Richmond Journal of Law & Technology Volume XXX, Issue 3

INTERNET OPENNESS AT RISK: GENERATIVE AI'S IMPACT ON DATA SCRAPING

Melany Amarikwa*

Cite as: Melany Amarikwa, *Internet Openness at Risk: Generative AI's Impact on Data Scraping*, 30 RICH. J.L. & TECH. 533 (2024).

* J.D. Candidate, 2024, University of Pennsylvania Law School; B.A., 2019, University of California, Berkeley. I am grateful to Professor Cynthia L. Dahl for her invaluable guidance and mentorship. I would also like to thank the editors of the Richmond Journal of Law & Technology for their efforts in bringing this article to publication.

ABSTRACT

Modern scraping practices—the automated extraction of data from online websites—by companies employing generative AI models threatens the foundational and essential openness of the internet. There are calls for regulating the use of scraping in generative AI models, but lawmakers, concerned about its impact on US global AI leadership, have failed to act. This article presents two legal frameworks aimed at regulating generative AI scraping. The adverse possession framework addresses property rights and allows for the use of copyrighted works where the author abandons or fails to claim their works. The public records framework addresses privacy rights and treats personal information made publicly available by the subject as a public record with context-based privacy exemptions. These frameworks seek to strike a balance between private interests in development and the public's interest in safeguarding its property and privacy rights.

I. INTRODUCTION

[1] When John McCarthy and Marvin L. Minsky coined the term artificial intelligence ("AI"),[1] they could never have predicted the troubles that AI would later bring.[2] Initially defined as a machine exhibiting behavior considered intelligent in comparison to human behavior,[3] AI later came to instill fear in the public.[4] With well-documented issues of AI bias and transparency, this fear stemmed from a well-founded lack of trust.[5]

[2] As the public learns more about the development of AI models, new issues emerge. Scraping, the automatized extraction of information from websites, is one such issue.[6] Once little more than an obscure business and research data practice, scraping is now synonymous with data breaches.[7]

[1] *See* John McCarthy et al., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, 27 AI MAGAZINE 12, 13 (2006).

[2] *See id.* at 12, 14.

[3] *See id.* at 13; *see also* Martijn Kuipers & Ramjee Prasad, *Journey of Artificial Intelligence*, 123 WIRELESS PERS. COMMC'N 3275, 3276 (2021).

[4] *See* Shira Ovide, *Why are we so afraid of AI?*, WASH. POST, https://www.washingtonpost.com/technology/2023/02/21/ai-polls-skeptics [perma.cc/F5ZB-F6ZU] (last updated Feb. 24, 2023, 1:54 PM).

[5] *See id.*

[6] *See* EJ Stanley, *What is Web Scraping? A Complete Guide*, FORTRA, https://www.fortra.com/resources/guides/what-is-web-scraping [perma.cc/27FK-KLUU] (last visited Mar. 4, 2024).

[7] *Cf. Hackers, User Rights, and Government Surveillance, in* THE REFERENCE SHELF: INTERNET LAW 95 (Grey House Publishing, 2020) ("Like all things from the early days of the Internet, hacking began with the impulse to create a free and open environment focused on innovation, to hardness the Internet's profound uniqueness for the good of all. But hacking developed a dark side.").

Bots are scraping articles,[8] books,[9] and social media conversations to help generative AI models learn how to write and respond to users' inquiries.[10] All of these scraping efforts can—and often do—occur without the authors' consent.[11]

[3] The knowledge of generative AI companies' practices of scraping and then using scraped data has been met with public outcry and a call for regulation.[12] Although lawmakers acknowledge the need for regulation, they fear restrictive regulations may lead the United States to lose its lead in the AI "arms race."[13] Consequently, Washington has been slow to move.

[4] Openness has been a fundamental characteristic of the internet since its inception nearly forty years ago.[14] In the context of modern US

[8] *See* 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 [perma.cc/4N5Y-ADV3].

[9] *See* Authors Guild v. Google, Inc., 770 F. Supp. 2d, 666, 670 (S.D.N.Y. 2011).

[10] *See e.g.,* Hard Fork, *Don't Scrape Me, Bro, the Activists Sabotaging Self-Driving Cars and How Reddit Beat a Rebellion*, N.Y. TIMES, at 54:09 (Aug. 11, 2023), https://www.nytimes.com/2023/08/11/podcasts/dont-scrape-me-bro-the-activists-sabotaging-self-driving-cars-and-how-reddit-beat-a-rebellion.html [perma.cc/3598-Y8XD] ("Reddit said, hey, we're uncomfortable with these large language models scraping our site. We are going to make changes to our API.").

