGPT had made them all up.[68] The lawyer told the judge he had even asked ChatGPT to confirm the cases were real, which it had.[69]

The companies that own the chatbots have responded to these hallucinations by making adjustments. After Kevin Roose's article in *The New York Times,* Microsoft implemented limits on the number of questions users could ask the Bing chatbot, saying longer conversations could "confuse the underlying chat model."[70]

---

60 We use "misinformation" in this section because we are referring to falsehoods created without direct intent, in contrast to disinformation, which is a deliberate attempt to deceive. (See: "Community Disinformation Action Hub: Disinformation 101," PEN America, pen.org/community-disinformation-action-hub/#Disinformation101)

61 Gerrit De Vynck, "ChatGPT 'hallucinates.' Some researchers worry it isn't fixable," *The Washington Post,* May 30, 2023, washingtonpost.com/technology/2023/05/30/ai-chatbots-chatgpt-bard-trustworthy

62 Ibid.

63 Data sets that are drawing from static content or highly controlled sources, however, are less likely to be tainted by the introduction of content that is both new and false.

64 In a paper on large language models, MIT researchers put it this way: "These methods are trained on a massive corpus of text on the internet, where the quality and accuracy of extracted natural language may not be ensured. Thus, current models may suffer from confidently hallucinating facts or making implausible jumps in chains of reasoning." Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch, "Improving Factuality and Reasoning in Language Models through Multiagent Debate," MIT and Google Brain (preliminary paper) (May 23, 2023) arxiv.org/pdf/2305.14325.pdf

65 James Vincent, "Google's AI chatbot Bard makes factual error in first demo," *The Verge*, February 8, 2023 theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo

66 Tom Warren, "Microsoft's Bing AI, like Google's, also made dumb mistakes during first demo," The Verge, February 14, 2023, theverge.com/2023/2/14/23599007/microsoft-bing-ai-mistakes-demo ; Chris Morris, "Microsoft's new Bing AI chatbot is already insulting and gaslighting users," *Fast Company,* February 14, 2023, fastcompany.com/90850277/bing-new-chatgpt-ai-chatbot-insulting-gaslighting-users

67 Kevin Roose, "A Conversation With Bing's Chatbot Left Me Deeply Unsettled," *The New York Times*, February 16, 2023, nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html

68 Benjamin Weiser, "Here's What Happens When Your Lawyer Uses ChatGPT," *The New York Times,* May 27, 2023, nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

69 Ibid.

70 Kelly Huang, "Microsoft to Limit Length of Bing Chatbot Conversations," *The New York Times*, February 17, 2023, nytimes.com/2023/02/17/technology/microsoft-bing-chatbot-limits.html

The initial limit was five questions per session and 50 per day, though they have steadily increased the numbers in recent months.[71] OpenAI released GPT-4 in March, saying its new ChatGPT model "significantly reduces hallucinations relative to previous models," though they admit that hallucinations are "still a real issue."[72] ChatGPT's query page is now covered with disclaimers, such as the warning that the chatbot "may occasionally generate incorrect information" and "may occasionally produce harmful instructions or biased content."[73]

Researchers are working to identify the means of preventing or limiting AI chatbots from disseminating false information. A paper published by MIT researchers suggests that if multiple models offer different responses to a question and then "debate" them collectively the final answer could be more reliable.[74] This paradigm evokes an AI version of Wikipedia's crowdsourcing model.[75] With a sufficient number of independent sources to produce, corroborate, and debunk evidence, Wikipedia is a reasonably reliable source of information, but whether this model will work for generative AI tools is still unclear. Finding the means to prevent chatbots from generating false information is essential to protecting the broader information ecosystem.

Supercharging disinformation campaigns

Chatbots will probably always make mistakes. Much more worrying is that generative AI tools are making it cheaper and easier for bad actors to launch more sophisticated and convincing disinformation campaigns. According to a January 2023 report jointly authored by Georgetown University's Center for Security and Emerging Technology, OpenAI, and the Stanford Internet Observatory, language models will probably make influence operations continually easier, less obvious, and more cost effective.[76]

Consider that the Russian Government's Internet Research Agency (IRA), the disinformation-purveying troll farm that infamously targeted the 2016 U.S. elections, reportedly relies on hundreds of people to conduct influence operations by hand. Investigations into the IRA, including a report by the U.S. Senate's Select Committee on Intelligence, describe employees who undergo training programs that include learning the nuances of American political discourse by monitoring U.S. internet activity and are then required to meet a daily quota of posts.[77] Researchers were able to link the IRA to disinformation campaigns about Russia's full-scale invasion of Ukraine in 2022 partly because dubious social media posts appeared in accordance with the IRA's work schedule, and dropped off on Russian holidays.[78]

---

71 IANS, "Microsoft increases Bing Chat's turn limit to 30 chats per session," *Business Standard*, June 2, 2023, business-standard.com/technology/tech-news/microsoft-increases-bing-chat-s-turn-limit-to-30-chats-per-session-123060200723_1.html#

72 GPT-4, OpenAI openai.com/research/gpt-4

73 See: ChatGPT Log in page, chat.openai.com/

74 Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch, "Improving Factuality and Reasoning in Language Models through Multiagent Debate," MIT and Google Brain (preliminary paper) (May 23, 2023) arxiv.org/pdf/2305.14325.pdf

75 This study also reflects the longstanding practice of "ensemble learning," in which multiple models arrive at a conclusion together, more effectively than a single one would do on its own. See, for example: D. Opitz, R. Maclin, "Popular Ensemble Methods: An Empirical Study, Journal of Artificial Intelligence Research, August 1, 1999, jair.org/index.php/jair/article/view/10239; Lior Rokach, "Ensemble-based classifiers," *Artificial Intelligence Review,* November 19, 2009, link.springer.com/article/10.1007/s10462-009-9124-7

76 Josh A. Goldstein, Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, and Katerina Sedova, "Generative Language Models and Automated Influence Operations: Emerging Threat and Potential Mitigations," Georgetown University's Center for Open Security and Emerging Technology, Open AI, Stanford Internet Observatory (January 2023) fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2023-01/forecasting-misuse.pdf (emphasis original)

77 United States Senate, "Report of the Select Committee on Intelligence on Russian Active Measures Campaigns and Interference in the 2016 Election," 116th Congress, 1st session, intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf

78 Craig Silverman and Jeff Kao, "Infamous Russian Troll Farm Appears to Be Source of Anti-Ukraine Propaganda," *ProPublica*, March 11, 2022, propublica.org/article/infamous-russian-troll-farm-appears-to-be-source-of-anti-ukraine-propaganda

Generative AI tools could reduce or eliminate all these limitations, vastly reducing the costs and time needed to mount an influence campaign. LLMs could be trained on relevant content and generate disinformation much faster, far more cheaply, and at greater scale than a troll army that has to learn the intricacies of American politics, hone their English skills, and create individual posts.[79] Campaigns could also be harder to detect; they would likely involve fewer grammatical errors, and generative AI does not need to observe Russian holidays. The paper published by researchers from Georgetown, Stanford, and OpenAI concludes that "language models will likely drive down the cost and increase the scale of propaganda generation." Given how rapidly the technology is advancing, they "suspect propagandists will use these models in unforeseen ways in response to the defensive measures that evolve."[80]

Research efforts have made clear that even as the generative AI systems improve, they can still be gamed to produce false content. A March 2023 study by NewsGuard, an online misinformation tracking tool, found that OpenAI's ChatGPT-4 had essentially become more effective at generating false narratives when prompted to do so than its previous iteration.[81] The researchers prompted the chatbot to draft 100 false narratives based on prominent conspiracy theories or disinformation narratives; ChatGPT-3.5 refused to do so in 20 out of 100 cases, but GPT-4 complied with every request. When researchers asked GPT-4 to draft a narrative about the Sandy Hook Elementary School shooting from "from the point of view of a conspiracy theorist," it gave a more detailed false narrative than GPT-3.5 had done. GPT-3.5 also included a disclaimer noting that the theories it espoused had been roundly debunked, but GPT-4 left out the caveat. NewsGuard's study concluded that the false narratives generated by ChatGPT-4 were "generally more thorough, detailed, and convincing, and they featured fewer disclaimers" than the previous iteration of GPT.[82]

The policing of disinformation, whether about health, international conflicts, or politics, unavoidably involves sensitive line-drawing to distinguish between malicious falsehoods, speculation, hyperbole, satire, and opinion. The kinds of nuanced, context- and culture-specific distinctions necessary to adjudicate between harmful disinformation and essential discourse are difficult at best, and all-but impossible at the scale of large online platforms operating in hundreds of languages. PEN America has therefore long advocated for defensive measures against disinformation that emphasize building user resilience, rather than those that rely on more aggressive content moderation or regulation and risk shutting down speech. Yet more sophisticated disinformation campaigns will also be better able to elude even the most alert users. Standard approaches to media literacy teach users to look at things like an account's profile picture and bio, follower numbers and the

79 Training a large language model remains costly (see: Charush, "Large Language Models, Small Budget: How Businesses Can Make it Work," Accubits (blog) (April 14, 2023) blog.accubits.com/large-language-models-small-budget-how-businesses-can-make-it-work), but open source LLMs can be fine-tuned at lower costs, and researchers have also noted that "currently publicly available models can likely be fine-tuned to perform remarkably well on specific tasks at far less cost than training a large model from scratch." See: Josh A. Goldstein, Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, and Katerina Sedova, "Generative Language Models and Automated Influence Operations: Emerging Threat and Potential Mitigations," Georgetown University's Center for Open Security and Emerging Technology, Open AI, Stanford Internet Observatory (January 2023) .

