epic.org

Electronic Privacy Information Center
1519 New Hampshire Avenue NW
Washington, DC 20036, USA

☎ +1 202 483 1140
🖶 +1 202 483 1248
🐦 @EPICPrivacy
🌐 https://epic.org

COMMENTS OF THE ELECTRONIC PRIVACY INFORMATION CENTER

to the

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

Request for Information Related to NIST's Assignments Under Sections 4.1, 4.5 and 11 of the Executive Order Concerning Artificial Intelligence

No. 2023-28232

February 2, 2024

TABLE OF CONTENTS

INTRODUCTION

The Electronic Privacy Information Center (EPIC) submits these comments in response to the National Institute of Standards and Technology's (NIST's) Request for Information (RFI) Related to NIST's Assignments Under Sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence.[1]

EPIC is a public interest research center in Washington, D.C., established in 1994 to secure the fundamental right to privacy in the digital age for all people through advocacy, research, and litigation.[2] We advocate for a human-rights-based approach to AI policy that ensures new technologies are subject to democratic governance.[3] Over the last decade, EPIC has consistently advocated for the adoption of clear, commonsense, and actionable AI regulations across the country.[4] EPIC has litigated cases against the U.S. Department of Justice to compel production of documents regarding "evidence-based risk assessment tools,"[5] against the U.S. Department of Homeland Security to produce documents about a program purported to assess the probability that an individual will commit a crime,[6] and against the National Security Commission on Artificial Intelligence (NSCAI) to enforce its transparency obligations under the Freedom of Information Act and the Federal Advisory Committee Act.[7] EPIC has also published extensive research on

[1] 88 Fed. Reg. 88368 (Dec. 21, 2023).
[2] *About Us*, EPIC, https://epic.org/about/ (2023).
[3] *See, e.g.*, *AI and Human Rights*, EPIC, https://epic.org/issues/ai/ (2023); *AI and Human Rights: Criminal Legal System*, EPIC, https://epic.org/issues/ai/ai-in-the-criminal-justice-system/ (2023); EPIC, Outsourced & Automated: How AI Companies Have Taken Over Government Decision-Making (2023), https://epic.org/outsourced-automated/ [hereinafter "Outsourced & Automated Report"]; Letter from EPIC to President Biden and Vice President Harris on Ensuring Adequate Federal Workforce and Resources for Effective AI Oversight (Oct. 24, 2023), https://epic.org/wp-content/uploads/2023/10/EPIC-letter-to-White-House-re-AI-workforce-and-resources-Oct-2023.pdf; EPIC, Comments on the NIST Artificial Intelligence Risk Management Framework: Second Draft (Sept. 28, 2022), https://epic.org/wp-content/uploads/2022/09/EPIC-Comments-NIST-RMF-09-28-22.pdf.
[4] *See, e.g.*, Press Release, EPIC, EPIC Urges DC Council to Pass Algorithmic Discrimination Bill (Sept. 23, 2022), https://epic.org/epic-urges-dc-council-to-pass-algorithmic-discrimination-bill/; EPIC, Comments to the Patent and Trademark Office on Intellectual Property Protection for Artificial Intelligence Innovation (Jan. 10, 2020), https://epic.org/wp-content/uploads/apa/comments/EPIC-USPTO-Jan2020.pdf; EPIC, Comments on the Department of Housing and Urban Development's Implementation of the Fair Housing Act's Disparate Impact Standard (Oct. 18, 2019), https://epic.org/wp-content/uploads/apa/comments/EPIC-HUD-Oct2019.pdf.
[5] *EPIC v. DOJ*, 320 F. Supp. 3d 110 (D.D.C. 2018), *voluntarily dismissed,* 2020 WL 1919646 (D.C. Cir. 2020), https://epic.org/foia/doj/criminal-justice-algorithms/.
[6] *See EPIC v. DHS – FAST Program*, EPIC, https://epic.org/documents/epic-v-dhs-fast-program/ (last visited Dec. 5, 2023).
[7] *EPIC v. NSCAI*, 419 F. Supp. 3d 82, 86, 95 (D.D.C. 2019), https://epic.org/documents/epic-v-ai-commission/.

emerging AI technologies like generative AI,[8] as well as the ways that government agencies develop, procure, and use AI systems around the country.[9]

As NIST considers ways to effectively carry out its responsibilities under Sections 4.1, 4.5, and 11 of Executive Order 14110, EPIC reemphasizes its call for NIST to implement actionable AI risk mitigation strategies with strong incentivize structures and accountability mechanisms—steps that will ensure that AI developers and deployers adopt the NIST AI Risk Management Framework ("AI RMF")[10] in its entirety.[11] At the same time, EPIC encourages NIST to view the risks of generative AI technologies and synthetic content as extensions of traditional AI and automated decision-making risks, not as qualitatively different risks requiring an entirely new framework. Many of the same AI risk management techniques at the core of NIST's AI RMF—including AI impact assessments,[12] regular AI accuracy testing,[13] and AI red-teaming efforts[14]—will be effective against the risks of generative AI technologies and the synthetic content they produce. Lastly, for certain risks of generative AI technologies, such as the risks imposed by AI hallucinations and deepfakes, the need for greater transparency, accountability, and data quality controls—including strong data minimization requirements during AI development—is even higher. To inform NIST's efforts in bolstering the AI RMF and establishing global consensus standards, EPIC has provided both a summary of key provisions under the draft European Union Artificial Intelligence Act (Section IV, *infra*) and EPIC's generative AI report (appended).

I. THE NIST AI RISK MANAGEMENT FRAMEWORK SHOULD APPLY TO ALL AI SYSTEMS, INCLUDING GENERATIVE AI

Responsive to Assignments 1–2

EPIC commends the ongoing efforts NIST has made to incorporate robust AI transparency, accountability, and oversight provisions into its AI RMF. In particular, EPIC supports NIST's decision to approach AI risk management broadly, encompassing both a wide array of AI risks

[8] EPIC, Generating Harms: Generative AI's Impact & Paths Forward (2023), https://epic.org/gai [hereinafter "EPIC Generative AI Report"].

[9] Outsourced & Automated Report; EPIC, Screened & Scored in the District of Columbia (2022), https://epic.org/wp-content/uploads/2022/11/EPIC-Screened-in-DC-Report.pdf [hereinafter "Screened & Scored Report"].

[10] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0) (2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf [hereinafter "NIST AI RMF"].

[11] *See* EPIC, Comments on the NIST Artificial Intelligence Risk Management Framework: Second Draft (Sept. 28, 2022), https://epic.org/wp-content/uploads/2022/09/EPIC-Comments-NIST-RMF-09-28-22.pdf.

[12] NIST AI RMF at 11, 36.

[13] *Id.* at 27–30, 35–36.

[14] NIST, AI RMF Playbook 31–32, 131, 200 (2023), https://airc.nist.gov/docs/AI_RMF_Playbook.pdf [hereinafter "NIST AI RMF Playbook"].

(errors, racial bias, environmental impacts, and more) and a wide range of AI actors (trade associations, researchers, end users, impacted individuals, and more).[15] And although EPIC has advocated for the inclusion of stronger accountability mechanisms and prohibitions on particularly egregious AI use cases like emotion recognition and one-to-many facial recognition,[16] NIST's AI RMF still includes several features—including AI impact assessments,[17] regular AI accuracy testing,[18] and AI red-teaming efforts[19]—that have informed the ways that EPIC approaches its own AI advocacy.

It is precisely because of the breadth of NIST's AI RMF that EPIC encourages NIST to extend existing provisions of the AI RMF to the risks and harms of generative AI technologies. All automated technologies—from simple algorithms to complex generative AI models—face significant accuracy and bias risks stemming from training data and data inputs.[20] While the specific form of errors due to inaccuracy and bias can differ between traditional automated systems and newer generative AI models,[21] the importance of data quality controls and AI testing procedures as ways to mitigate errors remains.[22] Further, many of the risk management strategies within the NIST AI RMF map cleanly onto the risks of generative AI. For example, the AI RMF states that "AI risk management efforts should consider that humans may assume that AI systems work—and work well—in *all settings*."[23] This framing extends not only to issues around perceived versus actual objectivity in AI systems compared to human decision-making, but also to perceived versus actual accuracy of synthetic content within leading foundation models. The risk

[15] *See id.* at 8–10, 12–18.
[16] *See* EPIC, Comments on the NIST Artificial Intelligence Risk Management Framework: Second Draft (Sept. 28, 2022), https://epic.org/wp-content/uploads/2022/09/EPIC-Comments-NIST-RMF-09-28-22.pdf.
[17] NIST AI RMF at 11, 36.
[18] *Id.* at 27–30, 35–36.
[19] NIST AI RMF Playbook at 31–32, 131, 200.
[20] *See, e.g.*, Oceane Duboust, *Unreliable Research Assistant? False Outputs from AI Chatbots Pose Risk to Science, Report Says,* Euronews (Nov. 20, 2023), https://www.euronews.com/next/2023/11/20/unreliable-research-assistant-false-outputs-from-ai-chatbots-pose-risk-to-science-report-s; Matt Burgess et al., *This Algorithm Could Ruin Your Life*, Wired (June 3, 2023), https://www.wired.co.uk/article/welfare-algorithms-discrimination; Stephanie Wykstra, *Government's Use of Algorithm Serves Up False Fraud Charges*, Undark (June 1, 2020), https://undark.org/2020/06/01/michigan-unemployment-fraud-algorithm/; Kashmir Hill, *Wrongfully Accused by an Algorithm*, N.Y. Times (June 24, 2020), https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.
[21] *See* Arun Shastri, *Generative AI Errs Differently Than Classical AI*, Forbes (Sept. 4, 2023), https://www.forbes.com/sites/arunshastri/2023/09/04/generative-ai-errs-differently-than-classical-ai/.
[22] *See* Kara Williams, *Assessing the Assessments: Comparing Risk Assessment Requirements Around the World*, EPIC Blog (Dec. 4, 2023), https://epic.org/impact-comparison/ (assessing AI impact assessment requirements around the world, which frequently include data quality assessments, performance audits, and bias testing); Mona Rakibe, *The Significance of Data Quality in the World of Generative AI*, Medium (June 21, 2023), https://mona-rakibe.medium.com/the-significance-of-data-quality-in-the-world-of-generative-ai-5f84eb524299.
[23] NIST AI RMF at 4 (emphasis in original).

measurement challenges included within the AI RMF—third-party software risks, emergent risks, the availability of reliable metrics, differing risks across the AI lifecycle, inscrutability, etc.[24]—also mirror challenges inherent to generative AI technologies: risks of third-party AI API integrations,[25] model drift and generative AI output degradation over time,[26] and so forth. And crucially, the same risks and harms exist even when AI companies market their AI systems as "low-risk" or "trustworthy"; NIST must ensure that its AI standards are clear and actionable as industry benchmarks by which judges, regulators, and end-users can determine an AI system's trustworthiness.

While EPIC has provided a discussion of specific risks inherent to generative AI in Section II, *infra*, the foundation for an effective AI risk management framework remains the same across models, use cases, and contexts: **transparency and accountability**. AI risk management takes considerable time and resources, so companies developing and deploying AI systems need strong incentives to implement proactive risk management strategies rather than take a wait-and-see approach to AI risk. Transparency and accountability add external pressure to these companies to carry out AI risk mitigation strategies in good faith. In fact, NIST's AI RMF already incorporates principles of transparency and accountability into AI risk management, stating, *inter alia*, that "[m]eaningful transparency provides access to appropriate levels of information based on the stage of the AI lifecycle and tailored to the role or knowledge of AI actors or individuals interacting with or using the AI system," and that "maintaining the provenance of training data and supporting attribution of the AI system's decisions to subsets of training data can assist with both transparency and accountability."[27]

These same transparency principles will be effective tools to address many of the risks inherent to generative AI technologies, including risks stemming from synthetic content. Providing durable methods for watermarking synthetic content for end-users is a form—albeit imperfect—of risk-mitigating transparency.[28] Disclosing the provenance of all data—including copyrighted

[24] *Id.* at 5–6.

[25] *See* Alex Akimov, *It's Critical to Regulate AI Within the Multi-Trillian-Dollar API Economy*, TechCrunch (Dec. 22, 2023), https://techcrunch.com/2023/12/22/its-critical-to-regulate-ai-within-the-multi-trillion-api-economy/.

[26] *See* Lauren Leffer, *Yes, AI Models Can Get Worse Over Time*, Sci. Am. (Aug. 2, 2023), https://www.scientificamerican.com/article/yes-ai-models-can-get-worse-over-time/; Benj Edwards, *Study Claims ChatGPT is Losing Capability, but Some Experts Aren't Convinced*, Ars Technica (July 19, 2023), https://arstechnica.com/information-technology/2023/07/is-chatgpt-getting-worse-over-time-study-claims-yes-but-others-arent-sure/.

[27] NIST AI RMF at 15–16.

[28] Makena Kell, *Watermarks Aren't the Silver Bullet for AI Misinformation*, Verge (Oct. 31, 2023), https://www.theverge.com/2023/10/31/23940626/artificial-intelligence-ai-digital-watermarks-biden-executive-order; Mehrdad Saberi et al., *Robustness of AI-Image Detectors: Fundamental Limits and Practical Attacks*, arXiv (Sept. 29, 2023) (preprint), https://arxiv.org/pdf/2310.00076.pdf; David Pierce, *Google Made a*

data—used to train and calibrate a foundation model is risk-mitigating transparency.[29] And publishing model cards alongside deployed machine-learning models detailing, e.g., intended uses cases, use contexts, limitations, and performance evaluation results is risk-mitigating transparency.[30] As the AI RMF reflects, it is essential that part of this transparency be an explicit disclosure of purpose of the system being used at all, combined with its risk-benefit analysis.

II. Certain Risks of Generative AI Warrant Additional Safeguards

Responsive to Assignments 1–2

Last year, EPIC published the first of its generative AI reports, *Generating Harms: Generative AI's Impact & Paths Forward*, appended below this comment.[31] Our report traces major risks and societal impacts of generative AI technologies using real case studies and research across academia and civil society, covering risks as far-ranging as misinformation, extortion, data security vulnerabilities, discrimination, copyright infringement, and environmental impacts. EPIC encourages NIST to incorporate measures to address all these generative AI risks within its AI RMF companion resource, and we would be happy to discuss generative AI risks and risk mitigation strategies further.

The following subsections highlight a small selection of generative AI risks that EPIC has encountered, taken both from our 2023 generative AI report and more recent work we have undertaken, to underscore the importance of strong AI risk mitigation strategies—including data risk mitigation strategies like data minimization—for reducing AI harm.

AI Hallucinations

In their seminal AI research article, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, Emily Bender et al. describe the growing number of AI language models as "system[s] for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any

Watermark for AI Images That You Can't Edit Out, Verge (Aug. 29, 2023), https://www.theverge.com/2023/8/29/23849107/synthid-google-deepmind-ai-image-detector.

[29] *See generally, e.g.*, Karl Werder et al., *Establishing Data Provenance for Responsible Artificial Intelligence Systems*, 13 ACM Transactions on Mgmt. Info. Sys. 1 (2022), https://doi.org/10.1145/3503488.

[30] *See* Margaret Mitchell et al., *Model Cards for Model Reporting* 220 (Proc. Conf. on Fairness, Accountability, & Transparency, 2019), https://arxiv.org/pdf/1810.03993.pdf.

[31] EPIC Generative AI Report.

reference to meaning."[32] In the three years since the article was published—and despite the growing popularity of commercial large language models (LLMs) like OpenAI's GPT-4, Alphabet's Bard, and Meta's LLaMA—this foundational issue within generative AI technologies remains unaddressed: because these systems operate by probabilistically stringing words and responses together, they can produce syntactically and grammatically correct responses that are substantively nonsensical. These nonsensical responses are what AI researchers sometimes call "hallucinations."[33]

The risks of these AI hallucinations are myriad but fall predominantly into three main categories: (1) **misinformation**, (2) **encoded bias**, and (3) **data security vulnerabilities**. At their core, AI hallucinations are instances of misinformation provided in response to user prompts, which may be woven into otherwise-accurate information. For example, in 2023, a law professor was included on an ChatGPT-generated "list of legal scholars who had sexually harassed someone," even when no such allegation existed.[34] As Princeton Professor Arvind Narayanan said in an interview with the Markup:

> "Sayash Kapoor and I call [natural language processing] a bullshit generator, as have others as well. We mean this not in a normative sense but in a relatively precise sense. We mean that it is trained to produce plausible text. It is very good at being persuasive, but it's not trained to produce true statements. It often produces true statements as a side effect of being plausible and persuasive, but that is not the goal."[35]

Even before considering encoded bias, AI hallucinations pose very real and very serious risks to end-users. Large language models have already produced scores of health misinformation around topics like vaccines and vaping.[36] As with the law professor mentioned above, they can

[32] Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 617 (Proc. Conf. on Fairness, Accountability, & Transparency, 2021), https://dl.acm.org/doi/pdf/10.1145/3442188.3445922.

[33] *See* The Politics of Everything, *The Great A.I. Hallucination*, New Repub. (May 10, 2023), https://newrepublic.com/article/172454/great-ai-hallucination-chatgpt.

[34] Pranshu Verma & Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused*, Wash. Post. (Apr. 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.

[35] Julia Angwin, *Decoding the Hype About AI*, Markup (Jan. 28, 2023), https://themarkup.org/hello-world/2023/01/28/decoding-the-hype-about-ai.

[36] *See, e.g.*, Michael DePeau-Wilson, *ChatGPT Quickly Authored 100 Blogs Full of Healthcare Disinformation*, MedPage Today (Nov. 13, 2023), https://www.medpagetoday.com/special-reports/features/107329; Tiffany Hsu & Stuart A. Thomspon, *Disinformation Researchers Raise Alarms About A.I. Chatbots*, N.Y. Times (June 20, 2023), https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.

produce harmful and defamatory statements. And when asked questions about U.S. elections, major AI chatbots developed by companies like Microsoft have replied with unfounded conspiracy theories and hallucinated political scandals.[37]

Beyond pure misinformation, AI hallucinations can facilitate discrimination through encoded bias within AI training data. As Emily Bender et al. state:

> "Biases can be encoded in ways that form a continuum from subtle patterns like referring to *women doctors* as if *doctor* itself entails not-woman or referring to *both genders* excluding the possibility of non-binary gender identities, through directly contested framings (e.g. *undocumented immigrants* v. *illegal immigrants* or *illegals*), to language that is widely recognized to be derogatory (e.g., racial slurs)."[38]

These encoded biases within LLMs have real world impacts as well. When LLMs produce hallucinations, they often reinforce historical biases encoded through language. For example, LLMs may exhibit gendered assumptions in their responses or inaccurately characterize someone based on racial, ethnic, or gendered stereotypes.[39] When used to produce documents like professional recommendation letters, LLMs tend to replicate gender bias and undervalue the qualifications of women.[40] And LLMs' generation of overtly abusive language can facilitate psychological and reputational harms—or reinforce second-order effects of abusive language, such as violence.[41]

Finally, AI hallucinations can serve as a vehicle by which malicious actors can understand data security vulnerabilities and extract personally identifiable information from an LLM.[42] Sometimes, AI hallucinations reveal sensitive information about training data directly, as was the

[37] *See* David Gilbert, *Microsoft's AI Chatbot Replies to Election Questions with Conspiracies, Fake Scandals, and Lies*, Wired (Dec. 15, 2023), https://www.wired.com/story/microsoft-ai-copilot-chatbot-election-conspiracy/.

[38] Bender et al., *supra* note 32, at 617 (emphasis in original).

[39] Similar issues abound in AI image generation models. *See* Nitasha Tiku et al., *These Fake Images Reveal How AI Amplifies Our Worst Stereotypes*, Wash. Post. (Nov. 1, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-generated-images-bias-racism-sexism-stereotypes/.

[40] Chris Stokel-Walker, *ChatGPT Replicates Gender Bias in Recommendation Letters*, Sci. Am. (Nov. 22, 2023), https://www.scientificamerican.com/article/chatgpt-replicates-gender-bias-in-recommendation-letters/.

[41] Bender et al., *supra* note 32, at 617.

[42] *Id.* at 618; *see also* Prasanth Aby Thomas, *Questions Raised as Amazon Q Reportedly Starts to Hallucinate and Leak Confidential Data*, ComputerWorld (Dec. 4, 2023), https://www.computerworld.com/article/3711467/questions-raised-as-amazon-q-reportedly-starts-to-hallucinate-and-leak-confidential-data.html.

case with Amazon Q, a generative AI assistant released in late 2023.[43] But they can also signpost the types of data vulnerable to malicious AI actors. As the number of data parameters used by a LLM increases, so too does a model's tendency to "output specific information from [its] training data."[44] And despite efforts to implement guardrails into the types of prompts and outputs allowed within LLMs, malicious actors can still readily bypass many of these restrictions through adversarial attacks like prompt injection,[45] making AI hallucinations and other unintended content generated by generative AI technologies an effective tool for accessing private and sensitive information.

As discussed below, the risks and harms of AI hallucinations fundamentally stem from the data used to train AI systems. Incorporating stronger data controls, including data minimization techniques and data audits, are still some of the most effective risk mitigation techniques for generative AI harms.