[11] *See* Deepa Seetharaman & Keach Hagey, *Outcry Against AI Companies Grows Over Who Controls Internet's Content*, WALL ST. J. (July 30, 2023, 5:30 AM), https://www.wsj.com/articles/outcry-against-ai-companies-grows-over-who-controls-internets-content-91d604c9 [perma.cc/6DB4-7QMH].

[12] *See* Mariam Salmanzadeh, *Web Scraping, What is it and Why Should I Care About it?*, INSIDE COMPLIANCE (Feb. 28, 2024), https://blogs.luc.edu/compliance/?p=5968 [perma.cc/S237-WAWU].

[13] *See* discussions *infra* Part III.

[14] *See* Reno v. ACLU, 521 U.S. 844, 886 (1997).

democracy, this openness plays a crucial role in civic engagement. However, the failure of the legislature to enact regulations addressing the scraping activities of generative AI models poses a significant threat to this foundational tenet, putting its continued existence in jeopardy.

[5] This Article aims to add to existing scholarship on scraping and offer frameworks that permit generative AI scraping while also balancing individuals' property and privacy interests. The Article is divided into three parts. Part II offers a technical background of scraping and popular generative AI models. First, it traces the evolution of scraping from its early uses to its current role as a crucial source of training data for generative AI models. It then offers an overview of the companies employing generative AI—namely OpenAI, Meta, Twitter, and Google. Finally, it explains each company's generative AI model and evaluates the information each company discloses regarding its training data.

[6] Part III provides an overview of public and private enforcement efforts aimed at mitigating the harms associated with training generative AI models using scraped data.

[7] In Part IV, the Article proposes two frameworks aimed at preserving the benefits of scraping while also protecting the public interest. The adverse possession framework permits the use of copyrighted works where the author abandons or fails to claim their works. The public records framework treats personal information made publicly available by the subject as a public record, with contextual privacy exemptions. Both frameworks are designed to uphold internet openness and facilitate the use of publicly available information by generative AI models, all while preserving individuals' privacy and property rights.

II. SCRAPING AND GENERATIVE AI

[8] Web scraping, the use of software tools to extract data from websites, dates back to the early days of the internet.[15] However, scraping's public image has evolved with the emergence of generative AI models. These AI models require large datasets for training, and internet scraping serves as a convenient method for obtaining data.[16] This trend is causing fundamental changes that may permanently transform the open landscape of the internet.

A. Scraping Background

[9] Data is the most valuable commodity in the modern era.[17] For decades, companies and individuals used scraping bots to extract data for various purposes, such as monitoring prices, gauging sentiment, and developing products.[18] Despite this long-standing acceptance, businesses are now taking proactive measures to stop scraping.[19] Even companies that use scraping methods are taking measures to safeguard their own data.[20]

[15] *See* GREG ELMER ET AL., *Scraping the First Person*, *in* COMPROMISED DATA: FROM SOCIAL MEDIA TO BIG DATA 114–15 (Ganaele Langlois et al. eds., Bloomsbury Publishing 2015) (citing Bernard Rieder, *Studying Facebook via Data Extraction: The Netvizz Application*, PROC. OF THE 5TH ANN. ACM WEB SCI. CONF. 346 (2015)).

[16] *See* Michael P. Goodyear, *Circumscribing the Spider: Trademark Law and the Edge of Data Scraping*, 70 KAN. L. REV. 295, 299–300 (2021).

[17] *Id.* at 295.

[18] *Id.* at 295–296

[19] *Id.* at 296.

[20] For example, Amazon deploys scraping bots to collect data from Walmart while simultaneously implementing measures to prevent bots from scraping its own data. *See id.* at 296, 302.

[10] After the creation of the internet, the first scraping bots known as "spiders, wanderers, crawlers, and worms" emerged.[21] These scraping bots collected information from all corners of the internet.[22] JumpStation, one of the first "crawler-based" search engines, launched in December 1993 to organize the increasing number of internet webpages.[23] Other early uses of scraping bots included gauging the size of the internet and identifying broken or dead links on servers.[24]

[11] There are three types of scraping methods. First, HTML, or screen, scraping involves writing a code that extracts data from a website's HTML.[25] Second, crawler scraping extracts HTML, but also crawls or searches websites via their hyperlinks.[26] For example, Googlebot is a crawler that indexes new websites.[27] Finally, API scraping involves using

[21] *See* Stacey Kimmel, *Robot-Generated Databases on the World Wide Web*, 19 DATABASE 40, 41 (1996); *see also* Martijn Koster, *Robots in the Web: threat or treat?,* NEXOR (Apr. 1995), https://webdoc.gwdg.de/ebook/aw/1999/webcrawler/mak/projects/robots/threat-or-treat.html [perma.cc/MA6D-WAW7] ("Robots have been operating in the World-Wide Web for over a year. In that time they have performed useful tasks, but also on occasion wreaked havoc on the networks. This paper investigates the advantages and disadvantages of robots, with an emphasis on robots used for resource discovery. New alternative resource discovery strategies are discussed and compared. It concludes that while current robots will be useful in the immediate future, they will become less effective and more problematic as the Web grows.").