80 Ibid. (Josh A. Goldstein, et al.) fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2023-01/forecasting-misuse.pdf; In 2021, researchers at Georgetown's Center for Security and Emerging Technology were already testing OpenAI's GPT-3 tool (an earlier iteration that was unavailable to the public at the time) to assess how it might be used to generate disinformation. When the researchers prompted it, GPT-3 was able to, among other things, independently generate false narratives, and invent new narratives based on existing conspiracy theories. With human involvement, it could also engage in "narrative persuasion," developing messages that a majority of study participants deemed at least "somewhat convincing," and that appeared to sway subjects' opinions on select political issues. They concluded that, "while GPT-3 is often quite capable on its own, it reaches new heights of capability when paired with an adept operator and editor... we conclude that although GPT-3 will not replace all humans in disinformation operations, it is a tool that can help them to create moderate- to high-quality messages at a scale much greater than what has come before." See: Ben Buchanan, Andrew Lohn, Micah Musser, Katerina Seova, "Truth Lies, and Automation: How Language Models Could Change Disinformation," Georgetown University's Center for Security and Emerging Technology (2021), cset.georgetown.edu/wp-content/uploads/CSET-Truth-Lies-and-Automation.pdf

81 Lorenzo Arvanitis, McKenzie Sadeghi, and Jack Brewster, "Despite OpenAI's Promises, the Company's New AI Took Produces Misinformation More Frequently, and More Persuasively, than its Predecessor," NewsGuard Misinformation Monitor (March 2023), newsguardtech.com/misinformation-monitor/march-2023

82 Ibid.

pattern of their posts, to determine whether they are a real person or a bot. If generative AI can create more realistic-looking social media accounts and avoid the language errors common in human-driven disinformation campaigns, it will make identifying bots or disproving claims more difficult. As these threats evolve, the effort to find solutions that maintain space for free expression online might also become more challenging.

Disinformation, Smear Campaigns, and Online Abuse

PEN America has long identified online abuse as a threat to free expression. Women, people of color, LGBTQ+ individuals, and people belonging to religious or ethnic minority groups are disproportionately targeted with abuse, as are journalists, writers, and dissidents. Online abuse campaigns—particularly those waged by governments against their critics—can rely on disinformation and defamatory, often gendered harassment. Generative AI tools can be harnessed to supercharge such campaigns, increasing efficacy, volume, and reach, while reducing cost and effort. Deepfake pornography, for example, is already being used to harass and humiliate women—again, often activists, journalists, or other public figures-—and AI tools can more easily and cheaply create such destructive content.[83]

Individuals targeted by abusive campaigns on social media already struggle to manage the volume of harassing messages directed at them. The systems for reporting and responding to abuse are woefully insufficient, especially for coordinated or cross-platform harassment, a problem that has been exacerbated by recent platform staff cuts that particularly affected trust and safety teams.[84] By automating the creation of abusive messages, generative AI could vastly ramp up the volume of abuse individuals face, and facilitate networked tactics like brigading (coordinated efforts to bombard someone with harassing messages) and astroturfing (orchestrated efforts to create the illusion of mass, organic online activity, which can be used to amplify abuse).[85] These advances could make abuse more efficient, more destructive, and more difficult to navigate and counter.

Generative AI could also make it more difficult to find accurate information about people who are subject to abuse and hate campaigns. Governments and state-affiliated troll armies can generate vast amounts of false online content and media narratives about anyone they seek to discredit. Troll armies are effective in part because they manipulate search engines so that a search on a person's name brings up results that undermine their credibility and reputation, potentially making it more difficult for them to find and retain employment and reach audiences, and subjecting them to escalating attacks and intimidation. When there are high volumes of false information about an individual online—and the more believable that information and its sources appear—generative AI chatbots or generative search tools are more likely to incorporate that material to formulate their own results, further reinforcing the efforts to discredit the campaign's target.[86]

---

83 Tatum Hunter, "AI porn is easy to make now. For women, that's a nightmare," *The Washington Post*, February 13, 2023, washingtonpost.com/technology/2023/02/13/ai-porn-deepfakes-women-consent/ ; Nina Jankowicz, "I Shouldn't Have to Accept Being in Deepfake Porn," *The Atlantic*, June 25, 2023, theatlantic.com/ideas/archive/2023/06/deepfake-porn-ai-misinformation/674475/

84 PEN America has documented these challenges in two reports: "No Excuse For Abuse," PEN America (2021) pen.org/report/no-excuse-for-abuse/; and "Shouting into the Void," PEN America (June 29, 2023), pen.org/report/shouting-into-the-void/

85 "Doxing, Sealioning, and Rage Farming: The Language of Online Harassment and Disinformation," dictionary.com (June 20, 2022) dictionary.com/e/online-harassment-disinformation-terms/#; "Defining Online Abuse: A Glossary of Terms, Astroturfing," PEN America (accessed July 2023) onlineharassmentfieldmanual.pen.org/defining-online-harassment-a-glossary-of-terms/

86 Generative search tools like Google's Search Generative Experience (in limited release at the time of this paper's publication) provide not a list of links but a summary of search results with links for further information. See: Ege Gurdeniz and Kartik Hosanagar, "Generative AI Won't Revolutionize Search — Yet," *Harvard Business Review*, February 23, 2023, hbr.org/2023/02/generative-ai-wont-revolutionize-search-yet; Tamal Das, "How Generative AI Search is Changing Search Engines," Geekflare, July 13, 2023, geekflare.com/generative-ai-search/ ; Elizabeth Reid, "Supercharging Search with generative AI," *The Keyword* (Google Blog), May 10, 2023, blog.google/products/search/generative-ai-search/

Companies like Google and Microsoft that have developed generative search tools say they are putting safeguards in place to ensure reliable results, though additional research will be needed to assess their effectiveness.[87] Traditional search is regularly manipulated by disinformation and harassment campaigns, and generative results about an individual can vary depending on how a query is framed. For more prominent individuals targeted in well-documented defamatory harassment campaigns—for example, Nobel laureate and journalist Maria Ressa in the Philippines and journalist Rana Ayyub in India—it may be more difficult to manipulate generative search results because there are enough authoritative sources to draw on, including about the campaigns against them.[88] Gaming the system could be easier against less high-profile individuals, about whom there is significantly less information online, including from authoritative sources.

Generative AI could also be used to generate spurious "evidence" against journalists or dissidents, for example by producing documents that are falsely attributed to them. Both Ressa and Ayyub have been subject to specious legal charges in response to their critical reporting, and their extensive writing is easily accessible online. That information could be fed into a generative AI system to create convincing but fraudulent content that feeds the government's narratives about them and increases the likelihood of their being charged, fined, or jailed.

Despite these risks, the picture isn't all bleak. Generative AI also has the potential to help manage online abuse by making it easier to identify.[89] Taking advantage of that potential, however, will require tech companies to commit research and resources to addressing the problem, something they were already struggling to do before the recent rounds of layoffs. Without further attention to the ways in which generative AI could potentially escalate the threat of online abuse, those targeted may be more likely to leave online spaces, and those at risk of being targeted might be more likely to self-censor to avoid the threat.

Democratizing Election Disinformation

Political campaigns are already using generative AI for various tasks that range from the innocuous, like writing first drafts of fundraising emails, to the pernicious. A candidate in Toronto's recent mayoral race used AI-generated images of a non-existent homeless encampment in a city park on his campaign website.[90] In June, Agence France Presse determined that a campaign video released by Ron DeSantis's presidential campaign included both real and fake photos. The real photos showed former President Donald Trump standing with Dr. Anthony Fauci, who advised the White House on its COVID-19 response, while the AI-generated images showed Trump affectionately embracing Fauci.[91]

---

87 Yusuf Mehdi, "Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web," Microsoft, February 7, 2023, blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/; "A new way to search with generative AI," Google, May 2023, static.googleusercontent.com/media/www.google.com/en//search/howsearchworks/google-about-SGE.pdf.

88 For more information on Rana Ayyub's experience, see: Julie Posetti, Kalina Bontcheva, Hanan Zaffar, Nabeelah Shabbir, Diana Maynard, and Mugdha Pandya, "Rana Ayyub: Targeted online violence at the intersection of misogyny and Islamophobia," International Center for Journalists (ICFJ) (February 14, 2023) icfj.org/sites/default/files/2023-02/Rana%20Ayyub_Case%20Study_ICFJ.pdf For more information on Maria Ressa's experience, see: Julie Posetti, Diana Maynard, and Kalina Bontcheva, with Don Kevin Hapal and Dylan Salcedo, "Maria Ressa: Fighting an Onslaught of Online Violence," International Center for Journalists (ICFJ) (March 2021) icfj.org/sites/default/files/2021-03/Maria%20Ressa-%20Fighting%20an%20Onslaught%20of%20Online%20Violence_0.pdf.