Disinformation, Harassment, Impersonation, and Extortion

Beyond errors and biased outputs encoded into generative AI technologies by negligence or design, the structure of consumer-facing generative AI tools also facilitates the intentional and malicious use of generative AI to produce misleading or harmful synthetic content. In fact, some of the earliest uses—or rather, misuses—of generative AI technologies are deepfakes[46]: realistic images or videos created using machine-learning algorithms to depict someone as saying or doing something they did not (often by replacing the likeness of one person with that of another).[47] Deepfakes are more than an isolated risk of generative AI; they can be used to facilitate or exacerbate many of the risks described throughout the NIST AI RMF and EPIC's *Generating Harms* report.[48] For example, deepfaked audio or video content has been used to harass former

[43] Thomas, *supra* note 42.
[44] *Id.*
[45] *See* Benj Edwards, *AI-Powered Bing Chat Spills its Secrets Via Prompt Injection Attack*, Ars Technica (Feb. 10, 2023), https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack/.
[46] The term, "deepfake," is a portmanteau of "deep learning" and "fake." The term was popularized by a Reddit user, @deepfakes, who posted the first viral deepfake video in 2017. *See* Moncarol Y. Wang, *Don't Believe Your Eyes: Fighting Deepfaked Nonconsensual Pornography with Tort Law*, 2022 U. Chi. Legal F. 415, 417–18 (2022).
[47] Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Cal. L. Rev. 1753, 1757 (2019). (defining deepfakes as the "full range of hyper-realistic digital falsification of images, video, and audio").
[48] EPIC Generative AI Report.

romantic partners with nonconsensual, synthetic sexual imagery,[49] intimidate journalists,[50] degrade celebrities,[51] and influence election turnout.[52]

Worse still, some of the most popular methods for reducing the risks of malicious AI use and harmful synthetic content—including AI watermarking—may be insufficient by themselves. Even when AI developers insert watermarks imperceptibly into the pixels or metadata of synthetic content, as is possible under technologies like Google's SynthID,[53] researchers out of the University of Maryland have found several ways to break existing watermarking methods—and even insert false watermarks into images.[54]

While NIST should explore ways to durably track and label synthetic content, robust AI impact assessment requirements, ongoing testing, and transparent disclosure of incident reports following data leaks or malicious use will still be necessary to meaningfully mitigate the risks of generative AI.

Model Collapse and the AI Feedback Loop

AI systems—including generative AI technologies—do not exist within a vacuum. Current methods for collecting, processing, and using AI training data are subject to the flaws and worldviews reflected in the sources of data used. And over the past few years, the internet has exploded with synthetic context. This content is not sequestered in internet walled gardens, but rather strewn about the web haphazardly[55]: dozens of news websites now use synthetic content,[56] AI-generated text and images are flooding online marketplaces like Amazon and Etsy,[57]

[49] *See* Matt Burgess, *Deepfake Porn is Out of Control*, Wired (Oct. 16, 2023), https://www.wired.com/story/deepfake-porn-is-out-of-control/.

[50] *See* Rana Ayyub, *I Was the Victim of a Deepfake Porn Plot Intended to Silence Me*, Huff. Post (Nov. 21, 2018), https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316.

[51] *See* Kat Tenbarge, *Explicit, AI-Generated Taylor Swift Images Continue to Proliferate on X, Instagram and Facebook*, NBC News (Jan. 30, 2024), https://www.nbcnews.com/tech/tech-news/explicit-ai-generated-taylor-swift-images-continue-proliferate-x-insta-rcna136193.

[52] *See* Tiffany Hsu, *New Hampshire Officials to Investigate A.I. Robocalls Mimicking Biden*, N.Y. Times (Jan. 22, 2024), https://www.nytimes.com/2024/01/22/business/media/biden-robocall-ai-new-hampshire.html.

[53] Pierce, *supra* note 28.

[54] Kell, *supra* note 28; Saberi et al., *supra* note 28.

[55] *See* James Vincent, *AI is Killing the Old Web, and the New Web Struggles to be Born*, Verge (Jue 26, 2023), https://www.theverge.com/2023/6/26/23773914/ai-large-language-models-data-scraping-generation-remaking-web.

[56] Matthew Cantor, *Nearly 50 News Websites are 'AI-Generated,' a Study Says. Would I Be Able to Tell?*, Guardian (May 8, 2023), https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study.

[57] *See* Elizabeth Lopatto, *I'm Sorry, But I cannot Fulfill This Request as it Goes Against OpenAI Use Policy*, Verge (Jan. 12, 2024), https://www.theverge.com/2024/1/12/24036156/openai-policy-amazon-ai-listings; Kaitlyn Tiffany, *AI-Generated Junk is Flooding Etsy*, Atlantic (June 15, 2023), https://www.theatlantic.com/technology/archive/2023/06/ai-chatgpt-side-hustle/674415/.

companies like Snap and Meta are turning to AI chatbots,[58] and major search engines like Google and Bing are weaving AI functionalities into their services.[59] In fact, a 2022 report from Europol estimates that as much as 90% of the internet will be AI-generated by 2026.[60]

Why does this explosion of synthetic content matter? Because AI developers still frequently rely on web-scraping to collect AI training data—and synthetic content used in AI training datasets can cause "irreversible defects in the resulting models."[61] Specifically, training AI models on synthetic data can cause **model collapse**: a "degenerative process whereby, over time, models forget the true underlying data distribution…. This process is inevitable, even for cases with almost ideal conditions for long-term learning."[62] Without robust training data controls and resilient techniques for identifying synthetic content, AI systems will only become less accurate, more biased, and less trustworthy over time.

With Executive Order 14110, NIST has the opportunity to strengthen its AI RMF with the key transparency, accountability, enforcement, data quality mechanisms necessary to mitigate the risks of AI hallucinations, malicious AI uses, AI model collapse, and so much more. AI developers need strong incentives to implement necessary risk management practices, and EPIC urges NIST to strengthen its AI RMF and companion resources with actionable guardrails and guidelines for both AI companies and enforcement agencies.

III. Data Minimization is Critical for Effective AI Risk Management

Responsive to Assignment 3

Section 11(b) of Executive Order 14110 directs NIST to engage in efforts to "advance responsible global technical standards for AI development and use," including "best practices regarding data capture, processing, protection, privacy, confidentiality, handling, and analysis."[63]

[58] *See* Alex Heath, *Snapchat is Releasing its AI Chatbot to Everyone for Free*, Verge (Apr. 19, 2023), https://www.theverge.com/2023/4/19/23688913/snapchat-my-ai-chatbot-release-open-ai.

[59] *See* Thomas Claburn, *AI is Chaing Search, for Better or For Worse*, Register (Jan. 30, 2024), https://www.theregister.com/2024/01/30/ai_is_changing_search/.

[60] Maggie H. Dupré, *Experts: 90% of Online Content will be AI-Generated by 2026*, Futurism (Sept. 18, 2022), https://futurism.com/the-byte/experts-90-online-content-ai-generated.

[61] Carl Franzen, *The AI Feedback Loop: Researchers Warn of 'Model Collapse' as AI trains on AI-Generated Content*, VentureBeat (June 12, 2023), https://venturebeat.com/ai/the-ai-feedback-loop-researchers-warn-of-model-collapse-as-ai-trains-on-ai-generated-content/; Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, arXiv (Cambridge Univ. Working Paper, 2023), https://www.cl.cam.ac.uk/~is410/Papers/dementia_arxiv.pdf.

[62] *Id.*

[63] 88 Fed. Reg. at 75223–24.

The inclusion of data regulations within responsible global AI standards is critical, as data serves as the foundation for many of the most common AI risks covered herein and throughout the NIST AI RMF. Poor data quality controls and unconstrained web scraping weave bias and inaccuracies (as well as illegal and disturbing material)[64] into AI training data, which are then reflected in AI outputs.[65] Data security vulnerabilities in AI models enable malicious actors to jailbreak generative AI technologies to access private information or manipulate model outputs for harmful ends.[66] And the sheer demand for AI training data has produced a global and exploitative labor industry for labeling commercial datasets.[67]

To mitigate the most prominent AI risks, NIST must lead on standards for collecting, processing, auditing, and using AI training data, and one proven method for mitigating data risks is **data minimization**.

Data minimization is a framework that requires companies to limit the collection, use, disclosure, and retention of personal information to that which is necessary for the purpose for which it was collected. Data minimization allows for certain appropriate necessary uses of personal information like to perform system maintenance, detect fraud, or protect against spam. In a data minimization framework, the onus is on a company to demonstrate the necessity and proportionality of the data processing it performs. With respect to AI, data minimization can ensure that datasets are used appropriately for legitimate, related, necessary purposes, such as facilitating audits of systems to ensure fairness or requiring that data be promptly deleted when no longer necessary.

This standard also stops companies from being incentivized from overcollection and excessive retention of sensitive personal information to train AI models. Invasive practices like data scraping allow companies to collect information, including photos and videos, online and use it to feed and train AI models. Serious privacy harms arise from the collection and use of sensitive information to feed AI systems. The collection and processing of certain types of sensitive

[64] *See* Ryan Heath, *Child Abuse Images Found in AI Training Data*, Axios (Dec. 20, 2023), https://www.axios.com/2023/12/20/ai-training-data-child-abuse-images-stanford.

[65] *See* NIST AI RMF at 38; Reva Schwartz et al., NIST, Toward a Standard for Identifying and Managing Bias in Artificial Intelligence 14–19 (NIST Special Pub. 1270, Mar. 16, 2022), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf.

[66] *See, e.g.*, Katyanna Quach, *OpenAI's GPT-4 Finally Meets its Match: Scots Gaelic Smashes Safety Guardrails*, Register (Jan. 31, 2024), https://www.theregister.com/2024/01/31/gpt4_gaelic_safety/; Todd Bishop, *Microsoft AI Engineer Says Company Thwarted Attempt to Expose DALL-E 3 Safety Problems*, GeekWire (Jan. 30, 2024), https://www.geekwire.com/2024/microsoft-ai-engineer-says-company-thwarted-attempt-expose-dall-e-3-safety-problem/; David Barry, *Microsoft's AI Data Leak Isn't the Last One We'll See*, Reworked (Sept. 29, 2023), https://www.reworked.co/information-management/microsofts-ai-data-leak-isnt-the-last-one-well-see/; Thomas, *supra* note 42.

[67] *See, e.g.*, Adrienne Williams et al., *The Exploited Labor Behind Artificial Intelligence*, Noema (Oct. 13, 2022), https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/.

information, like emotion-related data or biometric data, in AI systems are some of the most concerning uses of AI.[68] Data minimization prevents this type of collection and use and would accordingly disrupt the current business model that sustains harmful collection and use of personal information. EPIC encourages NIST to include a robust data minimization standard to ensure an effective AI Risk Management Framework because it is the strongest way to limit harmful uses and impacts of sensitive information within AI systems and prevent the most privacy invasive risks of AI systems.

IV. TRANSPARENCY & OVERSIGHT ARE CRITICAL FOR EFFECTIVE AI RISK MANAGEMENT

Responsive to Assignment 3

Allowing AI developers and deployers to implement internal procedures without transparency obligations or external review mechanisms risks rendering NIST's substantive AI risk management procedures meaningless. This is especially true for companies developing and deploying generative AI systems. In fact, transparency already serves as a core feature within the draft text of the European Union Artificial Intelligence Act ("EU AI Act").[69] As NIST begins to develop global AI consensus standards pursuant to Section 11(b) of Executive Order 14110, EPIC encourages the agency to incorporate key AI transparency accountability provisions already in place abroad. Between best practices, the EU AI Act, and several other laws in states around the US and countries around the world, companies using AI will have to increase their transparency – NIST's recommendations should be in line with that growing obligation.

Specifically, the EU AI Act addresses AI risks from a harm- and use-case-based framework. Depending on the level of risk posed by an AI use case to fundamental rights, public safety, and public health, the AI Act either (1) prohibits the development and deployment of AI systems for such a use case, (2) imposes mandatory obligations on AI developers, or (3) suggests a voluntary code of conduct. Under the EU AI Act, generative AI is neither regulated as a

[68] AI Now Institute, Data Minimization as A Tool For AI Accountability (2023), https://ainowinstitute.org/spotlight/data-minimization.

[69] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts* ("*Artificial Intelligence Act*"), COM/2021/0106(COD), https://drive.google.com/file/d/1xfN5T8VChK8fSh3wUiYtRVOKIi9oIcAF/view [hereinafter "EU AI Act"]. Note that this is not the final version of the regulation. The EU AI Act finished its lengthy negotiations phase between the Parliament and the Council and is now in the final stages of adoption. On January 22, 2024, Luca Bertuzzi, a journalist for E.U. news outlet, Euractiv, leaked a post-negotiation draft of the AI Act text, which is the most recent version of the proposal available to the public. This comment cites to that version of the draft. The document includes four columns; the first three columns are previous drafts, and the rightmost column contains the negotiated, most up-to-date text.

monolithic technology nor separated from other, non-generative AI systems. Therefore, if a generative AI model falls under a regulated high-risk use case—such as use by a judicial authority to research and interpret the law—the generative AI model is subject to the high-risk use case obligations.

The only obligations under the AI Act specific to generative AI models fall on the "providers" of generative AI models. Providers are natural or legal persons, public authorities, agencies, or other bodies that develop an AI system or general-purpose AI model, or that has one developed on its behalf, that puts such system into service or places the model on the European Union market whether for payment or free of charge.[70] Under Article 52(1), providers of AI systems that "generate synthetic audio, image, video or text content" shall mark the outputs of the AI system in a "machine-readable format" to ensure that the outputs are "detectable as artificially generated or manipulated."[71] This technical solution—commonly described as AI watermarking—is required to be "effective, interoperable, robust and reliable" and should be up to par with the "generally acknowledged state-of-the-art" technical standards.[72] This requirement does not apply when the AI system performs an assistive function for standard editing or does not substantially alter the input data (or semantics thereof) provided by the deployer of the AI system. This marking requirement also does not apply where authorized by law to "detect, prevent, investigate, and prosecute criminal offenses."[73]

Beyond this one provision, generative AI models are otherwise subject to the exact same regulatory scheme as non-generative AI models. This regulatory scheme includes privacy safeguards at every level of the AI system's lifecycle when the AI system's use is deemed a "high risk" use case. For example, the EU AI Act requires data training sets to be tailored to the intended use of the AI system,[74] error free to the extent possible,[75] and requires that the training sets be evaluated for bias[76]—requirements that align broadly with data minimization and data quality control requirements. Providers of AI are also required to (1) keep documentation regarding the creation of the data set, including the formulation of assumptions and what the data is supposed to measure and represent,[77] and (2) make automatic event logging technically possible when developing high-risk AI systems.[78] And before the high-risk AI systems are deployed or placed on

[70] EU AI Act at Art. 3(2).
[71] *Id.* at Art. 52(1).
[72] *Id. But see* Kell, *supra* note 54 (describing limitations on current AI watermarking techniques).
[73] EU AI Act at Art. 52(1).
[74] *Id.* at Art. 10(2).
[75] *Id.* at Art. 10(3).
[76] *Id.* at Art. 10(2)(fa).
[77] *Id.* at Art. 10(2)(d).
[78] *Id.* at Art. 12.

the market, providers must do a fundamental rights impact assessment. This impact assessment must include, among other certifications:

(1) A description of the intended use of the AI product;

(2) The time period within which the AI product will be deployed;

(3) The natural persons or groups likely to be affected by the product's intended use and the specific risk of harm to those people;

(4) A description of the risk mitigation procedures, including human oversight measures;

(5) Instructions for deployers on how to use the system appropriately; and

(6) Instructions on how to take corrective action if such risks materialize during the deployment of the product.[79]

Next, the EU AI Act requires post-deployment monitoring of high-risk AI systems to ensure that, as the system is deployed, it continues to comply with the regulations—and its provider and deployers continue to mitigate AI risks.[80] Specifically, providers have a duty to disable, remove, and/or recall AI systems from the market if there is a significant incident relating to public safety, public health, or fundamental rights.[81] Providers also have a duty to inform deployers, other downstream users, and the relevant regulatory authorities of such incidents.[82] It is critical that there are members of the regulatory infrastructure besides those who stand to profit from the ongoing use and growth of a system.

In addition to the fundamental rights impact assessment, providers and deployers are required to create and maintain a risk management system as a "continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system" that includes regular review and updating.[83] The risk management system must include the identification and analysis of known and reasonably foreseeable risks to fundamental rights, public safety, and public health when the AI system is used in its intended purpose, under conditions of reasonably foreseeable misuse, and based on data from post-market monitoring.[84] Special consideration is given to risks adversely affecting minors and "other vulnerable groups[.]"[85] In response to the identification of such risks, providers and deployers shall adopt "appropriate and targeted" risk management

[79] *Id.* at Art. 10(2)(d).
[80] *Id.* at Art. 61–68e.
[81] *Id.* at Art. 21.
[82] *Id.*
[83] *Id.* at Art. 9(2)(a).
[84] *Id.* at Art. 9(2)(b).
[85] *Id.* at Art. 9(8).

measures as far as technically feasible through adequate design and development.[86] Where risks cannot be eliminated, providers and deployers shall implement adequate mitigation and control measures.[87] Providers are expected to provide technical documentation and, in some cases, training to deployers to ensure that the AI system is used in its intended context to effectively mitigate risk. Providers must also test high risk AI systems to identify the most appropriate risk mitigation measures, which may include testing in real world conditions.[88]

Lastly, the EU AI Act stresses the importance of AI transparency across all dimensions of AI risk management. Several obligations center around transparency between providers and deployers; between providers and end-users who are natural persons; and between providers, deployers, and the general public. Providers of high-risk AI systems are required to give deployers technical documentation on how the AI system works, instructions on how to properly use the AI system, and illustrative examples of the risks and limitations of the AI system to ensure adequate comprehension.[89] As previously mentioned, providers of generative AI should ensure that end-users who are natural persons should be made aware that they are engaging with synthetic content.[90] And finally, providers of general-purpose AI models like foundation models—and, when deployers control AI input data, deployers of such models—must keep and publish a "sufficiently detailed" record of the content used to train the model so that parties with legitimate interests can enforce their rights.[91] These rights may include copyright interests and other rights related to the data collection process.[92]

The EU AI Act, while imperfect, underscores the value of (1) actionable transparency requirements for both data inputs and AI system outputs, (2) robust AI testing procedures and impact assessments, and (3) strict data controls like data minimization as tools for effectively managing AI risks. As NIST develops companion resources for the AI RMF and explores opportunities to develop global AI consensus standards pursuant to Executive Order 14110, EPIC urges NIST to incorporate these same AI risk management techniques and lead on global AI standards.

[86] *Id.* at Art. 9(3), 9(4)(a).
[87] *Id.* at Art. 9(4)(b).
[88] *Id.* at Art. 9(5)–(6).
[89] *Id.* at Art. 13.
[90] *Id.* at Art. 52(1).
[91] *Id.* at Art. 52c(2)(d).
[92] *Id.* at Art. 52c(2)(c), Recital 60j; *see also, e.g.*, Commission Regulation 2016/679 of Apr. 27, 2016, *on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)* O.J. (L 119.1), Art. 12–23. (including the right to access data, right to erasure, and right to rectification, among others).

Conclusion

EPIC welcomes NIST's efforts to bolster its responsible AI development and use standards in response to Executive Order 14110 and applauds NIST's decision to focus on the current risks and harms of generative AI technologies instead of imagined existential threats. NIST can and should use its statutory authority and recent Presidential mandate to:

1. Clarify and expand the applicability of existing AI risk management techniques under the AI RMF to generative AI technologies;

2. Bolster key provisions of the AI RMF—including transparency and disclosure requirements, AI testing requirements, and data controls—to incentivize compliance and guide enforcement against negligent AI developers and malicious AI actors;

3. Incorporate strong data minimization principles within the AI RMF and companion resources to mitigate myriad AI risks at the data source; and

4. Explore opportunities to incorporate strong AI risk management language from the EU AI Act, including specific language around AI transparency and accountability requirements, within global AI consensus standards.

We appreciate this opportunity to reply to NIST's RFI and are willing to engage with NIST further on any of the issues raised within our comment, including the centrality of data controls to AI risk management, the value and structure of effective AI red-teaming, and the emerging risks of generative AI. EPIC has also joined the U.S. AI Safety Institute Consortium (AISIC) and plans to engage further with responsible AI development and use standards therein. EPIC's recommendations align closely to the goals of Executive Order 14110 and the NIST AI RMF to increase the safety, equity, and reliability of AI technologies both now and long into the future.