[22] Koster*, supra* note 21.

[23] *See* Ansel Barrett, *What is Web Scraping and How Does It Work*, OCTOPARSE (Oct. 21, 2018), https://www.octoparse.com/blog/web-scraping-introduction [perma.cc/2UUL-5R5S]; *see also Web Scraping - The Comprehensive Guide for 2024*, CRAWLBASE (Mar. 23, 2023), https://crawlbase.com/blog/web-scraping-the-comprehensive-guide/ [perma.cc/MA6D-WAW7].

[24] *See* Kimmel, *supra* note 21.

[25] *See* ELMER, *supra* note 15, at 117.

[26] *Id.* at 119.

[27] *Id.*

the Application Program Interfaces ("API"), or complex phenomena that are made available by a website, to extract data.[28]

[12] Both HTML and API scraping allow for the extraction of information from webpages—though they differ in their technical approaches. HTML scraping involves the creation of custom code scripts tailored to each website.[29] This individualized approach makes HTML scraping technically demanding, inefficient, time-consuming, and prone to errors. In contrast, API scraping provides a more structured and often authorized channel for data access, with a more user-friendly format.[30]

B. The Evolution of Generative AI Models

[13] ChatGPT is far from the first chatbot but, rather, is an evolution of predecessor bots. The first chatbot, ELIZA, was created in 1964 by Joseph Weizenbaum, a German computer scientist at MIT.[31] ELIZA employed a keyword-based approach to generate responses, searching the text for relevant keywords and using set rules to produce appropriate responses.[32]

---

[28] *See id.* at 120.

[29] *See* Goodyear, *supra* note 16, at 298–99.

[30] API scraping is typically available to users by website operators. For example, Facebook's Graph API allows authorized users to "get data into and out of the Facebook platform." *Graph API: Overview*, META FOR DEVELOPERS, https://developers.facebook.com/docs/graph-api/overview [perma.cc/GH8P-DJ5W] (last visited Mar. 6, 2024). *See* Xiao, *infra* note 78, at 713 (describing "application programming interfaces ("APIs"), which are website-created tools specifically facilitating the scraping process.").

[31] *See* SIMONE NATALE, DECEITFUL MEDIA: ARTIFICIAL INTELLIGENCE AND SOCIAL LIFE AFTER THE TURING TEST 50–51 (2021).

[32] *See id.* at 52.

[14] Generative AI relies on large language models ("LLMs") capable of generating humanlike prose, answering questions, and producing content.[33] Scraping aids generative AI models by extracting content such as posts, articles, and texts from websites.[34] This data is then used to train the LLMs.[35] Unlike ELIZA, generative AI models are not restricted to set rules.[36] Rather, they can "think" intelligently and produce responses based in part on their training data.

[15] Although AI developers generally keep the specifics of their training datasets undisclosed,[37] it is recognized that most training data comes from scraping publicly accessible websites.[38] OpenAI, Meta, and Google acknowledge that they obtained their training data by scraping the internet—potentially including personal information.[39] Specifically, OpenAI stated that their training data includes scraped data, and ChatGPT

[33] *See* Sheera Frenkel & Stuart A. Thompson, *'Not for Machines to Harvest': Data Revolts Break Out Against A.I.*, N.Y. TIMES (July 15, 2023), https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html [https://perma.cc/7238-FMFS].

[34] *See* Paresh Dave, *Stack Overflow Will Charge AI Giants for Training Data*, WIRED (Apr. 20, 2023, 5:19 PM) https://www.wired.com/story/stack-overflow-will-charge-ai-giants-for-training-data [perma.cc/4KWD-XJ7G].

[35] *Id.*

[36] *See* George Lawton, *What is generative AI? Everything you need to know*, TECHTARGET, https://www.techtarget.com/searchenterpriseai/definition/generative-AI [perma.cc/4DWZ-8YQ8] (last updated Jan. 2024).

