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 Comments on Draft Documents Responsive to NIST's Assignments Under Executive Order 14110 (Sections 4.1, 4.5, and 11)

No. 2024-09824

June 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 Comments on Draft Documents Responsive to NIST's Assignments Under Sections 4.1, 4.5, and 11 of Executive Order 14110.[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 also published extensive research on emerging AI technologies like generative AI (GAI),[5] as well as the ways that government agencies develop, procure, and use AI systems around the country.[6]

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

[1] 89 Fed. Reg. 38097 (May 7, 2024).
[2] *About Us*, EPIC, https://epic.org/about/ (2024).
[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, Generating Harms: Generative AI's Impact & Paths Forward (2023), https://epic.org/gai [hereinafter "EPIC GenAI I Report"]; EPIC, Generating Harms II: Generative AI's New & Continued Impacts (2024), https://epic.org/wp-content/uploads/2024/05/EPIC-Generative-AI-II-Report-May2024-1.pdf [hereinafter "EPIC GenAI II Report"].
[6] 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"].

Framework ("AI RMF")[7] in its entirety.[8] At the same time, EPIC encourages NIST to view the risks of GAI technologies and synthetic content as extensions of traditional AI and automated decision-making risks, not as qualitatively different risks requiring entirely new forms of risk mitigation. Many of the same AI risk management techniques at the core of NIST's AI RMF—including AI impact assessments,[9] regular AI accuracy testing,[10] and AI red-teaming efforts[11]—will be effective against the risks of GAI technologies and the synthetic content they produce. The feedback EPIC has provided within this comment is meant to inform NIST's draft documents while at the same time placing technical solutions like watermarking within their broader sociotechnical context; technical standards cannot, by themselves, remedy the risks of GAI and other emerging AI technologies. To further inform NIST's efforts to categorize and address the risks of GAI, EPIC has also appended our two GAI reports to this comment: *Generating Harms* (2023) and *Generating Harms II* (2024). Both reports delve deeper into high-risk GAI use contexts and apply two leading typologies of privacy and AI harms that EPIC recommends as foundations for NIST's ongoing research—the typology from Danielle Citron's and Daniel Solove's *Privacy Harms* paper[12] and Joy Buolamwini's Taxonomy of Algorithmic Harms.[13]

I. NIST IS CORRECT TO FOCUS ON GENERATIVE AI'S PRESENT RISKS, BUT IT CAN FURTHER REFINE ITS RISK CATEGORIZATION

Responsive to NIST AI 600-1: Generative AI Profile

EPIC commends the detailed approach NIST has undertaken with NIST AI 600-1 to incorporate robust and actionable oversight mechanisms into its GAI risk management profile. While many of the risks of GAI mirror those of other AI and automated decision-making systems,

[7] 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"].
[8] *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.
[9] NIST AI RMF at 11, 36.
[10] *Id.* at 27–30, 35–36.
[11] NIST, AI RMF Playbook 31–32, 131, 200 (2023), https://airc.nist.gov/docs/AI_RMF_Playbook.pdf [hereinafter "NIST AI RMF Playbook"].
[12] Danielle K. Citron & Daniel J. Solove, *Privacy Harms*, 102 B.U. L. Rev. 793, 830–861 (2022).
[13] *See, e.g., Artificial Intelligence: Societal and Ethical Implications: Hearing before the U.S. House Comm. on Sci., Space & Tech.*, 116th Cong. (2019) (testimony of Joy Buolamwini), https://www.congress.gov/116/meeting/house/109688/witnesses/HHRG-116-SY00-Wstate-BuolamwiniJ-20190626.pdf; Lauren Smith, *Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making*, Future of Privacy Forum (Dec. 11, 2017), https://fpf.org/blog/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/.

EPIC supports NIST's decision to provide more detailed descriptions of risks and recommended actions for GAI technologies. Many GAI developers and deployers routinely downplay, deprioritize, or ignore the present risks of GAI systems—especially those targeting marginalized communities, the environment, and democracy—in favor of nebulous, potential existential threats.[14] Publishing specific and actionable guidelines for what AI companies should do today will ensure that the present risks of GAI to rights and safety are adequately mitigated while giving enforcement agencies, judges, and legislators a clearer understanding of where leading GAI companies are failing.

To facilitate actionable improvements to NIST AI 600-1, EPIC has chosen to limit our substantive feedback to Section 2 of the draft document, which outlines NIST's proposed typology of risks unique to or exacerbated by GAI. EPIC's recommendations, which build upon research from EPIC's two GAI reports,[15] fall into two broad categories:

1. Recommendations to unify NIST's overall approach to AI risk categorization,
2. Recommendations to improve NIST AI 600-1's current typology of risks, and
3. Recommendations for additional GAI risks currently absent from NIST AI 600-1.

NIST's AI guidance is most effective when it provides clear examples and guidelines for what AI developers and deployers should and should not do to mitigate AI harms. Without clear and actionable instructions for how AI actors should identify and mitigate each type of AI risk, NIST's AI standards do little to ensure AI harms are prevented—and may instead provide the language that negligent or malicious AI actors need to undermine attempts to enforce AI regulations in the future. EPIC has provided recommendations below to add important clarity to the agency's typology of generative AI risks to encourage meaningful industry compliance, and we would be happy to further discuss any risks or recommendations with NIST staff.

Unifying NIST's Approach to AI Risk Categorization

Currently, NIST uses three overlapping approaches to categorize AI risks and remedial approaches within the AI RMF. First, NIST recommends categorizing AI harms according to their target (Fig. 1).[16] Second, NIST recommends remedying those harms by ensuring AI systems are responsive to seven characteristics of trustworthy AI (Fig. 4).[17] And finally, NIST recommends pursuing AI risk management through four core functions—Govern, Map, Measure, and Manage (Fig. 5).[18] By including a typology of twelve risks unique to or exacerbated by GAI within NIST

[14] *See, e.g.*, Blake Richards et al., *The Illusion of AI's Existential Risk*, Noema (July 18, 2023), https://www.noemamag.com/the-illusion-of-ais-existential-risk/.
[15] *See generally* EPIC GenAI I Report; EPIC GenAI II Report.
[16] AI RMF at 5.
[17] *Id.* at 12.
[18] *Id.* at 20.

AI 600-1, NIST adds a fourth approach to categorize GAI risks without clearly establishing how this new typology interacts with the AI RMF's risk management approaches.

Fig. 1. Examples of potential harms related to AI systems. Trustworthy AI systems and their responsible use can mitigate negative risks and contribute to benefits for people, organizations, and ecosystems.

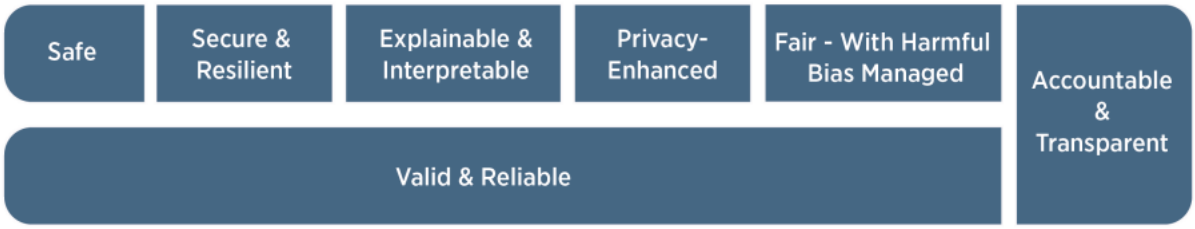


Fig. 4. Characteristics of trustworthy AI systems. Valid & Reliable is a necessary condition of trustworthiness and is shown as the base for other trustworthiness characteristics. Accountable & Transparent is shown as a vertical box because it relates to all other characteristics.

In its current state, NIST AI 600-1 provides much needed clarity over specific GAI risks but lacks a unifying vision for how GAI risks and their related actions fit into the broader AI RMF. As a result, it serves less as a companion resource to the AI RMF and more as a parallel risk management framework—one that may inject additional confusion to the AI risk management process or even undermine key guardrails within the AI RMF that can and should apply to GAI use cases. Many GAI risks are extensions of traditional AI risks, and EPIC encourages NIST to further clarify that AI developers and deployers should manage the unique GAI risks outlined in NIST AI 600-1 *alongside* the general AI risks presented within the AI RMF.

To support a unifying vision for risk categorization and mitigation within the AI RMF and NIST AI 600-1, EPIC recommends, at minimum, that NIST further sort the GAI risks outlined in NIST AI 600-1 according to whom and how they harm (see Fig. 1). For example, in EPIC's two GAI reports, we contextualize GAI harms within two preexisting typologies: the

typology of privacy harms coined by Professors Danielle Citron and Daniel Solove[19] and the taxonomy of algorithmic harms promoted by Dr. Joy Buolamwini and the Algorithmic Justice League.[20] Citron's and Solove's typology of privacy harms comprises the following impacts:

1. **Physical harms**;
2. **Economic harms**;
3. **Reputational harms**;
4. **Psychological harms**, including emotional distress and disturbance;
5. **Autonomy harms**, including coercion, manipulation, failure to inform, thwarted expectations, lack of control, and chilling effects;
6. **Discrimination harms**; and
7. **Relationship harms**.

Fig. 5. Functions organize AI risk management activities at their highest level to govern, map, measure, and manage AI risks. Governance is designed to be a cross-cutting function to inform and be infused throughout the other three functions.

[19] Citron & Solove, *supra* note 12, 830–861 (2022).
[20] *See, e.g., Artificial Intelligence: Societal and Ethical Implications: Hearing before the U.S. House Comm. on Sci., Space & Tech.*, 116th Cong. (2019) (testimony of Joy Buolamwini), https://www.congress.gov/116/meeting/house/109688/witnesses/HHRG-116-SY00-Wstate-BuolamwiniJ-20190626.pdf; Lauren Smith, *Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making*, Future of Privacy Forum (Dec. 11, 2017), https://fpf.org/blog/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/.

Dr. Buolamwini's taxonomy of algorithmic harms consists of the following impacts:

1. **Loss of opportunity**, including in the areas of hiring, employment, insurance, social benefits, housing, and education;
2. **Economic loss**, including in the areas of credit and differential prices of goods;
3. **Social stigmatization**, including loss of liberty, increased surveillance, stereotype reinforcement, and other dignitary harms.

By situating GAI risks within these frameworks and the three approaches NIST uses within the AI RMF—as well as potential frameworks for identifying and measuring GAI harms to organizations, the environment, and the public overall—NIST can clarify that GAI risk management is a manageable *extension* of the AI RMF that can be adopted efficiently alongside other AI risk management processes. **Above all, EPIC recommends clarifying how risks unique to or exacerbated by GAI interact with the overlapping categorization approaches used within the AI RMF.**

Refining GAI Risks within NIST AI 600-1

In addition to EPIC's feedback on the structure of GAI risk categorization within NIST AI 600-1, we also propose seven specific changes to the risk categories currently outlined within the draft document.

a. Confabulation

Currently, NIST's definitions of confabulation and information integrity encompass the same underlying risk: individuals may believe false or synthetic content is true and real. While the process and responsibility for producing synthetic falsehoods may differ between confabulations (inaccuracies due to development shortfalls) and information integrity (inaccuracies due to negligent or intentional end-user actions), the impact is often the same.

Given NIST's approach to risk categorization discussed above, EPIC recommends combining confabulation and information integrity under the umbrella term, "Information Manipulation." In EPIC's *Generating Harms* report, for example, we categorize confabulations as a form of misinformation—i.e., unintentionally false synthetic content—alongside other forms of harmful information manipulation like scams, disinformation campaigns, and cybersecurity threats that would be more akin to NIST's definition of information integrity. Our reasoning is simple: the impacts of confabulations and other false or misleading synthetic content are the same. Inaccurate synthetic content, regardless of its source, can degrade trust, crowd out accurate and valuable content online, and influence human behavior in harmful ways. As Princeton Professor Arvind Narayanan argues, GAI 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."[21] When GAI systems like LLMs are designed to persuade an audience of the plausibility of a statement regardless of its truth, they function in ways that are not meaningfully different—at least as far as harm is concerned—from a human acting to persuade an audience of the plausibility of a statement regardless of its truth. Combining confabulations and information integrity will streamline GAI risk management and better facilitate compliance without requiring NIST to ignore important nuances in how information manipulation risks should be managed; responsible AI actors should mitigate both the risk that GAI systems generate inaccurate outputs and the risk that users will misuse a GAI system.

b. Dangerous or Violent Recommendations

Currently, NIST's discussion of dangerous or violent GAI recommendations ignores a common reason for incitement, radicalism, and other threats: election manipulation. Although incitement to violence, threats, disinformation, and more exist outside of the election context, election manipulation supercharges the risk that synthetic content will be used to erode trust, undermine democratic institutions, and spark violent or otherwise harmful conspiracy theories.[22] For example, fake audio recordings of a Slovakian party leader discussing how to rig an election were posted on social media just before the national election—an election that Slovakian party went on to lose.[23] Even OpenAI's most recent LLM model, GPT-4o, can be easily manipulated to generate racist language and conspiracy theories. Just last week, for example, Radio-Canada investigators were able to manipulate GPT-4o into producing content promoting authoritarian fascism in Quebec, anti-vaccine conspiracy theory content, content promoting self-harm, and content encouraging political violence against minorities.[24]

Because of how frequently dangerous or violent synthetic content stems from attempts at political violence and election manipulation, EPIC recommends that NIST explicitly include references to election manipulation and political violence within its description of dangerous or violent recommendations. Providing greater detail on common use contexts like elections will enable AI actors to adopt stronger and more tailored safeguards for GAI risks. For more information on the GAI risks surrounding elections, see EPIC's *Generating Harms II* report appended below.

---

[21] Julia Angwin, *Decoding the Hype About AI*, Markup (Jan. 28, 2023), https://themarkup.org/hello-world/2023/01/28/decoding-the-hype-about-ai.
[22] *See, e.g.*, EPIC GenAI II Report at 1–8; Mekela Panditharatne & Noah Giansiracusa, *How AI Puts Elections at Risk—and the Needed Safeguards*, Brennan Ctr. for Just. (July 21, 2023), https://www.brennancenter.org/our-work/analysis-opinion/how-ai-puts-elections-risk-and-needed-safeguards.
[23] *See* Morgan Meaker, *Slovakia's Election Deepfakes Show AI is a Danger to Democracy*, Wired (Oct. 3, 2023), https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/.
[24] Nicholas De Rosa, *How the New Version of ChatGPT Generates Hate and Disinformation on Command*, CBC News (May 31, 2024), https://www.cbc.ca/news/science/chatgpt-disinformation-hate-artificial-intelligence-1.7220138.

c. Data Privacy

EPIC broadly supports NIST's incorporation of detailed data privacy risks within NIST AI 600-1, including data security risks due to leaks and adversarial attacks as well as risks inherent to the widespread practice of indiscriminate web scraping for use in the training, testing, and operation of GAI systems**. To strengthen NIST's approach to GAI data privacy risks, EPIC encourages NIST to adopt GAI risk management actions that explicitly endorse data minimization and purpose limitations for data collected to train, test, and operate GAI systems.**

Current GAI development practices are built atop a data maximalist foundation: AI developers are incentivized to collect more and more data to train models, regardless of data quality or relevance.[25] As NIST has stated within NIST AI 100-2e2023: "The performance of [GAI] text-to-image and language models scales with model size and data size and quality… Thus, it has become common for [GAI] foundation model developers to scrape data from a wide range of uncurated sources."[26] While indiscriminate data collection may be convenient and cost-effective, the failure to AI developers to impose meaningful data quality controls and filtering processes means not only that the risk of inaccurate outputs increases (due to inaccurate inputs), but also that the risk of accurate but sensitive information—such as personally identifiable information—may be disclosed unintentionally during an irrelevant user interaction or exposed following an adversarial attack. Even minimal data minimization and purpose limitation expectations with GAI risk management would go far to limit myriad GAI risks at the input and development stage.

d. Environmental

Currently, NIST's treatment of GAI environmental risks in NIST AI 600-1 is focused primarily on *resource intensity*, but GAI environmental risks may also stem from issues surrounding *resource allocation*.[27] If major AI developers reduce their global carbon footprint while concentrating server clusters or resource sourcing operations to a specific region, for example, the environmental impact of GAI may be catastrophic to that specific region and its inhabitants even while the global environmental impact of GAI declines. Relatedly, most GAI resource needs are concentrated among a few major AI companies, raising market entry barriers to smaller companies that may otherwise innovate in ways that reduce the environmental costs of GAI. **To address the harms that can emerge from disparate resource allocation even as GAI resource intensity decreases, EPIC recommends that NIST explicitly include disparate resource allocation and the risks to communities local to GAI server clusters and other resource-intensive GAI sites within its discussion of GAI environmental risks**.