Respectfully submitted,

/s/ Ben Winters
Ben Winters
EPIC Senior Counsel

/s/ Sara Geoghegan
Sara Geoghegan
EPIC Counsel

/s/ Grant Fergusson
Grant Fergusson
Equal Justice Works Fellow

/s/ Maria Villegas Bravo
Maria Villegas Bravo
EPIC Law Fellow

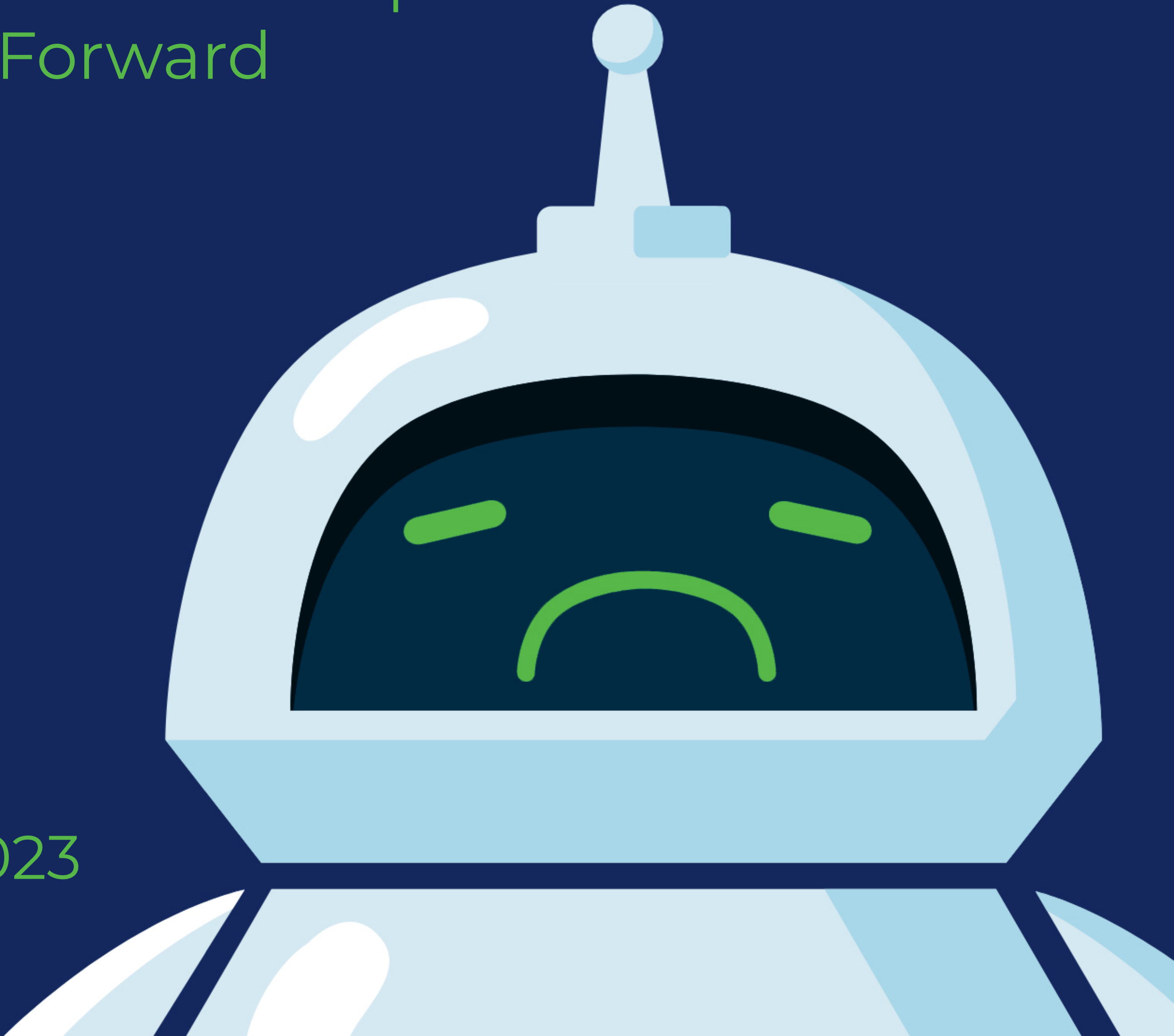
/s/ Kara Williams
Kara Williams
EPIC Law Fellow

ELECTRONIC PRIVACY
INFORMATION CENTER (EPIC)
1519 New Hampshire Ave. NW
Washington, DC 20036
202-483-1140 (tel)
202-483-1248 (fax)

epic.org
ELECTRONIC
PRIVACY
INFORMATION
CENTER

GENERATING HARMS

Generative AI's Impact & Paths Forward



MAY 2023

CONTRIBUTIONS BY

Grant Fergusson
Caitriona Fitzgerald
Chris Frascella
Megan Iorio
Tom McBrien
Calli Schroeder
Ben Winters
Enid Zhou

EDITED BY

Grant Fergusson, Calli Schroeder, Ben Winters, and Enid Zhou

Thank you to Sarah Myers West and Katharina Kopp for your generous comments on an earlier draft of the paper.

Notes on this Paper:
This is version 1 of this paper and is reflective of documented and anticipated harms of Generative AI as of May 15, 2023. Due to the fast-changing pace of development, use, and harms of Generative AI, we want to acknowledge that this is an inherently dynamic paper, subject to changes in the future.

Throughout this paper, we use a standard format to explain the typology of harms that generative AI can produce. Each section first explains relevant background information and potential risks imposed by generative AI, then highlights specifics harms and interventions that scholars and regulators have pursued to remedy each harm. This paper draws on two taxonomies of A.I. harms to guide our analysis:

1. Danielle Citron's and Daniel Solove's Typology of Privacy Harms, comprising physical, economic, reputational, psychological, autonomy, discrimination, and relationship harms;[1] and
2. Joy Buolamwini's Taxonomy of Algorithmic Harms, comprising loss of opportunity, economic loss, and social stigmatization, including loss of liberty, increased surveillance, stereotype reinforcement, and other dignitary harms.[2]

These taxonomies do not necessarily cover all potential AI harms, and our use of these taxonomies is meant to help readers visualize and contextualize AI harms without limiting the types and variety of AI harms that readers consider.

Table of Contents

Introduction

OpenAI's decision to release ChatGPT, a chatbot built on the Large Language Model GPT-3, last November thrust AI tools to the forefront of public consciousness. In the last six months, new AI tools used to generate text, images, video, and audio based on user prompts exploded in popularity. Suddenly, phrases like Stable Diffusion, Hallucinations, and Value Alignment were everywhere. Each day, new stories about the different capabilities of generative AI—and their potential for harm—emerged without any clear indication of what would come next or what impacts these tools would have.

While generative AI may be new, its harms are not. AI scholars have been warning us of the problems that large AI models can cause for years.[3] These old problems are exacerbated by the industry's shift in goals from research and transparency to profit, opacity, and concentration of power. The widespread availability and hype of these tools has led to increased harm both individually and on a massive scale. AI replicates racial, gender, and disability discrimination, and these harms are weaved inextricably through every issue highlighted in this report.

OpenAI and other companies' decisions to rapidly integrate generative AI technology into consumer-facing products and services have undermined longstanding efforts to make AI development transparent and accountable, leaving many regulators scrambling to prepare for the repercussions. And it is clear that generative AI systems can significantly amplify risks to both individual privacy and to democracy and cybersecurity generally. In the words of the OpenAI CEO, who indeed had the power not to accelerate the release of this technology, "I'm especially concerned that these models could be used for widespread misinformation…[and] offensive cyberattacks."

This rapid deployment of generative AI systems without adequate safeguards is clear evidence that self-regulation has failed. Hundreds of entities, from corporations to media and government entities, are developing and looking to rapidly integrate these untested AI tools into a wide range of systems. And this rapid rollout will have disastrous results without necessary fairness, accountability, and transparency protections built in from the beginning.

We are at a critical juncture as policymakers and industry around the globe are focusing on the substantial risks and opportunities posed by AI. There is an opportunity to make this technology work *for* people. Companies should be required to show their work, make it clear when AI is in use, and offer informed consent throughout the training, development, and use process.

One thread of public concern focuses on AI's "existential" risks—speculative long-term risks in which robots replace humans at work, socially, and ultimately taking over, a la "I, Robot." Some legislators on the state and federal level have begun to take the issue of addressing AI more seriously—however, it remains to be seen if their focus will be only on supporting companies with their development of AI tools and requiring marginal disclosure and transparency requirements. Enacting clear prohibitions on high-risk uses, addressing the easy spread of disinformation, requiring meaningful and proactive disclosures that facilitate informed consent, and bolstering consumer protection agencies are necessary to address the harms and risks specific to generative AI. This paper strives to provide a broad outline of different issues that the use of generative AI brings up, educate lawmakers and the public, and offer some paths forward to mitigate harm.

- Ben Winters, Senior Counsel

Turbocharging Information Manipulation

BACKGROUND AND RISKS

The widespread availability of free and low-cost generative AI tools facilitates the spread of high volumes of text, image, voice, and video content. Much of the content created by AI systems is likely benign or could be beneficial to specific audiences, but these systems will also facilitate the spread of extremely harmful content. For example, generative AI tools can and will be used to propagate content that is false, misleading, biased, inflammatory, or dangerous. As generative AI tools grow more sophisticated, it will be quicker, cheaper, and easier to produce this content—and existing harmful content can serve as the foundation to produce more. In this section, we consider five categories of harmful content that AI tools would turbocharge: Scams, Disinformation, Misinformation, Cybersecurity Threats, and Clickbait and Surveillance Advertising. Though we draw distinctions between disinformation (purposeful spread of false information) and misinformation (less purposeful spread or creation of false information), the spread of AI-generated content will blur this line for parties that use AI-generated content without first editing or factchecking it. Entities using AI-generated outputs without exercising due diligence should be held jointly responsible with the entity behind the generation of that output for the harm it causes.

CASE STUDY – ELECTION 2024

Products using GPT-4 and subsequent large language models can create quick and unique human-sounding "scripts" that can be distributed via text, email, print, or through an AI voice generator combined with AI video generators. These AI-generated scripts can be used to dissuade or scare voters—or spread misinformation about voting or elections. In 2022, for example, text messages were sent to voters in at least five states with purposefully wrong voting information. This type of election misinformation has become common in recent years, but generative AI tools will supercharge bad actors' ability to quickly spread believable election misinformation. Congress must enact legislation that protects against deliberate voter intimidation, deterrence, or interference through false or misleading information, as well as false claims of endorsement.

SCAMS

Scam phone calls, texts, and emails have long been out of control, harming the public in many ways. In 2021 alone, 2.8 million consumers filed fraud reports with the FTC, claiming more than $2.3 billion in losses, and nearly 1.4 million consumers filed identity theft reports.[4] Generative AI can accelerate the creation, personalization, and believability of these various scams using AI-generated text, voices, and videos. AI voice generation can also be used to mimic the voice of a loved one, calling to request immediately financial assistance for bail, legal help, or ransom.[5]

According to a 2022 report from EPIC and the National Consumer Law Center, there are over one billion scam robocalls made to American telephones each month, which led to nearly $30 billion in consumer losses between June 2020-21—most frequently targeting vulnerable communities like seniors, individuals with disabilities, and people in debt.[6] These scams are made at scale, and often use an automated voice speaking a script generated by a text generator like ChatGPT designed to pretend they're someone of authority to scare consumers into sending money. In 2022, estimated consumer losses increased to $39.5 billion,[7] with the FTC reporting more than $326 million lost from scam texts alone.[8]

Auto dialers, robo-texts, robo-emails, and mailers, combined with data brokers that sell lists of numbers or email addresses, enable entities to send out a massive number of messages at once. The same data brokers can sell lists of people as potential targets along with "insights" about their mental health conditions, religious beliefs, or sexuality that can be exploited. The degree of targeting that data brokers are allowed to use on individuals exacerbates AI-generated harm.

Text generation services also increase the likelihood of successful phishing scams and election interference by bad actors. This has already happened—in a 2021 study, researchers found phishing emails generated by GPT-3 were more effective than human-generated ones.[9] Generative AI can expand the pool of potentially effective fraudsters by aiding people with limited English skills in crafting natural and accurate-sounding emails that can then target employees, intelligence targets, and individuals in a way that makes it much more difficult to detect the scam.

DISINFORMATION

Bad actors can also use generative AI tools to produce adaptable content designed to support a campaign, political agenda, or hateful position and spread that information quickly and inexpensively across many platforms. This rapid spread of false or misleading content—AI-facilitated disinformation—can also create a cyclical effect for generative AI: when a high volume of disinformation is pumped into the digital ecosystem and more generative systems are trained on that information via reinforcement learning methods, for example, false or misleading inputs can create increasingly incorrect outputs.

The use of generative AI tools to accelerate the spread of disinformation could fuel efforts to influence public opinion, harass specific individuals, or affect politics and elections. The impacts of increased disinformation may be far-reaching and cannot be easily countered once spread; this is especially concerning given the risks disinformation poses to the democratic process.

MISINFORMATION

The phenomenon of inaccurate outputs by text-generating large language models like Bard or ChatGPT has already been widely documented. Even without the intent to lie or mislead, these generative AI tools can produce harmful misinformation. The harm is exacerbated by the polished and typically well-written style that AI generated text follows and the inclusion among true facts, which can give falsehoods a veneer of legitimacy. As reported in the Washington Post, for example, a law professor was included on an AI-generated "list of legal scholars who had sexually harassed someone," even when no such allegation existed.[10] As Princeton Professor Arvind Narayanan said in an interview with The Markup:

> Sayash Kapoor and I call it a bullshit generator, as have others as well. We mean this not in a normative sense but in a relatively precise sense. We mean that it is trained to produce plausible text. It is very good at being persuasive, but it's not trained to produce true statements. It often produces true statements as a side effect of being plausible and persuasive, but that is not the goal.[11]

AI-generated content implicates a broader legal issue as well: our trust in what we see and hear. As AI-generated media becomes more common, so too will circumstances where we are tricked into believing something fictional is real[12]—or that something real is fictional.[13] When individuals can no longer trust information and new information is generated faster than it can be checked for accuracy, what can they do? Information sources like Wikipedia could be overwhelmed with false AI-generated content. This can be harmful in targeted situations by inducing a target to act under the assumption that, e.g., their loved ones are in crisis.[14]

SECURITY

The same phishing concerns described above pose a security threat. Though chatbots cannot (yet) develop their own novel malware from scratch, hackers could soon potentially use the coding abilities of large language models like ChatGPT to create malware that can then be minutely adjusted for maximum reach and effect, essentially allowing more novice hackers to become a serious security risk. In fact, security professionals have noted that hackers are already discussing how to install malware and extract information from targets using ChatGPT.[15]

Generative AI tools could very well begin to learn from repeated exposure to malware and be able to develop more novel and unpredictable malware that evades detection by common security systems.

CLICKBAIT AND FEEDING THE SURVEILLANCE ADVERTISING ECOSYSTEM

Beyond misinformation and disinformation, generative AI can be used to create clickbait headlines and articles, which manipulate how users navigate the internet and applications. For example, generative AI is being used to create full articles, regardless of their veracity, grammar, or lack of common sense, to drive search engine optimization and create more webpages that users will click on. These mechanisms attempt to maximize clicks and engagement at the truth's expense, degrading users' experiences in the process. Generative AI continues to feed this harmful cycle by spreading misinformation at faster rates, creating headlines that maximize views and undermine consumer autonomy.

HARMS

- **Economic/Economic Loss**: Successful scams and malware can result in victims' direct economic loss through extortion, trickery, or gaining access to financial accounts. This can lead to long-term impacts on credit as well.
- **Reputational/Relationship/Social Stigmatization**: Misinformation and disinformation can generate and spread false or harmful information about an individual resulting in harm to their reputation in the community, potential damage to their personal and professional relationships, and impacts to their dignity.
- **Psychological—Emotional Distress**: Disinformation and misinformation can cause severe emotional harm as individuals navigate the impacts of false information being spread about them—in addition, many individuals face shame and embarrassment if they are the victim of scams and may feel manipulated or used in the context of clickbait and surveillance advertising.
- **Psychological—Disturbance**: The influx of false or misleading information and clickbait makes it difficult for individuals to carry on their daily activities online.
- **Autonomy**: The spread of misinformation and disinformation makes it increasingly difficult for individuals to make properly informed choices and the manipulative nature of surveillance advertising complicates the issue of choice even further.
- **Discrimination**: Scams, disinformation, misinformation, malware, and clickbait all prey on vulnerabilities of the "marks," including membership in certain vulnerable groups and categories (the elderly, immigrants, etc.).

EXAMPLES

- People used AI to call in fake bomb threats to public places like schools.[16]
- AI voice generators were used call people's loved ones, convincing them that their family member was in jail and desperately needed money for bail and legal assistance.[17]
- The Center for Countering Digital Hate tested Google's Bard chatbot to see if they would replicate 100 common conspiracy theories including Holocaust denial and saying the mass child murder tragedy at Sandy Hook was staged using "crisis actors." Bard pumped out text based on these lies 78 out of 100 times without context or disclosure.
- Unedited AI Spam was found by Vice reporters widely throughout the internet.[18]
- CNET, a tech news website, paused its use of AI and issued corrections in 41 out of the 77 stories that it published which had been written using an AI tool. The AI-written articles, which were designed to be viewed more on Google searches to increase ad revenue, contained inaccurate and misleading information.[19]
- Similarly, Buzzfeed reportedly published AI-written content, namely travel guides, with the aim to attract search traffic about different destinations. The quality of the results was uniformly reviewed as useless and unhelpful.[20]

INTERVENTIONS

- Enact a law that makes intimidating, deceiving, or deliberately misinforming someone about an election or candidate illegal (regardless of the means), such as the Deceptive Practices and Voter Intimidation Prevention Act.

  - Pass the American Data Privacy Protection Act. The ADPPA will limit the collection and use of personal information to that which is reasonably necessary and proportionate to the purpose for which the information was collected. Such limitation will limit personal information being used to profile users to target them with ads, phishing attempts, and other scams. The ADPPA will also restrict the use of personal data to train generative AI systems that can manipulate users.
- Promulgate an FTC Commercial Surveillance rule that sets a data minimization standard prohibiting out-of-context secondary uses of personal information, which would similarly prevent training generative AI systems using personal information collected for an unrelated purpose.

Harassment, Impersonation, and Extortion

BACKGROUND AND RISKS

Some of the earliest uses—or misuses—of generative AI technologies are deepfakes:[21] realistic images or videos created using machine-learning algorithms to depict someone as saying or doing something they did not—often by replacing the likeness of one person with that of another.[22] Deepfakes and other AI-generated content can be used to facilitate or exacerbate many of the harms listed throughout this report, but this section focuses on one subset: intentional, targeted abuse of individuals. AI-generated images and videos provide several ways for bad actors to impersonate, harass, humiliate, exploit, and blackmail others. For example, a deepfake video could show a victim praising a cause they detest or engaging in sexually explicit or otherwise humiliating acts. These images and videos can spread rapidly across the internet as well, making it difficult or impossible for victims, law enforcement, and other interested parties to identify the creator(s) and ensure harmful deepfakes are removed. Unfortunately, many victims of targeted deepfakes are left without recourse, and those who pursue recourse are often forced to identify and confront the perpetrators themselves.

The harms of synthetic media predate AI and machine learning. As far back as the 1990s, commercial photo editing software enabled users to alter appearances or swap faces in photos. However, modern deepfakes and other AI-generated synthetic content trace their roots to Google's 2015 release of TensorFlow, an open-source tool for building machine-learning models, and the viral spread of a 2017 deepfake created using such a tool.[23] To create these early deepfakes—many of which involved placing celebrities' faces onto the bodies of pornographic film actors—a creator had to build a machine-learning model (often, a generative adversarial network, or GAN) using a tool like TensorFlow, train it on various image, video, or audio files, and then instruct the model to map a specific person's features or voice onto another person's body.[24] The release of new generative AI services like Midjourney and Runway removed these technical hurdles, enabling anyone to quickly create AI-generated content by providing a few key images, a source video, or even text entries.

At its core, using AI-generated content to impersonate, harass, humiliate, exploit, or blackmail an individual or organization is frequently no different from doing the same using other methods. Victims of deepfake harms may still turn to existing criminal and civil remedies for fraud,[25] impersonation,[26] extortion,[27] and cyberstalking[28] to redress malicious uses of generative AI tools. However, generative AI raises novel legal issues and exacerbates harm in new ways, straining the ability of victims and regulators alike to use existing legal avenues to redress harm. For example, deepfake impersonations of deceased people—a phenomenon described as "ghostbots"—may not only implicate defamation law, but also cause emotional distress among a deceased individual's loved ones where false textual quotes may not.[29] These new legal issues fall into roughly three categories: issues involving malicious intent; issues involving privacy and consent; and issues involving believability.

CASE STUDY – SILENCING A JOURNALIST

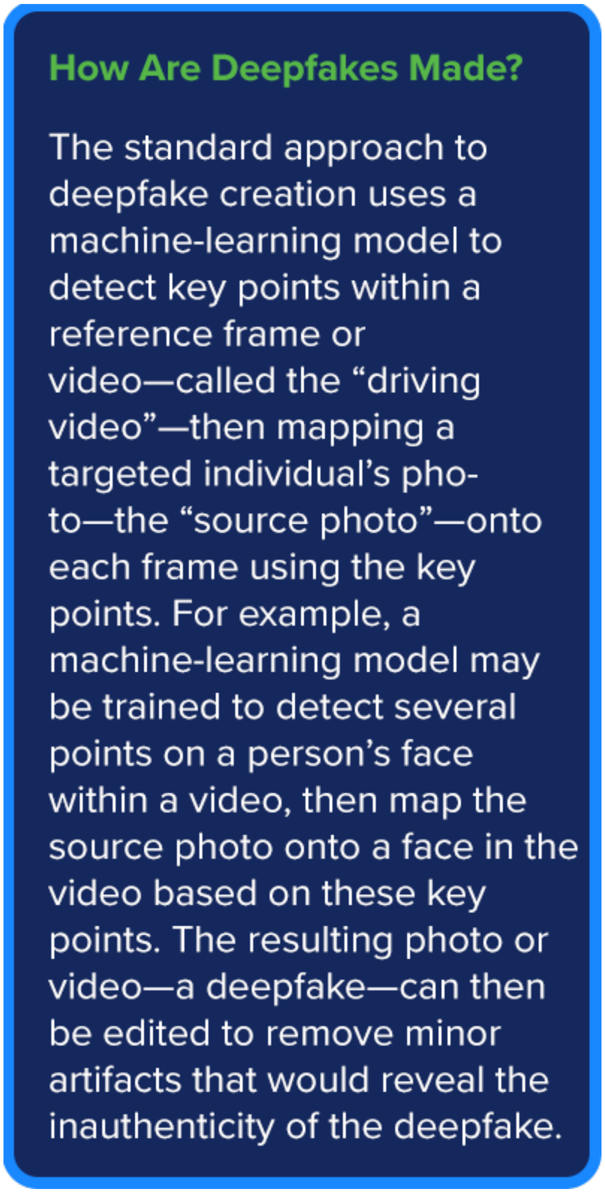
In April of 2018, Indian investigative journalist Rana Ayyub received an email from a source within the Modi government. A video of her engaging in sexual acts was going viral, leading to public humiliation and criticism from those who wanted to discredit her work. But it was a fake. Ayyub's likeness was inserted into a pornographic video using an early deepfake technology. As public scrutiny increased, her home address and cell phone information were leaked, leading to death and rape threats. This early video was circulated to harass, shame, and ostracize a vocal critic of the government – and for months, it succeeded.

MALICIOUS INTENT

A frequent malicious use case of generative AI to harm, humiliate, or sexualize another person involves generating deepfakes of nonconsensual sexual imagery or videos. These sexual deepfakes are some of the earliest and most common examples of deepfake technology, garnering widespread media attention.[30] However, many existing nonconsensual pornography laws limit liability to circumstances where content is published with an intent to harm.[31] Some malicious uses of generative AI no doubt meet this threshold, but many deepfake creators may *not* intend to harm the subject of a sexual deepfake; rather, they may create and circulate the deepfake without ever expecting the subject to see or be impacted by the content.