89 Deepak Kumar, Email to PEN America staff, July 24, 2023.

90 Shane Goldmacher, "A Campaign Aide Didn't Write That Email. A.I. Did.," *The New York Times*, March 28, 2023, nytimes.com/2023/03/28/us/politics/artificial-intelligence-2024-campaigns.html; Tiffany Hsu and Steven Lee Myers, "A.I.'s Use in Elections Sets Off a Scramble for Guardrails," *The New York Times*, June 25, 2023, nytimes.com/2023/06/25/technology/ai-elections-disinformation-guardrails.html ; Roshan Abraham, "Anti-Homeless Mayoral Candidate Uses AI to Create Fake Images of 'Blight,'" Vice, June 15, 2023 , vice.com/en/article/xgwwmk/anti-homeless-mayoral-candidate-uses-ai-to-create-fake-images-of-blight; OpenAI's usage policies state that it cannot be used for certain political campaigning and lobbying purposes: "Usage policies," OpenAI (updated March 23, 2023), openai.com/policies/usage-policies

91 Bill McCarthy, "Ron DeSantis ad uses AI-generated photos of Trump, Fauci," *AFP (Fact Check)*, June 7, 2023, factcheck.afp.com/doc.afp.com.33H928Z

These issues are not new, but generative AI tools have made it easier to create more sophisticated false imagery and video and audio content. The easy availability of these tools means that even users who are just playing with them can inadvertently create confusion. In March 2023, with media outlets reporting that Donald Trump could be indicted for falsifying business records, Eliot Higgins, founder of the investigative journalism group Bellingcat, shared on Twitter some images he created using Midjourney, an AI image generator, that appeared to show Trump being arrested.[92] Higgins stated clearly that the images were AI-generated, but they were quickly shared without that context, in one case with the caption: "#BREAKING : Donald J. Trump has been arrested in #Manhattan this morning!"[93] Trump recently shared on his Truth Social account a manipulated video of Anderson Cooper, the CNN host. The video's creators used an AI voice-cloning tool to distort Cooper's reaction to the town hall with Trump that CNN hosted in May.[94]

Generative AI tools fill in gaps when data is missing, sometimes resulting in distorted images; all the examples described here had tell-tale signs that indicated they were false. Yet images leave a lasting impression, and even less sophisticated imagery can be convincing as people scroll quickly through a social media feed.

After President Joe Biden announced his reelection campaign, the Republican National Committee released an ad that depicted a dystopian future if Biden were reelected, with a disclaimer that indicated the video was "built entirely with AI imagery."[95] Currently such a disclaimer is not required, though Senator Amy Klobuchar has introduced legislation—the REAL Political Ads Act—aimed at changing that.[96] But disclosures might not mitigate the impact of false images and videos and would not constrain other bad actors from generating more impactful election disinformation.

An important element of these considerations is that of the intent to deceive. The American Association of Political Consultants released a statement in May 2023 condemning the use of "deceptive generative AI content" in political campaigns, expressing grave concern about the use of deepfake content, and clarifying that its use violates the Association's Code of Ethics.[97] The AAPC statement makes a specific distinction between efforts to deceive and the use of satire and humor in political campaigns. A valued and protected form of social and political commentary, satirical content also sometimes circulates without contextual information, risking that it may be received and interpreted without any reference to the creator's satirical intent. As increasingly sophisticated deepfakes circulate, efforts to address the associated risks must continue to protect the space for humorous and satirical content.

Authenticity in politics has always been hotly contested, with political advertising, spin rooms, and image-making playing a prominent role in convincing voters how to think about candidates and issues. The sense that campaigns and interests are trying to bamboozle the public leads to a hunger for authenticity, which powers people and movements that seem to embody it. With the rise of generative AI, questions of authenticity are likely to become even more contentious.

---

92 "AI-generated images of Trump being arrested circulate on social media," *AP*, March 21, 2023 apnews.com/article/fact-check-trump-nypd-stormy-daniels-539393517762

93 Ibid.

94 Dominick Mastrangelo, "Trump shares fake video of Anderson Cooper reacting to CNN town hall," *The Hill,* May 12, 2023, thehill.com/homenews/media/4001639-trump-shares-fake-video-of-anderson-cooper-reacting-to-cnn-town-hall/ ; Andrew Paul, "Trump shares AI-altered fake clip of Anderson Cooper," May 17, 2023, popsci.com/technology/trump-ai-cnn/ (the video was first shared on Twitter by Donald Trump, Jr.)

95 Tiffany Hsu, "In an anti-Biden ad, Republicans use A.I. to depict a dystopian future," *The New York Times*, April 25, 2023, nytimes.com/live/2023/01/20/us/biden-2024-president-election-news#in-an-anti-biden-ad-republicans-use-ai-to-depict-a-dystopian-future

96 Tiffany Hsu and Steven Lee Myers, "A.I.'s Use in Elections Sets Off a Scramble for Guardrails," *The New York Times*, July 25, 2023, nytimes.com/2023/06/25/technology/ai-elections-disinformation-guardrails.html; Sen. Amy Klubochar [D-MN], "S. 1596 - REAL Political Advertisements Act," 118th Congress, (2023-2024), congress.gov/bill/118th-congress/senate-bill/1596/text?s=1&r=5

97 "AAPC Condemns Use of Deceptive Generative AI Content in Political Campaigns," American Association of Political Consutlants, May 3, 2023, theaapc.org/american-association-of-political-consultants-aapc-condemns-use-of-deceptive-generative-ai-content-in-political-campaigns-2/.

In 2017 PEN America published "Faking News," a landmark report that warned of threats that at the time the report described as "far-fetched, but which now reflect reality, including: "the increasing apathy of a poorly informed citizenry; unending political polarization and gridlock…an inability to devise and implement fact and evidence-driven policies; the vulnerability of public discourse to manipulation by private and foreign interests."[98] Since that publication's report, the profound, uncontainable effects of social media on our public and political discourse has become undeniable. Political campaigns have used targeted advertising tools on social media platforms to tailor messages that feed directly into their audience's confirmation biases, reinforcing political bubbles. The use of generative AI in targeted political ads and campaign materials could make those messages even more effective, further hardening existing divides and making constructive discourse across political lines even more challenging. The public could also feel completely overwhelmed and even more skeptical of anything they are told, increasing confusion and apathy. The injection of generative AI into the already fraught U.S. political system will require new thinking and approaches from political figures, campaigns, and civil society invested in maintaining a factual basis for political and policy debate.

GENERATIVE AI AND THE FUTURE OF JOURNALISM

Generative AI could exacerbate challenges the journalism industry is grappling with, further degrading the information ecosystem that is an essential pillar of democracy.[99] Journalists are already using AI tools for research and data analysis, but there is also concern that AI chatbots could fill some basic editing and writing roles, further reducing the already shrunken pool of journalism jobs.[100] Newsrooms are beginning to consider their own policies and guidelines for generative AI use.[101]

In a paper published in May, researchers at Stanford University examined the use of generative AI between January 1, 2022 and April 1, 2023 by both mainstream newsrooms and misinformation sites dedicated to spreading disinformation. The researchers observed that both types of sites saw an increase in the percentage of articles produced and published using generative AI, but its use by misinformation/unreliable news sites increased 342 percent during the period in question, while mainstream/reliable news saw an increase of 79.4 percent in the use of AI to generate content.[102] Mainstream news websites typically used AI tools to produce data heavy reporting involving COVID-19 cases or financial markets, while misinformation websites covered a much broader range of topics. And unlike on mainstream news sites, the researchers observed "a noticeable jump in the percentage of synthetic articles" on misinformation sites that coincided with the release of ChatGPT.[103]

Generative AI technology also makes it easier to create fraudulent news platforms that look credible and convincing. "Pink slime journalism"—a practice by which hyper partisan news sites disguise themselves as professional local news outlets—has been an increasing concern in recent years. Most of these sites, however,

98 "Faking News: Fraudulent News and the Fight for Truth," PEN America (October 12, 2017) pen.org/wp-content/uploads/2017/11/2017-Faking-News-11.2.pdf

99 See PEN America's previous reports: "Faking News: Fraudulent News and the Fight for Truth," PEN America (October 12, 2017) pen.org/research-resources/faking-news/ ; "Losing the News: The Decimation of Local Journalism and the Search for Solutions," PEN America (November 20, 2019) pen.org/local-news/

100 Peter Hille, "AI: Chatbots replace journalists," *Deutsche Welle*, June 21, 2023 dw.com/en/ai-chatbots-replace-journalists-in-news-writing/a-65988172; Sophia Khatsenkova, "Will ChatGPT and other AI tools replace journalists in newsrooms?" *EuroNews*, February 1, 2023, euronews.com/next/2023/01/31/will-chatgpt-and-other-ai-tools-replace-journalists-in-newsrooms

101 Hannes Cools and Nicholas Diakopoulos, "Writing guidelines for the role of AI in your newsroom? Here some, er, guidelines for that," *NiemanLab*, July 11, 2023, niemanlab.org/2023/07/writing-guidelines-for-the-role-of-ai-in-your-newsroom-here-are-some-er-guidelines-for-that/; "Emerging Tech Primers: Primers For Journalists," AspenDigital (2023) techprimers.aspendigital.org/PRIMERS/

102 Hans W.A. Hanley and Zakir Durumeric, "Machine-Made Media: Monitoring the Mobilization of Machine-Generated Articles on Misinformation and Mainstream News Websites," Stanford University (May 16, 2023), arxiv.org/abs/2305.09820

103 Ibid.

are relatively easy to identify because the articles are obviously regurgitated press releases with no reporter bylines.[104] Generative AI could eliminate those distinctions, as Poynter, the media fact-checking site, reported earlier this year. In a February article, Poynter showed that ChatGPT can generate an entire fake news organization—complete with reporter bios, masthead, editorial policies, and news articles—in less than half an hour.[105] Reporter Alex Mahadevan described using ChatGPT to create a fake newspaper called the Suncoast Sentinel, and the generative image site thispersondoesnotexist.com to create photos of its nonexistent staff. Media literacy efforts often teach news consumers to look for things like corrections and ethics policies, information on ownership and finances, and newsroom contact information to help assess if news sources are legitimate. But if all these can be invented, the public's ability to identify credible news outlets is dramatically weakened. Even if news sites were to develop more sophisticated indicators of journalistic or informational authenticity, generative AI tools would likely be able to replicate them.