[25] *See* EPIC GenAI II Report at 9–17.
[26] Apostol Vassilev et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations* 40, NIST AI 100-2e2023 (2024), https://doi.org/10.6028/NIST.AI.100-2e2023.
[27] *See* EPIC GenAI I Report at 40–43, 57–59.

e. Human-AI Configuration

EPIC greatly appreciates NIST's interest in addressing risks stemming from unexpected or otherwise improper interactions between GAI systems, human developers, and users. Beyond the interactions described within NIST AI 600-1, however, GAI systems can greatly impact the way we work; even seemingly innocuous decisions around how GAI systems operate—like rapidly generating editable elements within a visual artist's software program versus generating a finished image—can massively impact those whose work the GAI system is meant to automate, augment, facilitate, or replace. For example, several news outlets and websites have already reduced their workforces in favor of using ChatGPT to write articles.[28] Hiring managers have turned to GAI to help write job descriptions and draft interview questions, among other administrative tasks.[29] Lawyers are using GAI to do research, complete administrative tasks, and even draft contracts.[30] And doctors are even using GAI to automate aspects of patient visits.[31]

Given how easily small changes in human-GAI configurations can impact labor trends, EPIC encourages NIST to explicitly incorporate the labor risks of GAI—including but not limited to job automation instead of augmentation,[32] the devaluation of labor,[33] growing economic inequality,[34] and harmful content licensing practices[35]—into its discussion of the risks stemming from human-AI configuration. For more on the labor risks of GAI, see *Generating Harms* and *Generating Harms II* (appended)

f. Intellectual Property

Currently, NIST's handling of GAI IP risks within NIST AI 600-1 focus primarily on violations of intellectual property law and other nonconsensual uses of IP. However, the IP risks of GAI systems goes far beyond legal violations and nonconsensual uses. When GAI tools are used to mimic, reproduce, or otherwise draw from protected works, they spark a complex cascade of negative incentives. Not only are the IP owners suffering economically, but the value of non-

---

[28] *See, e.g.*, David Folkenflik, *'Wall Street Journal' Layoffs Continue, Despite Lucrative AI Deal and Record Profits*, NPR (May 30, 2024), https://www.npr.org/2024/05/30/nx-s1-4986419/wall-street-journal-layoffs; 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.
[29] *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.
[30] *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/.
[31] *See* 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.
[32] EPIC GenAI I Report at 46–47.
[33] *Id.* at 48–49.
[34] *Id.*
[35] EPIC GenAI II Report at 27–33.

GAI content creation goes down as well.[36] Creators whose work has been used to train GAI systems may face deep emotional and psychological impacts.[37] And increasingly, AI content licensing agreements may exclude content creators altogether due to online platforms' predatory terms of use policies, leaving creators without recourse.[38] **EPIC encourages NIST to expand the scope of the intellectual property risk in NIST AI 600-1 to incorporate these second-order effects of nonconsensual IP usage within GAI systems.** For more information on second-order IP harms, see *Generating Harms* and *Generating Harms II* (appended).

g. Obscene, Degrading, and/or Abusive Content

Currently, NIST's handling of obscene, degrading, and abusive synthetic content focuses primarily on the production of and access to harmful content without adequately describing the intent behind this synthetic content or the downstream effects of their use. Further describing these aspects of abusive GAI content is important because abusive uses of GAI technologies can go beyond the creation of obviously obscene, degrading, or abusive content. In *Generating Harms*, EPIC lists this category of GAI harm as "harassment, impersonation, and extortion": generating deepfakes can be abusive not only because they show the target in obscene, degrading, or otherwise abusive scenarios, but also because seemingly innocuous synthetic content—a synthetic video or audio message from a friend or family member, for example—may be used as a tool to force the targets of abuse into obscene, degrading, or otherwise exploitative circumstances. **EPIC recommends that NIST expand this section to include the use of non-obscene, degrading, or abusive synthetic content as a tool for obscene, degrading, abusive, or otherwise exploitative means.** Scams, blackmail, and extortion are some of the oldest and most common abuses of GAI technologies, yet NIST AI 600-1 is almost entirely silent on these risks.

AN ADDITIONAL RISK: DATA DEGRADATION

GAI companies have flooded the internet with synthetic content of varying quality—a tidal wave of AI text, imagery, and video that has demonstrably worsened the average user experience online. The process of drowning out reliably, non-synthetic content with low-quality synthetic content—a process that EPIC calls "data degradation"—raises two discrete risks. First, the rapid production and dissemination of low-quality synthetic content—incoherent articles to farm search engine clicks, confabulations, discriminatory outputs, and more—makes it increasingly difficult to find the information we want online.[39] Truthful and reliable information is lost within a sea of confabulations. Second, the growing wave of synthetic content being disseminated across the web risks creating a negative feedback loop for AI development: GAI systems tend to fall apart when trained on synthetic content, making it increasingly difficult for new market entrants to build GAI

[36] EPIC GenAI I Report at 36–37.
[37] *Id.* at 37–38.
[38] EPIC GenAI II Report at 27–33.
[39] *Id.*

models using traditional methods for collecting training data.[40] Specifically, training AI models on synthetic content 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."[41]

To counteract the degradation of new and forthcoming GAI models—as well as our shared online experience—EPIC urges NIST to add data degradation or model collapse as a unique risk of GAI within NIST AI 600-1.

II. WATERMARKING IS NOT SUFFICIENT TO MITIGATE THE RISKS OF SYNTHETIC CONTENT

Responsive to NIST AI 100-4: Reducing Risks Posed by Synthetic Content

EPIC has a longstanding interest in regulating synthetic content, which not only poses privacy and data risks tied to training, test, and input data, but also implicates risks connected to negligent or malicious uses of GAI.[42] In recent comments to NIST, for example, EPIC encouraged the adoption of AI watermarking standards as one way to increase transparency and accountability around the development and use of GAI.[43] As we wrote, "providing durable methods for watermarking synthetic content for end-users is a form—albeit imperfect—of risk-mitigating transparency."[44] However, some of the most popular methods for labeling or watermarking synthetic content are insufficiently durable to serve as reliable risk mitigation techniques. Even when AI developers insert watermarks imperceptibly into the pixels or metadata of synthetic

[40] *See* EPIC GenAI II Report at 18–26; EPIC, Comments on NIST's RFI Related to NIST's Assignments Under Sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence, 88 Fed. Reg. 88368 (Feb. 2, 2024), https://epic.org/wp-content/uploads/2024/02/EPIC-Comment-on-NIST-AI-Executive-Order-Mandates-RFI-02.02.24.pdf.
[41] 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/%7Eis410/Papers/dementia_arxiv.pdf.
[42] *See, e.g.*, EPIC GenAI I Report at 9–10, 24–32; EPIC GenAI II Report at 18–26; EPIC, Comments on the NTIA's Request for Comment Concerning Dual Use Foundation AI Models with Widely Available Model Weights, 89 Fed. Reg. 14059 (Mar. 27, 2024), https://epic.org/wp-content/uploads/2024/03/EPIC_Comment_NTIA_Dual_Use_Foundation_Models_with_Appendix.pdf.
[43] EPIC, Comments on NIST's RFI Related to NIST's Assignments Under Sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence, 88 Fed. Reg. 88368 (Feb. 2, 2024), https://epic.org/wp-content/uploads/2024/02/EPIC-Comment-on-NIST-AI-Executive-Order-Mandates-RFI-02.02.24.pdf.
[44] *Id.* at 5; 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), https://arxiv.org/pdf/2310.00076.pdf

content using tools like Google's SynthID, for example, research out of the University of Maryland has identified methods to remove them—and even insert false watermarks into images.[45]

Understanding the limitations of labeling and watermarking techniques is necessary to develop an effective synthetic content risk management process; AI developers cannot rely on imperceptibility and robustness alone when pursuing AI watermarking. **To better incorporate labeling and watermarking within NIST's broader approach to GAI risks, EPIC recommends further exploring how to foster greater transparency around the limitations of digital content transparency and how to counteract them. For example, in the table that starts on page 4 of NIST AI 100-4, EPIC encourages NIST to add a column listing common use and context limitations for each digital content transparency approach.** Adding information about limitations—without providing instructions for malicious users to circumvent a digital content transparency approach—will enable AI developers to better understand when to incorporate multiple transparency techniques within its AI risk management process, as well as how to complement digital content transparency with other AI risk management techniques.

III. Global Engagement on AI Standards Requires Global Engagement with Civil Society, Academia, and Impacted Communities

Responsive to NIST AI 100-5: A Plan for Global Engagement on AI Standards

EPIC has a long history of engaging with international AI standard-setting[46] and welcomes NIST's efforts to champion international coordination on responsible AI development and use standards. The focus of AI risk management has been and should continue to be on mitigating risks to human safety and rights; innovation, national security, and competitive advantages mean nothing if they come at the cost of our lives, livelihoods, rights, or opportunities. For these reasons, EPIC applauds NIST's efforts to pursue global engagement on AI standards, especially insofar as they foster greater transparency and diversity within the development of global AI standards.

To further foster effective and efficient coordination on global AI standards within NIST AI 100-5 and beyond, EPIC encourages NIST to engage deeply with stakeholders *beyond* government bodies and standards developing organizations. Civil society groups, academia, and impacted communities across different countries provide critical research and experiential insights that standards bodies, government agencies, and industry actors do not. For

[45] *See* Kell, *supra* note 44; Saberi et al., *supra* note 44; David Pierce, *Google Made a 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.

[46] *See, e.g.*, EPIC, Comments on NIST Request for Information on AI Standards, 84 Fed. Reg. 18490 (May 31, 2019), https://epic.org/wp-content/uploads/privacy/ai/NIST-RFI-EPIC%2020190531.pdf.

example, Data Privacy Brazil, a nonprofit civil society organization based in Brazil, has already published comprehensive research into AI regulatory interoperability across countries.[47] **To ensure the United States leverages the research advances achieved by organizations like Data Privacy Brazil, EPIC urges NIST to include more robust, frequent, and explicit engagement and feedback mechanisms with civil society, academia, and impacted communities as part of NIST AI 100-5. Additionally, EPIC encourages NIST to add an additional bullet point to Section 5.1: "Facilitate research access to AI datasets, models, and related materials to ensure standards development efforts are backed by robust research."** Providing access to underlying AI data, models, processes, and testing materials will be a crucial step to ensure that AI standards reflect the reality of AI development—and enable meaningful AI oversight long after AI standards have been published.

CONCLUSION

EPIC welcomes NIST's efforts to identify and address the complex risks of GAI and synthetic content pursuant to Executive Order 14110. GAI both imposes new risks beyond those outlined in the AI RMF and exacerbates risks already identified in the AI RMF. However, output- and data provenance-focused interventions are only a subset of the techniques needed to adequately mitigate the risks of new and emerging GAI technologies. To improve the effectiveness and longevity of its GAI draft documents, NIST can and should:

1. Further refine its approach to GAI risks, given the overlapping risk categorization approaches found within the AI RMF and the value of civil society and academic typologies for AI harms;

2. Expand the scope of the twelve risks currently included to add important clarity for how GAI risks emerge and interact;

3. Consider ways to discuss the limitations of digital content transparency approaches within NIST AI 100-4, such that AI actors can effectively pursue complementary and redundant risk management strategies for synthetic content; and

4. Engage more deeply with civil society, academia, and impacted communities throughout NIST AI 100-5 to minimize redundant research needs and maximize the effectiveness of global AI standards.

[47] *See* Bruno Biono et al., Data Privacy Brazil, *Key Themes in AI Regulation: The Local, Regional, and Global in the Pursuit of Regulatory Interoperability* (2023), https://www.dataprivacybr.org/wp-content/uploads/2024/05/position-paper-AI-ENG-Final.pdf.

EPIC appreciates this opportunity to reply to NIST's Request for Comments and are willing to engage with NIST further on any of the issues raised within our comment. 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/ Grant Fergusson
Grant Fergusson
Equal Justice Works Fellow

/s/ Enid Zhou
Enid Zhou
EPIC Senior Counsel

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

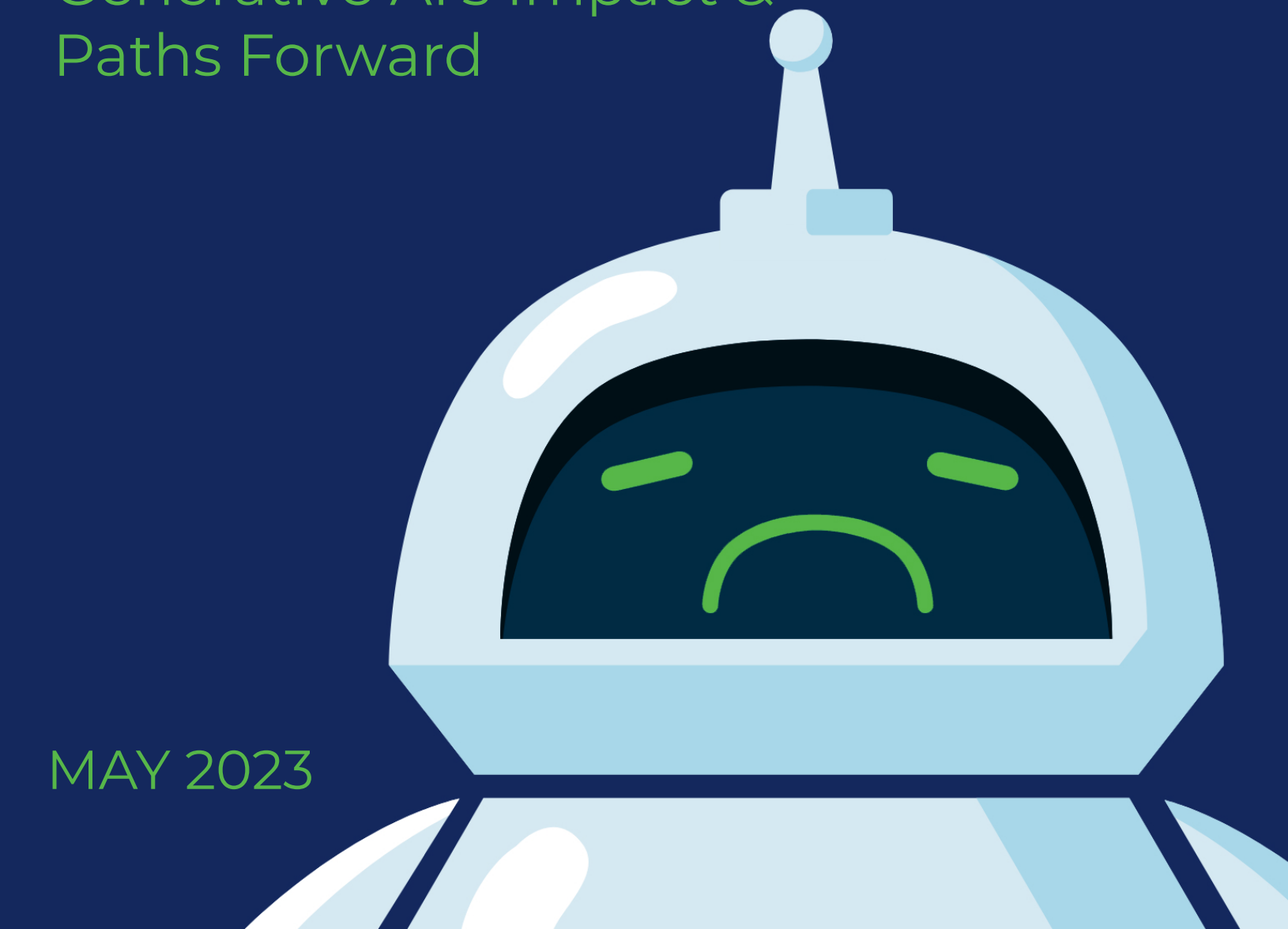
ELECTRONIC PRIVACY
INFORMATION CENTER (EPIC)
1519 New Hampshire Ave. NW
Washington, DC 20036
202-483-1140 (tel)
202-483-1248 (fax)

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.