Intent requirements permeate other criminal laws applicable to malicious uses of generative A.I as well. For example, the federal cyberstalking statute, 18 U.S.C. § 2261A, only applies to those who act "with the intent to kill, injure, harass, intimidate, or place under surveillance [with similar intent]." State impersonation statutes like California Penal Code § 528.5 similarly limit enforcement to those who impersonate another "for purposes of harming, intimidating, threatening, or defrauding another person." Using generative AI to intimidate, harass, defraud, or extort another person may fall within these criminal statutes, but creating harmful or sexual deepfakes for personal enjoyment or entertainment may not.

Lastly, divining the intent of a deepfake creator is made more difficult by a modern feature of many online platforms: user anonymity. When a victim becomes aware of a malicious deepfake as it spreads online—as happened to Journalist Rana Ayyub in 2018—it can be incredibly difficult, if not impossible, to track down the original creator to bring a lawsuit or criminal charges.



How Are Deepfakes Made?

The standard approach to deepfake creation uses a machine-learning model to detect key points within a reference frame or video—called the "driving video"—then mapping a targeted individual's photo—the "source photo"—onto each frame using the key points. For example, a machine-learning model may be trained to detect several points on a person's face within a video, then map the source photo onto a face in the video based on these key points. The resulting photo or video—a deepfake—can then be edited to remove minor artifacts that would reveal the inauthenticity of the deepfake.

PRIVACY AND CONSENT

Even when a victim of targeted, AI-generated harms successfully identifies a deepfake creator with malicious intent, they may still struggle to redress many harms because the generated image or video *isn't* the victim, but instead a composite image or video using aspects of multiple sources to create a believable, yet fictional, scene. At their core, these AI-generated images and videos circumvent traditional notions of privacy and consent: because they rely on public images and videos, like those posted on social media websites, they often don't rely on any private information. This feature of AI-generated content excludes certain traditional privacy torts, including intrusion upon seclusion and publication of private facts, which depend explicitly on the publication or intrusion upon *private* facts.[32] Other privacy torts, including false light, fare better because they only require plaintiffs to show that the creator knew or recklessly disregarded whether a reasonable person would find the AI-generated content highly offensive.[33] Still, these claims too face a difficult legal hurdle: the First Amendment.[34]

The generative nature of new AI tools like Midjourney and Runway places them at a difficult crossroads between free expression protections and privacy protections for deepfake victims. Many AI-generated photos and videos transform source material or include new content in ways that may be protected under the First Amendment, but they can *appear* to be real footage of the victim in embarrassing, sexual, or otherwise undesirable circumstances. This tension between free speech, privacy, and consent raises new and difficult legal questions for both private individuals and public figures like celebrities and politicians.

Consider the issue of consent. Many harmful AI depictions of private individuals use public source photos that victims post online. Victims may disapprove of the fictional, yet believable, photos and videos that generative AI tools produce of them, but existing legal claims may not provide the remedies these victims expect. Although the legal right of publicity originally protected the privacy and dignity of individuals, for example, some modern courts have focused their attention on the economic interest that a victim holds in their identity—namely, celebrities' economic interest in their public image, which others may appropriate for their own commercial gain.[35] These courts and similar state appropriation laws may not provide the easy legal remedy that victims expect when facing nonconsensual deepfakes; they may expect the victim to show some economic or physical injury in addition to their lack of consent, or they may expect the deepfake creator to have benefited financially. These laws and judicial interpretations did not develop with generative AI in mind, meaning that even AI harms that should be easy to remedy can become complex, costly, and confusing for victims. Of course, victims of malicious deepfakes and other AI-generated content can still pursue several other legal claims, such as defamation or negligent infliction of emotional distress, but the generative nature of new AI tools suggest that even these claims may face legal hurdles. The novelty and scalability of generative AI can be obstacles for victims of malicious deepfakes, even when their underlying legal claim is strong.

Defamation is yet another example of a legal claim made more challenging by generative AI. While private individuals may hold the creator of a defamatory deepfake liable so long as the depiction was false and harmed the victim, public figures like celebrities and politicians must overcome a higher First Amendment hurdle to get redress. In *New York Times Co. v. Sullivan*, for example, the Supreme Court held that public figures had to show that a defendant published defamatory material with actual malice—in other words, "with knowledge that it was false or with reckless disregard of whether it was false or not."[36] And in *Hustler Magazine, Inc. v. Falwell*, the Supreme Court applied the same standard to defeat a claim of intentional infliction of emotional distress.[37] However, the actual malice standard applied in these cases developed based on assumptions about what a reasonably prudent person could do to investigate and uncover the truth of information they receive. As generative AI tools grow more sophisticated, it will only become more difficult for individuals and press organizations to tell whether something is real or generated by AI, effectively raising the hurdle that public figures must overcome to redress harms caused by defamatory deepfakes.

Importantly, the malicious use of generative AI can impact everyone—private individuals and public figures alike. The distinction between private individuals and public figures within the law is far from clear, and both private individuals and public figures have successfully overcome the First Amendment, privacy, and consent hurdles discussed above.[38] These cases and the legal tests they implicate merely highlight legal assumptions that may not hold true when someone uses generative AI to impersonate, harass, defame, or otherwise harm others—legal assumptions that may impose barriers to redress and perpetuate AI-generated harm. While many traditional legal remedies may still be available for victims of malicious deepfakes and other generative AI harms, the novel legal questions that generative AI raises—as well as the potentially massive volume of violations that a publicly available generative AI tool can produce—will no doubt make these legal remedies harder to pursue and less effective in practice.

BELIEVABILITY

Deepfakes can impose real social injuries on their subjects when they are circulated to viewers who think they are real. Even when a deepfake is debunked, it can have a persistent negative impact on how others view the subject of the deepfake.[39] And the believability of AI-generated content can undermine victims' ability to pursue legal redress as well. The proliferation of generative AI and deepfakes undermines core assumptions about how legal fact-finding and the authentication of evidence occurs.[40] Currently, the bar for authenticating courtroom evidence is not particularly high.[41] All a claimant must show is that a reasonable juror could find in favor of authenticity or identification,[42] after which point the determination of authenticity is up to the jury.[43] In addition, many courts have adopted assumptions about the authenticity of aural and visual evidence that deepfakes undermine. For example, some courts recognize the "silent witness" theory of video authentication, wherein the existence of a recording speaks to the evidence's authenticity without the need for a human witness's observations.[44] Others assume the authenticity of evidence taken from press archives or government databases, both of which may be vulnerable to deepfakes.[45] As AI-generated content grows more common and more believable, courts and regulators alike will need to identify and adopt methods to determine whether images and videos are real and reconsider legal assumptions about the truth and value of evidence submitted at trial.

HARMS

- **Physical**: In some contexts, believable deepfakes of the victim seeming to engage in certain behaviors may put them at risk of physical harm and violence, for example, in cultures where publicly known sex acts would shame the family or in cultures where same-sex relationships are illegal.
- **Economic/Economic Loss**: Distribution of AI-generated fake images and videos that are pornographic in nature or touch on hot-button political or social topics could lead to job loss for the victim as well as trouble finding future employment.
- **Reputational/Relationship/Social Stigmatization**: Victims' standing in the community, intimate and professional relationships, and dignity could all be severely damaged or destroyed if, for example, deepfakes convinced others that person was cheating on a partner or engaging in illicit acts with minors.
- **Psychological**: Victims of these attacks often feel severely violated and may face feelings of hopelessness and fear that their lives have been destroyed.
- **Autonomy/Loss of Opportunity**: Deepfakes have already been weaponized to intentionally silence journalists, activists, and other vulnerable individuals and can lead to loss of opportunity and change in life circumstances if believed broadly. This can also contribute to a threat to democracy and social change.

- **Autonomy/Discrimination**: Deepfakes can easily be tools used to target already-vulnerable individuals belonging to marginalized groups or to make individuals appear to belong to marginalized groups—they also may reinforce negative attitudes about sex work and sex workers.

EXAMPLES

- The European Union's police force issued an official warning that "grim" criminal abuse using ChatGPT and other generative AI tools is here and growing.[46]
- A Twitch streamer made Deepfake porn of another Twitch streamer, imposing her face onto porn and passing it off as if it was her.[47]
- A TikTok user spoke out about digitally created nude photos of her shared on the internet. The photos were used to threaten and blackmail her.[48]
- Video game voice actors had their voice taken and used to train an AI to use their voice to harass and expose information about them, all without their knowledge or consent.[49]

INTERVENTIONS

- **Technological solutions** include deepfake detection software and methods for watermarking AI-generated content. These solutions may help victims, courts, and regulators identify AI-generated content, but the effectiveness of these solutions depends entirely on technical experts and responsible AI actors developing innovative detection and authentication tools faster than malicious AI developers can develop new, harder-to-detect AI tools.
- Many **longstanding legal tools** may still apply despite the novel features of generative AI tools and the legal challenges they impose. For example, deepfakes that exploit copyrighted content—potentially including photos that victims took of themselves[50]—may be vulnerable

to traditional **copyright claims**. Depending on the circumstances surrounding the AI-generated content, victims may also turn to various **tort claims** like defamation, false light, intentional infliction of emotional distress, and appropriation of name and likeness.[51] To circumvent the challenge of identifying anonymous creators, victims may be able to **sue the online platforms that host and circulate malicious AI-generated content** if the platforms—including the providers of AI tools like Midjourney and Runway— materially contributed to what makes the content harmful or otherwise illegal.[52] And several **criminal laws**, from criminal impersonation and fraud statutes to incitement to violence, could apply to claims involving the malicious circulation of AI-generated content.[53]

- Several **regulatory interventions** may further protect victims of deepfakes and other malicious uses of generative AI. While a general ban on deepfakes or generative AI tools may run afoul of the First Amendment,[54] expanding claims under **copyright law or privacy torts** to cover fictional depictions of victims created with reckless disregard to the content's impact on victims would go far to redress the harms caused by malicious uses of generative AI. **Criminal statutes** could also be updated or complemented with statutory language that captures the issues raised above, including language that lowers the intent required to hold someone liable for nonconsensual, AI-generated sexual depictions of another person. And given the difficulty in identifying believable deepfakes and authenticating evidence, the **Federal Rules of Evidence** may benefit from higher authentication standards to counteract possible deepfakes. Lastly, malicious deepfakes and other AI-generated content created for commercial purposes could be regulated by administrative agencies like the Federal Trade Commission and state Attorneys General Offices on the grounds that they are unfair and deceptive.[55]

Spotlight: Section 230

Section 230 of the Communications Decency Act says that a provider of an interactive computer service is not to be "treated as the publisher or speaker of information provided by" a third party.[56] Historically, companies—and courts—have taken an expansive view on what it means to treat a company as the publisher or speaker of information—basically, if the lawsuit had anything to do with third-party provided content, companies claimed Section 230 immunity. In recent years, courts have begun to cabin Section 230's reach, finding instead that companies can only claim Section 230 immunity if the basis for liability is dissemination of improper information that the company played no role in making improper.[57]

Generative AI tools do not get blanket immunity: Some commentators have framed the generative AI Section 230 debate as an all-or-nothing determination, with some proclaiming that generative AI tools receive Section 230 immunity[58] and others proclaiming they do not.[59] But judges in recent major court decisions have declined to apply Section 230 in such a broad manner. Instead, courts apply Section 230 on a claim-by-claim basis.[60] Thus, whether a company will get Section 230 protection will depend on the specific facts and legal obligations at issue, not simply whether they have deployed a generative AI tool.

Section 230 should not apply to some claims, like products liability claims, because they do not treat the company as the publisher or speaker of information: In the past, courts have applied Section 230 very broadly, largely by reading the provision to mean that a company is treated as the publisher or speaker whenever their allegedly unlawful activities involved the dissemination of third-party information. Courts have begun to backtrack on this and are recognizing that Section 230 does not protect against claims that target a company's own obligations not to cause harm.[61] Thus, claims that generative AI companies violated their own duties regarding the design of their service, the collection, use, or disclosure of information, and the creation of content should not be barred by Section 230.

For instance, generative AI companies will have difficulty using Section 230 to escape product liability claims—such as for negligent design or failure to warn—at least in the Ninth Circuit, where courts now recognize that such claims are based not on harm caused by third-party information but on a company's breach of their duty to design products that do not pose an unreasonable risk of injury to consumers.[62] Generative AI companies should also have to face claims that they violated privacy laws that limit how generative AI companies can collect, use, and disclose personal information because these laws impose duties on companies to respect the privacy interests of third-parties.

Generative AI companies will not get Section 230 protection when the tool is wholly responsible for creating the content: Generative AI companies could potentially face several different types of claims about the information that their tools generate. Section 230 provides companies with protection for legal claims based on information provided by another party—another "information content provider," in Section 230 lingo. An information content provider is defined as "any person or entity that is responsible, in whole or in part, for the creation or development of information" provided to the company.[63] So, a generative AI company does not get Section 230 protection if it is, itself, an information content provider of the information at issue—that is, if the company "is responsible, in whole or in part, for the creation or development of the information."

When a generative AI tool is alleged to have created new harmful content, such as when it "hallucinates" or makes up information that is not in its training data,[64] the legal claim is not based on third-party information and Section 230 should not apply. For example, when a generative AI tool makes up false and reputationally damaging information about an individual, the generative AI company will not be protected by Section 230 for, say, defamation or false light, because the company, and not any third party, is responsible for creating the false and reputationally damaging information that is the basis for the legal claim.

Generative AI companies will not get Section 230 protection when they materially contribute to the improper content: In some cases, generative AI companies will try to argue that the outputs at issue originated with a third party, either as user input or training data.[65] In such cases, courts will have to determine whether the company created or developed the information in part. The prevailing test is whether the company materially contributed to making the information improper.[66] Material contribution can include altering or summarizing third-party information to make it violate a law,[67] requiring or encouraging the third party to input information that violates a law,[68] or otherwise acting in a way that contributes to the illegality.

When a user asks that a generative AI tool create misinformation or a deepfake, or when a tool uses training data to create harmful content, the tool transforms inputs into harmful content and the company that deployed it should not be able to use Section 230 to avoid liability. The user inputs—the request for harmful information, the photos or videos of the target of a deepfake—are not themselves harmful or sufficient to create the harmful content. After all, that is why the user is using generative AI to create the content. The inputs are also unlikely, on their own, to be sufficient to form the legal basis for the claim against the generative AI company. In such scenarios, a company deploying a generative AI tool materially contributes to the improper information by transforming information that cannot form the basis for liability into information that can form the basis for liability.

If, on the other hand, a user asks a generative AI tool to simply repeat a defamatory statement that the user enters into the tool, or to repeat harmful information from other sources, the tool may not materially contribute to the harm and may, consequently, benefit from Section 230 protection.

Section 230 should not be an obstacle to holding companies accountable for harms caused by generative AI tools. Any new regulation or claim should be stated in terms of the generative AI company's obligations and the harm the tool itself caused by generating harmful content.

It is not clear that scraped training data is information "provided by" a third party: To obtain Section 230 protection, a company must show that the information that forms the basis for liability was "provided by" a third party. There is very little precedent on the question of when information has been "provided by" a third party.[69] To "provide" information can mean to supply it or make it available to another.[70] Generative AI companies would likely argue that publicly available information is made available for everyone to republish, including generative AI tools. But it is not at all clear that third parties intend to make their information available to generative AI tools simply by making their information viewable by a general audience on the internet—in fact, in many cases it is clearly the opposite.

The relationship between the internet company and the third-party information provider matters for determining whether the third party provided the information.[71] The types of services that Section 230 originally contemplated had users that directly provided information to the service, such as the Prodigy message boards that were the basis of the case that inspired Section 230.[72] Search engines and other types of services that third parties do not provide information directly to have also been found to enjoy some Section 230 protection,[73] but even these companies afford third parties some control over whether and to what extent their information is published or republished on their services. For example, websites can tell Google's search engine crawlers not to index their pages,[74] but there is no effective means to block an AI company from scraping their site.[75] The lack of control third parties have over the use of their information in generative AI tools, along with similar considerations described in [privacy section], could sway courts against finding that scraped data is "provided by" third parties.

Profits Over Privacy: Increased Opaque Data Collection

BACKGROUND AND RISKS

Generative AI tools are built on top of a variety of large, complex machine-learning models, which need a large amount of training data to function. For tools like ChatGPT, the data includes text scraped from across the internet. For products like Lensa or Stable Diffusion, the data includes photos and art. With generative AI's voracious need for data, many AI developers may scrape the web indiscriminately for data. While, in some cases, these developers attempt to sanitize their training data by filtering out protected work, explicit content, hate speech, or biased inputs, the practice of cleaning data is far from industry-standard. Without meaningful data minimization or disclosure rules, companies have an incentive to collect and use increasingly more (and more sensitive) data to train AI models. The excuse for collecting this data indiscriminately—increasing competition and innovation within the AI space—is harmful to the state of data privacy. This arms race narrative creates a justification for maximizing data collection just in case it provides some nebulous advantage later. In reality, these tools can be built with less data and without coercive and secretive data collection processes.

SCRAPING TO TRAIN DATA

Many generative AI tools use models built on data scraped from publicly available websites. This information often includes personal information posted on social media and other websites. People post information on social media and elsewhere for a variety of reasons: to allow potential employers to find them on LinkedIn; so that friends and acquaintances can find them on Facebook, Twitter, and Venmo; and so forth. These reasons have an important common feature: people post information on a website for the purpose of making that information viewable on that website. But sometimes, a person's personal information is made publicly available without their consent. Third parties might publish their photo or other information about them. A platform's confusing privacy settings may lead a person to accidentally make their information available. A software error[76] or design change[77] can also expose information that a person had set to be viewable only to a select few.

When companies scrape personal information and use it to create generative AI tools, they undermine consumers' control of their personal information by using the information for a purpose for which the consumer did not consent. The individual may not have even imagined their data could be used in the way the company intends when the person posted it online. Individual storing or hosting of scraped personal data may not always be harmful in a vacuum, but there are many risks. Multiple data sets can be combined in ways that cause harm: information that is not sensitive when spread across different databases can be extremely revealing when collected in a single place, and it can be used to make inferences about a person or population. And because scraping makes a copy of someone's data as it existed at a specific time, the company also takes away the individual's ability to alter or remove the information from the public sphere.

The privacy harms that follow from indiscriminate scraping of personal information for AI training data also create risks for online speech and the openness of the internet. As AI tools use people's personal information for more and more harmful purposes, people may become more hesitant to share any information on social media or sites that could potentially be scraped in the future, even if those sites promise to secure their data. They may be less likely to post photos of themselves, to participate in public debates on "the vast public forums of the internet"[78]—particularly social media—or to have social media profiles or personal websites that can be associated with them at all. Disincentivizing people from engaging in public discourse and interacting online will limit the usefulness of the internet as a whole and networking tools in particular.

Basic data minimization principles dictate that peoples' personal information should only be collected or used for the specific purpose for which each person provided the information. But there are currently no statutes that prohibit companies from scraping people's personal information and using it to train generative AI tools. Privacy laws in the U.S. exempt most publicly available information from regulation based on a concern that collection and use of this information is protected by the First Amendment. But lawmakers underestimate the significant countervailing privacy interests against allowing companies to indiscriminately scrape personal information.

People should be able to post public profile photos without fear that these photos will be used to create deepfakes of them or feed other abusive AI applications. Laws limiting the collection of publicly available personal information and/or its subsequent use would both protect people's interest in controlling their information and encourage people to continue to make information publicly available on the internet.

GENERATIVE AI USER DATA

Many generative AI tools require users to log in for access, and many retain user information, including contact information, IP address, and all the inputs and outputs or “conversations” the users are having within the app. These practices implicate a consent issue because generative AI tools use this data to further train the models, making their “free” product come at a cost of user data to train the tools. This dovetails with security, as mentioned in the next section, but best practices would include not requiring users to sign in to use the tool and not retaining or using the user-generated content for any period after the active use by the user.

GENERATIVE AI OUTPUTS

Generative AI tools may inadvertently share personal information about someone or someone’s business or may include an element of a person from a photo. Particularly, companies concerned about their trade secrets being integrated into the model from their employees have explicitly banned their employees from using it.

HARMS

- **Physical**: Individuals who may want to remove personal data for their own safety, such as domestic violence or stalking victims, may be unable to do so where data has been added to generative AI data sets and so may be at risk from their abusers.
- **Economic/Economic Loss**: Businesses whose trade secrets have been incorporated into training sets face potential economic loss.
- **Psychological**: Individuals unable to remove their personal data from training sets may face frustration or fear if the data could impact them negatively if spread.

- **Autonomy**: Individuals unable to block addition or force removal of their personal information from training sets demonstrably have lost control of their data.
- **Autonomy**: Individuals are often not informed, consulted, or given options about whether their personal data will be added to training datasets.
- **Autonomy/Loss of Opportunity**: Inability to remove data that is inaccurate or no longer accurate or make updates may result in incorrect outputs that then exacerbate as the incorrect information proliferates.

EXAMPLES

- The Italian Data Protection Authority began an enforcement action based on the EU's General Data Protection Regulation against OpenAI, banning the service in the country pending investigation. This led the company to institute some privacy disclosures and controls to the system.[79] Regulatory interest from bodies throughout the world will likely act as a catalyst to improved data protection behavior.
- Photos from private medical records were found in public database LAION-5B, which are used to make image generators.

INTERVENTIONS

- Enforce laws that prohibit unfair and deceptive trade practices, consent requirements for child users, and require justification for data processing.
- Enact laws and regulations that impose a data minimization standard that would limit use of personal data for generative AI training (e.g., the American Data Privacy Protection Act, the FTC commercial surveillance rule, and certain state privacy regulations)

- Support tools that are built using a limited and disclosed set of data.
- Adopt a strict data minimization standard by developers to help mitigate the privacy harms of creating, tweaking, and updating models to train AI. Data minimization is a standard that, depending on the precise definition, should only allow collection of personal data to the extent that it is necessary to carry out the service requested by the user. The tenets of data minimization are fundamentally at odds with the large-scale creation of generative AI datasets from public info without disclosure or consent.