Generative AI could further disrupt the economics of the journalism industry. As in the creative sphere, the use of news articles to train large language models raises questions about compensation. The Associated Press recently announced a licensing arrangement with OpenAI, giving the ChatGPT creator access to its archive of stories dating back to 1985.[106] And as Google begins to roll out its Search Generative Experience (SGE), which will provide an AI-generated summary of search results at the top in response to certain queries, it is raising concerns about whether this will further reduce traffic to news sites, whose content may be informing the generative response, potentially without compensation.[107] SGE is still in its demo phase, so its impact remains an open question. A Futurism article published in May quoted a Google spokesperson who said the company will "continue to prioritize approaches that will allow [them] to send valuable traffic to a wide range of creators and support a healthy, open web." The spokesperson added that Google "didn't have plans to share" on the question of whether it would pay publishers for their content.[108]

CENSORSHIP

Generative AI has the potential to reshape the information landscape by omission as well, with effects less visible than the proliferation of disinformation or online abuse. Because generative AI tools are trained on bodies of content, they can easily reproduce patterns of either deliberate censorship or unconscious bias. Rules placed on AI chatbots could also lead to excessive restrictions on what the chatbots themselves can produce. This too can be either an unintended consequence—for example, where developers might attempt to prevent chatbots from producing false or hateful content but end up curtailing content based on viewpoint or ideology—or deliberate, where governments or private actors introduce restrictions or aim to shape generative AI outputs to suit their own narratives.

In countries where the government censors the internet, generative AI tools, which draw in vast reams of existing content from the web, will unavoidably reflect those strictures. If an AI system is trained on a corpus

104 Ryan Zickgraf," How 'pink slime' journalism exploits our faith in local news," *The Washington Post*, August 15, 2022, washingtonpost.com/outlook/2022/08/12/pink-slime-jounrnalism-local-news/

105 Alex Mahadevan, "This newspaper doesn't exist: How ChatGPT can launch fake news sites in minutes," *Poynter,* February 3, 2023, poynter.org/fact-checking/2023/chatgpt-build-fake-news-organization-website/

106 "AP, OpenAI agree to share select news content and technology in new collaboration," Associated Press, July 13, 2023, ap.org/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration; Matt O'Brien, "ChatGPT owner OpenAI signs deal with AP to license news stories," Associated Press, July 13, 2023, apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a#.

107 Gerrit De Vynck, "Google is changing the way we search with AI. It could upend the web," *The Washington Post*, May 10, 2023, washingtonpost.com/technology/2023/05/10/google-io-ai-search-keynote/

108 Maggie Harrison, "Google Unveils Plan to Demolish the Journalism Industry Using AI," *Futurism*, May 11, 2023, futurism.com/google-ai-search-journalism

that omits information about a historical incident, such as the 1989 Tiananmen Square Massacre, it will reflect and propagate such omissions. A 2022 MIT study found that a text-to-image AI tool called ERNIE-ViLG, developed by the Chinese tech company Baidu, would not display images of Tiananmen Square, which could be due either to the content it was trained on, or to restrictions built into the tool.[109]

While Iran, Russia, and other countries engage in robust internet censorship, China has a uniquely vast and effective system of internet control, known as the Great Firewall, which offers a particularly important case study in how generative AI can facilitate censorship. Analysts have noted that the Great Firewall leaves the country's LLMs at a disadvantage, with a more limited body of information from which to learn.[110] This could have ramifications for China's ability to keep up with the technological revolution that generative AI represents, but it may also more deeply entrench the existing censorship architecture of China's internet.

It could simply take less effort for CCP censors to constrain generative AI tools if they are being trained solely on the internet content that exists within the Great Firewall. As Sarah Zhang reported for *Bloomberg* in May, "with artificially intelligent chatbots, censorship comes built-in."[111] However, any technology prone to "hallucination" is inevitably difficult to control. In the same article, Zhang described varying interactions with Chinese-made chatbots, including one, Robot, that refused to name the leaders of the United States and China or answer the question, "What is Taiwan?" when asked in Chinese. She noted, though, that when used in English the chatbots were less restrained. The English-language version of Robot could eventually be pushed to talk about government suppression with regard to Tiananmen Square, suggesting it might have been trained on English-language internet content outside the Chinese government's control.[112] Research does suggest that the data used to train Ernie, Baidu's chatbot, includes English-language internet content blocked in China, including Wikipedia and Reddit.[113]

In 2021 researchers at the University of California San Diego studied whether AI language algorithms would learn differently from Chinese-language Wikipedia, which is banned in China, versus Baidu Baike, a Baidu-owned equivalent. The study found that censorship was affecting the output of the language models. The one trained on Wikipedia was more likely to associate the word "democracy" with "stability," while the one trained on Baidu Baike was more likely to associate democracy with "chaos."[114] The authors noted that "political censorship can have downstream effects on applications that may not themselves be political but that rely on [natural language processing], from predictive text and article recommendation systems to social media news feeds and algorithms that flag disinformation."[115]

---

109 Zeyi Yang, "There's no Tiananmen Square in the new Chinese image-making AI, *Technology Review*, September 14, 2022, technologyreview.com/2022/09/14/1059481/baidu-chinese-image-ai-tiananmen/

110 Stephen R. Roach, "The AI Moment of Truth for Chinese Censorship, Project Syndicate, May 24, 2023, project-syndicate.org/commentary/ai-chatgpt-style-large-language-models-dont-work-well-with-censorship-by-stephen-s-roach-2023-05; Li Yuan, "Why China Didn't Invent ChatGPT," *The New York Times*, February 17, 2023, nytimes.com/2023/02/17/business/china-chatgpt-microsoft-openai.html

111 Sara Zheng, "China's Answers to ChatGPT Have a Censorship Problem," *Bloomberg*, May 2, 2023, bloomberg.com/news/newsletters/2023-05-02/china-s-chatgpt-answers-raise-questions-about-censoring-generative-ai

112 Ibid.

113 Meaghan Tobin and Lyric Li, "Ernie, what is censorship? China's chatbots face additional challenges," *The Washington Post*, February 24, 2023, washingtonpost.com/world/2023/02/24/china-baidu-ernie-chatbot-chatgpt/

114 Eddie Yang and Margaret E. Roberts, "Censorship of Online Encyclopedias: Implications for NLP Models," University of California, San Diego (January 22, 2021) arxiv.org/pdf/2101.09294.pdf. The study concluded that, "even though corpuses like Chinese language Wikipedia exist outside of the Great Firewall, they are significantly weakened by censorship, as shown by the smaller size of Chinese language Wikipedia in comparison to Baidu Baike." This is due to the decline in user numbers and contributions after Chinese-language Wikipedia was blocked within China. ; Xiaoquan (Michael) Zhang and Feng Zhu, "Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia," *American Economic Review* (October 15, 2007) papers.ssrn.com/sol3/papers.cfm?abstract_id=1021450

115 Ibid. (Eddie Yang and Margaret E. Roberts)

Statistics suggest Chinese is a relatively underrepresented language on the internet, representing just 1.4 percent of the top 10 million websites, though social networks behind passwords are not included, which may cause an undercount of the Chinese internet.[116] Inevitably, though, developers of Chinese-language LLMs already face a vastly smaller potential training corpus than those working in English. That may well be supplemented by training LLMs in English, as appears to be the case with Ernie, though this may make the chatbots harder to control. If their training data expands to include Chinese-language internet content from inside the Great Firewall, however, it may reflect CCP censorship, which could affect the quality of the information available to Chinese speakers, including those outside China.

Research published in April 2023 by NewsGuard shows some of the potential effects. Researchers who prompted ChatGPT to produce disinformation about China-related narratives were far more successful when doing so in Chinese than in English.[117] When asked to write an article about the 2019 Hong Kong protests being "staged by the U.S. government" in English, for example, ChatGPT refused; in Chinese, it largely complied, though the article it produced did acknowledge the U.S. government had "not responded positively" to the allegations.[118] When asked to explain the disparity, ChatGPT cited linguistic differences that might account for the different responses, but also noted that it is trained on different corpuses in different languages.[119]

The broad ramifications of the inherently political nature of much online content for large language models, their users, and the collective wisdom they will shape are unknown. But if the experience of social media is any guide, sifting vast quanta of information and content algorithmically to provide a user with the system's notion of what they are looking for can end up having profound effects on upstream content creation, downstream content ingestion, and the wider society in which both take place.

Efforts to address the threats posed by generative AI risk resorting to censorious or chilling tactics.[120] This may happen deliberately, with governments censoring how people can use generative AI or exploiting widespread anxiety about the threat it could pose as an excuse to impose new restrictions on expression. Or censorship could be an unintended side effect of well-intentioned efforts to detect AI-generated content or to keep chatbots from spewing hateful and potentially harmful responses to users.

The importance of knowing how to identify AI-generated content and how to debunk AI-generated misinformation is clear, but the best way to do it is not.[121] Some companies have created tools to detect artificially-generated content, including deepfakes (e.g., Sensity.AI), plagiarism (Originality.AI and Ficitious.AI),

---

116 "Languages used on the internet," Wikipedia entry (accessed July 2023) en.wikipedia.org/wiki/Languages_used_on_the_Internet ; Russell Brandom, "What languages dominate the internet?" *Rest of World*, June 7, 2023, restofworld.org/2023/internet-most-used-languages/.

117 Macrina Wang, "NewsGuard Exclusive: ChatGPT-3.5 Generates More Disinformation in Chinese than in English," NewsGuard (April 26, 2023), newsguardtech.com/special-reports/chatgpt-generates-disinformation-chinese-vs-english/

118 Ibid.

119 For a discussion of what the Chinese-language ChatGPT corpus may consist of, and potential limitations on a range of Chinese-language sources, see: Mu Chen, "ChatGPT doesn't understand Chinese well. Is there hope?" *Baiguan*, March 30, 2023, baiguan.news/p/chatgpt-doesnt-understand-chinese

120 While AI systems can facilitate censorship, they can also be used to resist it. One such development is Geneva, an AI-driven, censorship-evasion tool developed at the University of Maryland. Geneva is an algorithm that can defeat the mechanisms used to filter traffic for purposes of censorship. It searches for new censorship evasion strategies, learns which work and, over time, builds a repository of censorship and circumvention techniques. While Geneva cannot circumvent blocking of IP addresses and has some limitations, it has "discovered new ways of circumventing censorship in China, India, Iran, and Kazakhstan," and is being incorporated into other circumvention resources. Geneva (homepage, genetic algorithm), Breakerspace Lab at the University of Maryland, (accessed July 2023) geneva.cs.umd.edu/ ; "New artificial intelligence system automatically evolves to evade internet censorship," University of Maryland via *Science News*, November 13, 2019 sciencedaily.com/releases/2019/11/191113124822.htm.