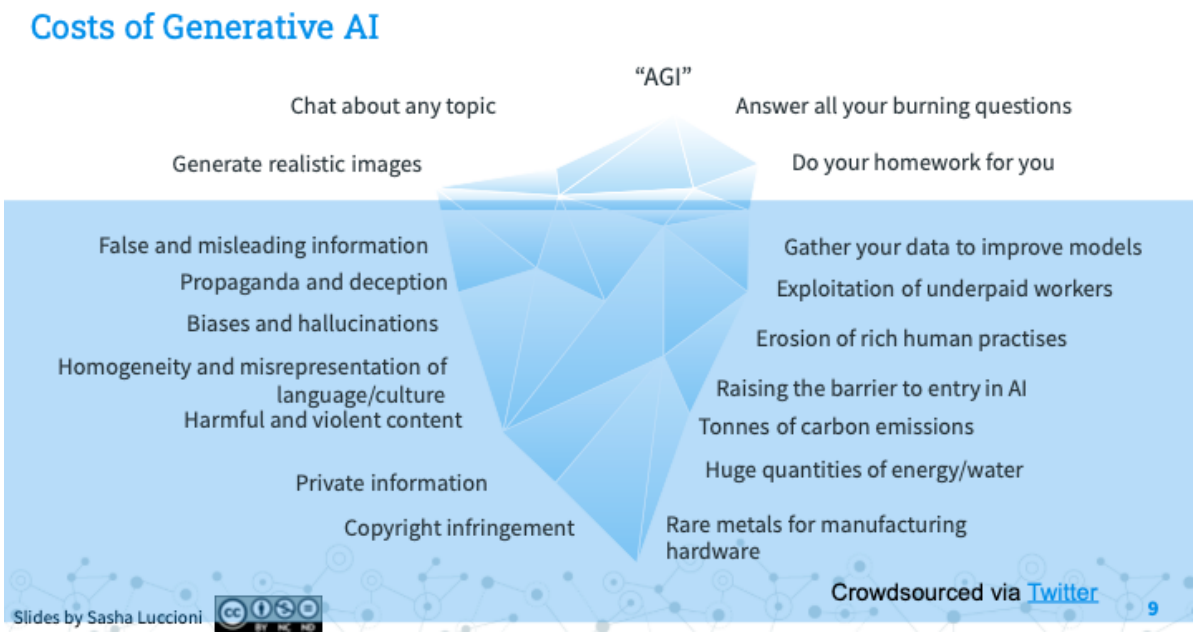


Figure 1: Credit: Sasha Luccioni

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

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.

> 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

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.

Examples	Harms: 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.

epic.org

ELECTRONIC PRIVACY INFORMATION CENTER

MAY 2024

GENERATING HARMS

Generative AI's New & Continued Impacts

CONTRIBUTIONS BY

Grant Fergusson
Sara Geoghegan
Calli Schroeder
Maria Villegas Bravo

EDITED BY

Chris Frascella, Tom McBrien, Calli Schroeder, Maria Villegas Bravo, Kara Williams, and Enid Zhou

Notes on this Paper:

This paper builds on the structure and issues raised by *Generating Harms: Generative AI's Impact & Paths Forward*, our report released in May 2023.

We have expanded into a sequel report to address additional harm areas that have become clearer in the intervening year. As generative AI continues to develop, we anticipate there may be future editions as well.

While we have largely maintained the typology of harms from the first report (modeled from Danielle Citron and Daniel Solove's Typology of Privacy Harms and Joy Buolamwini's Taxonomy of Algorithmic Harms), there are some issue areas that presented specific harm types outside of those categories. We have identified these where necessary.

Table of Contents

Introduction

In the year since EPIC issued the first Generating Harms Report,[1] generative AI has exhibited near-unfettered growth, expanding across every industry and policy discussion worldwide. Though proponents of generative AI continue to tout these systems as having unlimited and world-altering capabilities, many have challenged this narrative. The public is not only aware of the spread of generative AI: They actively distrust and dislike it,[2] especially because of its impact on privacy.[3] Enforcement bodies have held multiple hearings on generative AI and are rapidly moving forward with targeted regulations. Technologists, civil rights leaders, advocates, and other experts continue to draw attention to existing laws and principles that should curb generative AI. In this report, we intend to add to that effort.

One marked shift that we have seen in the past year lies in the conversation about generative AI harms. Despite industry efforts to pivot harms discussions to hypothetical future existential risks of generative AI, the focus is now squarely on the present: Who are these systems harming, what are the harms, and what can we do about it? The effort to distract us with ideas of a dystopian future has not stopped us from keeping up the pressure to change our dystopian present.

Emphasizing the real-life harms perpetuated by generative AI serves several purposes. It raises basic questions about the fitness and safety of the technology, prompting discussion of why we are permitting unproven technology into sensitive areas of our lives. It forces companies developing and selling generative AI to be accountable for the consequences of their technology's largely unchecked spread. It pushes authorities to look at how generative AI harms their constituents and what they can and should do to stop that harm.

This report expands on our previous work by looking at additional areas of generative AI harms and some of the efforts taken to counter them. The areas of harm that we identified in our initial report remain relevant, with multiple examples of both real-life harms and active pushback against generative AI in those spaces. However, additional harm categories have emerged that merit examination. In addition, a year of expanded generative AI development and discussion has produced several proposed and implemented efforts to counter generative AI harms. This report sheds light on both of these aspects of generative AI.

— Calli Schroeder, EPIC Senior Counsel and Global Privacy Counsel

Endangering Elections

BACKGROUND AND RISKS

Through election-related harms, generative AI can alter the futures of nations. Problems like misinformation, disinformation, foreign influence, and security issues are not new. But generative AI supercharges those problems to a degree that presents a genuine threat to democracy.[4] AI can generate convincing deceptive text, audio, and video content that often takes a great deal of time and technical knowledge to debunk. In addition, the volume of election-related harms enabled by generative AI turns problems that were once minor or manageable into devastating attacks that erode trust in elections. Think of it like a building that previously had the occasional rock thrown at it now being hit by a wrecking ball.

Election-related harms can range from degrading trust in elections to manipulating voter behavior to creating security threats. If elections are successfully altered via AI-generated content, the harms expand to empowering parties who otherwise would not have won and altering policy and political decisions for years to come. In this section, we look at three ways generative AI is damaging elections: mis/disinformation, foreign influence, and security infrastructure. Unless these harms are dealt with swiftly, we will see a massive erosion of trust in the security and accuracy of elections, leading to widespread mistrust and societal unrest with potentially catastrophic consequences.

MISINFORMATION AND DISINFORMATION

Political campaigns are often plagued by rumors, scandals, and falsehoods about candidates, campaigns, and political parties—what we categorize as

misinformation and disinformation. Typically, misinformation refers to the spread and creation of unintentionally false information while disinformation is purposeful spread of false information. However, the speed at which AI-generated content spreads blurs this distinction, as people share information without first editing or fact-checking it.[5] Whether false information is created and spread purposefully or unknowingly, the impact is the same.

AI-generated mis- and disinformation can be polished and highly convincing, often mixing accurate and inaccurate information. Even when not deliberate or prompted by a user intending to produce disinformation, generative AI systems may spontaneously generate inaccurate information when they are trained on data that is not reviewed for accuracy prior to its addition to the training dataset or where the algorithm draws the wrong conclusions and connections internally. In some cases, the incorrect information may stem from the phrasing of the end user's prompt and be incorporated into the generated content. Generative AI systems have more than enough content in their training databases to produce wildly inaccurate election information, as there have been years of discussion, conspiracies, and deception related to voting machine security, mail voting, "stolen" election narratives, and biographical information and theories about candidates that these systems can draw from.

AI-generated mis- and disinformation can take many forms, with many repercussions. Audio, video, and images may warp public perception of candidates, their policies, their actions, or broad election practices (such as deepfake images of discarded mail ballots), producing scandals based on entirely false information. The volume of generated content may convince voters that there is broad consensus on a political matter or widespread acceptance of a falsified scandal. AI may be combined with targeted influence campaigns using demographic information held by data brokers,

allowing highly tailored and manipulative targeting of individuals or groups with similar characteristics. AI-generated content may flood party representatives with comments from fake "constituents" or put out false poll information to push parties to change their political stances in response to the "will of the people." Voters could be disenfranchised through AI-generated content, spreading false information about where and how to vote, eligibility requirements, or false risks of voting.

Once created, this AI-generated content will spread like wildfire across social media, in chat groups, and possibly even through news outlets. Reports have demonstrated that mis- and disinformation tend to have *more* engagement than factual information.[6] The risk is even higher in areas without reliable local news, where AI-generated content and websites will take advantage of the gap to seed mis- and disinformation.[7]

The impact of mis- and disinformation is particularly damaging when it is released without adequate time to effectively debunk it. For instance, fake audio recordings of a Slovakian party leader discussing how to rig the election was posted on social media just days before the election—that party then lost the election.[8] The challenge of confirming whether content is real cuts both ways. We have already begun to see politicians claim that actual photos, videos, and audio recordings of them were AI generated.[9] When voters have no clear way to determine when information is true or fake, they are much more vulnerable to manipulation, election interference, and other propaganda that may affect how they vote.

In addition to concerns about accuracy around candidates and political stances, AI-generated mis- and disinformation could be used to directly disenfranchise voters. For example, two days before the primary in New Hampshire, robocalls went out using AI-generated audio of President Biden's voice to urge Democrat voters to "save their votes" for November

and not vote in the primary.[10] Similar ploys could tell voters incorrect poll or voting time information or attempt to intimidate voters by claiming voting would put them at risk. For instance, Jacob Wohl and Jack Burkman made robocalls to Black New Yorkers claiming that their personal information would be sent to law enforcement, debt collectors, and other authorities if they voted by mail.[11] Mimicking the voice or image of a trusted figure through generative AI makes these types of attacks even more potentially damaging.

FOREIGN INFLUENCE

Generative AI's ability to produce content that sounds like it comes from a local speaker of the target language is a boon to foreign actors attempting to influence elections. Textual tells, odd ways of forming sentences, or phrasing have often been giveaways that a piece of content may come from a foreign actor. AI services smooth over these signs, producing content that is indistinguishable from how a local would speak or write it. New York City mayor Eric Adams has already employed this tactic, using AI to generate robocall messages of his voice speaking multiple languages.[12] Further, if a foreign actor wants to make audio or video content to make themselves more convincing, generative AI can create speech in a local accent. Because it is very difficult to confirm where content has been AI generated, much less what human individual prompted the creation, authorities have a very difficult time enforcing against foreign actors who would take advantage of this technology.[13]

SECURITY AND PERSONAL SAFETY

Elections are often a time of dramatically increased communications to voters from candidates, political parties, and interest groups, creating an opportunity for bad actors to solicit personal and sensitive information from individuals under the guise of needing it to process voting information or financial support. For example, someone may thank an individual for their

"donation" to a cause. When the individual responds that they made no such donation, that person may then request bank information in order to cancel the transaction. Generative AI systems could be used to generate phishing attacks on individuals, or hackers could embed malware in AI-generated content. While these risks have already been addressed in our initial Generated Harms report under "Turbocharging Information Manipulation," the unique circumstances of elections may make individuals more likely to believe authority figures are contacting them and that they should respond. In addition, generative AI could be used to directly target election officials for phishing or doxing attacks, endangering election security even further.[14]

HARMS

- **Economic/Economic Loss**: Scams, phishing attacks, and malware may result in direct economic loss for individuals through fraudulent donation funds, gaining access to financial accounts, and other instances of manipulation and deception. Successful attacks in this area may also have long-term impacts on credit.
- **Reputational/Relationship/Social Stigmatization**: Mis- and disinformation about individuals connected to elections (candidates, their families, party leaders, etc.) can permanently impact their public image and interpersonal relationships, potentially on multiple levels—local, national, and international.
- **Psychological**: The individuals implicated in mis- and disinformation may face severe emotional harm, shame, and embarrassment due to the spread of false information, recordings, and images. In addition, individuals tricked by scams or mis- and disinformation campaigns that are later debunked may be embarrassed or feel manipulated and used.

- **Autonomy**: The wide spread of mis- and disinformation and the difficulty of confirming the truth in a timely and effective manner impacts individuals' ability to make a properly informed and non-manipulated decision regarding their vote and other actions.
- **Discrimination**: Many of the harms listed above are specifically targeted at marginalized or vulnerable communities.[15]
- **Societal**: The effects of election-specific generative AI harms erode trust in government and basic voting institutions, threatening civic participation and democracy.

EXAMPLES

- Microsoft has already tracked a marked increase in campaigns from both Russia[16] and China[17] to influence U.S. elections using AI-generated content.
- New Hampshire residents were bombarded three days before the primary with robocalls featuring an AI-generated audio mimicking President Biden's voice, instructing them not to vote in the primary.[18]
- Deepfake videos of U.S. politicians have been widely circulated. This includes a video of "President Biden" reinstating the draft,[19] which was viewed more than 8 million times on Twitter, and a video of "Senator Elizabeth Warren" saying Republicans should be barred from voting.[20]
- Internationally, harmful AI-generated content has thus far included a video of Moldova's president supporting a Russia-friendly political party,[21] audio of Slovakia's liberal party leader discussing vote rigging,[22] and videos of female Bangladeshi opposition politicians wearing a bikini and in a swimming pool (offending many who observe the conservative dressing customs of the Muslim-majority country).[23]
- Politicians are now arguing that authentic text, audio, and video footage of them has been AI-generated, including former President

Trump dismissing attack ads featuring videos of public gaffes,[24] a Taiwanese politician denying the accuracy of video of him entering a hotel with a woman,[25] and an Indian politician denouncing audio of him accusing his party of illegally accumulating $3.6 billion.[26]

- Several[27] analyses[28] have demonstrated how social media algorithms amplify mis- and disinformation over factual information.

INTERVENTIONS

- Enact a law that would make generating deliberate mis- or disinformation regarding elections, candidates, or political parties illegal. Some examples of this approach are the proposed Deceptive Practices and Voter Intimidation Prevention Act and South Carolina's H4660.[29]
- Build on the FCC's decision to outlaw robocalls[30] containing AI-generated voices to combat mis- and disinformation campaigns impersonating politicians. Federal agencies could also mandate that political ads either be barred from using AI-generated content or be forced to make clear disclosures where they are used—including information put out by paid influencers.
- Mandate that generative AI companies adopt baseline precautions to counter AI disruption of elections, such as preventing the technology from generating images or impersonations of known political figures.[31]
- Require AI developers to implement filters for known election falsehoods and continuously update those filters as more falsehoods are developed and spread. This includes both filtering outputs of generative AI systems and constantly checking the training datasets to ensure accuracy.
- Require social media companies to proactively develop frameworks to counter AI-generated election risks. Approaches may include

modifying their algorithms to ensure mis- and disinformation are not promoted or enabling a trust verification on fact-checked information. Some technology companies have already signed on to a voluntary framework as a first step.[32]

Eroding Privacy

BACKGROUND AND RISKS

While several harm categories in our initial report mentioned privacy as a risk factor, none examined how existing generative AI systems and practices damage individual and societal privacy. Generative AI consumes vast amounts of data, from data used to create training datasets, to data created while monitoring user inputs, to the platform's own pattern forming and outputs. Unless carefully and consistently curated, that data will include personal data, sensitive data, and inferred information about individuals. People whose data is processed by generative AI may be unable to get the data removed and, in many cases, may be entirely unaware that their personal data is being used by the system at all. In this section, we look at three key areas of privacy harm: maximalist data use, scraping to train data, and data security issues.

MAXIMALIST DATA USE

Generative AI is built on a data maximalist approach: AI developers are incentivized to collect more and more data to train models. Often, developers will use methods like data scraping to indiscriminately collect information from the internet to feed and train their models with little regard to the quality or accuracy of the data. The National Institute of Standards and Technology ("NIST") has explained:

> The performance of GenAI text-to-image and language models scales with model size and dataset size and quality. For example, scaling laws indicate that training a 500 billion parameter models would require 11 trillion tokens of training data. Thus, it has

> become common for GenAI foundation model developers to scrape data from a wider range of uncurated sources.[33]

Some developers may use filters or parameters to attempt to sanitize or clean the data, but the industry lacks meaningful oversight or requirements to ensure this.