Increasing Data Security Risk

BACKGROUND AND RISKS

The Identity Theft Resource Center estimated a record-breaking 1,862 data breaches occurred in 2021;[80] with another 1,802 in 2022.[81] Beyond the inherent privacy harms of a breach, there can be severe downstream impacts as well. A Government Accountability Office report indicated that victims have "lost job opportunities, been refused loans, or even been arrested for crimes they did not commit as a result of identity theft."[82] Yet these harms do not appear on the victim's bank statement or credit report, and can be nearly impossible to control where a Social Security Number (SSN) is used; by virtue of its unique and unchangeable nature, the SSN serves as a powerful identifier for both government and private sector entities. To make matters worse, a stolen SSN, unlike a stolen credit card, cannot be effectively cancelled or replaced.[83] Criminals in possession of SSNs can open new financial accounts and perpetrate identity theft because many financial institutions rely on SSNs to verify transactions.[84] Unsurprisingly, research by the Bureau of Justice Statistics indicates that identity theft can result in severe distress.[85]

The threat landscape has gotten worse in recent years as well, with the introduction of ransomware-as-a-service, malware-as-a-service, and other proxy services by which hackers-for-hire have productized methods for unauthorized access to data.[86] We should expect to see continued examples of purchasable tools by which data can be accessed, encrypted, and/or manipulated without authorization.

Just as every other type of individual and organization has explored possible use cases for generative AI products, so too have malicious actors. This could take the form of facilitating or scaling up existing threat methods, for example drafting actual malware code,[87] business email compromise attempts,[88] and phishing attempts.[89] This could also take the form of new types of threat methods, for example mining information fed into the AI's learning model dataset[90] or poisoning the learning model data set with strategically bad data.[91] We should also expect that there will be new attack vectors that we have not even conceived of yet made possible or made more broadly accessible by generative AI.

HARMS

- **Physical**: Where an individual is the victim of identity theft, they may face arrest for crimes they have not committed.
- **Economic/Economic Loss**: Victims may lose job opportunities or be refused loans as the result of identity theft and destruction of credit.
- **Reputational/Social Stigmatization**: Identity theft can cause severe reputational damage and malware can also be used to reveal sensitive information about an individual, resulting in additional social harms.
- **Psychological**: Victims of these attacks may face embarrassment and fear as well as feelings of helplessness, anger, and more due to the results of such attacks.
- **Autonomy**: Loss of identity control, financial control, image, and more may accompany these attacks.
- **Discrimination**: Scams often target historically vulnerable groups, such as the elderly.

EXAMPLES

- ChatGPT suffered a massive data breach, exposing user information and prompt history.[92]
- Samsung banned use of AI after secure information was found leaked to ChatGPT by employees.[93]
- OpenAI is allowing users to get information from users through plug-ins that let the chatbot get new sources of information like Expedia and Instacart.[94]

INTERVENTIONS

- If companies invest in employee training and patching known vulnerabilities, they can mitigate some of the risk of generative AI super-charging existing threat methods. However, risks related to the use of the AI model itself will require distinct solutions, including but not limited to those outlined in NIST's AI Risk Management Framework[95] and those required in the proposed American Data Privacy and Protection Act (ADPPA).[96]

Confronting Creativity: Impact on Intellectual Property Rights

BACKGROUND AND RISKS

Intellectual property law (IP law) encompasses copyrights, patents, trademarks, and trade secrets. Loosely, copyrights protect original works in any medium of expression (think books, music, theater, and artwork), patents protect inventions, trademarks protect any words or symbols used to identify the source of specific goods and services, and trade secrets protect proprietary business information.[97] Each of the branches of IP law contain specific rights for the creators and owners of a work over that work—for example, controls over how that work is used or preventing others from claiming the work as theirs. While all areas of IP law have challenged generative AI use and generation of works, copyright has been by far the most frequently invoked.

The extent and effectiveness of legal protections for intellectual property have been thrown into question with the rise of generative AI. Generative AI trains itself on vast pools of data that often include IP-protected works. As stated in a recent open letter from the Center for Artistic Inquiry and Reporting, “AI-art generators are trained on enormous datasets, containing millions upon millions of copyrighted images, harvested without their creator's knowledge, let alone compensation or consent. This is effectively the greatest art heist in history."[98] The systems trained on these works may then learn to mimic specific styles, as has already occurred in several cases.[99] Several artists whose style has been copied have expressed deep frustration, anger, and dismay over their work being mimicked, noting that the AI is profiting off the work they have put in to develop distinct styles, impacting their livelihood, and reducing deeply personal work to an algorithm. In the words of Nick Cave, an artist confronted with a song generated "in the style of" Nick Cave, "This song is bullshit, a grotesque mockery of what it is to be human." [100]

Questions about how far IP protections extend in the realm of generative AI can be categorized into either the input or output cycle of these systems.

CASE STUDY – WHAT'S IN A VOICE?

An AI-generated song, billed as a "collaboration" between Drake and the Weeknd, popped up on Spotify, Tidal, Apple Music, and YouTube, quickly collecting enough listens and views over the weekend to appear on music charts by Monday. The song was generated by scraping multiple samples of both artists' voices and music, creating a realistic-sounding new hit. It was taken down after multiple copyright claims by Universal Music Group, leading to questions about whether an original song can be copyrighted and what protections exist for the artists whose voices and musical styles are being cloned.

INPUTS

Generative AI systems can produce extremely detailed and adaptable content because they are trained on enormous amounts of data scraped from across the internet. The type of data taken in will vary depending on the system type. For example, AI art generators will scrape art and images, translating information about their key features into code that is then reviewed by those systems for patterns, relationships, and rules, then used to generate responses to user prompts. Because these systems' outputs become more "accurate" or responsive with more data, many are programmed to scrape their preferred content type continuously and automatically.

These vast datasets nearly always contain protected works. The entities using the datasets to create a generative AI system rarely, if ever, have permission or license from the creators and owners of artistic works to use them. In fact, many artists have openly stated that they do not want their work going into systems that may make them obsolete.[101]

There is serious and ongoing debate over whether generative AI tools should be permitted to use protected works without a license. Some argue that such use constitutes fair use, an exception to some copyright protections with a very limited scope of application. Fair use often depends on the use of copyrighted material. For instance, a research or non-profit group using the content may have a better fair use claim than a company intending to sell the work generated using the original work. The extent to which fair use may apply to generative AI is still unsettled law.

OUTPUTS

End-users of generative AI have already attempted to claim ownership over the outputs of generative AI tools, including several who have attempted to file for copyrights with the United States Copyright Office.[102] The rising use of generative AI to create creative works and subsequent copyright filing attempts has been significant enough to prompt the Copyright Office to launch a new AI initiative.[103]

Statements from the U.S. Copyright Office so far have mandated that a work cannot receive copyright protections unless it contains "creative contribution from a human actor,"[104] noting that copyright may only protect material that is "the product of human creativity."[105] While some have argued that the prompt constitutes sufficient "human creativity" to result in IP protections for the resulting work, the Copyright Office disagrees, comparing a prompt to "instructions to a commissioned artist—they identify what the prompter wishes to have depicted, but the machine determines how these instructions are implemented in its output."[106]

This distinction becomes more complex when a portion of the work is AI-generated and a portion is human-generated.[107] Copyright may be applied to work that contains or builds off AI-generated work, but the copyright will apply solely to the human-authored aspects.[108]

HARMS

The harms to individual creators of works and to the artist community are substantial.

- **Economic/Economic Loss**: Legal harms in this space likely include infringement on works and right of use, along with questions about ownership of AI-generated products, as discussed above. Infringement on works would likely stem from generative AI outputs. These include unauthorized reproduction (which may be the case of the AI-generated work is too similar to original pieces) and derivative work (work which contains too many original elements from the initial piece, often seen in reproductions, condensations, or abridgments of original work). Right of use would relate to generative AI input, whether the generative AI systems have a license to use original work in learning sets or whether this could fall under an exception, such as Fair Use.
- **Economic/Economic Loss**: Creators and owners may face severe economic harms as demand for their work shrinks when similar work can be easily and cheaply generated. These harms likely would

manifest in lack of opportunity and hiring (as many creator jobs and commissions are replaced with generative AI) and infringement on economic gain from originally created work (missing out on licensing or selling work since buyers are using replicated work instead). This could also lead to a sharp drop in professional artists if there is rising fear that AI-generated work will make it too difficult to make a living as a creator. Finally, the influx of AI-generated work will impact the market for that work.

- **Reputational**: Creators may face reputational harm as well. It is entirely possible for fans to confuse AI-generated work with that of the inspiring creator, which is a problem when the creator has no part in that work and cannot give input regarding use, quality, or direction. Work generated in a specific creator's style or voice may be used to promote causes the creator disagrees with or may be of low quality, both of which could cause reputational damage.
- **Psychological**: Several artists have expressed pain, sadness, anger, and more regarding their work being used and replicated by generative AI. In many cases, artists' work is deeply personal, making copies and exploitation of that work deeply personal as well.
- **Loss of Opportunity/Relationship/Dignitary:** If creators and their work are not protected from exploitation and copying via generative AI, it will likely result in fewer artists putting in the time and effort to develop their own distinct art styles, leading to an overall drop in the creator community and volume of all creative works by humans.

EXAMPLES

- Several artists have found that their work has already been used to train AI without their permission, in some cases leading to AI that can convincingly replicate their precise artistic style when asked.[109]

- As noted in the case study above, this extends to AI-generated songs "performed" in exact mimics of artists' vocal tones and musical styles.[110]
- Artists have found AI copying their style or modifying their work in ways that make it seem as if it supports hateful messages, as with one artist who found the alt-right using AI-generative tools to create express offensive worldviews in her artistic style.[111]
- There are current examples of individuals attempting to claim copyright over work generated by AI tools.[112]
- Three artists have filed a class action in the Northern District of California against generative AI image companies for using their artwork without consent or compensation to build the training sets that inform the platforms.[113]
- Getty Images has started legal proceedings against Stability AI for copying and processing millions of its images for training sets without license.[114]

INTERVENTIONS

- AI developers could be forced to license any IP included in training data for generative AI technology, which would prevent indiscriminate, continuous content scraping across the internet.
- Customers could be obligated to perform a form of due diligence to confirm whether models were trained with any protected content prior to using the model.
- AI tools could be forced to acquire creator permission before generating art "in the style" of any specific creator.
- Academic researchers at the University of Chicago developed a tool called Glaze that introduces nearly imperceptible elements to artwork that are designed to disrupt generative AI's ability to scrape information about the artwork and add it to the training data.[115]

- Shutterstock is putting together the option for creators to opt out of their work being used in AI training sets and has established a contributor fund to compensate creators if their IP is used in training sets.[116]
- DeviantArt has implemented a metadata tag for images shared on the web that contains a warning to AI systems not to scrape the tagged content.[117]

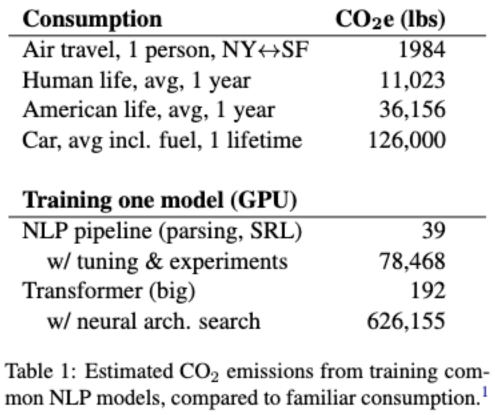
Exacerbating Effects of Climate Change

BACKGROUND AND RISK

The planet is hurtling toward ongoing climate disasters.[118] Climate change has already caused hundreds of thousands to millions of deaths, billions of dollars in economic damage, and mass species extinction.[119] In the future, every tenth of a degree of warming that we are able to avoid will mean millions of saved lives, avoidance of enormous economic loss, and a chance at a livable future.[120] Eventually, by choice or by necessity, our society will evaluate every industry and activity in terms of its resource and carbon cost.

Into this high-risk situation crashes the growing field of generative AI, which brings with it direct and severe impacts on our climate: generative AI comes with a high carbon footprint and similarly high resource price tag, which largely flies under the radar of public AI discourse.

Training and running generative AI tools requires companies to use extreme amounts of energy and physical resources. Training one natural language processing model with normal tuning and experiments emits, on average, the same amount of carbon that seven people do over an entire year.[121]



Consumption	CO_2e (lbs)
Air travel, 1 person, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experiments	78,468
Transformer (big)	192
w/ neural arch. search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

[122]

A comparison of pounds of carbon dioxide produced by everyday people/objects and generative-AI-related tasks

AI models take an enormous amount of carbon to produce, and this trend is not likely to meaningfully improve given the incentives in the industry. Much AI research, especially at large tech companies that effectively control the space, focuses on accuracy or related measures at the expense of all other considerations.[123] A good portion of AI research seeks to "buy" better results by investing exponentially more money over time for a linear increase in accuracy, disregarding other externalities like resource costs.[124] These costs are physical requirements for many AI systems: the relationship between model performance and model complexity is at best logarithmic, so for a linear gain in performance, an exponentially larger and more resource-inefficient model is required.[125] While some AI researchers have begun focusing on efficiency, whether for cost-cutting or environmental reasons, there is no reason to think that large tech companies will abandon their quest for accuracy any time soon.

Meanwhile, the data centers that AI developers use to train and host generative AI models have high energy costs and massive carbon footprints. Though some of this energy may come from renewable resources, data centers' energy consumption is still concerning for several reasons. First, many regions that house data centers still use carbon-intensive energy sources to generate electricity.[126] Second, even when renewable energy is available, it may be better allocated to heat a family's home, power a greenhouse, or further other socially important goals, rather than train an AI model—but this tradeoff is generally not examined or discussed.[127]

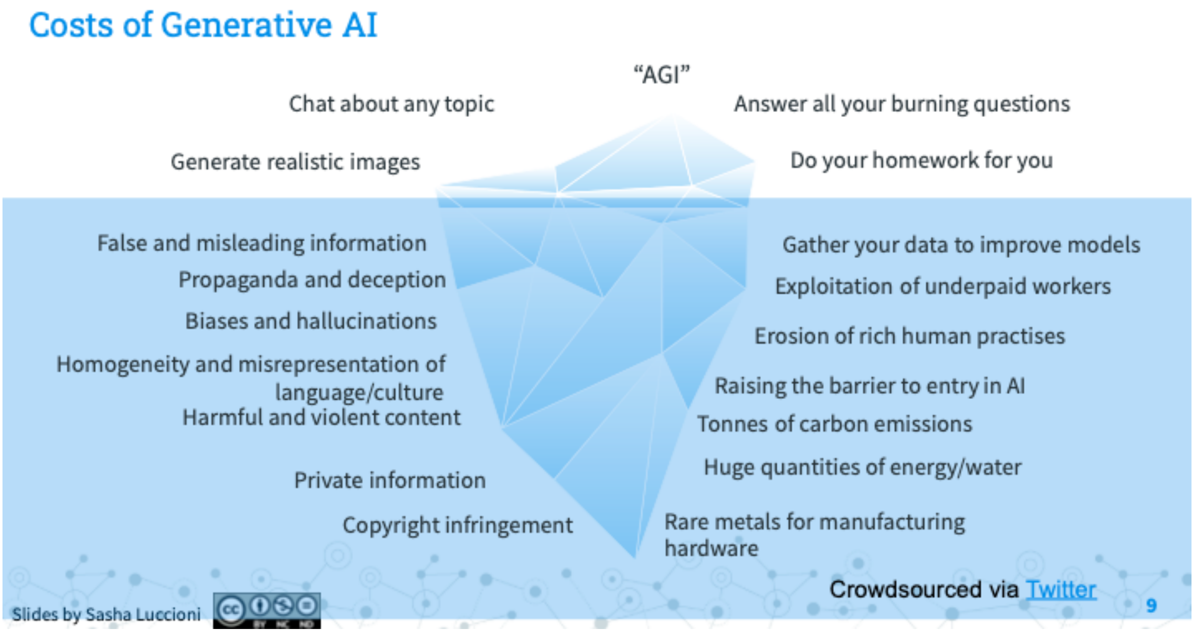
These data centers are also resource-intensive in unsustainable ways. Many tech firms draw from public water supplies to cool their centers in the middle of drought-prone areas—a practice that has led to public backlash.[128] "New research suggests training for GPT-3 alone consumed 185,000 gallons (700,000 liters) of water. An average user's conversational exchange with ChatGPT basically amounts to dumping a large bottle of fresh water out on the ground, according to a new study."[129] These technologies also rely heavily on minerals that are procured under violent and exploitative conditions.[130]

HARMS

- **Physical**: Severe environmental changes will result in substantial physical harms to people globally (drought, natural disasters, etc.).
- **Economic/Economic Loss**: The economic resources required to counter environmental harms or to run generative AI as-is are significant.
- **Autonomy**: So many limited resources going to large companies using them for generative AI necessarily means that others will have less access and suffer shortages.

EXAMPLES

- Sasha Luccioni, the Climate Lead at HuggingFace, evaluates environmental and societal impact of Generative AI – she highlights "tonnes of carbon emissions," "huge quantities of energy/water," and "Rare metals for manufacturing hardware" in her Iceberg model of Generative AI costs.



Costs of Generative AI

"AGI"

Chat about any topic

Answer all your burning questions

Generate realistic images

Do your homework for you

False and misleading information

Propaganda and deception

Biases and hallucinations

Homogeneity and misrepresentation of language/culture

Harmful and violent content

Private information

Copyright infringement

Gather your data to improve models

Exploitation of underpaid workers

Erosion of rich human practises

Raising the barrier to entry in AI

Tonnes of carbon emissions

Huge quantities of energy/water

Rare metals for manufacturing hardware

Crowdsourced via Twitter

Slides by Sasha Luccioni

9

Figure 1: Credit: Sasha Luccioni

INTERVENTIONS

- Because of environmental disruption, the Dutch Government imposed a nine-month moratorium on large data centers in the country in order to stand up regulations.[131]
- Tech companies should be required to track and publish the amount of energy and resources their models and data centers are using.
- Conferences should require tracking of resources to develop and run a system.
- Academic researchers should be given equitable access to computational resources. As of now, academics do not have sufficient access to understand the specifics of how modern AI tools work and what resources they require. This knowledge is hoarded inside large tech companies. Without this knowledge and access, the focus is kept on profit/accuracy, not environmental concerns. Sunshine is the best disinfectant, and academics who understand computer science are a useful window to let the sunshine in.

Labor Manipulation, Theft, and Displacement

BACKGROUND AND RISKS

Recent clickbait headlines playing into fears and hype trumpet that generative AI is coming for people's jobs.[132] While generative AI will disrupt the way certain industries work, it is still too early to see how this technology will impact labor markets and integrate into existing work.

When it comes to labor and market dominance, large tech companies like Apple, Meta, Amazon, Google, and Microsoft employ much of the AI research and development industry. These companies are directing this specialized workforce to develop commercial AI products that can be used for private profit rather than public benefit.

Major tech companies have also been the dominant players in developing new generative AI systems because training generative AI models requires massive swaths of data, computing power, and technical and financial resources. Their market dominance has a ripple effect on the labor market, affecting both workers within these companies and those implementing their generative AI products externally. With so much concentrated market power, expertise, and investment resources, these handful of major tech companies employ most of the research and development jobs in the generative AI field. The power to create jobs also means these tech companies can slash jobs in the face of economic uncertainty. And externally, the generative AI tools these companies develop have the potential to affect white-collar office work intended to increase worker productivity and automate tasks.

GENERATIVE AI IN THE WORKPLACE

The development of AI as whole is changing how companies design their workplace and business models. Generative AI is no different. Time will tell whether and to what extent employers will adopt, implement, and integrate generative AI in their workplaces—and how much it will impact workers.

Still, early signs suggest that generative AI will change white-collar work. Many white-collar workers have already begun to embrace generative AI to help with daily tasks like drafting presentations, marketing materials, speeches, emails, conducting research, and even coding. A Fishbowl survey found that 43% of working professionals have used generative AI tools to complete tasks at work and 70% of those respondents do so without their boss's knowledge.[133] News outlets and websites have used ChatGPT to write whole or part of articles.[134] Hiring managers are turning to generative AI to help write job descriptions and draft interview questions, among other administrative tasks.[135] Lawyers are using generative AI to do research, administrative tasks, and even draft contracts.[136] And in medicine, physicians are using generative AI for research and summarizing patient visits.[137]

The proliferation of generative AI has also created a demand for workers with experience using the tools and entirely new jobs built around these tools. According to a ResumeBuilder.com study, nine out of ten surveyed companies are currently seeking workers with ChatGPT experience.[138] The rise in generative AI tools has also created a growing demand for "prompt engineers"—people who train AI chatbots to test and improve answers or otherwise facilitate better prompt-inputs for large language models like ChatGPT.[139] In fact, there is already a prompt database where people can sell their own prompts to produce better results.[140]

Not all employers are jumping on the generative AI bandwagon. Some workplaces are cautious to quickly adopt this technology due to concerns about reliability, as the technology sometimes responds to prompts with misinformation or wrong answers. Other employers have expressed concern over security risks and restricted employee use. Workplaces like J.P. Morgan, Chase & Co., Bank of America, Citigroup, and Verizon prohibited employees from using ChatGPT.[141] Samsung banned generative AI tools after employees uploaded sensitive data to ChatGPT, expressing concern that the data transmitted is stored on external servers where it is difficult to retrieve or delete and could be leaked to others.[142]

The overall effect of generative AI on the economy remains to be seen. Some experts have said unregulated and freely deployed generative AI can cause harm to competition, push down wages, and lead to excessive automation and inequality.[143] But when discussing potential risks of generative AI on labor, there needs to be a distinction between whether generative AI tools lead to automation or augmentation of job roles. Since the 1980s, a significant portion of income inequality has been driven by automation.[144] When generative AI is used for automation, potential risks include job loss as well as the devaluation of labor and heightened economic inequality.