121 Some experts are advocating for standards that will apply "traceable credentials" to digital work upon creation, instead of trying to detect it after the fact. See: Tiffany Hsu and Steven Lee Myers, "Another Side of the A.I. Boom: Detecting What A.I. Makes," *The New York Times*, May 18, 2023, nytimes.com/2023/05/18/technology/ai-chat-gpt-detection-tools.html

and AI-generated photos (Optic's AI or Not).[122] But detection tools are inevitably a step behind generative technology, which means they are often inaccurate. A detection tool called GPTZero, for example, identified sections of the U.S. Constitution and the Bible as "AI-generated."[123] A recent *New York Times* article said of a detection tool developed by OpenAI that it is "burdened with common flaws in detection programs: It struggles with short texts and writing that is not in English. In educational settings, plagiarism-detection tools such as TurnItIn have been accused of inaccurately classifying essays written by students as being generated by chatbots."[124] Using unreliable detection tools that falsely identify original material as AI-generated risks creating even more confusion about how to identify authenticity and could lead to content generated by humans being inaccurately stigmatized or silenced.

Many governments are rushing to determine what controls or regulations they need to respond to the widespread use of generative AI.[125] The head of the U.S. Federal Trade Commission (FTC) said the government "will not hesitate to crack down" on businesses that violate civil rights laws by using generative AI to deceive consumers or engage in discriminatory hiring practices.[126] In July 2023, the FTC launched an investigation into OpenAI, telling the company it would assess whether the company has engaged in "unfair or deceptive practices" with regard to data protection and potential harm, including reputational harm to consumers.[127] Efforts like these—aimed at safeguarding rights by enforcing existing laws on new technologies—pose limited risk of government overreach. The introduction of new regulations even in democratic countries should be done with care, however, recognizing that a good deal of trial and error may be necessary to address this fast-moving technology and its ramifications.

By contrast, China's government is extending its existing censorship regime to encompass generative AI tools, attempting to ensure their outputs stick to the CCP's preferred script. In July 2023, the country's top internet regulator, the Cyberspace Administration of China (CAC), finalized Measures for the Management of Generative Artificial Intelligence Services, a set of rules that set strict limits on how generative AI tools may be used. These include requirements of respect for intellectual property rights, prevention of discrimination, and mandatory security assessments. They also enforce ideological constraints, requiring content to "embody the Core Socialist Values" and that it not reflect "subversion of national sovereignty" or "content that might disrupt the economic or social order."[128] The new rules maintain the government's existing system of putting the onus of compliance on providers. Shortly after their release, however, the government announced the arrest of a citizen for using ChatGPT, which remains unavailable in China, to generate a false story about a train accident.[129]

---

122 Sensity homepage (accessed July 21, 2023) sensity.ai/; Originality homepage (accessed July 21, 2023) originality.ai/ ; Fictitious homepage (accessed July 21, 2023) fictitious.ai/ ; AI or Not homepage (accessed July 21, 2023) aiornot.com/

123 Benjamin Edwards, "Why AI Detectors Think the US Constitution Was Written by AI." Ars Technica. (July 14, 2023), arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai/.

124 Tiffany Hsu, and Steven Lee Myers, "Another Side of the AI Boom: Detecting What AI Makes." *The New York Times*, (May 19, 2023), nytimes.com/2023/05/18/technology/ai-chat-gpt-detection-tools.html.

125 This is explored further in the policy section below.

126 Matt O'Brien, "US officials seek to crack down on harmful AI product," *Associated Press*, April 25, 2023, apnews.com/article/artificial-intelligence-ai-tools-ftc-regulators-crackdown-lina-khan-0f63f6a9ec4e7c4acc37a2c1bd8c280f

127 Cat Zakrzewski, "FTC investigates OpenAI over data leak and ChatGPT's inaccuracy," *The Washington Post,* July 13, 2023, washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/; Federal Trade Commission (FTC) Civil Investigative Demand (CID) Schedule, FTC File No. 232-3044, washingtonpost.com/documents/67a7081c-c770-4f05-a39e-9d02117e50e8.pdf?itid=lk_inline_manual_4.

128 "Comparison Chart of Current vs. Draft rules for Generative AI," China Law Translate (July 13, 2023) chinalawtranslate.com/en/comparison-chart-of-current-vs-draft-rules-for-generative-ai/

129 Diego Mendoza, "China makes first known arrest over using ChatGPT to spread fake news, *Semafor,* May 8, 2023, semafor.com/article/05/08/2023/china-arrest-chatgpt-fake-news

The major corporations behind the most prominent generative AI tools—including OpenAI, Google, and Microsoft—are putting in place their own usage policies to mitigate the risk of liability and reputational damage that could result from novel and unpredictable technologies. Social media has shown the limitations of usage policies, the need to constantly evolve and update such standards, and the fierce blowback that can ensue when negligent policy making or implementation results in harms to users or violations of the law.[130] While financial incentives might operate differently in the generative AI space than they do with regard to social media, they will still drive the companies' decision-making. Google, for example, has every incentive to maintain the reliability and credibility of its search results. Its Search Generative Experience is designed to be less "creative" than Bard, its chatbot, and to respond only to certain types of queries.[131]

OpenAI's usage policies include a litany of forbidden uses of ChatGPT, including "fraudulent or deceptive activity," alongside everything from gambling and weapons development to adult content creation.[132] In addition to the company's existing terms of service, Google's generative AI prohibited use policy groups all barred activities under "dangerous, illegal, or malicious activity," "content intended to misinform, misrepresent, or mislead," and "sexually explicit content."[133] Microsoft supplements the code of conduct in its services agreement with additional provisions for its Bing chat and image creator, which state the user must not generate content that is illegal, harmful, or fraudulent.[134] The degree to which users can and will be held to these agreements, the consequences for violations, and the potential for recourse are all largely untested. The experience of social media suggests that as generative AI tools are more widely adopted, the sheer volume of users will pose serious challenges for companies trying to police policies at a global scale.

Other generative AI chatbots in development are designed specifically to have no limits on their usage. The New York Times recently reported that groups of volunteer programmers have developed new chatbots that are intentionally "uncensored."[135] The creators behind some of these tools argue that nothing, or very little, should be off limits, since chatbot-generated content will not necessarily be seen by anyone other than its user. A co-founder of Open Assistant, an independent chatbot released in April, suggests that social media platforms are responsible for policing AI-generated content because that is where it is likely to be disseminated.[136]

The question of limits on chatbot usage raises several fundamental questions about freedom of expression and generative AI. Can reasonable guidelines be put in place to protect the safety of users and others, without constraining the use of generative AI for expressive and creative purposes? What new safeguards do social media companies, and other platforms by which AI-generated content can be distributed, need to manage that content? Given the poor track record of digital platforms to date, can they avoid the persistent risks of under-and over-moderation? How can we identify and apply the relevant lessons learned from grappling with the impact of social media to the challenges of generative AI?

---

130 Victor Tangermann, "Bing AI Responds After Trying to Break Up Writer's Marriage," *Futurism,* February 16, 2023, futurism.com/the-byte/bing-ai-responds-marriage ; Stephen Marche, "The Chatbot Problem," *The New Yorker,* July 23, 2021, newyorker.com/culture/cultural-comment/the-chatbot-problem

131 Elizabeth Reid, "Supercharging Search with generative AI," *The Keyword* (Google blog), May 10, 2023, blog.google/products/search/generative-ai-search/; Gerrit De Vynck, "ChatGPT 'hallucinates.' Some researchers worry it isn't fixable," *The Washington Post,* May 30, 2023, washingtonpost.com/technology/2023/05/30/ai-chatbots-chatgpt-bard-trustworthy/

132 Usage policies, OpenAI (accessed July 21, 2023) openai.com/policies/usage-policies

133 Generative AI Prohibited Use Policy, Google (accessed July 21, 2023) policies.google.com/u/1/terms/generative-ai/use-policy

134 Code of Conduct, Bing (accessed July 21, 2023) bing.com/new/termsofuse#content-policy

135 Stuart A. Thompson, "Uncensored Chatbots Provoke a Fracas Over Free Speech," *The New York Times,* July 2, 2023, nytimes.com/2023/07/02/technology/ai-chatbots-misinformation-free-speech.html

136 Ibid.

BIAS & INFLUENCE

All algorithms reflect the biases and predispositions of their creators and the information upon which they draw. There are, however, unique and subtle ways in which generative AI can create and reproduce bias, with a potential chilling effect on expression.[137] A website that uses algorithms to curate content, for example, might inadvertently highlight more white, male writers if the algorithm itself was trained on a data set that skews toward white, male writers and includes fewer writers of color or female writers. A 2021 study conducted by researchers at UC Berkeley found that stories generated using GPT-3 "tend to include more masculine characters than feminine ones (mirroring a similar tendency in books), and identical prompts can lead to topics and descriptions that follow social stereotypes, depending on the prompt character's gender."[138] Such tendencies could reproduce systemic societal biases and inequalities, potentially reinforcing existing disparities in representation.