This approach exposes individuals to exorbitant privacy risks and violations. Uncontrolled data scraping will pick up every piece of personal information that can be found online. This includes items that a person may consider public or benign (name, employer, age), more sensitive data that a person may want more tightly protected (address, relationships, location), and highly sensitive data that could expose or endanger an individual (sexuality, health information, religion, political affiliation). It may also include false or illegal personal data, such as revenge porn images or deepfakes of the individual. Most data scraping does not have meaningful checks in place to ensure that the collected data is accurate, high-quality, or even legal (as has been demonstrated by high-profile investigations showing child sexual abuse material ("CSAM") in several datasets).[34] Processing these massive amounts of data also take a significant environmental toll due to the energy and water required for processing data and cooling components.

Not only is indiscriminate scraping potentially harmful on an individual level, the practice also contradicts fundamental data privacy principles like data minimization and purpose limitation. Data minimization requires anyone using personal data to limit its collection, use, disclosure, processing, and retention to only the information that is reasonably necessary to furnish the product or service requested. Purpose limitations, a necessary part of a strong data minimization rule, require that the data only be used for the purpose for which the data was collected. For example, if someone provides their email address to verify their account when making a purchase online,

they would likely expect to receive email communications about the status of their order, since that is connected to the service they have requested. However, if they were to post a status and a photo of their children on social media, they would not expect that the photo and text would be scraped by an AI developer and used to train an AI system. Not only is that use wholly unrelated to the purpose of the post, it involves a company they have never willingly interacted with taking and using personal data with no notice to the individual or opportunity for them to refuse. Similarly, a person might interact with a chatbot on a website seeking medical or mental health advice. That user would not expect that their sensitive health information could be used to train the bot and this purpose would contradict assumptions of confidentiality and discretion around this sensitive information.

Generative AI practices contradict several other generally recognized data privacy principles like accountability, accuracy, and transparency. When developers indiscriminately scrape publicly available websites for data to train their models, they may have no actual knowledge of the quality or legality of the data collected. Scraping lacks meaningful oversight to ensure that the data is accurate or without bias. Data collection and training dataset creation practices for generative AI are largely opaque, happening in black boxes without transparency. Without universal requirements to limit this, developers will build, train, and deploy a generative AI system without any independent review, such as third-party audits to test for accuracy and bias or independent privacy impact assessments. California has attempted to correct this with its newest regulations, but the impacts will remain to be seen because the rules have not yet gone into effect.

SCRAPING TO TRAIN DATA

Generative AI systems are often built using training datasets comprising data scraped indiscriminately from the web. That means the data used to

feed generative AI systems can be any information that is publicly available online. Most companies developing these models do not release detailed information about the data sets they have used, "but these data sets inevitably include some sensitive personal information, such as addresses, phone numbers, and email addresses."[35] This may also include data released through breach, illegally shared information like CSAM or revenge porn, or information revealed through doxing. Training datasets have already been shown to include confidential and privileged information that should not have been publicly available, such as private medical record photos.[36]

Often, generative AI systems lack filters to prevent this type of data from being used to build or train models. Foundation models make "heavy use of unsupervised learning" during the pre-training stage when they are created.[37] This means information that was never meant to be public could be used to train a system, which poses serious threats to privacy, security, and accuracy. When a generative AI system inputs this type of information, it may generate a result that includes the very information that fed the system. For example, data from a breach that reveals a consumer's social security number could be scraped and used in a generative AI system that generates a response that includes the consumer's social security number. Images that were scraped and fed into the system may generate substantially similar images, revealing intimate details about the person in the scraped image. The lack of oversight and limitations on the collection of publicly available personal information poses a serious privacy threat because sensitive personal information may be revealed in a generative AI system's output.

The European Union has begun to break ground on regulating the use of web scraping because of its threat to the fundamental rights to privacy and data protection. For example, the Dutch Data Protection Authority ("DPA") recently published guidelines discussing web scraping.[38] The DPA

concluded that web scrapers and web crawlers almost inevitably capture personal data, including special categories of data because of the sheer breadth of data on the internet (including data leaked from data breaches)[39] In most cases, this data scraping violates European law because there is no lawful basis of processing of the data, nor is there notification to the data subjects that their data has been processed.[40] The Artificial Intelligence Act ("AI Act" or "EU AI Act"), set to come into force in 2026, also strictly prohibits algorithms that scrape images off the internet or CCTV footage for the purpose of creating or expanding facial recognition databases.[41]

Finally, the vast amount of data consumed by these systems can lead to extremely revealing inferences. Generative AI systems are built to detect patterns in how information is connected and constructed, from basic sentence structure to much more revealing information. Once those patterns are built into a system, it may solidify inferred information that may be highly sensitive, like analyzing communication patterns to reveal romantic relationships or tracking movement patterns that would reveal an individual's home address.

DATA SECURITY ISSUES

The excessive collection, use, and retention of personal information in generative AI systems makes them ripe for data security issues. As explained previously, generative AI systems rely on massive amounts of data that are retained indefinitely. This includes both the data collected for training datasets and data input into the AI system by users. Security concerns with this second data collection avenue has prompted bans from scores of federal agencies[42] and private businesses, particularly across medicine, finance, journalism, security, and other industries dealing with sensitive confidentiality requirements.[43] The persistent threat of a breach increases when data is excessively collected or retained; data that is

deleted after it is no longer needed cannot be subject to breach. These indefinitely retained troves of personal information are vulnerable to attacks by bad actors, either through breach or adversarial machine-learning techniques. One such technique is prompt injection, which can get an AI system to reveal its raw training data, exposing any personal data contained in the dataset. There are two types of prompt injection attacks: direct and indirect. A direct prompt injection happens when an actor inputs text intending to alter the behavior of the Large Language Model ("LLM").[44] An indirect prompt injection involves an attacker that manipulates the data used in a LLM to remotely inject system prompts without directly interacting with the application.[45] One example of an indirect prompt injection is when an actor changes information on a webpage that the LLM will read when generating a prompt, manipulating the outcome. NIST explained how many parts in the process can be susceptible:

> These security and privacy challenges include the potential for adversarial manipulation of training data, adversarial exploitation of model vulnerabilities to adversely affect the performance of the AI system, and even malicious manipulations, modifications or mere interaction with models to exfiltrate sensitive information about people represented in the data, about the model itself, or proprietary enterprise data.[46]

These attacks may allow either direct access to raw data contained in the datasets, which will almost certainly include personal data, or may basically "short circuit" the generative AI system to provide pieces of personal data in its outputs.[47] Generative AI provides more opportunities for bad actors to manipulate data and prompts to manipulate outputs. Without meaningful limitations, these opportunities threaten the security of the data used to build and train a system.

HARMS

- **Physical**: Generative AI systems may reveal data that could put a person at risk, whether real or perceived. For example, sensitive information, such as a person's email address or home address may be revealed to stalkers, abusers, or other bad actors.
- **Reputational/Relationship/Social Stigmatization**: Generative AI can reveal a person's sensitive information, which may result in damage to reputation or social stigmatization. A person's sexuality may be inferred where sexual images, information related to sexual behaviors, or location information were fed to an LLM.
- **Economic**: Businesses whose trade secrets have been incorporated into training sets or individuals whose economic information has been incorporated face potential economic injuries.
- **Psychological**: Individuals may suffer from anxiety or fear due to the lack of control over removing their personal data from training sets and may fear consequences from the ways in which the data is used. People may also be angry, frustrated, or feel exploited that their information has been used to feed a for-profit LLM, even where the data is anonymized.
- **Autonomy**: Individuals cannot control the collection and use of their personal information, including whether it is used to train datasets.
- **Discrimination**: Biased data can be scraped and fed into an LLM, which will likely then produce biased results built from historic discrimination.

EXAMPLES

- ChatGPT falsely accused a law professor of sexual harassment by including his name on a generated list of professors that had sexually harassed someone, citing a non-existent news article.[48]
- Researchers at Google forced ChatGPT to reveal some of its training data, revealing an email address and phone number. The researchers said that when instructed to repeat a single word forever, the bot often released personal information and raw training data.[49]
- A recent report[50] found that "an organization can expect around 660 daily prompts to ChatGPT for every 10,000 users, with source code being the most frequently exposed type of sensitive data, posted by 22 out of 10,000 enterprise users and generating, on average, 158 incidents monthly. This is ahead of regulated data (on average, 18 incidents), intellectual property (on average, four incidents), and posts containing passwords and keys (on average, four incidents) every month."[51]
- A ChatGPT vulnerability may have revealed users' payment-related information and titles from users' chat histories.[52]
- Microsoft, a strong proponent of generative AI and the developer of one of the most popular generative AI systems, accidentally incorporated two employees' back up computers—including passwords, encryption keys, and Teams threads—into its training data.[53]

INTERVENTIONS

- Enact a general, comprehensive federal privacy law that limits the collection, use, and retention of personal information to that which is reasonably necessary to fulfill the service requested. This will include

purpose limitations to ensure that personal information is not used in an out-of-context way to train generative AI systems.

- NIST has suggested several mitigation techniques to minimize harms from prompt injections, while explaining that these measures do not provide full immunity to all attacker techniques: training for alignment, prompt instruction and formatting techniques, detection techniques, reinforcement learning from human feedback, filtering retried inputs, an LLM moderator, interpretability-based solutions.[54]
- Enforce laws that prohibit unfair and deceptive trade practices, mandate consent requirements for child users, and require justification for data processing.
- Build support tools by only 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.

Data Degradation

BACKGROUND AND RISKS

Generative AI companies have flooded the digital public square with content of varying quality that has demonstrably worsened the experience of daily internet use. The influx of synthetic content has fundamentally stripped the utility from the internet as a service, leaving end users locked in to an ever-worsening system. This phenomenon mirrors what Cory Doctorow describes as "enshittification," which we refer to as "data degradation" in this context.[55] The data degradation process is gradual with consistent, identifiable steps. First, generative AI companies hook businesses, governments, and users on the idea that they will make life better and easier by performing some simple tasks (like text generation) well and claiming that this success can extend to unlimited other use cases. Second, the companies abuse the primary creators and audiences of synthetic content to provide value to business clients—in this case, by flooding search engine results to brute force search engine optimization and game social media algorithms. Third, and finally, the AI company will claw back more value for themselves by cutting costs, downgrading the quality of their offerings and exploiting the business clients. Software that was once open source will be taken off the market and deemed proprietary;[56] previously curated datasets will be taken over by datasets built by web scrapers. This process ends in a flood of inaccurate, non-sensical, and low value synthetic content that takes over the digital ecosystem so completely that it is no longer possible to meaningfully sort the good from the bad.

This data degradation problem is two-pronged, partly because the quality of AI-generated content varies wildly. On one hand, much of the content being generated is low quality—incoherent sentence structures, inaccurate data,

discriminatory outputs, and a penchant for hallucinations.[57] Generating low quality content creates a feedback loop where the system that scrapes the internet and its own outputs for new training data ingests the low quality content, leading to eventual model decay and continually worsening output quality.[58] On the other hand, synthetic content that is indistinguishable from authentic content is distorting reality and overwhelming human-generated content. Convincing and higher-quality outputs make humans unable to distinguish between authentic content and generated content, which not only erodes trust but also means people will unknowingly amplify AI generated content to the detriment of human created content. This lack of trust and ability to differentiate content carries over into settings that require objective, authentic evidence, like trials, mediations, election campaigns, and more. Every corner of the internet, and beyond,[59] has been infiltrated by generative AI to the detriment of everyday people.

A FULL-SCALE INVASION

The actual scale of the generative AI problem is impossible to quantify, but more and more of the internet is being taken over by synthetic content. Generative AI developers and deployers are seeing unprecedented growth. In February 2023, ChatGPT boasted 100 million monthly users after only two months in business—a feat that took Facebook almost four and a half years after launch.[60] Nine months later, ChatGPT reported over 100 million weekly users after only a year in business.[61] Over two million developers have taken advantage of OpenAI's model with ChatGPT and Whisper's (a speech recognition tool) application programming interfaces ("API"), which allow entities to use and iterate on software for the entity's own use.[62] In total, the top 50 generative AI models received over 24 *billion* site views in the course of 12 months.[63] Retail platforms like Amazon[64] and Etsy[65] are being flooded by AI generated listings, and generative AI is creating full websites in seconds.[66]

Generative AI companies are not the only culprits. Platforms like social media websites and search engines are boosting synthetic content across the internet. For example, AI generated news articles are topping search engine results, often beating out authentic news outlets, including those whose content was scraped to create the synthetic article.[67] In a preprint study by the Stanford Internet Observatory, researchers found that Facebook's recommendation algorithm pushes AI generated content because AI-generated content appears to generate more engagement.[68] In fact, one of the most viewed pieces of content on Facebook in Q3 of 2023 (boasting 40 million views and almost 2 million interactions) was an AI generated image.[69]

One reason that the actual scale of the synthetic content issue is hard to calculate is that there are currently no consistent requirements on generative AI companies or platforms to label or otherwise distinguish synthetic content from authentic content. In the United States, many of the proposals around labelling AI-generated content are tied to election-related content.[70] In addition, NIST[71] and several state legislatures have proposed more general watermarking requirements, but none have been enacted as of the release of this publication.[72] The European Union's AI Act requires any fully synthetic or partially manipulated content to be labelled and obvious to the user as AI-generated before or during interaction with the content.[73] However, the AI Act will not come into force until 2026. The United Kingdom Information Commissioner's Office is also exploring possible watermarked or other labeling requirements for generative AI content.[74]

Watermarking may be a helpful first step, but cannot be a full solution to the problem of identifying AI content since it does not address audio or text based synthetic content and can likely be copied, manipulated, removed, or otherwise defeated.[75] Researchers at the University of Maryland have

already found ways to break watermarking methods and insert false watermarks onto images.[76] Few other options for identifying generated content have been put forth, so watermarking remains the most commonly proposed labelling method. Other methods of checking for synthetic material (such as counting fingers or checking image consistency) are already losing utility.[77] Some companies advertise services where AI algorithms detect synthetic content,[78] but this technology is still in its infancy and is easy to fool. Even OpenAI threw in the towel and decommissioned its own synthetic content detection software.[79] Without methods to reliably identify synthetic content, the full scale of the issue remains unknown and the true extent of the harms stemming from generative AI remains unaddressed.

AN OUROBOROS OF DATA DEGRADATION

LLMs and other generative AI algorithms continuously evolve by incorporating new data into the model's training set and "learning." To satisfy the enormous quantity of data required to build and maintain training datasets, AI companies use web crawlers to scrape images, text, and all other types of data from the internet.[80] This data is incorporated into the training dataset and the algorithm "learns" common patterns, data structure, type, classification, and other categories from it. The patterns and information "learned" from the constantly-updated training data in turn show up in the generated outputs. While the data collection method itself comes with several issues, this section of the report will focus on model collapse.

The increasing volume of AI generated content means more and more of the data collected by web scrapers will be from AI feeding AI. Humans may be able to distinguish where content is clearly AI generated, obviously incorrect, or incomprehensible. However, web scrapers have no such discernment. Accuracy and quality issues in training datasets could be

addressed by using smaller amounts of carefully curated and fact-checked data or employing human review before data is added to a training dataset; however, these responsible practices take resources that generative AI developers thus far seem unwilling to expend.

Where scraped data is automatically put into training datasets with no quality checks or human oversight, the models deteriorate rapidly until all new outputs mirror the incomprehensible, inaccurate, and hallucinatory inputs.[81] This deterioration creates a never-ending spiral of ever worsening inputs and outputs until the AI model is entirely useless, and the ouroboros of data degradation is complete.

REALITY DISTORTION

The more common AI generated content becomes, the harder it is to discern what is authentic and what is synthetic content. With images, synthetic content is often identified once viewers count fingers, look at ears and other complex body parts, and use the relative scale of objects to discern authenticity. Models like DALL-E initially struggled to generate these intricate, complex features, giving synthetic images a number of "tells." However, these synthetic content giveaways are quickly disappearing. Generative AI systems are consuming exponential amounts of content, including more and more synthetic content as the volume generated grows. This is leading to a situation where generative AI systems either quickly spiral into model collapse or begin to produce much higher quality outputs that are increasingly difficult to distinguish from real images.

Furthermore, the increase in volume of generated content has not prompted a similar increase in the public's ability or inclination to verify every piece of content they see. The lack of information verification means false information spreads quickly and easily, while true information can be dismissed as AI generated.[82] This lack of certainty creates a warped reality

where no one is fully confident in fundamental facts and news, experts, and even our own eyes and ears can no longer be trusted.