JOB AUTOMATION INSTEAD OF AUGMENTATION

There are both positive and negative aspects to the impact of AI on labor. A White House report states that AI "has the potential to increase productivity, create new jobs, and raise living standards," but it can also disrupt certain industries, causing significant changes, including job loss.[145] Beyond risk of job loss, workers could find that generative AI tools automate parts of their jobs—or find that the requirements of their job have fundamentally changed.

The impact of generative AI will depend on whether the technology is intended for automation (where automated systems replace human work) or augmentation (where AI is used to aid human workers). For the last two decades, rapid advances in automation have resulted in a "decline in labor share, stagnant wages[,] and the disappearance of good jobs in many advanced economies."[146] AI used exclusively for automation could exacerbate these negative trends.[147]

Some studies suggest that AI may lead to reduced hiring if the technology replaces many routine tasks previously performed by workers.[148] But other studies suggest that AI could create new opportunities—particularly in high-skilled jobs—and increase worker productivity.[149] Proponents of generative AI say that technology like ChatGPT can automate repetitive tasks and free more time for people to focus on complex or creative tasks. However, employers trying to reduce costs, maximize profits, and increase shareholder value are more likely to prioritize AI technology that automates rather than augments work.

While it is still too early to determine whether AI will either significantly devalue or fully replace workers, preliminary research shows that generative AI does impact job-related tasks. According to research by OpenAI, "80% of the U.S. workforce could have at least 10% of their work tasks affected" by large language models, and this effect is projected to span all wage levels across industries.[150] The OpenAI paper also found that "approximately 19% of workers may see at least 50% of their tasks impacted."[151]

A Goldman Sachs report states that generative AI could impact as much as 300 million jobs.[152] Generative AI could substitute a quarter of current work, with white-collar workers in administrative and legal sectors most likely to be affected.[153] The Goldman Sachs report also shows that AI will impact the labor market more generally, but the report emphasizes that the impact depends greatly on the technology's capabilities and how it is adopted.

DEVALUATION OF LABOR & HEIGHTENED ECONOMIC INEQUALITY

Technological advancement to accelerate productivity, automate jobs, and increase profitability by reducing costs began way before the generative AI boom. Historically, automation is one of the clearest factors in wage decline. According to a White House report, much of the development and adoption of AI is intended to automate rather than augment work.[154] The report notes that a focus on automation could lead to a less democratic and less fair labor market.[155]

Consider the potential labor impacts that generative AI is having in the software engineering industry, where many start-ups are using GPT-4 to spend less on human programmers.[156] While generative AI will not replace all software engineers anytime soon, it will impact the accessibility of learning code, how much programmers' services cost, and how in-demand human programmers are.[157] Beginner coders could benefit from using generative AI to help them learn code, but more experienced programmers may find the value of their labor decrease with increased competition.

In 2021, OpenAI CEO Sam Altman predicted that there would be an "unstoppable" technological AI revolution where the price of many types of labor "will fall toward zero once sufficiently powerful AI 'joins the workforce.'"[158] Altman elaborates that, since labor is the driving cost of the supply chain, AI performing tasks will lower the cost of goods and services.[159] He acknowledged that, if public policy does not adapt to such a predicted revolution, "most people will end up worse off than they are today." This prediction shows how the CEO of the leading generative AI company is viewing the future—a future where economic inequality is accelerated by AI.

In addition, generative AI fuels the continued global labor disparities that exist in the research and development of AI technologies. Outsourcing labor to subcontractors in the Global South for the benefit of the Global North is a problem inherent in the tech industry—and the entire global economic ecosystem more broadly. Labor that is deemed simple and routine is often outsourced to locations where workers are forced into terrible working conditions with low wages. The AI supply chain reflects and reproduces the inequities of imperial colonialism, where the Global North, wielding greater economic power, profit from the proliferation of AI technology while excluding the Global South.[160]

The development of AI has always displayed a power disparity between those who work on AI models and those who control and profit from these tools.[161] Overseas workers training AI chatbots or people whose online content has been involuntarily fed into the training models do not reap the enormous profits that generative AI tools accrue.[162] Instead, companies exploiting underpaid and replaceable workers or the unpaid labor of artists and content creators are the ones coming out on top. The development of generative AI technologies only contributes to this power disparity, where tech companies that heavily invest in generative AI tools benefit at the expense of workers. For instance, OpenAI is projected to make $1 billion in revenue by 2024.[163]

But collective worker action around AI is growing. For example, over 150 content moderation and data label employees in Africa recently voted to unionize.[164] Even more, the Writers Guild of America went on strike partly over the studios refusal to negotiate on banning the use of AI to generate scripts and using the writers' written work to train AIs.[165]

HARMS

- **Economic/Economic Loss/Loss of Opportunity**: Outsourcing labor to generative AI may lead to job loss and job replacement on a global scale, including impacting jobs that are currently being outsourced to other countries.
- **Autonomy/Loss of Opportunity**: Entire industries may be affected by workplace demands for generative AI, meaning that those specializing in certain industries may be unable to find work and have to shift to new venues, possibly meaning education, training, and experience in that field is "wasted."

EXAMPLES

- Sama, a San Francisco-based firm, hires workers in Uganda, Kenya, and India to label data for tech companies like Microsoft, Meta, and Google.[166] OpenAI outsourced work to Sama where Sama paid Kenyan workers less than $2 per hour to label data to help make ChatGPT less toxic.[167] The company subjected workers to traumatic content moderation practices where workers read and labeled textual descriptions of hate speech, violence, and sexual abuse.[168] Workers became mentally scarred by the distressing nature of the work, with one Sama worker describing it as "torture."[169] The traumatizing nature of the work led Sama to eventually end its relationship with OpenAI in February 2022, ceasing work eight months early.[170]
- In a survey of 1,000 U.S. business leaders by ResumeBuilder.com, half of companies surveyed are using ChatGPT while 30% plan to and 48% have replaced workers with ChatGPT.[171]

- In January 2023, BuzzFeed said it would use ChatGPT to create quizzes and personalized content for its readers and employees expressed concern about whether this move would lead to a reduction in the workforce.[172] At that time Buzzfeed contended that it remains "focused on human-generated journalism in its newsroom,"[173] but since then BuzzFeed shut down its news division as part of a 15% reduction of its workforce.[174]

INTERVENTIONS

REDISTRIBUTE POWER AND PROFITS AMONG ALL PARTICIPANTS

- Workplaces should not use generative AI as a means to cheapen labor costs and devalue workers' contributions. In fact, wages should be increased to match the increased worker productivity from generative AI.
- Tech companies should include voices and give decision making power to those who are actually working on the development and training of generative AI, especially those in the Global South. Major companies need to elevate the involvement and participation of workers to ensure equity.
- Technology vendors and service providers should invest in AI research and development that improves worker productivity rather than replacing job functions.
- If workplaces benefit economically from generative AI, companies should share in the profits with those whose labor they benefit from instead of concentrating it among shareholders and top earners.

INVEST IN PEOPLE

- Employers should invest in training and job transition services where they are training workers in new skills for jobs that have been changed by generative AI.

- Employers should invest in training where there is a growing demand for new jobs that have been created by generative AI (e.g., prompt engineers, machine managers, AI auditors, and AI trainers).
- Companies should implement policy programs where there is a commitment to invest in training to retain labor rather than to cut costs by reducing staffing in favor of generative AI technology.
- Companies, local and federal government, and other public-private programs should make commitments to invest in resources that help workers displaced by generative AI find alternative jobs.

INVEST IN COMPLEMENTARY AI

- Workplaces should be investing and implementing generative AI that augments and complements work rather than replaces work.
- Tech companies should invest in AI research and development that improves worker productivity rather than replacing job functions.
- Policymakers should regulate and redirect generative AI research to develop technology for public-interest use cases rather than for primarily commercial-use cases.

Spotlight: Discrimination

Artificial intelligence and other automated decision-making systems have long been deployed in opaque and unaccountable ways that harm individuals and exacerbate existing biases. Because AI is trained on historical data and often used by the resource controlling actor (hiring company, landlord, government benefits agency), Black people, women, individuals with disabilities, and poor people are hardest hit. And the harm isn't trivial—algorithmic systems have landed innocent Black men in jail, given lower credit limits to women, higher interest rates to graduates of Historically Black or Latino Colleges, and prevented people from receiving interviews or job offers.

An image generator is more likely to show a woman when you ask it to generate an image of a cleaner, and white men when you ask it to generate an image of a boss. Google's Bard text generator has replicated dangerous conspiracy theories. It recommended conversion therapy for gay people, generated text saying that Trans people are "groomers," and generated text claiming that major parts of the holocaust was fabricated.[175]

Generative AI is not appropriate for use in determinations for important life opportunities, but the public must remain vigilant in identifying inappropriate use of AI for these purposes – such as a Chatbot that stands as an arbiter for people on criminal justice or social services websites.

Discrimination is at the heart of every risk outlined in this paper, and the negative effects of security breaches, privacy violations, and environmental impacts will be felt most closely by marginalized communities.

The Potential Application of Products Liability Law

BACKGROUND AND RISK

Products liability is an area of law that developed throughout the twentieth century to respond to the harms that mass-produced products can impose at scale on society. This area of law focuses on three main harms: defectively designed products, defectively manufactured products, and defectively marketed products. Products liability law is characterized by two elements: (i) its adaptability and ability to evolve to address new types of products and harms, and (ii) its concern with distinguishing blameworthily harmful products from products which did harm but could not have been designed, manufactured, or marketed differently—or those that were simply used in an unreasonable way.

Like manufactured items like soda bottles, mechanized lawnmowers, pharmaceuticals, or cosmetic products, generative AI models can be viewed like a new form of digital products developed by tech companies and deployed widely with the potential to cause harm at scale. For example, generative AI products can cause harm to people's reputations by defaming them, directly abuse or facilitate abuse against people, violate intellectual property rights, and violate consumers' privacy.

Products liability evolved because there was a need to analyze and redress the harms caused by new, mass-produced technological products. The situation facing society as generative AI impacts more people in more ways will be similar to the technological changes that occurred during the twentieth century, with the rise of industrial manufacturing, automobiles, and new, computerized machines. The unsettled question is whether and to what extent products liability theories can sufficiently address the harms of generative AI.

So far, the answers to this question are mixed. In *Rodgers v. Christie* (2020), for example, the Third Circuit ruled that an automated risk model could not be considered a product for products liability purposes because it was not "tangible personal property distributed commercially for use or consumption."[176] However, one year later, in *Gonzalez v. Google*, Judge Gould of the Ninth Circuit argued that "social media companies should be viewed as making and 'selling' their social media products through the device of forced advertising under the eyes of users."[177] Several legal scholars have also proposed products liability as a mechanism for redressing harms of automated systems.[178] As generative AI grows more prominent and sophisticated, their harms—often generated automatically without being directly prompted or edited by a human—will force courts to consider the role of products liability in redressing these harms, as well as how old notions of products liability, involving tangible, mechanized products and the companies that manufacture them, should be updated for today's increasingly digital world.[179]

HARMS

- **Physical**: Generative AI may produce false information about individuals, leading to physical violence and danger, or could hold information that individuals are trying to delete for their own safety.

- **Economic/Economic Loss/Loss of Opportunity**: Generative AI leads to loss of income for artists whose style is mimicked and may cause job loss or shrinking of opportunities for work in certain industries, with no redress for individuals absent some form of liability.
- **Reputational/Relationship/Social Stigmatization**: Spread of incorrect information about an individual can severely damage their reputation, relationships, and dignity.
- **Psychological**: Impacts of generative AI harms may cause emotional distress, fear, helplessness, frustration, and other serious emotional harm.

INTERVENTIONS

- Scholars, policymakers, and plaintiffs' attorneys should explore ways that common law and statutory products liability law regimes can apply to redress generative AI harms. Products liability law itself may prevail, or instead a new doctrine based on some of its tenets, but either way, private people must have a remedy when they are harmed by generative AI.

Exacerbating Market Power and Concentration

BACKGROUND AND RISK

Developing, training, using, and maintaining generative AI tools is a resource intensive endeavor. In addition to the environmental costs discussed in the Environmental Impact section above, generative AI tools cost an incredibly large amount of money and computing resources to develop and maintain. To maintain the underlying computing power necessary to run ChatGPT, for example, experts have estimated that OpenAI must spend roughly $700,000 a day,[180] leading to an estimated $540 million loss for OpenAI[181] in 2022 alone. To compensate for this loss, OpenAI sought and received an investment from Microsoft of over $10 billion dollars, which included critically necessary and expensive cloud hosting services using Microsoft Azure.[182] OpenAI is reported to be seeking additional investments of $100 billion dollars as well. And it will cost Alphabet an estimated $20 million in computing costs to train its massive, 540-billion parameter language model, PaLM (Pathways Language Model).[183]

The astronomical cost of large-scale AI models means that only the biggest tech companies can handle—and afford—both the rapidly expanding needs of maintaining and controlling the models and the public relations and lobbying needs that recent generative AI advances require. One example of the entrenched power influencing public opinion about generative AI is the prevalence of the word, “foundational,” used to describe large models like GPT-4 and LAION B-5, among others. As the AI Now Institute explains, the term “foundational” was introduced by Stanford University in early 2022, in the wake of the publication of an article listing the many existential harms associated with large language models. Calling these models “foundational’’ aimed to equate them (and those espousing them) with unquestionable scientific advancement, a steppingstone on the path to “artificial general intelligence” (AGI)—another fuzzy term evoking science-fiction notions of replacing or superseding human intelligence. By describing generative AI tools as foundational scientific advances, tech companies and AI evangelists frame the wide-scale adoption of generative AI as inevitable.

In addition, many of the leading generative AI tools, as well as the training methods and cloud computing services that support them, are owned and maintained by a select few tech companies, including Amazon, Google, and Microsoft. The dominance of these few companies in not only developing generative AI but also providing the underlying tools and services that generative AI requires, further concentrates a market that, despite promoting “open source” technologies, is captured by a few powerful companies with opaque AI development methods and incentives to restrict competition.

HARMS

- **Economic/Economic Loss/Loss of Opportunity**: Concentration of power in only a few large companies means that any individual who either does not wish to work for those specific companies or has been rejected by those companies may be unable to work in the generative AI industry altogether.
- **Autonomy**: Big Tech’s monopoly over generative AI lessens the ability for competitors to develop or for others to have access to necessary resources.

- **Autonomy**: Choice is necessarily limited with fewer actors in the space.
- **Autonomy/Discrimination**: Any problems with data quality will be exacerbated through re-use and spread among the few dominant players.
- **Discrimination**: Any discriminatory behaviors in hiring or the workplace by Big Tech companies will more directly and strongly impact the field due to the limited opportunity to change employer or protest treatment and remain working in the field.

EXAMPLES

- The Wall Street Journal illustrates how the generative AI “race” will make Google and Microsoft richer and “even more powerful.”[184]
- The Federal Trade Commission is launching an inquiry into the Business Practices of Cloud Computing Providers.

INTERVENTIONS

- Enact laws that provide additional resources to and bolster authorities of Antitrust enforcers.
- Advocates and commentators should explicitly tie the data and computing resource advantage in coverage of the industry.
- Reform merger guidelines to reflect how consolidation of data advantages is considered in Antitrust reviews.
- Advocates and reporters should refrain from emboldening the “Arms Race” dynamic with China propagated largely from interested industry actors.

Recommendations

LEGISLATIVE AND REGULATORY

- Enact a law that makes intimidating, deceiving, or deliberately misinforming someone about an election or candidate illegal (regardless of the means), such as the Deceptive Practices and Voter Intimidation Prevention Act.
- Pass the American Data Privacy Protection Act – The ADPPA will limit the collection and use of personal information to that which is reasonably necessary and proportionate to the purpose for which the information was collected. Such limitation will limit improper secondary uses of personal data, such as cross-site tracking and targeting/profiling based on sensitive data. The ADPPA will also restrict the use of personal data to train generative AI systems that can manipulate users.
- Provide additional resources to antitrust law enforcement agencies to adequately monitor and take enforcement action against violations related to concentration of the data and computer markets.
- Impose a data minimization standard through legislative or regulatory means that would limit the use of personal information for generative AI training.
- Enact legislation that requires both government and commercial use of AI to be provably nondiscriminatory and proactively transparent by mandating audits and impact assessments—and prohibit manipulative or otherwise unacceptably risky uses. Both the White House AI Bill of Rights and the National Institute of Standards & Technology's AI RMF provide helpful frameworks for these requirements.

- Do not provide broad immunity (under Section 230 or otherwise) for companies or operators of Generative AI tools.
- Do not provide legislative or regulatory exemptions for copyright infringement when images are used in AI training.
- Do not invest more money in the development of AI without dedicating comparable resources to evaluation professionals, control mechanisms, and enforcement capabilities.
- Ensure that entities using AI outputs are held jointly responsible with entities behind the generation of those outputs for the harm that the entity using AI has caused with those outputs.

ADMINISTRATIVE AND ENFORCEMENT

- Continue to use existing consumer protection authorities, including Unfair and Deceptive Acts or Practices (FTC) and Unfair, Deceptive, or Abusive Acts or Practice (CFPB) authorities, to protect against manipulative, deceptive, and unfair AI practices.
- Establish standards through advisory opinions and policy statements for evaluating intellectual property and other claims relating to generative AI (e.g., copyright, trademark, etc.).
- Require the publication of environmental footprints of Generative AI models and their use.
- Secure injunctive relief to halt the operation of generative AI systems that lack necessary safeguards (as seen in Italy using GDPR).
- Promulgate rules that require both government and commercial uses of AI to be provably nondiscriminatory and proactively transparent by mandating audits and impact assessments—and prohibit manipulative or otherwise unacceptably risky uses. Both the White House AI Bill of Rights and the National Institute of Standards & Technology's AI RMF provide helpful frameworks for these requirements.

PRIVATE ACTOR BEHAVIOR

- Entities considering using generative AI procurement should critically examine whether these tools are appropriate.
- Entities should proactively document the data lifecycle and implement data audit trails.
- Individuals, companies, and research teams should develop tools to detect protected information within training models – like Glaze.
- Develop tools to detect deepfakes and make those tools easily accessible and usable by the public to help debunk deepfakes quickly.
- Watermark any protected documents or images to prevent or limit their use in the training of AI models.
- Publish the data sources, training sets, and logic of AI systems.
- Limit the scope of permissible external uses and modifications of generative AI models (including through API access).
- Limit permissible uses of Generative AI to low-risk settings.
- Determine and publish environmental footprints for Generative AI models and their use.
- Employers should invest in training workers in new skills for jobs that have been changed by generative AI.
- Employers should invest in training where there is a growing demand for new jobs that have been created by generative AI (e.g., prompt engineers, machine managers, AI auditors, and AI trainers).
- Companies should invest in training to retain labor rather than cutting costs by reducing staff in favor of generative AI technology.
- Companies, governments, and public-private programs should commit resources to helping workers displaced by generative AI find alternative jobs.

- Technology vendors and service providers should invest in AI research and development that improves worker productivity rather than replacing job functions.
- Technology companies should include the voices of, and give decision-making power to, those who are actually working on the development and training of generative AI, especially those in the Global South. Major companies need to elevate the involvement and participation of workers to ensure equity.
- Share profits with those whose labor helped build the systems rather than concentrating it among shareholders and top earners.
- Wages should be increased to match the increased worker productivity from generative AI. Workplaces should not use generative AI as a means to cheapen labor costs and devalue workers' contributions.

Appendix of Harms

> We shouldn't regulate AI until we see some meaningful harm that is actually happening – once we see that there is real meaningful harm. What is the real problem? There is not even $1,000 in damage.
>
> Microsoft Chief Economist Michael Schwarz, April 2023

Algorithmic harms exist today—and have been around for a long time. However, with the introduction of generative AI tools like ChatGPT, the scope and severity of algorithmic harms have exploded. In addition to unique harms posed by violations of data privacy and algorithmic systems, generative AI will accelerate the disintegration of trust in authoritative sources of information, exacerbate existing harms like IP theft and impersonation, and undermine existing legal protections for those harmed.

This Appendix is meant to give you a better sense of the universe of harms that generative AI is causing or exacerbating right now. However, AI is innovating at a rapid pace and new examples of harm emerge every day, so encapsulating every potential harm that generative AI could cause would be impossible. This appendix serves as a snapshot of pressing and real harms caused by generative AI today, rather than a comprehensive analysis of all possible harms.

This appendix includes:

1. Definitions for many common harms caused by generative AI.
2. Examples of real-world harms caused by generative AI.
3. A Table comparing the harms implicated by each example.

COMMON AI HARMS

1. **Physical Harms:** These are harms that lead to bodily injury or death, which may include acts by AI companies that facilitate or encourage physical assault.
2. **Economic Harms:** These are harms that cause monetary losses or decrease the value of something, which may include the harms of fraudulent transactions conducted by those using AI to impersonate a victim.
3. **Reputational Harms:** These harms involve injuries to someone's reputation within their community, which may in turn result in lost business or social pariahdom.
4. **Psychological Harms:** These harms include a variety of negative—and legally cognizable—mental responses, such as anxiety, anguish, concern, irritation, disruption, or aggravation. Danielle Citron and Daniel Solove place these harms within two categories: emotional distress or disturbance.
5. **Autonomy Harms:** These harms restrict, undermine, or otherwise influence people's choices and include acts like coercion, manipulation, failing to inform someone, acting in ways that undermine a user's choices, and inhibiting lawful behavior.
6. **Discrimination Harms:** These are harms that entrench or exacerbate inequality in ways that disadvantage certain people based on their demographics, characteristics, or affiliations. Discrimination harms often lead to other types of AI harms.
7. **Relationship Harms:** These harms involve damaging personal or professional relationships in ways that negatively impact one's health,

wellbeing, or functioning in society. Often, these harms damage relationships by degrading trust or damaging social boundaries.

8. **Loss of Opportunity:** Related to economic, reputational, discrimination, and relationship harms, loss of opportunity is an especially common AI harm in which AI-mediated content or decisions serve as a barrier to individuals accessing employment, government benefits, housing, and educational opportunities.
9. **Social Stigmatization and Dignitary Harms:** Related to reputational, discrimination, and relationship harms, these harms undermine individuals' sense of self and dignity through, e.g., loss of liberty, increased surveillance, stereotype reinforcement, or other negative impacts on one's dignity.