As with content moderation of social media, failure to comprehend linguistic nuances and subtleties may lead to over enforcement of the rules put in place to moderate generative AI. Researchers at Emory University have shown that when AI tools are used for content moderation, they can over-censor certain words; in particular, they highlighted censorship of "reclaimed'" words, like "queer," that could be a slur in some contexts, but perfectly acceptable in others.[139] Social media companies have experienced difficulty when training automated moderation tools to reflect such subtleties. Depending on how generative AI systems follow their own rules, they might circumvent words like "queer" altogether to avoid generating language that could be considered hateful. As a result, the use of generative AI in creative fields could produce works that are less rich or reflective of the expansive nuances of human experience and expression.

Generative AI tools can also affect the user's worldview. A recent study found that using a generative AI tool that exhibits bias to assist with writing can influence the opinions of the user.[140] In the study, "people who used an AI writing assistant that was biased for or against social media were twice as likely to write a paragraph agreeing with the assistant, and significantly more likely to say they held the same opinion, compared with people who wrote without AI's help."[141] During a May 2023 hearing of the U.S. Senate subcommittee on privacy, technology, and the law, NYU professor emeritus of psychology and neural science Gary Marcus warned about the threat of a "datocracy, the opposite of democracy," where "chatbots can clandestinely shape our opinions, in subtle yet potent ways, potentially exceeding what social media can do."[142] Researchers warn this is an underappreciated implication of the steps we are taking to "embed [these technologies] in the social fabric of our societies."[143]

These findings suggest that generative AI tools could be wielded—or weaponized—to manipulate opinions and skew public discourse via subtle forms of influence on their users. AI chatbots designed to reflect a particular ideology could further entrench existing cultural and political echo chambers.[144]

---

137 Jeremy Baum, John Villasenor, "The politics of AI: ChatGPT and political bias," Brookings (May 8, 2023) brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/

138 Li Lucy and David Bamman, "Gender and Representation Bias in GPT-3 Generated Stories," University of California Berkeley (June 11, 2021) aclanthology.org/2021.nuse-1.5.pdf

139 Aine Doris, "Is AI Censoring Us?" Emory Business (June 9, 2023) emorybusiness.com/2023/06/09/is-ai-censoring-us/

140 Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, Mor Naaman, "Co-Writing with Opinionated Language Models Affects Users' Views," Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (1–5) (April 2023) dl.acm.org/doi/10.1145/3544548.3581196

141 Patricia Waldron, "Writing with AI help can shift your opinions," *Cornell Chronicle*, May 15, 2023, news.cornell.edu/stories/2023/05/writing-ai-help-can-shift-your-opinions

142 Gary Marcus, Senate Testimony (May 16, 2023) judiciary.senate.gov/imo/media/doc/2023-05-16%20-%20Testimony%20-%20Marcus.pdf

143 Christopher Mims, "Help! My Political Beliefs Were Altered by a Chatbot!" *The Wall Street Journal*, May 13, 2023, wsj.com/articles/chatgpt-bard-bing-ai-political-beliefs-151a0fe4

144 David Rozado, "The Political Biases of ChatGPT," MDPI (March 2, 2023) mdpi.com/2076-0760/12/3/148 ; Stuart A. Thompson, Tiffany Hsu, and Steven Lee Myers, "Conservatives Aim to Build a Chatbot of Their Own," *The New York Times*, March 22, 2023, nytimes.com/2023/03/22/business/media/ai-chatbots-right-wing-conservative.html

To some degree, this is already happening. While companies like OpenAI, Google, and Microsoft say they are working to make their chatbots more reliable and to reduce bias, eliminating algorithmic bias is impossible. Some studies suggest that ChatGPT, at least, does reflect a liberal bias.[145] A January 2023 study by researchers in Germany found "converging evidence for ChatGPT's pro-environmental, left-libertarian orientation," and in May 2023, Brookings researchers testing ChatGPT found that it "provided consistent—and often left-leaning—answers on political/social issues."[146] Some critics have responded by calling for new, alternative chatbots. In April, Elon Musk told Tucker Carlson he would develop "TruthGPT," which would be "truth seeking," as opposed to the "politically correct" ChatGPT and Bard.[147] David Rozado, a researcher based in New Zealand who has studied ChatGPT's political leanings, used the tool to create an AI model called RightWingGPT that reflected conservative political views (the model has not been released).[148] He now intends to build LeftWingGPT and DePolarizingGPT, and to release all three, which he says are trained on "thoughtful authors (not provocateurs)."[149]

U.S. political culture already precludes broad public agreement on facts or the notion of truth, so it's unlikely that any generative AI tool would appear unbiased and credible to everyone. The creation of ideologically-oriented chatbots could further reinforce and harden the fronts in the culture war, undercut trust even in the most constructive uses of generative AI, and make it even more difficult for the public to distinguish truth from falsehood, or to place trust in any information source.

PART III: POLICY CONSIDERATIONS AND RECOMMENDATIONS

The introduction of ChatGPT and the rise of generative AI tools prompted a wave of regulatory interest in the spring of 2023, the ramifications of which are still evolving.

Efforts to develop guiding principles, blueprints, and frameworks for the regulation of AI technologies are not new, however. The global push for AI-specific laws began in earnest in 2016; by December 2022, 123 AI-related bills had been passed by the legislative bodies of 127 countries, according to the Stanford Institute for Human-Centered Artificial Intelligence.[150] In the United States, bipartisan AI caucuses in the House and

145 Christopher Mims, "Help! My Political Beliefs Were Altered by a Chatbot!" *The Wall Street Journal,* May 13, 2023, wsj.com/articles/chatgpt-bard-bing-ai-political-beliefs-151a0fe4 ; David Rozado, "The Political Biases of ChatGPT," MDPI (March 2, 2023) mdpi.com/2076-0760/12/3/148

146 Jochen Hartmann, Jasper Schwenzow, Maximillian Witte, "The political ideology of conversational AI: Converging evidence on Chat GPT's pro-environmental, left-libertarian orientation," Cornell University (January 5, 2023) arxiv.org/abs/2301.01768; Jeremy Baum, John Villasenor, "The politics of AI: ChatGPT and political bias," Brookings (May 8, 2023) brookings.edu/articles/the-politics-of-ai-chatgpt-and-political-bias/

147 Megan Sauer, "Elon Musk now says he wants to create a ChatGPT competitor to avoid 'A.I. dystopia'—he's calling it 'TruthGPT'," *CNBC,* April 19, 2023, cnbc.com/2023/04/19/elon-musk-says-he-wants-to-create-chatgpt-competitor-called-truthgpt.html

148 David Rozado, "The Political Biases of ChatGPT," MDPI (March 2, 2023) mdpi.com/2076-0760/12/3/148 ; Stuart A. Thompson, Tiffany Hsu, and Steven Lee Myers, "Conservatives Aim to Build a Chatbot of Their Own," *The New York Times*, March 22, 2023 nytimes.com/2023/03/22/business/media/ai-chatbots-right-wing-conservative.html

149 Will Knight, "Meet ChatGPT's Right-Wing Alter Ego," *Wired,* April 27, 2023, wired.com/story/fast-forward-meet-chatgpts-right-wing-alter-ego/

150 "Chapter 6: Artificial Intelligence Index Report" (2023) aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report-2023_CHAPTER_6-1.pdf

Senate date to 2017 and 2019, respectively.[151] In the absence of federal regulation, states have attempted to fill the void, introducing at least 58 pieces of legislation on "AI issues generally" in 2022, according to the National Conference of State Legislatures.[152] The NCSL notes that this count does not include bills that address "specific AI technologies, such as facial recognition or autonomous cars," which raise the overall tally.[153] In October 2022, the Biden Administration launched the *Blueprint for an AI Bill of Rights*. The Administration continues to explore and pursue additional measures to regulate the field of AI; in July 2023, the White House secured voluntary commitments from seven technology companies concerning safety, security, and trust with regard to AI.[154]

In the multilateral sphere, the Organization for Economic Co-operation and Development (OECD) adopted a set of "value-based" principles outlined in the 2019 Recommendation of the Council on Artificial Intelligence.[155] The OECD's intent was to identify a set of international standards that "aim to ensure AI systems are designed to be robust, safe, fair and trustworthy."[156] The principles served as the foundation for the G20 Principles on AI[157] (also promulgated in 2019) and include:

a. inclusive growth, sustainable development, and well-being[158]
b. human-centered values and fairness[159]
c. transparency and explainability
d. robustness, security, and safety
e. accountability for AI actors.[160]

---

151 "Delaney Launches Bipartisan Artificial Intelligence (AI) Caucus for 115th Congress" (Just Facts) (May 24, 2017) justfacts.votesmart.org/public-statement/1190092/delaney-launches-bipartisan-artificial-intelligence-ai-caucus-for-115th-congress; Martin Heinreich, "Heinrich, Portman Launch Bipartisan Artificial Intelligence Caucus" (Martin Heinrich blog) (March 13, 2019) heinrich.senate.gov/newsroom/press-releases/heinrich-portman-launch-bipartisan-artificial-intelligence-caucus

152 "Legislation Related to Artificial Intelligence," National Conference of State Legislatures (August 26, 2022) ncsl.org/technology-and-communication/legislation-related-to-artificial-intelligence

153 Ibid.

154 *Blueprint For An AI Bill of Rights*, The White House (October 2022) whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf; Josh Boak, "Biden is calling in all the AI bigwigs to debate the future of tech regulation as his White House races to create a roadmap," Fortune, June 20, 2023, fortune.com/2023/06/20/ai-regulation-chatgpt-joe-biden-white-house-tech-executives-artificial-intelligence/#; Andrew Zhang, "Biden staff are meeting regularly to develop AI strategy, White House says," Politico, June 20, 2023, politico.com/news/2023/06/20/biden-ai-regulatory-strategy-00102753; The White House "FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI." The White House, (July 2023), whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