The low quality of synthetic content ruins institutions that rely on truthful, authentic information like legal investigations, academic research, and journalism. If a bad actor sends the Federal Bureau of Investigation a deepfake of a person saying they are going to bomb a bank, the FBI will waste valuable resources evaluating the veracity of the bomb threat and could begin to erroneously investigate a person who was impersonated by the bad actor. This ability to impersonate other people and recreate evidence in painstaking detail creates reasonable doubt in a jury's mind, whether the evidence itself is authentic or synthetic. Humans also over rely on visual evidence, leaving juries more likely to believe visual synthetic content such as deepfakes. Even for non-legal contexts, like journalism, synthetic content impedes the spread of truth and creates doubt and confusion for the investigative process. Finally, many types of academic research rely heavily on data pulled from the internet for sources, analysis, and more. If there are hallucinations and an increased volume of inaccurate information, the data found by researchers is rendered useless because it introduces noise and skews results.

HARMS

- **Economic/Loss of Opportunity**: Journalists, musicians, actors, artists, and other humans engaged in intellectual property creation are losing employment opportunities to generative AI. The ever-lowering quality of outputs also causes economic harms to the companies using these systems as they must correct the unusable outputs.
- **Psychological**: The volume of AI generated content and its widely varied quality makes individuals less able to distinguish between authentic content and GAI content. This causes frustration and

helplessness when trying to ascertain the veracity of online content such as videos of political candidates or divisive news stories.

- **Reputation**: The continued degradation of both training datasets and generative AI outputs leads to more and more inaccuracy and hallucinations that incorporate real individuals' data. This wrongly links people to damaging acts and behaviors in AI generated content and adequately countering this false information is nearly impossible.
- **Environmental**: The volume of electricity and water for cooling required to run generative AI systems will only increase as they are made to process more and more data, in training and in outputs.
- **Autonomy/Loss of Consumer Choice**: Because of the network effects of the internet and the sheer scale of generative AI garbage created, individuals cannot reasonably avoid generative AI content nor systems. End users will eventually be left with virtually no alternatives untouched by generative AI if this continues.
- **Autonomy/Loss of Service Quality**: The degrading quality of generative AI content means that the overall quality of online platforms and services is falling as well.
- **Autonomy/Behavior Manipulation**: Algorithmic feeds, buzzword-laden headlines generated by AI, and other generative AI integrations actively effect what information individuals can find, allowing the companies employing generative AI to manipulate human behavior[83] by distorting the lens through which they take in information.
- **Statutory Harms/Constitutional**: The inability to discern between authentic content and synthetic content leads to the degradation of the criminal justice system by calling into question the veracity of every piece of evidence. Evidence presented at trial or used as "probable cause" for investigations will be continuously called into

question and authenticating that evidence will take significant time and resources.

- **Societal**: The inability to differentiate real from generated information leads to expanding mistrust of any evidence, calling into question historic events, current events, and our own perception.

EXAMPLES

- Google search results have become so clogged with AI spam that they have had to modify the ranking systems specifically to downrank generative AI content.[84]
- AI generated "obituary spam" has become a chronic problem, flooding the internet with strange obituaries filled with search keywords and often entirely incorrect. In some cases, the obituaries were generated for still-living people, causing panic with misinformation.[85]
- Studies on model collapse caused by AI generated content in training datasets reveal "irreversible defects in the resulting models" across all forms of generative models.[86]
- Low-quality scam books generated by AI and capitalizing on newly released human-penned books are flooding Amazon and confusing consumers into incorrect purchases.[87]
- AI-generated news and information sites are flooding the internet with inaccurate information with no human oversight.[88]
- Many newsrooms have shifted to using AI-generated content that is often full of errors, partly or wholly plagiarized, and badly written.[89]
- In several instances, generative AI has hallucinated entire court cases that people then cited in court briefs without reviewing whether those cases existed.[90]

INTERVENTIONS

- Enforce existing consumer protection and product safety laws on generative AI systems and their outputs. This includes, but is not limited to, product liability standards, defamation laws, intellectual property rights, and criminal enforcement for non-consensual deepfakes.
- Institute synthetic content identification requirements mandating that generative AI models label content as fully synthetic or otherwise manipulated to end users. Several regulations at the state, federal, and international level have included watermarking requirements to address the identification problem. However, watermarking is an incomplete and flawed solution that is not viable for audio and text-based content. Furthermore, researchers have already found ways to remove, copy, or otherwise defeat watermarking software like Google's SynthID. Regulators must engage in a multifactor approach to identify synthetic content.[91]
- Require generative AI companies to regularly review and curate the data being added to their training datasets. One way to enforce this would be to remind generative AI companies that they have strict liability for any illegal content that could be scraped and added to datasets, like CSAM. The prospect of prison time for irresponsible data scraping practices may incentivize better practices.

AI Content Licensing

BACKGROUND AND RISKS

Since our first generative AI report, *Generating Harms*, the debate over AI copyright and the online intellectual property landscape has reached a fever pitch. While some lawsuits had already emerged in 2022—mainly focused on breaches of contract[92]—2023 and 2024 saw an explosion of generative AI lawsuits filed by artists and content creators alleging copyright and privacy violations.[93] Most targeted the largest AI developer in the market: OpenAI.

While the core practice of training generative AI models like GPT-4, Midjourney, and Sora on unlicensed, copyrighted works remains unchanged, the prominence of AI-generated content skyrocketed. Plans to displace actors, writers, and other entertainment workers with AI sparked major labor strikes from the Writers Guild of America ("WGA") and the Screen Actors Guild-American Federation of Television and Radio Artists ("SAG-AFTRA").[94] Even when AI-powered chatbots spout misinformation, companies and the public sector have widely adopted these chatbots.[95] And across the web, AI-generated content is rapidly becoming the most common form of content we see.[96]

Amidst this changing AI landscape, we are returning to a topic we recommended in last year's report: content licensing. As we described in *Generating Harms*, AI developers rarely, if ever, have permission to use copyrighted content to train their AI models. Today, unlicensed web scraping remains common.[97] Given these trends, one would think formal content licensing agreements would be strong contractual

protections for content creators. However, recent examples of AI content licensing agreements between AI developers and companies like Shutterstock,[98] the Associated Press,[99] and Reddit[100] suggest that AI content licensing may raise new concerns that undermine the benefits they promise.

There are three main issues emerging from today's AI content licensing practices: (1) restricting competition, (2) avoiding judicial review, and (3) exploiting the division between content creation and content ownership.

RESTRICTING COMPETITION

Most major AI developers have already scraped the web for publicly accessible content, but the value of web scraping has gone down tremendously since generative AI tools were first released due to a curious feature of generative AI development: generative AI models appear to collapse when trained on AI-generated content, generating less and less accurate content (as discussed in our "data degradation" section above).[101] To increase AI model accuracy today, AI developers are competing for the last vestiges of purely human-developed content online.

While licensing agreements have been proposed as a way to make sure content creators have adequate control over this new AI training paradigm, *exclusive* content licensing arrangements can be used as a tool to further entrench major AI developers. For example, if a major content provider signed an exclusive licensing deal with an AI developer like OpenAI, that provider's content—news articles, images, videos, etc.—would be unavailable to other competitors unless they risk unlicensed web scraping themselves. Even if competitors did try to scrape or otherwise access the licensed content, major AI developers with exclusive licenses would be incentivized to enforce their exclusive licenses against competitors. The

result: smaller AI developers will not be able to compete against larger AI developers, creating an anticompetitive environment that may increase consumer costs and decrease the quality of generative AI products and services.

AVOIDING JUDICIAL REVIEW

Thus far, major AI developers have raked in billions of investment dollars by training their AI models on content scraped indiscriminately from the public internet.[102] Until recently, the practice of training AI systems via web scraping has remained a legal gray area. It is still being actively debated whether AI developers are protected by fair use, an exception to copyright protections, or liable for infringing copyright law.

As of this writing, there are over a dozen copyright lawsuits against AI developers like OpenAI, each arguing, among other things, that the core business model behind generative AI development is illegal under copyright law.[103] A single court ruling that AI developers violated the law could undermine core assumptions about the future of generative AI; without easy access to large amounts of data, current generative AI training practices are unsustainable.

Given the current litigation landscape around generative AI, content licensing agreements implicate the same risk as overreliance on judicial settlements: AI developers may pursue licensing as a way to foreclose active or potential litigation before a court ruling comes down. While content licensing agreements are a net positive outcome for the involved copyright owners, avoiding court rulings means that web scraping and other unlicensed uses of copyrighted material by AI systems will remain broadly unsettled law.

EXPLOITING THE DIVISION OF CONTENT CREATION AND OWNERSHIP EXPLOITATION

Lastly, the recent Reddit-Google AI licensing agreement raises the specter of one final risk: those profiting from AI content licensing may not be the content creators themselves.[104] Across the internet, millions of people post text, images, audio, and video on online platforms like Reddit, Instagram, TikTok, and the platform formerly known as Twitter every day.[105] But most platforms include broad licensing arrangements in their Terms of Service where any user content posted on a platform is available for licensing without the content creator's express consent.[106] Because of these broad licensing arrangements, Reddit and other online platforms have full rights to license what users post to AI developers without users having a say. Even worse, artists may have their work posted on sites like Reddit by *other* users. A content creator who actively avoids online platforms or rejects the licensing arrangements in these platforms' Terms of Service may still be impacted.

HARMS

- **Economic Loss/Competition**: Exclusive and otherwise anticompetitive AI content licensing arrangements will lead to anticompetitive effects in the AI market, reducing the quality of products and services while increasing costs.
- **Economic Loss**: Artists who have provided their content to an aggregator like Shutterstock—or had their content posted on an online platform like Reddit—without any conditions on that content being licensed to AI developers, will not receive any money from lucrative AI content licensing deals the platforms are engaging in.

- **Economic Loss/Demand**: Artists who manage to avoid their content being licensed by others may see decreased demand for their work, as AI models continue to reproduce art in their style at low cost.
- **Reputational:** As generative AI models become more sophisticated, consumers will find it more difficult to differentiate between an artist's real work and AI-generated content designed to look similar to their distinctive styles. On the flip side, legitimate artists may be wrongfully accused of using generative AI models, such that demand for their work decreases.
- **Psychological:** Many artists have expressed pain, sadness, and anger about their creative works being used to train AI models without their consent or knowledge.
- **Autonomy/Lack of Control**: Under many online platforms' Terms of Service, an artists' work may still be licensed to AI developers when posted by users other than the artist, even if the artist otherwise avoids AI content licensing.
- **Behavior**: Facing sophisticated generative AI tools and few avenues for lucrative content licensing arrangements, fewer artists and other content creators will continue to invest the time and resources into creating and publicly sharing non-AI content.[107] To a lesser extent, the threat of AI being trained on social media posts may chill users' behavior online as well.
- **Autonomy**: For artists and other content creators incorporated—consensually or not—into AI content licensing agreements, the terms of such agreements will restrict what legal recourse artists can pursue against AI developers themselves and what options content creators have to protect their own work from new AI developers and others interested in mimicking their work.

EXAMPLES

- Artists have found their works or artistic style explicitly recreated in major generative AI models.[108]
- Reddit and Google have signed a $60 million licensing deal for Reddit user content. Because Reddit's terms of service require users to give Reddit broad licensing rights, users will have no control or compensation from the deal.[109]
- OpenAI has been negotiating content licensing agreements with news organizations even as it faces a series of copyright lawsuits from the industry—sometimes even with the very plaintiffs suing them.[110]

INTERVENTIONS

- Prohibit AI developers from entering into exclusive AI content licensing agreements, such that competing developers would be unable to compete.
- Prohibit online content platforms like Reddit, Instagram, and X/Twitter from licensing user-generated content without explicit, affirmative user consent and/or sharing profits with users.[111]
- Require AI developers to acquire the explicit, affirmative consent of artists before permitting their AI models to mimic and/or wholly recreate an artist's style.
- Content creators can use technological tools like Glaze[112] and Nightshade[113] to disrupt AI developers' ability to train models on the content—tools that may be removed when a content creator explicitly agrees to license their content as AI training data.
- Corporate owners of content licensing rights can implement an opt-in process—or less optimally, an opt-out process—to give creators

control over whether their content is included in AI content licensing agreements.

- While an imperfect solution, require AI developers to add labels or watermarks to AI-generated content to reduce the likelihood that laypeople misconstrue AI-generated content as real or vice versa.[114]

Remedies and Enforcement

Since the *Generating Harms* report, we have seen countless harms come to fruition, including scams involving fraudulent firms sending fake Digital Millennium Copyright Act ("DMCA") notices,[115] the viral spread of AI-generated war propaganda,[116] embedded racial bias in AI resume scanners,[117] AI-generated child pornography,[118] using AI to identify military targets,[119] and more.[120] In the wake of those harms and increasing calls to counter them before generative AI becomes so ingrained in society that it cannot be slowed, a slew of proposals for how to combat generative AI harms have emerged. We want to draw attention to these efforts and what more can be done. The proposals and actions we have seen thus far typically fall into one of three categories: enforcement, regulations and policies, and technical and private-sector remedies.

ENFORCEMENT

Generative AI is a multi-faceted technology with several use and risk areas, making it challenging to regulate absent a high level, AI specific regulation. However, the novelty of the technology does not mean it is exempt from existing consumer protection, civil rights, and product safety laws. Many enforcement bodies have been using their authority to address generative AI harms that fall within their jurisdiction. In the absence of a federal regulation directly addressing generative AI harms, we anticipate continued efforts to enforce existing regulations applicable to these harms.

The Federal Trade Commission ("FTC") has used its authority to launch an inquiry into five companies providing generative AI services, looking to

ensure there are not competition and antitrust violations taking place.[121] The FTC has also finalized a rule barring AI impersonations of governments or businesses,[122] proposed expanding the impersonation fraud rule to address individuals and tech companies engaged in AI deepfakes and voice cloning,[123] and launched an inquiry into Reddit's deals licensing data to AI companies for model training.[124]

The Federal Communications Commission ("FCC") issued a Declaratory Ruling confirming that calls made with AI-generated deepfake voices are illegal under the Telephone Consumer Protection Act.[125] The U.S. Copyright Office repeatedly denied registration of AI-generated artwork, explaining in two[126] decisions[127] and in a subsequent statement of policy[128] that only content created by "creative contribution from a human actor" could be granted copyright registration.

Finally, some private companies and individuals have taken enforcement into their own hands. Several media outlets, including The New York Times,[129] The Intercept, and more, have sued generative AI companies for copyright violations.[130] Getty Images,[131] Universal Music Group,[132] various authors,[133] and several artists[134] are doing the same. These cases are not yet decided, but the snowball effect of lawsuits indicates both immense dissatisfaction with the technology and the lack of clarity around how existing laws apply.

REGULATION AND POLICIES

FEDERAL

While several federal AI laws have been proposed (20 in the most recent legislative session alone),[135] the U.S. does not currently have a federal law on AI. Because the U.S. also lacks a comprehensive federal privacy law, even the most basic protections regarding AI system use of personal data are absent. Though the FTC,[136] FCC,[137] Consumer Financial Protection

Bureau,[138] and other agencies have made attempts to fill gaps by applying various consumer protection rules to AI, these efforts are no substitute for a law designed specifically to address AI's many harms.

It is always a challenge to regulate technology where the full scope of use and harms is yet unclear. However, building a framework of consistent and comprehensive protections and rights would not only address the issues we see in AI, but put in place guardrails for future technologies. For example, the EU was able to pass its AI Act fairly quickly by building on established rights, principles, and frameworks within its General Data Protection Regulation.[139] The U.S. has actively proposed many privacy bills—55 in the 118th Congress alone—including broad privacy bills and bills specific to health, financial, children's, biometric, and other privacy matters.[140] However, until privacy or AI laws are actually passed, the protections in this area remain a patchwork.

While not a regulation, the Executive Order "Safe, Secure, and Trustworthy Development and use of Artificial Intelligence" contains a number of measures to regulate federal use of AI technology, directs the Attorney General to mitigate algorithmic discrimination and other civil rights violations tied to AI, provide guidance on AI use, require AI developers to share information on training and safety tests with the government, and promote privacy rights.[141] The Executive Order also includes measures on hiring AI experts and training existing employees on AI issues, developing guidance on technical AI labeling and content authentication, setting testing and transparency standards, and more.