REAL EXAMPLES OF HARM

1. **Suicide:** ChatGPT encouraged an individual to commit suicide.
2. **Impersonation:** Scammers used generative AI to trick a woman into thinking her daughter was kidnapped, demanding $1,000,000 in return for her release.
3. **Deepfakes:** A prominent investigative reporter was ridiculed online after a pornographic deepfake of her went viral online.
4. **Defamation:** ChatGPT falsely included a law professor on a list of professors accused of sexual assault.
5. **Sexualization:** Lensa, an AI image generation application, portrayed women—particularly Asian and Black women —in a hypersexualized manner regardless of the source photos provided.
6. **Threats of Physical Harm:** An individual used ChatGPT to designate whether a person originating from different countries of origin should be tortured or not.
7. **Misinformation:** In Turkey's election, generative AI was used to spread over 150 unwarranted claims of terrorism by a presidential candidate.

8. **Copyright Infringement:** Parts of artists' work are routinely mimicked or duplicated by AI image generators, including commercially protected art.

9. **Labor Disputes:** Studios have threatened to use generative AI to replace striking writers, undermining labor negotiations.

10. **Data Breaches:** A viral generative AI tool's lax security practices and maintenance of personal data led to personal information like name, prompts, and email are exposed.

Harms

Examples	Physical	Economic	Reputational	Psychological	Autonomy	Discrimination	Relationship	Loss of Opportunity	Dignitary
Suicide	✓		✓	✓	✓				
Impersonation		✓	✓	✓	✓				
Deepfakes		✓	✓	✓	✓	✓	✓	✓	✓
Defamation			✓	✓			✓	✓	✓
Sexualization			✓	✓	✓	✓			✓
Threat of Physical Harm	✓	✓	✓	✓	✓	✓		✓	
Misinformation	✓	✓	✓	✓	✓			✓	
Copyright Infrigement		✓	✓	✓	✓			✓	
Labor Disputes		✓	✓	✓	✓		✓	✓	
Data Breaches		✓	✓	✓	✓				✓

References

[1] Danielle K. Citron & Daniel J. Solove, *Privacy Harms*, 102 B.U. L. Rev. 793 (2022).

[2] *See* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 Proc. Mach. Learning Rsch. 1 (2018); Gender Shades Project, http://gendershades.org/overview.html.

[3] *See, e.g.*, Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, Proc. 2021 ACM Conf. on Fairness, Accountability, & Transparency 610 (2021).

[4] Press Release, FTC, New Data Shows FTC Received 2.8 Million Fraud Reports from Consumers in 2021 (Feb. 22, 2022), https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0.

[5] Pranshu Verma, *They Thought Loved Ones Were Calling for Help. It was an AI Scam.,* The Washington Post (Mar. 5, 2023), https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/; Erielle Reshef, *Kidnapping Scam Uses Artificial Intelligence to Clone Teen Girl's Voice, Mother Issues Warning,* ABC News (Apr. 13, 2023), https://abc7news.com/ai-voice-generator-artificial-intelligence-kidnapping-scam-detector/13122645/.

[6] *See* National Consumer Law Center & EPIC, Scam Robocalls: Telecom Providers Profit (2022), https://epic.org/documents/scam-robocalls-telecom-providers-profit/.

[7] *See* TrueCaller, 2022 U.S. Spam & Scam Report (2022), https://www.truecaller.com/blog/insights/truecaller-insights-2022-us-spam-scam-report (noting that "[t]he total money lost to scams is also comparable to the entire child care budget of $39 billion for the American Rescue Plan Act. If phone scam fraud was somehow eliminated, the amount saved could fund federally subsidized child care across the U.S. for a full year to help families and employers."). The same source reported $29.8 billion in actual consumer losses in 2021 and $19.7 billion in losses in 2020, an increase of nearly $10 billion every year since 2019.

[8] Reported losses from text scams more than doubled from $131M to $330M between 2021 and 2022. FTC Consumer Sentinel Network, Fraud Reports by Contact Method, Reports & Amount Lost by Contact Method (2023), https://public.tableau.com/app/profile/federal.trade.commission/viz/FraudReports/FraudFacts ("Losses & Contact Method" tab selected, with quarters 1 through 4 checked for 2021, 2022).

[9] Lily Hay Newman, *AI Wrote Better Phishing Emails Than Humans in a Recent Test*, Wired (Aug. 7, 2021), https://www.wired.com/story/ai-phishing-emails/.

[10] Pranshu Verma & Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused,* Wash. Post (Apr. 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.
[11] Julia Angwin, *Decoding the Hype About AI*, Markup (Jan. 28, 2023), https://themarkup.org/hello-world/2023/01/28/decoding-the-hype-about-ai.
[12] *See, e.g.*, James Vincent, *The Swagged-out Pope is an AI Fake—and an Early Glimpse of a New Reality*, Verge (Mar. 27, 2023), https://www.theverge.com/2023/3/27/23657927/ai-pope-image-fake-midjourney-computer-generated-aesthetic.
[13] Danielle Citron and Robert Chesney have called attempts to sow distrust in real information using the specter of generative AI—and the increasing success that perpetrators would have as the public grows more aware of generative AI—the "Liar's Dividend." Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Cal. L. Rev. 1753,22 1785–86 (2019).
[14] *See* Ashley Belanger, *Thousands Scammed by AI Voices Mimicking Loved Ones in Emergencies*, Ars Technica (Mar. 6, 2023), https://arstechnica.com/tech-policy/2023/03/rising-scams-use-ai-to-mimic-voices-of-loved-ones-in-financial-distress/.
[15] *OpwnAI: Cybercriminals Starting to Use ChatGPT,* Check Point Rsch. (Jan. 6, 2023), https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-use-chatgpt/.
[16] Joseph Cox, A Computer Generated Swatting Service is Causing Havoc Across America, Vice (Apr. 13, 2023), https://www.vice.com/en/article/k7z8be/torswats-computer-generated-ai-voice-swatting**.**
[17] Pranshu Verma, *They Thought Loved Ones Were Calling for Help. It was an AI Scam.,* Wash. Post (Mar. 5, 2023), https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/.
[18] Matthew Gault, *AI Spam is Already Flooding the Internet and It Has an Obvious Tell*, Vice (Apr. 24, 2023), https://www.vice.com/en/article/5d9bvn/ai-spam-is-already-flooding-the-internet-and-it-has-an-obvious-tell.
[19] Igor Bonifacic, *CNET Had to Correct Most of its AI-Written Articles*, Engadget (Jan. 25, 2023), https://www.engadget.com/cnet-corrected-41-of-its-77-ai-written-articles-201519489.html.
[20] Jay Peters, *BuzzFeed is Using AI to Write SEO-Bait Travel Guides*, Verge (Mar. 30, 2023), https://www.theverge.com/2023/3/30/23663206/buzzfeed-ai-travel-guides-buzzy.
[21] The term, "deepfake," is a portmanteau of "deep learning" and "fake." The term was popularized by a Reddit user, @deepfakes, who posted the first viral deepfake video in 2017. *See* Moncarol Y. Wang, *Don't Believe Your Eyes: Fighting Deepfaked Nonconsensual Pornography with Tort Law*, 2022 U. Chi. Legal F. 415, 417–18 (2022).
[22] *See* Citron & Chesney, *supra* note 1313, at 1757 (defining deepfakes as the "full range of hyper-realistic digital falsification of images, video, and audio").
[23] *See* Wang, *supra* note 2121.

[24] *See* Anna Yamaoka-Enerklin, *Disrupting Disinformation: Deepfakes and the Law*, 22 N.Y.U. J. Legis. & Pub. Pol'y 725, 731 (2020).
[25] *See, e.g.*, Restatement (Second) of Torts § 525 (1977) (fraudulent misrepresentation); Colo. Code Regs. § 18-5-113 (2016).
[26] *See, e.g.*, Cal. Penal Code § 528.5; Haw. Rev. Stat. Ann. § 711-1106.6; La. Rev. Stat. § 14:73.10; Miss. Code Ann. § 97-45-33; N.Y. Penal Law § 190.25; R.I. Gen Laws § 11-52-7.1; Tex. Penal Code § 33.07.
[27] *See, e.g.*, 18 U.S.C. § 873; D.C. Code § 22-3252 (2019).
[28] *See, e.g.*, 18 U.S.C. § 2261A.
[29] *See* Edina Harbinja et al., *Governing Ghostbots*, 48 Comp. L. & Sec. Rev. 105791 (2023).
[30] *See, e.g.*, Kat Tenbarge, *Hundreds of Sexual Deepfake Ads Using Emma Watson's Face Ran on Facebook and Instagram in the Last Two Days*, NBC News (Mar. 7, 2023), https://www.nbcnews.com/tech/social-media/emma-watson-deep-fake-scarlett-johansson-face-swap-app-rcna73624; William Turton & Matthew Justus, *"Deepfake" Videos Like That Gal Gadot Porn are Only Getting More Convincing—and More Dangerous*, Vice (Aug. 27, 2018), https://www.vice.com/en/article/qvm97q/deepfake-videos-like-that-gal-gadot-porn-are-only-getting-more-convincing-and-more-dangerous.
[31] *See 46 States + DC + One Territory Now Have Revenge Porn Laws*, Cyber Civ. Rts. Initiative, https://www.cybercivilrights.org/revenge-porn-laws/ (last visited May 15, 2023); Orin S. Kerr, Computer Crime Law 245–47 (4th ed. 2018); *State Revenge Porn Policy*, EPIC, https://epic.org/state-revenge-porn-policy/.
[32] *See* Restatement (Second) of Torts § 652B, 625D.
[33] *Id.* § 652E.
[34] *See, e.g.*, Cass R. Sunstein, *Falsehoods and the First Amendment*, 33 Harv. J. L. & Tech. 387 (2020).
[35] *See White v. Samsung Elecs. Am., Inc.*, 971 F.2d 1395, 1398 (9th Cir. 1992); *Carson v. Here's Johnny Portable Toilets, Inc.*, 698 F.2d 831, 835 (6th Cir. 1983).
[36] 376 U.S. 254, 280 (1964).
[37] 485 U.S. 46, 50–52 (1988).
[38] *See, e.g.*, *Bollea v. Gawker Media, LLC*, No. 522012CA012447, 2016 WL 4073660 (Fla. Cir. Ct. June 8, 2016) (Hulk Hogan awarded $140 million against Gawker on privacy grounds).
[39] *See* Jeffrey T. Hancock & Jeremy N. Bailenson, *The Social Impact of Deepfakes*, 24 Cyberpsych., Behav., & Soc. Networking 149, 150 (2021).
[40] *See* Riana Pfefferkorn, *"Deepfakes" in the Courtroom*, 29 Pub. Int. L.J. 245, 259–74 (2020); Danielle C. Breen, *Silent No More: How Deepfakes will Force Courts to Reconsider Video Admission Standards*, 21 J. High Tech. L. 122, 132, 150–53 (2021).
[41] *United States v. Gagliardi*, 506 F.3d 140, 151 (2d Cir. 2007) (citing *United States v. Dhinsa*, 243 F.3d 635, 658 (2d Cir. 2001)).
[42] *United States v. Workinger*, 90 F.3d 1409, 1415 (9th Cir. 1996).

[43] *United States v. Vayner*, 769 F.3d 125, 130 (2d Cir. 2014).
[44] Pfefferkorn, *supra* note 4040, at 260.
[45] *Id.*; *see also, e.g.*, Fed. R. Evid. 902(4)(A) ("A copy of an official record" is self-authenticating "if the copy is certified as correct by... the custodian or another person authorized to make the certification.")
[46] Cox, *supra* note 1616.
[47] Samantha Cole, *'You Feel So Violated': Streamer QTCinderella Is Speaking Out Against Deepfake Porn Harassment*, Vice (Feb. 13, 2023), https://www.vice.com/en/article/z34pq3/deepfake-qtcinderella-atrioc; *Deepfake Porn Booms in the Age of A.I.*, NBC News (Apr. 28, 2022), https://www.nbcnews.com/now/video/deepfake-porn-booms-in-the-age-of-a-i-171726917562.
[48] Chandler Treon, *'Please Stop': Tiktoker Frightened After Being Harassed with AI-Generated Nudes of Herself*, Yahoo News (May 3, 2023), https://news.yahoo.com/please-stop-tiktoker-frightened-being-182335652.html.
[49] Joseph Cox, *Video Game Voice Actors Doxed and Harassed in Targeted AI Voice Attack*, Vice (Feb. 13, 2023), https://www.vice.com/en/article/93axnd/voice-actors-doxed-with-ai-voices-on-twitter.
[50] Citron & Chesney, *supra* note 13, at 1793; Megan Farokhmanesh, *The Debate on Deepfake Porn Misses the Point*, Wired (Mar. 1, 2023), https://www.wired.com/story/deepfakes-twitch-streamers-qtcinderella-atrioc-pokimane/.
[51] Citron & Chesney, *supra* note 1313, at 1793–94.
[52] *See, e.g.*, *Fair Hous. Council v. Roommates.com, LLC*, 521 F.3d 1157, 1168 (9th Cir. 2008) (holding that website that contributes materially to the alleged illegality of user content is not shielded from liability under 47 U.S.C. § 230).
[53] Citron & Chesney, *supra* note 1313, at 1803.
[54] *See United States v. Alvarez*, 567 U.S. 709, 719 (2012) (plurality) (concluding "falsity alone" could not remove expression from First Amendment protection).
[55] Citron & Chesney, *supra* note 1313, at 1805–06.
[56] 47 U.S.C. § 230(c)(1).
[57] The Supreme Court issued an opinion in *Gonzalez v. Google*, No. 21-1333, vacating the Ninth Circuit's decision but otherwise refused to weigh in on the proper test for Section 230 protection. The decision essentially leaves in place the status quo, where courts of appeals have been steadily converging on a test that is increasingly skeptical of industry arguments for Section 230 protection. The emerging test does not perfectly encapsulate EPIC's position on Section 230, but we follow precedent in this section. EPIC has argued that Section 230(c)(1) simply means that internet companies are not to be treated the same, for liability purposes, as the third parties who publish information on their services. Our test will often generate the same outcome as the test that is emerging in the circuit courts, but with less room for judicial discretion. *See* Brief for EPIC as Amicus Curiae in

Support of Neither Party, *Gonzalez v. Google LLC*, 143 S. Ct. 80 (2022) (No. 21-1333), https://epic.org/wp-content/uploads/2022/12/EPIC-Amicus-Gonzalez-v.-Google-1.pdf.

[58] *See* Jess Miers, *Yes, Section 230 Should Protect ChatGPT and Other Generative AI Tools*, TechDirt (Mar. 17, 2023), https://www.techdirt.com/2023/03/17/yes-section-230-should-protect-chatgpt-and-others-generative-ai-tools/.

[59] *See* Cristiano Lima, *AI Chatbots Won't Enjoy Tech's Legal Shield, Section 230 Authors Say*, Wash. Post (Mar. 17, 2023), https://www.washingtonpost.com/politics/2023/03/17/ai-chatbots-wont-enjoy-techs-legal-shield-section-230-authors-say/.

[60] *Henderson v. Source for Public Data, L.P.*, 53 F. 4th 110, 125 (4th Cir. 2022); *Lemmon v. Snap, Inc.*, 995 F.3d 1085 (9th Cir. 2021).

[61] *See, e.g.*, *Roommates.com*, 521 F.3d at 1168*; HomeAway v. City of Santa* Monica, 918 F.3d 676 (9th Cir. 2019); *Airbnb, Inc. v. City and County of San Francisco*, 217 F. Supp. 3d 1066 (N.D. Cal. 2016); *Lemmon*, 995 F.3d at 1085.

[62] *Lemmon*, 995 F.3d at 1092 (9th Cir. 2021); *see also A.M v. Omegle.com*, 614 F. Supp. 3d 814, 819–21 (D. Or. 2022).

[63] 47 U.S.C. § 230(f)(3).

[64] *See, e.g.,* Pranshu Verma and Will Oremus, *ChatGPT Created a Sexual Harassment Scandal and Named a Real Law Prof as the Accused,* Wash. Post (Apr. 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/; *see also* Eugene Volokh, *Communications Can Be Defamatory Even If Readers Realize There's a Considerable Risk of Error, Volokh Conspiracy* (Mar. 31, 2023), https://reason.com/volokh/2023/03/31/communications-can-be-defamatory-even-if-readers-realize-theres-a-considerable-risk-of-error/.

[65] *See* Miers, *supra* note 5858.

[66] This test originated in *Roommates.com*, 521 F.3d at 1157, and its latest significant articulation is found in *Henderson*, 53 F.4th at 110.

[67] *See, e.g.*, *Henderson*, 53 F.4th at 125.

[68] *Roommates.com*, 521 F.3d at 1168.

[69] The leading case on this issue is *Batzel v. Smith*, 333 F.3d 1018 (9th Cir. 2003), which concerned whether a person who sent an email to a listserv moderator with no intention of having the email posted online could be said to have "provided" information as contemplated by Section 230. A split Ninth Circuit panel determined that information is "provided by" a third party "when a third person or entity that created or developed the information in question *furnished it to the provider or user* under circumstances in which a reasonable person in the position of the service provider or user would conclude that the information was provided for publication on the Internet or other 'interactive computer service.'" *Id.* at 1034 (emphasis added). *Batzel* could be read to require direct furnishing of information to the defendant or, at the very least, some sort of relationship wherein the defendant could form a reasonable basis to believe that the third party intend for the defendant to publish the information.

[70] Merriam-Webster Dictionary, *Provide* (2023), https://www.merriam-webster.com/dictionary/provide.
[71] See discussion of *Batzel v. Smith*, *supra* note 6969.
[72] *Stratton Oakmont, Inc. v. Prodigy Servs.*, 23 Media L. Rep. (BNA) 1794 (N.Y. Sup. Ct. 1995).
[73] *See, e.g.*, *O'Korley v. Fastcase, Inc.*, 831 F.3d 352 (6th Cir. 2016).
[74] Google, *Block Search Engine Indexing with noindex*, Search Central (Feb. 20, 2023), https://developers.google.com/search/docs/crawling-indexing/block-indexing.
[75] Available technical tools to block scrapers, such as robot.txt flags, IP blockers and CAPTCHAs, can be bypassed by those determined enough to collect the data, which is why companies have attempted to use legal tools such as breach of contract and the Computer Fraud and Abuse Act to stop unauthorized scraping. *See* Kyle R. Dull & Julia B. Jacobson, *LinkedIn's Data Scraping Battle with hiQ Labs Ends with Proposed Judgment*, National Law Review (Dec. 19, 2022), https://www.natlawreview.com/article/linkedin-s-data-scraping-battle-hiq-labs-ends-proposed-judgment.
[76] For instance, in May 2018, Facebook made public the posts of as many as 14 million users that thought they were only sharing with their friends or a smaller group. Kurt Wagner, *Facebook Says Millions of Users Who Thought They Were Sharing Privately with Their Friends May Have Shared with Everyone Because of a Software Bug*, Vox (June 7, 2018), https://www.vox.com/2018/6/7/17438928/facebook-bug-privacy-public-settings-14-million-users. A few weeks later, Facebook unblocked users who had been previously blocked by other users, allowing the newly unblocked users to view content they should not have been per- mitted to view. Kurt Wagner, *Facebook's Year of Privacy Mishaps Continues—This Time with a New Software Bug that 'Unblocked' People*, Vox (July 2, 2018), https://www.vox.com/2018/7/2/17528220/facebook-soft-ware-bug-block-unblock-safety-privacy.
[77] For example, the FTC's 2011 consent order against Facebook was based, in part, on Facebook's decisions to change privacy settings to make public information users had previously set to private. *See* FTC, *Facebook Settles FTC Charges That It Deceived Consumers By Failing To Keep Privacy Promises* (Nov. 29, 2011), https://www.ftc.gov/news-events/news/press-releases/2011/11/facebook-settles-ftc-charges-it-deceived-consumers-failing-keep-privacy-promises.
[78] *Packingham v. North Carolina*, 137 S. Ct. 1730, 1735 (2017) (quoting *Reno v. American Civil Liberties Union*, 521 U.S. 844, 868 (1997)).
[79] Ryan Browne, *Italy Became the First Western Country to Ban ChatGPT. Here's What Other Countries are Doing*, CNBC (Apr. 4, 2023), https://www.cnbc.com/2023/04/04/italy-has-banned-chatgpt-heres-what-other-countries-are-doing.html; Natasha Lomas, *ChatGPT Resumes Service in Italy After Adding Privacy Disclosures and Controls*, TechCrunch (Apr. 28, 2023), https://techcrunch.com/2023/04/28/chatgpt-resumes-in-italy/.