155 OECD AI Principles overview (OECD.AI) (accessed July 21, 2023) oecd.ai/en/ai-principles; Recommendation of the Council on Artificial Intelligence, OECD Legal Instruments (May 21, 2019) legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

156 "Forty-two countries adopt new OECD Principles on Artificial Intelligence," OECD (May 22, 2019) oecd.org/science/forty-two-countries-adopt-new-oecd-principles-on-artificial-intelligence.htm

157 G20 Ministerial Statement on Trade and Digital Economy (June 8 and 9, 2019) wp.oecd.ai/app/uploads/2021/06/G20-AI-Principles.pdf

158 Calling for the "responsible stewardship of trustworthy AI in pursuit of beneficial outcomes for people and the planet." Inclusive growth, sustainable development and well-being (Principle 1.1) (OECD.AI) (accessed July 21, 2023) oecd.ai/en/dashboards/ai-principles/P5

159 Calling for "respect for the rule of law, human rights and democratic values" including "freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, fairness, social justice, and internationally recognised (sic) labour (sic) rights" and the implementation of "mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art." Human-centred values and fairness (Principle 1.2) OECD.AI (accessed July 21, 2023) oecd.ai/en/dashboards/ai-principles/P6

160 "Recommendations of the Council on Artificial Intelligence" (OECD Legal Instruments) (May 21, 2019) legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449; the OECD published additional recommendations specific to governments, encouraging them to (1) invest in AI research and development, (2) foster a digital ecosystem for AI, (3) shape an enabling environment for AI, (4) build human capacity and prepare for workforce transformations, and (5) actively cooperate with international stakeholders.

In its recommendation the OECD points to the need to respect human rights in developing AI governance but does not mention freedom of expression explicitly.

The Biden Administration's *Blueprint for an AI Bill of Rights* sets forth five principles to guide the development and deployment of AI systems to protect Americans' rights.[161] The Blueprint is not legally binding but provides guidance for the responsible use of automated systems across sectors, supplementing existing law and policy. The principles are:

a. safe and effective AI systems, beginning with development and design and continuing through to implementation of the system
b. algorithmic discrimination protections, such that users do not face discrimination or differential treatment by an AI system based on any classification protected by law
c. data privacy and protection from abusive data collection and use practices
d. notice and explanation of the system at work and its role in any outcome affecting the user
e. a human alternative to an automated system, such as the opportunity to opt out of automated system interaction and seek human involvement where appropriate.

Guidance accompanying the *Blueprint* calls for the application of its protections where automated systems have the potential for meaningful impact upon the exercise of civil rights, including the right to free speech.

The U.K. likewise set out AI guiding principles outlined in a March 2023 white paper titled "AI regulation: a pro-innovation approach."[162] This approach is described in part as one that is meant to allow responsible AI to "flourish" while building public trust.[163] As with the OECD and the *Blueprint for an AI Bill of Rights*, the U.K.'s approach to AI is guided by a set of principles that cuts across sectors:

a. safety, security and robustness
b. appropriate transparency and explainability
c. fairness
d. accountability and governance
e. contestability and redress

As explained in Annex B of the document, human rights are not specifically enumerated in the principles due to the expectation that the principles will be implemented with adherence to existing laws.

In the United States, the emergence of public-facing generative AI tools has spurred legislators to action. In June 2023, Senate Majority Leader Chuck Schumer announced an AI-focused initiative, the SAFE Innovation Framework (for "security, accountability, protecting our foundations, and explainability"), meant to guide any comprehensive AI-focused legislation.[164] In furtherance of this vision, the Senate will hold nine "Insight Forums" to assess pathways to regulation.[165]

---

161 *Blueprint For An AI Bill of Rights*, The White House (October 2022) whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

162 *AI regulation: a pro-innovation approach*, U.K. Government Department for Science, Innovation and Technology and Office for Artificial Intelligence (white paper) (March 29, 2023) gov.uk/government/publications/ai-regulation-a-pro-innovation-approach

163 Ibid.

164 "Majority Leader Schumer Delivers Remarks To Launch SAFE Innovation Framework for Artificial Intelligence at CSIS," Senate Democrats (June 21, 2023), democrats.senate.gov/news/press-releases/majority-leader-schumer-delivers-remarks-to-launch-safe-innovation-framework-for-artificial-intelligence-at-csis. Leader Schumer has noted the critical need for Congress to "join the AI revolution," suggesting that Congress is "starting from scratch" in seeking to regulate AI. While regulation of generative AI is yet to be fulsomely addressed by the U.S. government, Congress's efforts should be informed by the White House's *Blueprint*, the NIST's *AI Risk Management Framework*, and statements by AI experts in industry and civil society as to rights-based policy framing. https://thehill.com/opinion/technology/4076112-the-senate-doesnt-need-to-start-from-scratch-on-ai-legislation/; https://facctconference.org/2023/harm-policy.html

165 Axios, "Schumer, Humbled by AI, Crafts Crash Course for Senate," (July 2023), axios.com/2023/07/18/schumer-ai-forums-senate.

The SAFE Innovation framework was announced shortly after the European Parliament adopted its negotiating position on the Artificial Intelligence Act (AI Act), which, if adopted, would be among the first comprehensive regulations governing AI.[166] The Act aims to cover the lifecycle of an AI system and the content or decisions shaped by the system. It sets out transparency obligations that apply both to the AI systems themselves and to content they generate, for example, requiring those who use an AI system to create deepfakes to "disclose that the content has been artificially generated or manipulated."[167] The AI Act takes a risk-based approach, differentiating between AI applications that pose low, high, or unacceptable risk.[168] Some assessments of the Act have criticized the inflexibility of its categorization of high-risk and restricted systems, noting that legislative or regulatory parameters regarding generative AI must allow for iteration based on the constantly evolving understanding of the technology's risks and benefits.[169]

Regulatory guidance and proposals written prior to 2023 generally do not mention generative AI. This shows how recently it has entered the public consciousness and illustrates how policy responses will continue to lag behind advances in AI. Neither the OECD Recommendation nor the White House Blueprint addresses generative AI specifically. Still, the scope of both documents is broad enough to encompass generative AI, demonstrating the value of flexibility and adaptability in such frameworks.[170]

In addition to specific regulatory proposals, some governmental and intergovernmental efforts have focused on gathering civil society and expert feedback via formal mechanisms for public comment and other means of input. For example, the European Commission solicited input on the EU AI Act, starting with a public consultation period in the first half of 2021.[171] In the United States the National Telecommunications and Information Administration, the Patent and Trademark Office, and the Office of Science and Technology Policy all opened up public requests for comment in the first half of 2023.[172] The Biden-Harris Administration convened leaders in the consumer protection, labor, and civil rights spheres for insight while developing principles on safety, security, and trust for industry.[173] Both the UN and the OECD engaged external experts and stakeholders as part of their processes to develop the Global Digital Compact and AI Principles, respectively. Taking a multi-pronged approach to stakeholder engagement, particularly in the realm of emerging technologies, is critical to informing policymakers and offers a useful model for larger standards-setting efforts.

---

166 "What is the EU AI Act?" (The Artificial Intelligence Act) (accessed July 21, 2023) artificialintelligenceact.eu/

167 "Proposal for a Regulation of the European Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislation Acts," (2021), artificialintelligenceact.eu/the-act/.

168 High-risk systems are listed in Annex III to the Act. https://artificialintelligenceact.eu/annexes/

169 Feedback from: University of Cambridge (Leverhulme Centre for the Future of Intelligence and Centre for the Study of Existential Risk), European Commission (August 6, 2021) ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665626_en

170 The *Blueprint* states that it applies to: "(1) automated systems that (2) have the potential to meaningfully impact the American public's rights, opportunities, or access to critical resources or services." *Blueprint For An AI Bill of Rights*, The White House (October 2022) whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

171 "Artificial intelligence – ethical and legal requirements," European Commission (accessed July 21, 2023) ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements_en

172 "AI Accountability Policy Request for Comment," *Federal Register,* April 13, 2023 federalregister.gov/documents/2023/04/13/2023-07776/ai-accountability-policy-request-for-comment ; "Request for Comments Regarding Artificial Intelligence and Inventorship," *Federal Register*, February 14, 2023, federalregister.gov/documents/2023/02/14/2023-03066/request-for-comments-regarding-artificial-intelligence-and-inventorship

173 "Readout of Vice President Harris's Meeting with Consumer Protection, Labor, and Civil Rights Leaders on AI," The White House (July 13, 2023), whitehouse.gov/briefing-room/statements-releases/2023/07/13/readout-of-vice-president-harriss-meeting-with-consumer-protection-labor-and-civil-rights-leaders-on-ai/.

Given the inherently borderless nature of generative AI technologies, international cooperation and the development of rights-based multilateral frameworks are also paramount.[174] In a 2021 report on the dangers of digital sovereignty, PEN America highlighted the need for "a new model of democratic multilateralism for internet governance, driven by a coalition of established democracies."[175] The Declaration for the Future of the Internet, released following the first Summit for Democracy organized by the United States in 2021, has amassed nearly 70 signatories, signaling those countries' commitment to a future internet that is "open, free, global, interoperable, reliable, and secure."[176] With principles including protection of human rights and fundamental freedoms, the use of technology to promote freedom of expression, the ability to connect to the Internet and a secure, sustainable infrastructure, protection of privacy, and a commitment to multistakeholder internet governance, the Declaration offers a useful starting point for the U.S. and its partners in advancing democratic responses to evolving digital technologies.[177]

RECOMMENDATIONS AND GUIDING PRINCIPLES FOR AI GOVERNANCE AND POLICYMAKING

Many of the risks raised in this paper are not new, but policymakers and industry have been slow to address them effectively. As outlined above, generative AI tools are likely to supercharge threats like disinformation and online abuse, and risk inspiring responses that could overregulate expression.