The Executive Order establishes many practical and responsible requirements around AI that should serve as a baseline for the AI industry and any users of AI systems: use policies, regular testing and audits, transparency around process and training, and ensuring compliance with existing regulations on consumer protection, federal tool standards, and civil

rights. Though it is not as stable as a regulation—after all, Executive Orders may be overturned by later administrations. Hopefully it lays a foundation of knowledge and practice that can be built upon in later policy.

STATE

Several states have put forth regulations that would either directly address or tangentially extend to AI systems (generally through comprehensive consumer privacy laws). Of the six AI-related laws that went into effect in 2023, five were comprehensive privacy laws and one was an AI-specific regulation addressing use of AI in hiring.[142] Comprehensive privacy laws often impact how AI systems function, affecting how AI systems collect, share, and use personal data, providing individual rights, like opting out of automated processing decisions or profiling, or mandating regular audits and assessments. AI companies often behave as if the technology is somehow beyond or exempt from existing consumer protection and other laws. This attitude has permeated AI company actions to the extent that Attorneys General have had to explicitly state that AI systems must comply with existing laws.[143]

States have already enacted several laws that directly address AI harms. While many of the broad state privacy laws also affect AI systems, some states have pushed for much more AI-specific regulation. Utah's Artificial Intelligence Policy Act mandates clear and conspicuous disclosures where a person is interacting with an AI system rather than a human and creates the Office of Artificial Intelligence Policy to implement further rules on AI.[144] New York, Maryland, and Illinois have laws regarding use of AI in employment interviews.[145] California has a similar law mandating disclosure when a person is interacting with a bot to incentivize a sale or transaction or influence a vote.[146] Other enacted state laws address use of algorithms and predictive models for insurance,[147] the use of deepfakes in elections,[148] and

use of assessment mechanisms related to prescriptions for vision correction.[149]

The many proposed state laws (54 that are currently active as of this drafting) address AI risks including disclosures on AI training datasets,[150] AI in elections,[151] an AI Bill of Rights,[152] AI use in publishing,[153] and many, many more.[154] The state law process tends to be more agile than federal regulation-making and so may be able to address specific AI harms more quickly. However, we do not yet know how many of the proposed state laws will pass and, even where they are passed, they will only protect residents of the applicable state, not individuals across America.

INTERNATIONAL

Similar to what we have seen in the U.S., the international approach to AI is varied. Some countries have doubled down on existing regulations that touch on data use and consumer safety, some have adopted a "wait and see" approach, some have issued vague policy statements, and some have moved forward with regulation.[155] Brazil, Canada, and China all have draft laws in ready-to-pass state, building on years of work and existing privacy and data regulations within the countries (noting that China also has already-enacted regulations on recommendation algorithms,[156] deepfakes,[157] and generative AI in particular[158]).[159] The EU's Artificial Intelligence Act, passed in March 2024, looks at AI systems through a product safety lens, categorizing systems into prohibited, high-risk, and low or no-risk systems with varying requirements according to the risk level.[160] The Act is broad, covering all forms of AI and adding in some measures related to generative AI late in the debate process.[161] It remains to be seen whether others will adopt a risk-based approach in proposed regulations or attempt different structures.

TECHNICAL AND PRIVATE-SECTOR REMEDIES

Some proposals to address generative AI harms try to approach the problem at the source with technical remedies linked directly to the AI-generated content or the systems. For example, several regulations[162] and discussions have addressed watermarking and labeling images created by AI, including pledges by companies to implement these measures.[163] This technology is currently an imperfect fix—not only is it largely inapplicable to text and audio creations,[164] but the marks may be spoofed or removed by bad actors once the public is aware of them.[165]

Another step to ensure basic product safety and fitness, along with establishing more trust and transparency in generative AI systems, is implementing regular, independent audits and assessments of those systems.[166] This would require clarity and transparency from the companies and independent auditors and assessors with the technical expertise to analyze multiple areas of the systems (the algorithms, the training data, the outputs, layers of decision criteria, and legal and regulatory exposures).[167]

Finally, the generative AI industry could set its own industry standards for quality and consumer protection until specific regulations are in place. Some generative AI companies have already signed on to voluntary commitments on managing identified risks.[168] Similarly, individual companies have implemented policies and restrictions on their technology to head off some common problems (for example, Google paused its image Gemini AI system's ability to generate images of humans after reports of issues).[169] However, both voluntary commitments and individual company practices and policies can be easily changed,[170] so this does not adequately address generative AI harms long-term.

CONCLUSION

The continuous, expanding damages of generative AI are still being explored and discovered, but are in no way slowing down. While industry is still focused on the technology's theoretical potential, people are facing real-world harms to their privacy, civil rights, livelihood, and sanity. It is never too late to take action against harmful technology, but it becomes harder the longer we wait and the more that technology imbeds itself throughout society. By highlighting the impacts and risks of the technology, we hope to shift the conversation from AI's theoretical potential to its actual harms and how we can protect individuals.

References

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

[2] *Poll Shows Overwhelming Concern About Risks From AI as new Institute Launches to Understand Public Opinion and Advocate for Responsible AI Policies,* A.I. Pol'y. Inst. (Aug. 2023), https://theaipi.org/poll-shows-overwhelming-concern-about-risks-from-ai-as-new-institute-launches-to-understand-public-opinion-and-advocate-for-responsible-ai-policies/ (Revealing that 72% of voters want to slow down AI development, 62% are primarily concerned about AI, and 82% don't trust tech executives to regulate AI); Alec Tyson & Emma Kikuchi, *Growing public concern about the role of artificial intelligence in daily life,* Pew Rsch. Ctr. (Aug. 28, 2023), https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/ (52% of respondents are more concerned than excited by AI and only 10% are more excited than concerned).

[3] Steven Overly, *What really worries people about AI,* Politico (Feb. 29, 2024), https://www.politico.com/newsletters/digital-future-daily/2024/02/29/what-really-worries-people-about-ai-00144224 (79% of experts and 66% of the public worry that AI will have a negative impact on privacy).

[4] Mekela Panditharatne & Noah Giansiracusa, *How AI Puts Elections at Risk—And the Needed Safeguards,* The Brennan Ctr. for Just. (July 21, 2023), https://www.brennancenter.org/our-work/analysis-opinion/how-ai-puts-elections-risk-and-needed-safeguards.

[5] The World Economic Forum notes misinformation and disinformation as the most severe global risk in the next two years, "Global Risks Report 2024: Insight Report," World Econ. at F. 8 (Jan. 2024), https://www.weforum.org/publications/global-risks-report-2024/.

[6] Jeff Allen, *Misinformation Amplification Analysis and Tracking Dashboard,* Integrity Inst. (Oct. 13, 2022), https://integrityinstitute.org/blog/misinformation-amplification-tracking-dashboard.

[7] Mekela Panditharatne, *'News Deserts' Could Impact Midterm Elections,* Brennan Ctr. For Just. (Oct. 31, 2022), https://www.brennancenter.org/our-work/analysis-opinion/news-deserts-could-impact-midterm-elections.

[8] Morgan Meaker, *Slovakia's Election Deepfakes Show AI Is a Danger to Democracy,* Wired (Oct. 3, 2023), https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/ (note that Slovakia has a legal bar on media or politicians discussing politics for 48 hours prior to polls opening, making addressing the audio even more difficult).

[9] Pranshu Verma & Gerrit De Vynck, *AI is destabilizing 'the concept of truth itself' in 2024 election,* Wash. Post (Jan. 22, 2024), https://www.washingtonpost.com/technology/2024/01/22/ai-deepfake-elections-politicians/; Nilesh Christopher, *An Indian politician says scandalous audio clips are AI deepfakes. We had them tested,* Rest of World (Jul. 5, 2023), https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/.

[10] John Hendel, *AI-generated Biden robocall linked to Texas companies, officials say,* Politico (Feb. 6, 2024), https://www.politico.com/news/2024/02/06/robocalls-fcc-new-hampshire-texas-00139864#:~:text=The%20calls%20included%20an%20artificial,to%20participate%20in%20their%20primary.

[11] Lola Fadulu, *2 Men Fined $1.25 Million for Robocall Scheme to Suppress Black Vote,* N.Y. Times (Apr. 9, 2024), https://www.nytimes.com/2024/04/09/nyregion/robocalls-black-voters-wohl-burkman.html.

[12] Emma Fitzsimmons & Jeffery Mays, *Since When Does Eric Adams Speak Spanish, Yiddish and Mandarin?,* N.Y. Times (Oct. 20, 2023), https://www.nytimes.com/2023/10/20/nyregion/ai-robocalls-eric-adams.html.

[13] *See* EPIC et al., Comments on FTC Rule on Impersonation of Government, Businesses, and Individuals (SNPRM) (Apr. 30, 2024), https://epic.org/documents/epic-and-partner-organizations-comments-on-ftc-rule-on-impersonation-of-government-businesses-and-individuals-snprm/.

[14] "Risk in Focus: Generative A.I. and the 2024 Election Cycle," Cybersecurity and Infrastructure Sec. Agency (Jan. 18, 2024), https://www.cisa.gov/sites/default/files/2024-01/Consolidated_Risk_in_Focus_Gen_AI_ElectionsV2_508c.pdf.

[15] Spencer Overton, *Overcoming Racial Harms to Democracy from Artificial Intelligence*, Iowa L. Rev. (Forthcoming) (Mar. 14, 2024), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4754903.

[16] *Microsoft finds Russian influence operations targeting U.S. election have begun,* Reuters (April 17, 2024), https://www.reuters.com/world/us/microsoft-finds-russian-influence-operations-targeting-us-election-have-slowly-2024-04-17.

[17] Dustin Volz, *China is Targeting U.S. Voters and Taiwan With AI-Powered Disinformation,* Wall St. J. (April 5, 2024), https://www.wsj.com/politics/national-security/china-is-targeting-u-s-voters-and-taiwan-with-ai-powered-disinformation-34f59e21.

[18] Maggie Astor, *Behind the A.I. Robocall That Impersonated Biden: A Democratic Consultant and a Magician,* N.Y. Times (Feb. 27, 2024), https://www.nytimes.com/2024/02/27/us/politics/ai-robocall-biden-new-hampshire.html.

[19] Stuart A. Thompson, *Making Deepfakes Gets Cheaper and Easier Thanks to A.I.,* N.Y. Times (Mar. 12, 2023), https://www.nytimes.com/2023/03/12/technology/deepfakes-cheapfakes-videos-ai.html.

[20] Aleks Phillips, *Deepfake Video Shows Elizabeth Warren Saying Republicans Shouldn't Vote,* Newsweek (Feb. 27, 2023), https://www.newsweek.com/elizabeth-warren-msnbc-republicans-vote-deep-fake-video-1784117.

[21] Madalin Necsutu, *Moldova Dismisses Deepfake Video Targeting President Sandu,* Balkan Insight (Dec. 29, 2023), https://balkaninsight.com/2023/12/29/moldova-dismisses-deepfake-video-targeting-president-sandu/.

[22] Morgan Meaker, *Slovakia's Election Deepfakes Show AI Is a Danger to Democracy,* Wired (Oct. 3, 2023), https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/.

[23] *Pakistanis, Bangladeshi politicians are new targets of deepfake, 90 per cent of videos online are pornographic,* Trib. India (Dec. 14, 2023), https://www.tribuneindia.com/news/trending/from-rashmika-mandanna-to-bangladeshi-politician-filmed-in-a-bikini-90-per-cent-of-deepfake-videos-online-are-pornographic-571782.

[24] Matt Novak, *Donald Trump Falsely Claims Attack Ad Used AI to Make Him Look Bad,* Forbes (Dec. 4, 2023), https://www.forbes.com/sites/mattnovak/2023/12/04/donald-trump-falsely-claims-attack-ad-used-ai-to-make-him-look-bad/.

[25] Weber Lai, *Deepfakes pose risk for the election,* Taipei Times (Dec. 11, 2023), https://www.taipeitimes.com/News/editorials/archives/2023/12/11/2003810443.

[26] Nilesh Christopher, *An Indian politician says scandalous audio clips are AI deepfakes. We had them tested,* Rest of World (Jul. 5, 2023), https://restofworld.org/2023/indian-politician-leaked-audio-ai-deepfake/.

[27] Allen, *supra* note 6.

[28] Jim Fournier, *How algorithms are amplifying misinformation and driving a wedge between people,* The Hill (Nov. 10, 2021), https://thehill.com/changing-america/opinion/581002-how-algorithms-are-amplifying-misinformation-and-driving-a-wedge/.

[29] Deceptive Practices and Voter Intimidation Prevention Act of 2021, S.1840, 117th Cong. (2021), https://www.congress.gov/bill/117th-congress/senate-bill/1840; H.4660, 125th Sess. (S.C. 2024), https://www.scstatehouse.gov/sess125_2023-2024/bills/4660.htm.

[30] Ali Swenson, *AI-generated voices in robocalls can deceive voters. The FCC just made them illegal,* Assoc. Press (Feb. 8, 2024), https://apnews.com/article/fcc-elections-artificial-intelligence-robocalls-regulations-a8292b1371b3764916461f60660b93e6; *FCC makes AI-Generated Voices in Robocalls Illegal,* Federal Communications Commission (Feb. 8, 2024), https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal; Fed. Trade Comm'n, Trade Rule on Impersonation of Government and Businesses, Supplemental Notice of Proposed Rulemaking; Request for Public Comment, 89 Fed. Reg. 15,072 (Mar. 1, 2024), https://www.federalregister.gov/documents/2024/03/01/2024-03793/trade-regulation-rule-on-impersonation-of-government-and-businesses.

[31] Matt O'Brien, *AI image-generator Midjourney blocks images of Biden and Trump as election looms,* Assoc. Press (Mar. 13, 2024), https://apnews.com/article/midjourney-ai-imagegenerator-biden-trump-deepfakes-bc6c254ddb20e36c5e750b4570889ce1.

[32] Matt O'Brien and Ali Swenson, *Tech companies sign accord to combat AI-generated election trickery,* Assoc. Press (Feb. 16, 2024), https://apnews.com/article/ai-generated-election-deepfakes-munich-accord-meta-google-microsoft-tiktok-x-c40924ffc68c94fac74fa994c520fc06.

[33] Apostol Vassilev, et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2023 (Jan. 2014), https://doi.org/10.6028/NIST.AI.100-2e2023 at 40.

[34] Chad de Guzman and Will Henshall, *As Tech CEOs Are Grilled Over Child Safety Online, AI Is Complicating the Issue,* TIME (Feb. 2, 2024), https://time.com/6590470/csam-ai-tech-ceos/; David Thiel, *Investigation Finds AI Image Generation Models Trained on Child Abuse,* Stan. Univ. (Dec. 20, 2023), https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse.

[35] Vassilev, et al., *supra* note 33 at 3.

[36] Benj Edwards, *Artist finds private medical record photos in popular AI training data set,* ArsTechnica (Sept. 21, 2022), https://arstechnica.com/information-

technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/.

[37] Vassilev, et al., *supra* note 33 at 36.

[38] Autoriteit Persoonsgegevens, Richtlignen scraping door private organisaties en particulieren (May 1, 2024), https://www.autoriteitpersoonsgegevens.nl/uploads/2024-05/Handreiking%20scraping%20door%20particulieren%20en%20private%20organisaties.pdf.

[39] *Id.*

[40] *Id*.

[41] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, European Commission, COM(2021) 206 final, 2021/0106 at § 5e, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206 [Hereinafter "EU AI Act"].

[42] Jason Koebler, *National Archives Bans Employee Use of ChatGPT,* 404 Media (May 1, 2024), https://www.404media.co/national-archives-bans-employee-use-of-chatgpt/; Rebecca Heilweil, *More federal agencies join in temporarily blocking or banning ChatGPT,* Fedscoop (Jan. 9, 2024), https://fedscoop.com/more-federal-agencies-join-in-temporarily-blocking-or-banning-chatgpt/.

[43] Jonathan Gillham, *Company AI Policy Examples and Template—Who Has Banned ChatGPT,* Originality.ai (Apr. 18, 2024), https://originality.ai/blog/ai-policy; Andrea Park, *Two-thirds of top 20 pharmas have banned ChatGPT-and many in life sci call AI 'overrated,' survey finds,* Fierce Pharma (Apr. 19, 2024), https://www.fiercepharma.com/marketing/two-thirds-top-20-pharmas-have-banned-chatgpt-and-many-life-sci-call-ai-overrated-survey.