[80] *See Record Number of Data Breaches in 2021*, IAPP Daily Dashboard (Jan. 25, 2022), https://iapp.org/news/a/record-number-of-data-breaches-in-2021/ (citing to ITRC report which estimated "1,862 breaches last year, up 68% from the year prior, and exceeded 2017's previous record of 1,506").
[81] *See* Identity Theft Resource Center (ITRC), 2022 Data Breach Report 2 (Jan. 2023), https://www.idtheftcenter.org/publication/2022-data-breach-report/.
[82] U.S. Gov't Accountability Off., GAO-14-34, Agency Responses to Breaches of Personally Identifiable Information Need to be More Consistent 11 (2013), http://www.gao.gov/assets/660/659572.pdf.
[83] *See id.* at 13.
[84] *See* Soc. Sec. Admin., Identity Theft and Your Social Security Number 1 (2021), https://www.ssa.gov/pubs/EN-05-10064.pdf ("A dishonest person who has your Social Security number can use it to get other personal information about you. Identity thieves can use your number and your good credit to apply for more credit in your name. Then, when they use the credit cards and don't pay the bills, it damages your credit. You may not find out that someone is using your number until you're turned down for credit, or you begin to get calls from unknown creditors demanding payment for items you never bought. Someone illegally using your Social Security number and assuming your identity can cause a lot of problems.")
[85] *See* Erika Harrell, Bureau of Just. Stat., Dep't of Just., Victims of Identity Theft, 2018 11 (Apr. 2020), https://bjs.ojp.gov/content/pub/pdf/vit18.pdf; Danielle K. Citron & Daniel Solove, *Risk and Anxiety: A Theory of Data Breach Harms*, 96 Tex. L. Rev. 737, 745 ("Knowing that thieves may be using one's personal data for criminal ends may produce significant anxiety.").
[86] *See, e.g.*, Cyber. & Infrastructure Sec. Agency (CISA), DarkSide Ransomware: Best Practices for Preventing Business Disruption from Ransomware Attacks, Alert Code AA21-131A (July 7, 2021), https://www.cisa.gov/news-events/cybersecurity-advisories/aa21-131a (describing one example of ransomware-as-a-service); Kaspersky, *Malware-as-a-service (MaaS)*, Encyclopedia by Kaspersky, https://encyclopedia.kaspersky.com/glossary/malware-as-a-service-maas/ (last visited May 15, 2023) (defining the term "malware-as-a-service"); Brian Krebs, *Giving a Face to the Malware Proxy Service 'Faceless'*, Krebs on Security (Apr. 18, 2023), https://krebsonsecurity.com/2023/04/giving-a-face-to-the-malware-proxy-service-faceless/ (describing a malware proxy service).
[87] *See, e.g.*, Elias Groll, *ChatGPT Shows Promise of Using AI to Write Malware*, CyberScoop (Dec. 6, 2022), https://cyberscoop.com/chatgpt-ai-malware/ ("'If not ChatGPT, then a model in the next couple years will be able to write code for real world software vulnerabilities,' [Dolan-Gavitt, an assistant professor in the Computer Science and Engineering Department at New York University.] added…. Benjamin Tan, a computer scientist at the University of Calgary, said he was able to bypass some of ChatGPT's

safeguards by asking the model to produce software piece by piece that, when assembled, might be put to malicious use. 'It doesn't know that when you put it all together it's doing something that it shouldn't be doing,' Tan said.").

[88] *See, e.g.*, Crane Hassold, *Executive Impersonation Attacks Targeting Companies Worldwide*, Abnormal Blog (Feb. 16, 2023), https://abnormalsecurity.com/blog/midnight-hedgehog-mandarin-capybara-multilingual-executive-impersonation ("Using widely available marketing technology and highly accurate translation apps, attackers can rapidly scale their efforts, maximizing their reach and wreaking havoc across the globe. And because many translation tools now use machine learning to improve context, such as translating the meaning of a sentence rather than each word individually, they're much easier to manipulate for nefarious purposes.")

[89] *See, e.g.*, Center for Strategic & International Studies, *A Conversation on Cybersecurity with NSA's Rob Joyce*, YouTube (Apr. 11, 2023), https://youtu.be/MMNHNjKp4Gs?t=530 (8:50 mark) (NSA Dir. of Cybersecurity Rob Joyce describing ChatGPT as able to optimize the workflow of bad actors seeking to use zero day exploits, and for malicious foreign actors to "craft very believable native-language English text that could be part of your phishing campaign or part of your interaction with a person or your ability to build a backstory—all the things that will allow you to do those activities, or even malign influence.").

[90] *See, e.g.*, Kate Park, *Samsung Bans Use of Generative AI Tools like ChatGPT After April Internal Data Leak*, TechCrunch (May 2, 2023), https://techcrunch.com/2023/05/02/samsung-bans-use-of-generative-ai-tools-like-chatgpt-after-april-internal-data-leak/; Dan Milmo and Agencies, *Italy's Privacy Watchdog Bans ChatGPT Over Data Breach Concerns*, Guardian (Apr. 1, 2023), https://www.theguardian.com/technology/2023/mar/31/italy-privacy-watchdog-bans-chatgpt-over-data-breach-concerns.

[91] *See, e.g.*, Marcus Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It*, Harvard Kennedy School Belfer Center for Science and International Affairs (Aug. 2019), https://www.belfercenter.org/publication/AttackingAI.

[92] *See, e.g.*, Eduard Kovacs, *ChatGPT Data Breach Confirmed as Security Firm Warns of Vulnerable Component Exploitation*, SecurityWeek (Mar. 28, 2023), https://www.securityweek.com/chatgpt-data-breach-confirmed-as-security-firm-warns-of-vulnerable-component-exploitation/.

[93] *See, e.g.*, Mark Gurman, *Samsung Bans Staff's AI Use After Spotting ChatGPT Data Leak*, Bloomberg (May 1, 2023), https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak.

[94] *See, e.g.*, Mitchell Clark & James Vincent, *OpenAI is Massively Expanding ChatGPT's Capabilities to Let It Browse the Web and More*, Verge (Mar. 23, 2023),

https://www.theverge.com/2023/3/23/23653591/openai-chatgpt-plugins-launch-web-browsing-third-party.

[95] Nat'l Institute of Standards & Tech. (NIST), Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI-100-1 (Jan. 2023), https://doi.org/10.6028/NIST.AI.100-1.

[96] *American Data Privacy and Protection Act Fact Sheet*, EPIC, https://epic.org/documents/american-data-privacy-and-protection-act-fact-sheet/ (last visited May 15, 2023) (e.g., algorithmic impact assessments).

[97] *See, e.g.*, *Intellectual Property Law*, Georgetown Law (May 2023), https://www.law.georgetown.edu/your-life-career/career-exploration-professional-development/for-jd-students/explore-legal-careers/practice-areas/intellectual-property-law/; *Explore the Four Areas of IP Law*, Suffolk Law (May 2023), https://www.suffolk.edu/law/academics-clinics/what-can-i-study/intellectual-property/intellectual-property-law-basics-certificate/explore-the-four-areas-of-ip-law.

[98] Open Letter, Ctr. for Artistic Inquiry, Restrict AI Illustration from Publishing: An Open Letter (May 2, 2023), https://artisticinquiry.org/AI-Open-Letter.

[99] Case Study Footnote: Chris Willman, *AI-Generated Fake 'Drake'/'Weeknd' Collaboration, 'Heart on My Sleeve,' Delights Fans and Sets Off Industry Alarm Bells,* Variety (Apr. 17, 2023), http://variety.com/2023/music/news/fake-ai-generated-drake-weeknd-collaboration-heart-on-my-sleeve-1235585451/; *see also* Will Knight, *Algorithms Can Now Mimic Any Artist. Some Artists Hate It.*, Wired (August 19, 2022), https://www.wired.com/story/artists-rage-against-machines-that-mimic-their-work/; Sarah Andersen, *The Alt-Right Manipulated My Comic. Then A.I. Claimed It.,* N.Y. Times (December 31, 2022), https://www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-algorithim-took-my-work.html; Vanessa Thorpe, *'ChatGPT Said I did Not Exist': How Artists and Writers are Fighting Back Against AI,* Guardian (March 18, 2023), https://www.theguardian.com/technology/2023/mar/18/chatgpt-said-i-did-not-exist-how-artists-and-writers-are-fighting-back-against-ai; Rachel Metz, *These Artists Found Out Their Work Was Used to Train AI. Now They're Furious,* CNN Business (October 21, 2022), https://www.cnn.com/2022/10/21/tech/artists-ai-images/index.html.

[100] *See, e.g.,* Nick Cave, *Issue #218,* Red Hand Files (January 2023), https://www.theredhandfiles.com/chat-gpt-what-do-you-think/.

[101] Metz, *supra* note 9999.

[102] U.S. Copyright Office, Libr. of Cong., *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence,* 88 Fed. Reg. 16190, 16191 (March 16, 2023), https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence#footnote-8-p16191.

[103] Press Release, U.S. Copyright Office, Copyright Office Launches New Artificial Intelligence Initiative (March 16, 2023), https://www.copyright.gov/newsnet/2023/1004.html.

[104] U.S. Copyright Office, Decision Affirming Refusal of Registration of a Recent Entrance to Paradise 2–3 (Feb. 14, 2022), https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf.
[105] U.S. Copyright Office & Libr. of Cong., *supra* note 102102.
[106] *Id.* at 16192.
[107] *See, e.g.,* U.S. Copyright Office, Cancellation Decision re: Zarya of the Dawn (Reg. No. VAu001480196) 2 (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf.
[108] 17 U.S.C. § 103(b).
[109] Metz, *supra* note 9999; Thorpe, *supra* note 9999; Knight, *supra* note 9999; Cave, *supra* note 100100.
[110] Willman, *supra* note 9999.
[111] Andersen, *supra* note 9999.
[112] U.S. Copyright Office & Libr. of Cong., *supra* note 102102.
[113] *See* Class Action Complaint and Demand for Jury Trial, *Andersen et al. v. Stability AI Ltd. et al.,* No. 3:23-cv-00201-WHO, (N.D. Cal. Jan. 13, 2023), https://stablediffusionlitigation.com/pdf/00201/1-1-stable-diffusion-complaint.pdf.
[114] Taylor Dafoe, *Getty Images Is Suing the Company Behind Stable Diffusion, Saying the A.I. Generator Illegally Scraped Its Content,* ArtNet (Jan. 17, 2023), https://news.artnet.com/art-world/getty-images-suing-stability-ai-stable-diffusion-illegally-scraped-images-copyright-infringement-2243631.
[115] Natasha Lomas, *Glaze Protects Art from Prying AIs,* TechCrunch (Mar. 17, 2023), https://techcrunch.com/2023/03/17/glaze-generative-ai-art-style-mimicry-protection/.
[116] *Shutterstock Datasets and AI-generated Content: Contributor FAQ*, Shutterstock (Mar. 20 2023), https://support.submit.shutterstock.com/s/article/Shutterstock-ai-and-Computer-Vision-Contributor-FAQ?language.
[117] Kyle Wiggers, *DeviantArt Provides a Way for Artists to Opt Out of AI Art Generators,* TechCrunch (Nov. 11, 2022), https://techcrunch.com/2022/11/11/deviantart-provides-a-way-for-artists-to-opt-out-of-ai-art-generators/.
[118] *See generally* Hans-Otto Pörtner et al., Intergovernmental Panel on Climate Change, Climate Change 2022: Impacts Adaptation and Vulnerability (2022), https://report.ipcc.ch/ar6/wg2/IPCC_AR6_WGII_FullReport.pdf [hereinafter "IPCC Report"].
[119] IPCC Report at 9–11.
[120] IPCC Report at 13–14.
[121] Emma Strubell et al., *Energy and Policy Considerations for Deep Learning in NLP*, Proc. 57th Ann. Meeting Ass'n for Comp. Linguistics 3645, 3645 (2019).
[122] *Id.*
[123] Roy Schwartz et al., *Green AI*, 63 Commc'ns ACM 54, 56 (2020).
[124] *Id.*
[125] *Id.*

[126] Strubell et al., *supra* note 121121, at 3645.
[127] *Id.*
[128] Amba Kak & Sarah Myers West, AI Now Institute, 2023 Landscape: Confronting Tech Power 100 (2023), https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf [hereinafter "AI Now Report"].
[129] Mack DeGeurin, *'Thirsty' AI: Training ChatGPT Required Enough Water to Fill a Nuclear Reactor's Cooling Tower, Study Finds*, Gizmodo (May 4, 2023), https://gizmodo.com/chatgpt-ai-water-185000-gallons-training-nuclear-1850324249.
[130] AI Now Report at 99.
[131] *Dutch Call a Halt to New Massive Data Centres, While Rules are Worked Out*, DutchNews (Feb. 17, 2022), https://www.dutchnews.nl/2022/02/dutch-call-a-halt-to-new-massive-data-centres-while-rules-are-worked-out/.
[132] *See e.g.*, Stephen Thomas, *Who Will You Be After ChatGPT Takes Your Job*, Wired (Apr. 21, 2023), https://www.wired.com/story/status-work-generative-artificial-intelligence/; Greg Ip, *The Robots Have Finally Come for My Job,* Wall St. J. (Apr. 5, 2023), https://www.wsj.com/articles/the-robots-have-finally-come-for-my-job-34a69146; Jyoti Mann, *Sam Altman Admits OpenAI is 'A Little Bit Scared' of ChatGPT and Says It Will 'Eliminate' Many Jobs*, Insider (Mar. 18, 2023), https://www.businessinsider.com/sam-altman-little-bit-scared-chatgpt-will-eliminate-many-jobs-2023-3; Steven Greenhouse, *US Experts Warn AI Likely to Kill off Jobs—and Widen Wealth Inequality*, Guardian (Feb. 8, 2023), https://www.theguardian.com/technology/2023/feb/08/ai-chatgpt-jobs-economy-inequality.
[133] *70% of Workers Using ChatGPT At Work Are Not Telling Their Bosses; Overall Usage Among Professionals Jumps to 43%*, Fishbowl (Feb. 1, 2023), https://www.fishbowlapp.com/insights/70-percent-of-workers-using-chatgpt-at-work-are-not-telling-their-boss/.
[134] *See e.g.*, Katie Notopoulos, *A Tech News Site Has Been Using AI to Write Articles, So We Did the Same Thing Here,* Buzzfeed News (Jan. 12, 2023), https://www.buzzfeednews.com/article/katienotopoulos/cnet-articles-written-by-ai-chatgpt-article; Connie Guglielmo, *CNET Is Experimenting With an AI Assist. Here's Why*, CNET (Jan. 16, 2023), https://www.cnet.com/tech/cnet-is-experimenting-with-an-ai-assist-heres-why/; Ryan Ermey, *ChatGPT Wrote Part of This Article—It Didn't Go Great*, CNBC (Jan. 26, 2023), https://www.cnbc.com/2023/01/26/chatgpt-wrote-part-of-this-article-it-didnt-go-great.html; Noor Al-Sibai & Jon Christian, *BuzzFeed is Quietly Publishing Whole AI-Generated Articles, Not Just Quizzes*, Futurism (Mar. 30, 2023), https://futurism.com/buzzfeed-publishing-articles-by-ai.
[135] *See* Kevin Travers, *How ChatGPT is Changing the Job Hiring Process, From the HR Department to Coders*, CNBC (Apr. 8, 2023), https://www.cnbc.com/2023/04/08/chatgpt-is-being-used-for-coding-and-to-write-job-descriptions.html.

[136] *See, e.g.*, Chris Morris, *A Major International Law Firm is Using an A.I. Chatbot to Help Lawyers Draft Contracts: 'It's Saving Time at All Levels'*, Fortune (Feb. 15, 2023), https://fortune.com/2023/02/15/a-i-chatbot-law-firm-contracts-allen-and-overy/.
[137] Belle Lin, *Generative AI Makes Headway in Healthcare*, Wall St. J. (Mar. 21, 2023), https://www.wsj.com/articles/generative-ai-makes-headway-in-healthcare-cb5d4ee2.
[138] *9 in 10 Companies That are Currently Hiring Want Workers with ChatGPT Experience*, Resume Builder (Apr. 17, 2023), https://www.resumebuilder.com/9-in-10-companies-that-are-currently-hiring-want-workers-with-chatgpt-experience/.
[139] Britney Nguyen, *AI 'Prompt Engineer' Job Can Pay up to $375,000 a Year and Don't Always Require a Background in Tech*, Insider (May 1, 2023), https://www.cnbc.com/2023/04/05/chatgpt-is-the-newest-in-demand-job-skill-that-can-help-you-get-hired.html.
[140] PromptBase, https://promptbase.com/.
[141] Alyssa Lukpa, *JPMorgan Restricts Employees From Using ChatGPT*, Wall St. J. (Feb. 22, 2023), https://www.wsj.com/articles/jpmorgan-restricts-employees-from-using-chatgpt-2da5dc34.
[142] Gurman, *supra* note 9393.
[143] *See* Daron Acemoglu, *Harms of AI* 49 (Nat'l Bureau of Econ. Rsch., Working Paper No. 29247, 2021), https://www.nber.org/system/files/working_papers/w29247/w29247.pdf.
[144] Daron Acemoglu & Pascual Restrepo, Tasks, Automation, and the Rise in US Wage Inequality 35 (2022), http://pascual.scripts.mit.edu/research/taskdisplacement/task_displacement.pdf.
[145] White House, The Impact of Artificial Intelligence on the Future of Workforces in the European Union and the United States of America 15 (2022), https://www.whitehouse.gov/wp-content/uploads/2022/12/TTC-EC-CEA-AI-Report-12052022-1.pdf.
[146] Daron Acemoglu, *Automation Shouldn't Always be Automatic: Marking Artificial Intelligence Work for Workers and the World*, OECD: The Forum Network (Nov. 6, 2020), https://www.oecd-forum.org/posts/automation-shouldn-t-always-be-automatic-making-artificial-intelligence-work-for-workers-and-the-world.
[147] *Id.*
[148] Daron Acemoglu et al., *AI and Jobs: Evidence from Online Vacancies* 3 (Nat'l Bureau of Econ. Rsch., Working Paper No. 28257, 2022), https://www.nber.org/system/files/working_papers/w28257/w28257.pdf.
[149] David H. Autor, *Why Are There Still So Many Jobs? The History and Future of Workplace Automation*, 29 J. Econ. Persp. 3 (2015), https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.29.3.3.
[150] Tyna Eloundou et al., *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models*, arXiv (Mar. 23, 2023), https://arxiv.org/pdf/2303.10130.pdf.

[151] *Id.*
[152] Jan Hatzius et al., Goldman Sachs, The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani) 1 (2023), https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf.
[153] *Id.* at 1, 6-7.
[154] White House, *supra* note 145145.
[155] White House, *supra* note 145145.
[156] Chloe Xiang, *Startups Are Already Using GPT-4 to Spend Less on Human Coders*, Motherboard (Mar. 20, 2023), https://www.vice.com/en/article/jg5xmp/startups-are-already-using-gpt-4-to-spend-less-on-human-coders.
[157] *Id.*
[158] *Moore's Law for Everything*, Sam Altman's Blog (Mar. 16, 2021), https://moores.samaltman.com/.
[159] *Id.*
[160] Alan Chan et al., *The Limits of Global Inclusion in AI Development*, arXiv (Feb. 2, 2021), https://arxiv.org/pdf/2102.01265.pdf.
[161] Wendy Liu, *AI Is Exposing Who Really Has Power in Silicon Valley*, Atlantic (Mar. 27, 2023), https://www.theatlantic.com/technology/archive/2023/03/open-ai-products-labor-profit/673527/.
[162] *Id.*
[163] Jeffrey Dastin et al., *Exclusive: ChatGPT Owner OpenAI Projects $1 Billion in Revenue by 2024*, Reuters (Dec. 15, 2022), https://www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15/.
[164] Billy Perrigo, *150 African Workers for ChatGPT, TikTok and Facebook Vote to Unionize at Landmark Nairobi Meeting*, Time (May 1, 2023) https://time.com/6275995/chatgpt-facebook-african-workers-union/
[165] Alissa Wilkinson, *The Looming Threat of AI to Hollywood, and Why It Should Matter to You*, Vox (May 2, 2023) https://www.vox.com/culture/23700519/writers-strike-ai-2023-wga.
[166] Billy Perrigo, *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per House to Make ChatGPT Less Toxic*, Time (Jan. 18, 2023), https://time.com/6247678/openai-chatgpt-kenya-workers/.
[167] *Id.*
[168] *Id.*
[169] *Id.*
[170] *Id.*
[171] Resume Builder, *supra* note 138138.

[172] Alexandra Bruell, *BuzzFeed to Use ChatGPT Creator OpenAI to Help Create Quizzes and Other Content*, Wall St. J. (Jan. 26, 2023), https://www.wsj.com/articles/buzzfeed-to-use-chatgpt-creator-openai-to-help-create-some-of-its-content-11674752660.
[173] *Id.*
[174] Jacklyn Diaz & Madj Al-Waheidi, *BuzzFeed Shutters Its Newsroom as the Company Undergoes Layoffs*, NPR (Apr. 21, 2023), https://www.npr.org/2023/04/20/1171056620/buzzfeed-news-shut-down-media-layoffs.
[175] *Misinformation on Bard, Google's New AI Chatbot,* Ctr. for Countering Digit. Hate (Apr. 5, 2023), https://counterhate.com/research/misinformation-on-bard-google-ai-chat/.
[176] 795 F. App'x 878, 879–80 (3d Cir. 2020) (quoting Restatement (Third) of Torts: Products Liability § 19(a) (Am. L. Inst. 1998)).
[177] 2 F.4th 871, 938 (9th Cir. 2021) (Gould, J., concurring in part).
[178] *See, e.g.*, Karni A. Chagal-Feferkorn, *Am I an Algorithm or a Product? When Products Liability Should Apply to Algorithmic Decision-Makers*, 60 Stan. L. & Pol'y Rev. 61 (2019) (distinguishing between algorithmic products that should be subject to products liability and thinking algorithms that should not).
[179] *Cf.* Catherine M. Sharkey, *Products Liability in the Digital Age: Online Platforms as "Cheapest Cost Avoiders"*, 73 Hastings L.J. 1327 (2022).
[180] Hasan Chowdhury, *ChatGPT Cost a Fortune to Make with OpenAI's Losses Growing to $540 Million Last Year, Report Says*, Insider (May 5, 2023), https://www.businessinsider.com/openai-2022-losses-hit-540-million-as-chatgpt-costs-soared-2023-5.
[181] *Id.*
[182] *See* Cade Metz & Karen Weise, *Microsoft to Invest $10 Billion in OpenAI, the Creator of ChatGPT*, N.Y. Times (Jan. 23, 2023), https://www.nytimes.com/2023/01/23/business/microsoft-chatgpt-artificial-intelligence.html.
[183] *See ChatGPT and More: Large Scale AI Models Entrench Big Tech Power*, AI Now Institute (Apr. 11, 2023), https://ainowinstitute.org/publication/large-scale-ai-models.
[184] Christopher Mims, *The AI Boom That Could Make Google and Microsoft Even More Powerful*, Wall St. J. (Feb. 11, 2023), https://www.wsj.com/articles/the-ai-boom-that-could-make-google-and-microsoft-even-more-powerful-9c5dd2a6.