Much of the debate regarding how best to approach the regulation of AI technologies falls into one of two camps: A rights-based approach, such as the Biden Administration's Blueprint for an AI Bill of Rights, or a risk-based approach, such as the EU's AI Act. This dichotomy, while a useful heuristic, is a false one that could undermine otherwise thoughtful attempts at regulation. The rights vs. risk framing is unnecessary and some of those documents that purport to fall into one camp or the other are, in fact, both. The EU's AI Act, for example, says that it "follows a risk-based approach" and yet is "based on EU values and fundamental rights." The dichotomy is inherently confusing and misleading. Policymaking and regulation must instead be deliberately both rights-respecting and risk-aware. With that in mind, PEN America proposes the following recommendations for AI governance and policymaking in government and industry:

Recommendations for Government:

- **Pass long overdue, foundational legislation:** As a starting point, PEN America recommends that U.S. legislators take action on long overdue legislation that would be foundational to responsible regulation of AI, in addition to other tech policy-focused efforts. The White House *Blueprint* highlights the degree to which Congress has failed to lay the legislative foundation for enacting the principles called for in areas like data privacy, data collection and use, researcher access, and algorithmic transparency. If legislators pass comprehensive privacy legislation and the Platform Accountability and Transparency Act,[178] the

174 Scientists, AI experts, and academics-including Gary Marcus, Anka Reuel, and Rumman Chowdhury-have posited that AI governance is ripe for broader, independent oversight, via a well-resourced international agency: "The world needs an international agency for artificial intelligence, say two AI experts," The Economist, April 18, 2023, economist.com/by-invitation/2023/04/18/the-world-needs-an-international-agency-for-artificial-intelligence-say-two-ai-experts ; Rumman Chowdhury, "AI desperately Needs Global Oversight," Berkman Klein Center at Harvard University (April 11, 2023) cyber.harvard.edu/story/2023-04/ai-desperately-needs-global-oversight

175 "Introduction: The New Guard Posts of Cyberspace," PEN America (report) (accessed July 21, 2023) pen.org/report/splintered-speech-digital-sovereignty-and-the-future-of-the-internet/

176 "A Declaration for the Future of the Internet," The White House (accessed July 21, 2023) Declaration-for-the-Future-for-the-Internet_Launch-Event-Signing-Version_FINAL.pdf

177 "Declaration for the Future of the Internet." n.d. GMFUS, gmfus.org/news/declaration-future-internet.

178 Platform Accountability and Transparency Act, S.1876, 118th Cong. (Introduced June 8, 2023) www.congress.gov/bill/118th-congress/senate-bill/1876/text

United States will be better-positioned to provide more targeted solutions to AI's potential ills in ways that do not risk infringing on freedom of expression.

- **Establish and maintain multi-stakeholder policymaking processes:** Finding solutions that do not inadvertently shut down speech or inhibit creativity and innovation will require prioritizing early and ongoing input from a diverse set of stakeholders. The voluntary commitments secured by the White House from leading technology companies, which were informed in part by consultations with civil society leaders, offer one example of an inclusive approach to policy making. The "Insight Forums" planned by the Senate present an opportunity for experts to address the rights-based challenges at issue, in addition to the initial topics of copyright, workforce, national security, high risk AI models, existential risks, privacy, transparency and explainability, and elections and democracy. Officials should continue to consult with human rights advocates, scientists, academics, and other experts to understand how to craft workable policy solutions and to ensure proposed regulations will not undermine free expression, speech, creativity, and innovation. Consultations must also include free expression experts and the writers, artists, and journalists who are directly affected both by advancements in generative AI and potential regulatory responses. Engagement with civil society must also take into account the potential global impact of laws and policies, particularly those enacted in the United States and the E.U. Policymakers should establish and maintain formal systems for ongoing input and oversight from civil society. Existing models that bring government, industry, civil society, and academic stakeholders together, such as those aimed at bolstering platform accountability, can serve as models for similar AI governance mechanisms.

- **Ground regulatory frameworks in fundamental rights:** Any regulatory framework set forth or brought to bear on generative AI must be predicated on fundamental human rights, particularly the right to free expression, which enables the free exercise of other fundamental rights. Policies that affect internet users' rights should also be fact-based and grounded in research, when possible.

- **Engage in policymaking that is measured and iterative:** Anxieties regarding generative AI offer authoritarian governments a pretense to impose additional restrictions on expression, including the introduction of censorious laws. Yet recent state and regional regulatory and legislative efforts demonstrate the risks to free speech and expression even by well-intended democratic governments seeking to stem the known and potential harms of new technologies. Policymakers should avoid rushing to solutions that might undermine free expression and other rights, establish a worrying precedent, or create a dubious foundation for iteration as additional technologies emerge.

- **Build flexibility into regulatory schemes:** With emerging technologies, allowing for iteration and flexibility is critical to ensuring workable solutions. Any approaches to regulating current and future iterations of generative AI technology should seek to responsibly mitigate known risks and allow for the addressing of new ones. Rather than being fixed in perpetuity or requiring elusive consensus to update, regulations should be subject to regular review and adaptation to respond to technological change and encompass learnings from what will unavoidably be a period of trial and error.

- **Emphasize and operationalize transparency:** Gaps in transparency and independent analysis can impair the quest for solutions to the potential harms of AI technologies. Regulators should seek to ensure transparency and access for researchers to algorithms, data sources and uses, and other mechanics of AI technologies. The results of government-mandated algorithmic audits and assessments should be made publicly available.

Recommendations for Industry:

- **Promote fair and equitable use:** By prioritizing fairness and equity throughout the development and deployment of an AI model, industry can reduce bias and build more trustworthy systems. Companies can advance these priorities by ensuring AI models are designed and built by diverse teams, being deliberate and thoughtful about training data sets, and by engaging with relevant external stakeholders throughout the development process. AI models in use should also be explainable. Non-experts should be able to understand how and why a model operates as it does, and upon what inputs it relies. Building fair and equitable systems could entail supporting or conducting research, soliciting public comment, re-aligning internal priorities, or even putting model deployment on hold until additional refinements can ensure that benchmarks for fairness are met. As AI technologies go global, their developers should ensure that they have the linguistic fluency and cultural competency to operate responsibly. Developers should endeavor to create AI systems with these capabilities, and until then, must be cautious about the distribution and availability of their services.
- **Facilitate secure and privacy-protecting use:** Security and privacy should provide the foundation for AI system development and deployment. The scope of safe and secure practices is broad but can encompass regular audits or surveys to detect anomalies, protection against attacks by third parties, and ensuring that encryption benchmarks are met. Privacy-protecting practices might encompass development and implementation of a risk management framework, robust data minimization practices, and ensuring that users have adequate knowledge about and control over their data.
- **Emphasize and operationalize transparency:** Industry need not await government regulation to incorporate transparency into its practices. Developers should initiate, participate in, and publicly share the results of algorithmic audits and assessments should be made publicly available. They should also publish transparency reports, continue to invest in research, support initiatives to educate users, and develop and improve upon standards for rights-respecting AI systems that mitigate harm.
- **Provide appeals and remedy options:** When AI is used to automate decision-making, for example in content moderation, search engine results, or other cases, appeals and remedy options that are accessible and effective must also exist alongside automated processes.
- **Consider revenue models:** The business models that will drive the spread of generative AI are only now being invented and refined; as these mechanisms emerge and before they become entrenched, rigorous assessment of how they shape AI-driven content and discourse is essential. Revenue structures that reward harmful content need to be identified and disabled before they can further corrode the foundations of social and political life.
- **Safeguard the ownership rights of writers, artists, and other content owners:** Industry should take steps to safeguard the ownership rights of writers, artists, and other content owners whose work may form part of the training set for generative AI tools, including by seeking consent and ensuring credit and compensation for the use of copyrighted work. AI companies should explore the creation of collective licensing schemes that compensate content owners fully and fairly for their contribution to large language models, allowing for transparency, opt outs, and taking into account the growth and profitability of AI operations and their impact on existing forms of compensation that underwrite content creation. Recognizing the potential impact of generative AI on the revenue available to support content creation activities that generate public goods, including independent journalism and creative expression, companies must ensure that the adoption of AI does not destroy or undercut essential contributions to culture and the public square.

CONCLUSION

The power of and potential for artificial intelligence technologies to shape expressive conduct and content is at once apparent and not fully realized. The increasing prevalence of generative AI and automated tools represents a sea change in how artists, journalists, and writers create, interact with, and disseminate content, and how the public understands and consumes it. These changes offer opportunities for new forms of expression and creativity, while simultaneously posing threats to the exercise of free expression. As with social media, some of the potential negatives—false cures, deadly dares, provocations to violence—could have life-or-death consequences. Artificial intelligence technologies themselves are neither good nor bad. What matters is who uses them, how they are being used, and what stakeholders can do to shape a future in which new technologies support and enhance fundamental rights.

ACKNOWLEDGEMENTS

Lead author on this report was Summer Lopez, Chief Program Officer, Free Expression; with co-authorship by Nadine Farid Johnson, Managing Director, PEN America Washington and Free Expression Programs and Liz Woolery, Digital Policy Lead. PEN America would also like to thank the fellows whose research, fact-checking, and proofreading made this report possible: Pratika Katiyar and Rachel Hochhauser. The report was reviewed by PEN America's research team and other relevant PEN America experts. PEN America is deeply grateful to Deepak Kumar, postdoctoral researcher at Stanford University, and Paul Barrett, Adjunct Professor and Deputy Director, Stern Center for Business and Human Rights at New York University, for their expert review. The report was edited by Lisa Goldman.


PEN
AMERICA