[44] *Id.* at 40.

[45] *Id.* at 44–45.

[46] Vassilev, et al., *supra* note 33 at 1.

[47] Tiernan Ray, *ChatGPT can leak training data, violate privacy, says Google's DeepMind,* ZDNet (Dec. 4, 2023), https://www.zdnet.com/article/chatgpt-can-leak-source-data-violate-privacy-says-googles-deepmind/.

[48] Vishwam Sankaran, *ChatGPT Cooks Up Fake Sexual Harassment Scandal And Names Real Law Professor As Accused*, Indep. (Apr. 6, 2023), https://www.independent.co.uk/tech/chatgpt-sexual-harassment-law-professor-

b2315160.html; 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/.

[49] Milad Nasr et al., "Extracting Training Data from ChatGPT," arXiv (Nov. 28, 2023), https://arxiv.org/abs/2311.17035; Milad Nasr et al., Extracting Training Data from ChatGPT, Not Just Memorization https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html?ref=404media.co) (summary of the academic research paper); Beatrice Nolan, *Google Researchers Say They Got OpenAI's ChatGPT To Reveal Some Of Its Training Data With Just One Word*, Bus. Insider (Dec. 4, 2023), https://www.businessinsider.com/google-researchers-openai-chatgpt-to-reveal-its-training-data-study-2023-12.

[50] Cloud and Threat Report 2024, Netskope (Aug. 2023), https://www.netskope.com/netskope-threat-labs/cloud-threat-report/cloud-and-threat-report-2024.

[51] Paolo Passeri, *The Risk of Accidental Data Exposure by Generative AI is Growing,* Infosecurity Mag. (Aug. 16, 2023), https://www.infosecurity-magazine.com/blogs/accidental-data-exposure-gen-ai/.

[52] James Coker, *ChatGPT Vulnerability May Have Exposed Users' Payment Information*, InfoSecurity Mag. (Mar. 29, 2023), https://www.infosecurity-magazine.com/news/chatgpt-vulnerability-payment/.

[53] 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/.

[54] *See* Vassilev, et al., *supra* note 33 at 44, 48-49.

[55] Cory Doctorow, *The 'Enshittification' of TikTok,* Wired (Jan. 23, 2023), https://www.wired.com/story/tiktok-platforms-cory-doctorow/.

[56] Steven J. Vaughan-Nichols, *ChatGPT, how did you get here? It was a long journey through open source AI*, The Reg. (Mar. 24, 2023), https://www.theregister.com/2023/03/24/column/.

[57] Generative AI hallucinations occur when the system fully invents information to fit a prompt, such as in the multiple cases where lawyers have submitted documents to the court that contain made-up cases that fit the fact pattern they were searching for. *See, e.g.,* Dan Mangan, *Judge sanctions lawyers for brief written by A.I with fake citations,* CNBC (Jun. 22, 2023), https://www.cnbc.com/2023/06/22/judge-sanctions-lawyers-whose-ai-written-filing-contained-fake-citations.html; Leyland Cecco, *Canada lawyer*

under fire for submitting fake cases created by AI chatbot, The Guardian (Feb. 29, 2024), https://www.theguardian.com/world/2024/feb/29/canada-lawyer-chatgpt-fake-cases-ai.

[58] Ina Fried and Scott Rosenberg, *AI could choke on its own exhaust as it fills the web,* Axios (Aug. 28, 2023), https://www.axios.com/2023/08/28/ai-content-flood-model-collapse.

[59] While this report focuses mainly on the internet, the harms from generative AI span the entire digital ecosystem. Robocalls, advertising seen in real life such as billboards, TV shows and movies, mobile apps, and several other venues are equally bombarded with synthetic content.

[60] Krystal Hu, *ChatGPT sets record for fastest-growing user base-analysis*, Reuters (Feb. 2, 2023), https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/; Jon Porter, *ChatGPT continues to be one of the fastest-growing services ever,* The Verge (Nov. 6, 2023), https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference.

[61] Porter, *supra* note 60.

[62] *Id.*

[63] Sujan Sarkar, *AI Industry Analysis: 50 Most Visited AI Tools and Their 24B+ Traffic Behavior,* Writerbuddy (Sept. 2023), https://writerbuddy.ai/blog/ai-industry-analysis.

[64] John Herrman, *The Junkification of Amazon*, N.Y. Mag. (Jan. 30, 2023), https://nymag.com/intelligencer/2023/01/why-does-it-feel-like-amazon-is-making-itself-worse.html.

[65] Kaitlyn Tiffany, *AI-Generated Junk Is Flooding Etsy*, The Atlantic (Jun. 15, 2023), https://www.theatlantic.com/technology/archive/2023/06/ai-chatgpt-side-hustle/674415/.

[66] James Vincent, *AI is being used to generate whole spam sites*, The Verge (May 2, 2023), https://www.theverge.com/2023/5/2/23707788/ai-spam-content-farm-misinformation-reports-newsguard.

[67] Joseph Cox, *Google news Is Boosting Garbage AI-Generated Articles,* 404 Media (Jan. 18, 2024), https://www.404media.co/google-news-is-boosting-garbage-ai-generated-articles/; Jason Koebler, et al., *We Need Your Email Address,* 404 Media (Jan. 26, 2024) https://www.404media.co/why-404-media-needs-your-email-address/.

[68] Renee DiResta & Josh A. Goldstein, "How Spammers and Scammers Leverage AI-Generated Images on Facebook for Audience Growth," arXiv (Mar. 19, 2024), https://arxiv.org/abs/2403.12838.

[69] *Id.* at 7.

[70] Public Citizen, *Tracker: State Legislation on Deepfakes in Elections* (Nov. 20, 2023), https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/; *see e.g.* N.Y. Elec. § 14-106(5)(b) (2024) (requiring disclosure of the use of generative AI in relation to elections. For visual synthetic content, the following disclosure needs to be clearly and easily legible: "This (image, video, or audio) has been manipulated." For synthetic audio such as a phone call or radio, the disclosure must be made before playing the audio.)

[71] *Synthetic Content*, NIST (Apr. 29, 2024), https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence-2.

[72] Public Citizen, *supra* note 70.

[73] EU AI Act at Art. 50; Titus Wu, *California Lawmakers Push for Watermarks on AI-Made Photo, Video,* Bloomberg Law (Jan. 26, 2024), https://news.bloomberglaw.com/artificial-intelligence/california-lawmakers-push-for-watermarks-on-ai-made-photo-video; Susan Haigh, *Connecticut Senate passes wide-ranging bill to regulate AI. But its fate remains uncertain,* Assoc. Press (Apr. 24, 2024), https://apnews.com/article/artificial-intelligence-ai-connecticut-regulation-b004b4477ac20cc365317edff9f7351b.

[74] Information Commissioner's Office, *Generative AI third call for evidence: accuracy of training data and model outputs* (Apr. 12, 2024), https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-third-call-for-evidence/.

[75] EPIC, Comments on the NTIA Request for Comment: Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights (Mar. 27, 2024), https://epic.org/wp-content/uploads/2024/03/EPIC_Comment_NTIA_Dual_Use_Foundation_Models_with_Appendix.pdf.

[76] 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.

[77] Gerrit De Vynck, *The AI deepfake apocalypse is here. These are the ideas for fighting it.,* Wash. Post (Apr. 5, 2024), https://www.washingtonpost.com/technology/2024/04/05/ai-deepfakes-detection/.

[78] Jonathan Gillham, *ContentAtScale AI Content Detection Review*, Originality.ai (Dec. 22, 2023), https://originality.ai/blog/contentatscale-ai-content-detection-review.

[79] Francisco Pires, *OpenAI Sunsets Generative AI Text Detection Tool*, tom'sHARDWARE (Jul. 26, 2023), https://www.tomshardware.com/news/openai-sunsets-generative-ai-text-detection-tool.

[80] Gary Drenik, *Data Privacy and ownership To Remain Key Concerns In Web Scraping Industry Next Year,* Forbes (Dec. 18, 2023), https://www.forbes.com/sites/garydrenik/2023/12/18/data-privacy-and-ownership-to-remain-key-concerns-in-web-scraping-industry-next-year/.

[81] *See* EPIC, Comments to NIST on Information Related to NIST's Assignments under Sections 4.1, 4.5, and 11 of the Executive Order Concerning Artificial Intelligence (Feb. 2, 2024), https://epic.org/wp-content/uploads/2024/02/EPIC-Comment-on-NIST-AI-Executive-Order-Mandates-RFI-02.02.24.pdf (describing value and limitations of watermarking); 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-executiveorder; 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 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; 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.

[82] Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy Democracy, and National Security*, 107 Cal. L. Rev. 1753 (2019), https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security.

[83] Neil Richards, *Why Privacy Matters* 35–37 (2022) ("But the important lesson of [Target using data analytics to infer pregnancy status to tailor ads to individuals] is not actually about the power of human information analytics to find surprising correlations like the one between lotion and pregnancy. Instead, the real lesson is about the power those insights confer to control human behavior. The reason Target wants to know about pregnancy is because Target wants consumers to buy as much as possible of everything they sell at their big box stores—not just diapers and baby clothes, but lawn furniture and underwear, wine and electronics.")

[84] David Pierce, *Google is starting to squash more spam and AI in search results,* The Verge (Mar. 5, 2024), https://www.theverge.com/2024/3/5/24091099/google-search-high-quality-results-spam-ai-content.

[85] Mia Sato, *The unsettling scourge of obituary spam,* The Verge (Feb. 12, 2024), https://www.theverge.com/24065145/ai-obituary-spam-generative-clickbait.

[86] Shumailov et al., *supra* note 81.

[87] Andrew Limbong, *Authors push back on the growing number of AI 'scam' books on Amazon,* NPR (Mar. 13, 2024), https://www.npr.org/2024/03/13/1237888126/growing-number-ai-scam-books-amazon; Kate Knibbs, *Scammy AI-Generated Book Rewrites Are Flooding Amazon,* Wired (Jan. 10, 2024), https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/.

[88] McKenzie Sadeghi et al., *Tracking AI-enabled Misinformation: 811 'Unreliable AI-Generated News' Websites (and Counting), Plus the Top False Narratives Generated by Artificial Intelligence Tools,* NewsGuard (Apr. 29, 2024), https://www.newsguardtech.com/special-reports/ai-tracking-center/; *Proliferating 'news' sites spew AI-generated stories,* France24 (Mar. 11, 2024), https://www.france24.com/en/live-news/20240311-proliferating-news-sites-spew-ai-generated-fake-stories; Stuart Thompson, *A.I.-Generated Content Discovered on News Sites, Content Farms and Product Reviews,* N.Y. Times (May 19, 2023), https://www.nytimes.com/2023/05/19/technology/ai-generated-content-discovered-on-news-sites-content-farms-and-product-reviews.html; Matthew Cantor, *Nearly 50 news websites are 'AI-generated', a study says. Would I be able to tell?,* The Guardian (May 8, 2023), https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study.

[89] Victor Tangermann, *Gizmodo and Kotaku Staff Furious After Owner Announces Move to AI Content,* Futurism (June 30, 2023), https://futurism.com/gizmodo-kotaku-staff-furious-ai-content; Jade Drummond, *Newsrooms around the world are using AI to optimize work, despite concerns about bias and accuracy,* The Verge (Sep. 28, 2023), https://www.theverge.com/2023/9/28/23894651/ai-newsroom-journalism-study-automation-bias.

[90] Benjamin Weiser & Jonah Bromwich, *Michael Cohen Used Artificial Intelligence in Feeding Lawyer Bogus Cases,* N.Y. Times (Dec. 29, 2023), https://www.nytimes.com/2023/12/29/nyregion/michael-cohen-ai-fake-cases.html; Dan Mangan, *Judge sanctions lawyers for brief written by A.I with fake citations,* CNBC (Jun. 22, 2023), https://www.cnbc.com/2023/06/22/judge-sanctions-lawyers-whose-ai-written-filing-contained-fake-citations.html; Leyland Cecco, *Canada lawyer under fire for submitting fake cases created by AI chatbot,* The Guardian (Feb. 29, 2024), https://www.theguardian.com/world/2024/feb/29/canada-lawyer-chatgpt-fake-cases-ai.

[91] *Supra* note 81.

[92] *See, e.g.*, *Doe 1 et al. v. GitHub, Inc. et al.*, 672 F. Supp. 3d 837 (N.D. Cal. 2023); *Doe 3 et al. v. GitHub, Inc. et al.*, No. 4:22-CV-07074, (N.D. Cal. Nov. 10, 2022).

[93] *See* Jonathan Gillham, *OpenAI and ChatGPT Lawsuit List*, Originality.ai (May 1, 2024), https://originality.ai/blog/openai-chatgpt-lawsuit-list;

[94] *See* Matt Scherer, *The SAG-AFTRA Strike is Over, But the AI Fight in Hollywood is Just Beginning*, CDT (Jan. 4, 2024), https://cdt.org/insights/the-sag-aftra-strike-is-over-but-the-ai-fight-in-hollywood-is-just-beginning/.

[95] *See, e.g.*, Maria Yagoda, *Airline Held Liable for its Chatbot Giving Passenger Bad Advice—What This Means for Travelers*, BBC (Feb. 23, 2024), https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know; *Introducing New AI Experiences Across Our Family of Apps and Devices*, Meta Blog (Sept. 27, 2023), https://about.fb.com/news/2023/09/introducing-ai-powered-assistants-characters-and-creative-tools/; Colin Wood, *After giving wrong answers, NYC chatbot to stay online for testing,* State Scoop (Apr. 3, 2024), https://statescoop.com/nyc-mayor-eric-adams-chatbot-wrong-answers/; Jessica Nix, *AI-Powered World Health Chatbot Is Flubbing Some Answers,* Bloomberg (Apr. 18, 2024), https://www.bloomberg.com/news/articles/2024-04-18/who-s-new-ai-health-chatbot-sarah-gets-many-medical-questions-wrong.

[96] *See* Jules Roscoe, *A 'Shocking' Amount of the Web is Already AI-Translated Trash, Scientists Determine*, Vice (Jan. 17, 2024), https://www.vice.com/en/article/y3w4gw/a-shocking-amount-of-the-web-is-already-ai-translated-trash-scientists-determine; Fried & Rosenberg, *supra* note 58.

[97] *See, e.g.*, Bryson Masse, *OpenAI Launches Web Crawling GPTBot, Sparking Blocking Effort by Website Owners and Creators*, VentureBeat (Aug. 8, 2023), https://venturebeat.com/ai/openai-launches-web-crawling-gptbot-sparking-blocking-effort-by-website-owners-and-creators/.

[98] Emma Roth, *OpenAI's DALL-E Will Train on Shutterstock's Library for Six More Years*, The Verge (July 11, 2023), https://www.theverge.com/2023/7/11/23791528/openai-shutterstock-images-partnership.

[99] Matt O'Brien, *ChatGPT-Maker OpenAI Signs Deal with AP to License News Stories*, Assoc. Press (July 13, 2023), https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[100] Anna Tong et al., *Exclusive: Reddit in AI Content Licensing Deal with Google*, Reuters (Feb. 21, 2024), https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/.

[101] *See* 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-ofmodel-collapse-as-ai-trains-on-ai-generated-content/; Shumailov et al., *supra* note 81.

[102] *See* Hayden Field & Kif Leswing, *Generative AI 'FOMO' is Driving Tech Heavyweights to Invest Billions of Dollars in Startups*, CNBC (Mar. 30, 2024), https://www.cnbc.com/2024/03/30/fomo-drives-tech-heavyweights-to-invest-billions-in-generative-ai-.html.

[103] *See* Gillham, *supra* note 93.

[104] *See* Annelise Gilbert, *Google-Reddit AI Deal Heralds New Era in Social Media Licensing*, BL (Mar. 7, 2024), https://news.bloomberglaw.com/ip-law/google-reddit-ai-deal-just-the-start-for-social-media-licensing.

[105] *See, e.g.*, Bell Wong, *Top Social Media Statistics and Trends of 2024*, Forbes Advisor (May 18, 2023), https://www.forbes.com/advisor/business/social-media-statistics/.

[106] *See, e.g.*, *Reddit User Agreement*, Reddit (Sept. 25, 2023), https://www.redditinc.com/policies/user-agreement-september-25-2023.

[107] *See, e.g.*, Kaitlyn Nguyen, *AI is Causing Student Artists to Rethink Their Creative Career Plans*, KQED (Apr. 26, 2023), https://www.kqed.org/arts/13928253/ai-art-artificial-intelligence-student-artists-midjourney.

[108] *See* Will Knight, *Algorithms Can Now Mimic Any Artist. Some Artists Hate It*, Wired (Aug. 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 (Dec. 31, 2022), https://www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-algorithim-took-my-work.html; Nick Cave, *Issue #218,* The Red Hand Files (Jan. 2023), https://www.theredhandfiles.com/chat-gpt-what-do-you-think/; Beatrice Nolan, *Artists say AI image generators are copying their style to make thousands of new images – and it's completely out of their control,* Bus. Insider (Oct. 17, 2022), https://www.businessinsider.com/ai-image-generators-artists-copying-style-thousands-images-2022-10.

[109] Gilbert, *supra* note 104.

[110] *See* Benjamin Mullin, *Inside the News Industry's Uneasy Negotiations with OpenAI*, N.Y. Times (Dec. 29, 2023), https://www.nytimes.com/2023/12/29/business/media/media-openai-chatgpt.html.

[111] Shutterstock has already adopted a content contributor fund approach, where they will directly compensate content creators if their content was used to develop generative AI models. *See AI-Generated Content on Shutterstock: Contributor FAQ*, Shutterstock Contributor Support (Apr. 18, 2024),

https://support.submit.shutterstock.com/s/article/Shutterstock-ai-and-Computer-Vision-Contributor-FAQ?language=en_US.

[112] Shawn Shan et al., *About the Glaze Project*, Sand Lab at U. Chi., https://glaze.cs.uchicago.edu/aboutus.html (last visited May 1, 2024).

[113] Shawn Shan et al., *What is Nightshade?*, Sand Lab at U. Chi., https://nightshade.cs.uchicago.edu/whatis.html (last visited May 1, 2024).

[114] *Supra* note 81.

[115] Kevin Purdy, *Fake AI law firms are sending fake DMCA threats to generate fake SEO gains,* Ars Technica (April 4, 2024), https://arstechnica.com/gadgets/2024/04/fake-ai-law-firms-are-sending-fake-dmca-threats-to-generate-fake-seo-gains/.

[116] *See, e.g.,* Miles Klee & Nikki McCann Ramirez, *AI Has Made the Israel-Hamas misinformation Epidemic Much, Much Worse,* Rolling Stone (Oct. 27, 2023), https://www.rollingstone.com/politics/politics-features/israel-hamas-misinformation-fueled-ai-images-1234863586/; Isabelle Frances-Wright & Moustafa Ayad, *Misleading and manipulated content goes viral on X in Middle East conflict,* Inst. for Strategic Dialogue (April 14, 2024), https://www.isdglobal.org/digital_dispatches/misleading-and-manipulated-content-goes-viral-on-x-twitter-in-middle-east-conflict-iran-israel-strikes; Amanda Hoover, *Hulu Shows Jarring Anti-Hamas Ad Likely Generated With AI,* Wired (Jan. 30, 2024), https://www.wired.com/story/hulu-anti-hamas-ad-generative-ai/.

[117] Leon Yin, Davey Alba, and Leonardo Nicoletti, *OpenAI's GPT is a Recruiter's Dream Tool. Tests Show There's Racial Bias,* Bloomberg (Mar. 7, 2024), https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination.

[118] Clayton Vickers, *Law enforcement struggling to prosecute AI-generated child pornography, asks Congress to act,* The Hill (Mar. 13, 2024), https://thehill.com/homenews/house/4530044-law-enforcement-struggling-prosecute-ai-generated-child-porn-asks-congress-act/.

[119] Bethan McKernan and Harry Davies, *'The machine did it coldly': Israel used AI to identify 37,000 Hamas targets,* The Guardian (Apr. 3, 2024), https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes.

[120] Will Oremus, *Hackers competed to find AI harms. Here's what they found.,* Wash. Post (Apr. 4, 2024), https://www.washingtonpost.com/politics/2024/04/04/hackers-competed-find-ai-harms-heres-what-they-found/.

[121] *FTC Launches Inquiry into Generative AI Investments and Partnerships,* Fed. Trade Comm'n (Jan. 25, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships.

[122] Rule on Impersonation of Government and Businesses, 16 C.F.R. § 461 (2024).

[123] K.C. Halm et al., *Who's Liable for Deepfakes? FTC Proposes to Target Developers of Generative AI Tools in Addition to Fraudsters,* Davis Wright Tremaine LLP (Feb. 22, 2024), https://www.dwt.com/blogs/artificial-intelligence-law-advisor/2024/02/ftc-targets-tech-companies-for-generative-ai-fraud.

[124] Alex Weprin, *Reddit Says FTC Inquiring About Deals to License Data and Train AI Models,* The Hollywood Rep. (Mar. 15, 2024), https://www.hollywoodreporter.com/business/digital/reddit-ftc-investigation-license-data-train-ai-models-1235853771/.

[125] *FCC Makes AI-Generated Voices in Robocalls Illegal,* Fed. Commc'n. Comm'n. (Feb. 8, 2024), https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal.

[126] U.S. Copyright Office Review Board, *Decision Affirming Refusal of Registration of a Recent Entrance to Paradise* (Feb. 14, 2022), https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf.

[127] U.S. Copyright Office, *Cancellation Decision re: Zarya of the Dawn (VAu001480196)* (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf.

[128] U.S. Copyright Office, *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence,* 88 FR 16190 (Mar. 16, 2023), https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.

[129] Michael M. Grynbaum & Ryan Mac, *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work,* N.Y. Times (Dec. 27, 2023), https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

[130] Emilia David, *The Intercept, Raw Story, and AlterNet sue OpenAI and Microsoft,* The Verge (Feb. 28, 2024), https://www.theverge.com/2024/2/28/24085973/intercept-raw-story-alternet-openai-lawsuit-copyright.

[131] James Vincent, *Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement,* The Verge (Feb. 6, 2023), https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion.

[132] Emilia David, *Universal Music sues AI company Anthropic for distributing song lyrics,* The Verge (Oct. 19, 2023), https://www.theverge.com/2023/10/19/23924100/universal-music-sue-anthropic-lyrics-copyright-katy-perry.

[133] Cat Zakrzewski et al., *OpenAI prepares to fight for its life as legal troubles mount,* Wash. Post (Apr. 9, 2024), https://www.washingtonpost.com/technology/2024/04/09/openai-lawsuit-regulation-lawyers/.

[134] Shanti Escalante-De Mattei, *Artists Are Suing Artificial Intelligence Companies and the Lawsuit Could Upend Legal Precedents Around Art,* Art in America (May 5, 2023), https://www.artnews.com/art-in-america/features/midjourney-ai-art-image-generators-lawsuit-1234665579/.

[135] Artificial Intelligence Legislation Tracker, Brennan Ctr. For Just. (last updated April 1, 2024), https://www.brennancenter.org/our-work/research-reports/artificial-intelligence-legislation-tracker.

[136] Press Release, *FTC Launches Inquiry into Generative AI Investments and Partnerships,* Fed. Trade Comm'n. (Jan. 25, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships; Press Release, *FTC Proposes New Protections to Combat AI Impersonation of Individuals,* Fed. Trade Comm'n. (Feb. 15, 2024), https://www.ftc.gov/news-events/news/press-releases/2024/02/ftc-proposes-new-protections-combat-ai-impersonation-individuals.

[137] Declaratory Ruling, "Implications of Artificial Intelligence Technologies on Protecting Consumers from Unwanted Robocalls and Robotexts," Fed. Commc'n Comm'n. FCC-24-17, Doc. No. 23-362 (Feb. 8. 2024).

[138] Press Release, *CFPB Issues Guidance on Credit Denials by Lenders Using Artificial Intelligence,* Consumer Fin. Prot. Bureau (Sept. 19, 2023), https://www.consumerfinance.gov/about-us/newsroom/cfpb-issues-guidance-on-credit-denials-by-lenders-using-artificial-intelligence/.

[139] Reid Blackman & Ingrid Vasiliu-Feltes, *The EU's AI Act and How Companies Can Achieve Compliance,* Harv. Bus. Rev. (Feb. 22, 2024), https://hbr.org/2024/02/the-eus-ai-act-and-how-companies-can-achieve-compliance.

[140] US Federal Privacy Legislation Tracker, Int'l Assoc. of Priv. Pro. (last updated Mar. 2024), https://iapp.org/resources/article/us-federal-privacy-legislation-tracker/.

[141] *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,* Biden White House (Oct. 30, 2023),

https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/; Kara Williams, *Summary: What Does Biden's Executive Order on Artificial Intelligence Actually Say?,* EPIC (Nov. 7, 2023), https://epic.org/summary-what-does-bidens-executive-order-on-artificial-intelligence-actually-say/.

[142] Katrina Zhu, *The State of State AI Laws: 2023,* EPIC (Aug. 3, 2023), https://epic.org/the-state-of-state-ai-laws-2023/.

[143] Press Release, "AG Campbell issues Advisory Providing Guidance On How State Consumer Protection And Other Laws Apply To Artificial Intelligence," Off. of the Mass. Att'y Gen. (April 16, 2024), https://www.mass.gov/news/ag-campbell-issues-advisory-providing-guidance-on-how-state-consumer-protection-and-other-laws-apply-to-artificial-intelligence.

[144] SB 149, Artificial Intelligence Amendments, Utah, 2024 General Session, https://le.utah.gov/~2024/bills/sbillint/SB0149.pdf.

[145] Local Law 2021/144, Automated Employment Decision Tools, New York, Ch. 5, Title 20, Admin Code of City of New York, https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9; Artificial Intelligence Video Interview Act 820 Ill. Comp. Stat. 42/1 (2020), https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68; Use of Facial Recognition Services Prohibited -- Consent by Applicant , Md. Lab. & Empl. § 3-717 (2023).

[146] Bolstering Online Transparency Act (BOT), Cal. Bus. & Prof. Code § 17940 *et seq.* (2023)

[147] Restrict Insurers' Use of External Consumer Data, Col. Rev. Stat. 10-3-1104.9 (2021).

[148] Public Citizen, *Tracker: State Legislation on Deepfakes in Elections* (Nov. 20, 2023), https://www.citizen.org/article/tracker-legislation-on-deepfakes-in-elections/; *see e.g.* N.Y. Elec. § 14-106(5)(b) (2024) (requiring disclosure of the use of generative AI in relation to elections. For visual synthetic content, the following disclosure needs to be clearly and easily legible: "This (image, video, or audio) has been manipulated." For synthetic audio such as a phone call or radio, the disclosure must be made before playing the audio.)

[148] *Synthetic Content*, NIST (Apr. 29, 2024), https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence-2.

[149] Ga. Code Ann. § 31-12-12 (2023).

[150] California AB-2013, https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013.

[151] Florida SB 850, https://m.flsenate.gov/Bill/850/; New York Assembly Bill A7904A, https://www.nysenate.gov/legislation/bills/2023/A7904/amendment/A.

[152] New York Assembly Bill A8126, https://www.nysenate.gov/legislation/bills/2023/A8129.

[153] New York Assembly Bill A8098, https://www.nysenate.gov/legislation/bills/2023/A8098/amendment/original; New York Assembly Bill A8158, https://www.nysenate.gov/legislation/bills/2023/A8158.

[154] US State-By-State AI Legislation Snapshot, Bryan Cave Leighton Paisner (last updated Jan. 26, 2024), https://www.bclplaw.com/en-US/events-insights-news/2023-state-by-state-artificial-intelligence-legislation-snapshot.html.

[155] Mikhail Klimentov, *From China to Brazil, here's how AI is regulated around the world,* Wash. Post (Sept. 3, 2023), https://www.washingtonpost.com/world/2023/09/03/ai-regulation-law-china-israel-eu/.

[156] China "Provisions on the Management of Algorithmic Recommendations in Internet Information Services," translation available at https://www.chinalawtranslate.com/en/algorithms/.

[157] China "Provisions on the Administration of Deep Synthesis Internet Information Services," translation available at https://www.chinalawtranslate.com/en/deep-synthesis/.

[158] China "Interim Measures for the Management of Generative Artificial Intelligence Services," translation available at https://www.chinalawtranslate.com/en/generative-ai-interim/.

[159] Brazil Bill 2338/2023, covered by Rob Rodrigues, et al., *Brazilian Lawmaker Introduces Bill to Allow AI as Inventor,* Lexology (Feb. 29, 2024), https://www.lexology.com/library/detail.aspx?g=6f1183c4-7670-4208-bf64-5a5d1221ff47; Canada Bill C-27, First Session, Forty-fourth Parliament, https://www.parl.ca/DocumentViewer/en/44-1/bill/C-27/first-reading; China "Measures for the Management of Generative Artificial Intelligence Services (Draft for Comment)," translation available at https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/.

[160] EU AI Act, *supra* note 41.

[161] Maria Villegas Bravo, *What U.S. Regulators can Learn from the EU AI Act,* EPIC (Mar. 22, 2024), https://epic.org/what-u-s-regulators-can-learn-from-the-eu-ai-act/.

[162] EU AI Act Art. 50(2) ("Providers of AI systems, including GPAI systems, generating synthetic audio, image, video or text content, shall ensure the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated."); Xuezi Dan & Yan Luo, *Labeling of AI Generated Content: New Guidelines Released in China,* Lexology (Aug. 25, 2023), https://www.lexology.com/library/detail.aspx?g=656b7b0d-9b82-4fa9-9b50-34d6ceeb4ba9.

[163] Nihal Krishan, *AI watermarking could be exploited by bad actors to spread misinformation. But experts say the tech still must be adopted quickly,* FedScoop (Jan. 3, 2024), https://fedscoop.com/ai-watermarking-misinformation-election-bad-actors-congress/; Kris Holt, *Meta plans to ramp up labeling of AI-generated images across its platforms,* Engadget (Feb. 6, 2024), https://www.engadget.com/meta-plans-to-ramp-up-labeling-of-ai-generated-images-across-its-platforms-160234038.html.

[164] Bob Gleichauf & Dan Geer, *Digital Watermarks Are Not Ready for Large Language Models,* LawFare (Feb. 29, 2024), https://www.lawfaremedia.org/article/digital-watermarks-are-not-ready-for-large-language-models.

[165] Nihal Krishan, *AI watermarking could be exploited by bad actors to spread misinformation. But experts say the tech still must be adopted quickly,* FedScoop (Jan. 3, 2024), https://fedscoop.com/ai-watermarking-misinformation-election-bad-actors-congress/; Kat Tenbarge & Kevin Collier, *Big Tech says AI watermarks could curb misinformation, but they're easy to sidestep,* NBC News (Mar. 19, 2024), https://www.nbcnews.com/tech/tech-news/watermark-deepfake-solution-ai-misinformation-cant-stop-de-rcna137370.

[166] Press Release, *Hickenlooper Proposes AI Auditing Standards, Calls for Protecting Consumer Data, Increasing Transparency,* U.S. Sen. Hickenlooper for Colo. (Feb. 5, 2024), https://www.hickenlooper.senate.gov/press_releases/hickenlooper-proposes-ai-auditing-standards-calls-for-protecting-consumer-data-increasing-transparency/.

[167] Mark Dangelo, *Auditing AI: The emerging battlefield of transparency and assessment,* Thomson Reuters (Oct. 25, 2023), https://www.thomsonreuters.com/en-us/posts/technology/auditing-ai-transparency/.

[168] *Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI,* Biden White House (Jul. 21, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/.

[169] Sabrina Ortiz, *Why Google just banned Gemini from generating images of people,* ZDNet (Feb. 22, 2024), https://www.zdnet.com/article/why-google-just-banned-gemini-from-generating-images-of-people/; *see also* Gerrit De Vynck, *AI companies agree to limit election 'deepfakes' but fall short of ban,* Wash. Post (Feb. 13, 2024), https://www.washingtonpost.com/technology/2024/02/13/google-ai-elections-deepfakes-open/; Matt O'Brien, *AI image-generator Midjourney blocks images of Biden and Trump as election looms,* Assoc. Press (Mar. 13, 2024), https://apnews.com/article/midjourney-ai-imagegenerator-biden-trump-deepfakes-bc6c254ddb20e36c5e750b4570889ce1.

[170] Sam Biddle, *OpenAI quietly deletes ban on using ChatGPT for "military and warfare,"* The Intercept (Jan. 12, 2024), https://theintercept.com/2024/01/12/open-ai-military-ban-chatgpt/.