[37] *See* CONG. RSCH. SERV., R47569, GENERATIVE ARTIFICIAL INTELLIGENCE AND DATA PRIVACY: A PRIMER 4 (2023) [hereinafter CRS GENERATIVE AI REPORT].

[38] *See id*.

[39] *See* Dave, *supra* note 34.

has been trained on a combination of "licensed content, publicly available content and content created by human A.I. trainers."[40]

[16] In summary, ChatGPT's prominence pushed internet scraping into the spotlight and exposed AI models' data scraping practices.[41] This public awareness marked a pivotal shift in the perception and handling of publicly available data, with a growing inclination toward restricting scrapers' access.[42] Companies on the internet previously made data publicly available and generated revenue through ads.[43] However, the current business model is shifting toward safeguarding data on private websites,[44] making it accessible only to registered or paying users.[45]

[40] *See* Frenkel & Thompson, *supra* note 33.

[41] *See id.*; *see also* NATALE, *supra* note 31, at 52 (describing "chatbot" as "a computer program able to interact with users via a natural language database").

[42] *See* Frenkel & Thompson, *supra* note 33; *contra* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1199 (9th Cir. 2022) ("[A] defining feature of public websites is that their publicly available sections lack limitations on access; instead, those sections are open to anyone with a web browser. In other words, applying the 'gates' analogy to a computer hosting publicly available webpages, that computer has erected no gates to lift or lower in the first place.").

[43] *See* Frenkel & Thompson, *supra* note 33.

[44] *Id.*

[45] *See id.* ("Brandon Duderstadt, the founder and chief executive of Nomic, an A.I. company [states] 'Previously, the thought was that you got value from data by making it open to everyone and running ads. Now, the thought is that you lock your data up, because you can extract much more value when you use it as an input to your A.I.'").

C. Generative AI Companies

[17] OpenAI's launch of ChatGPT revolutionized the technology industry.[46] Following the success of ChatGPT, several other companies entered the generative AI field. First, Meta developed Llama 2, an "open source large language model,"[47] which like ChatGPT is trained on publicly available data.[48] The datasets used to train Llama include CommonCrawl, C4, GitHub, Wikipedia, Gutenberg and Books3, ArXiv, and StackExchange, with over half of the training data coming from CommonCrawl.[49] CommonCrawl offers users "a copy of the Internet,"[50] serving as one of the largest and most widely used repositories of scraped data.[51] Notably, Meta explicitly states that Llama 2's training data excludes "data from Meta's [own] products or services" and that efforts were made

[46] *See* Melany Amarikwa, *Generative AI Will Not Solve Algorithmic Bias*, *in* A PROMETHEAN MOMENT: TOWARDS AN UNDERSTANDING OF GENERATIVE AI AND ITS IMPLICATIONS ON BIAS 6 (2023).

[47] *See Discover the power of Llama*, META, https://ai.meta.com/llama/ (last visited Mar. 10, 2024).

[48] *See* HUGO TOUVRON ET AL., LLAMA: OPEN AND EFFICIENT FOUNDATION LANGUAGE MODELS 1 (2023), https://arxiv.org/pdf/2302.13971.pdf [perma.cc/J7MU-953J] (stating that Llama "use[s] publicly available data, making our work compatible with open-sourcing, while most existing models rely on data which is either not publicly available or undocumented (e.g. 'Books – 2TB' or 'Social media conversations')).").

[49] *See id.* at 2.

[50] *Frequently asked questions*, COMMON CRAWL, https://commoncrawl.org/faq [perma.cc/7CLJ-JFMQ] (last visited Mar. 10, 2024).

[51] *See* JAY M. PATEL, GETTING STRUCTURED DATA FROM THE INTERNET 278 (2020) ("When we take the common crawl data cumulatively, across monthly crawls since 2008, it represents one of the largest publicly accessible web crawl data corpuses on a petabyte scale, and this is one major reason why it's been used so widely in academia and the industry.").

to "remove data from certain sites known to contain a high volume of personal information about private individuals."[52]

[18] Second, Twitter[53] developed Grok amidst Elon Musk's concerns about AI companies scraping Twitter's data.[54] A distinct feature of Grok is its willingness to answer "spicy" questions often avoided by other AI models, which generally prioritize safety measures.[55] Moreover, Twitter does not provide information on the dataset used to train Grok.[56] The only information provided states that Grok was trained using "the Internet up to Q3 2023 and the data provided by AI Tutors."[57] It is also unclear whether Twitter's platform or its user data contributed to Grok's training.[58]

---

[52] Jonathan Gillham, *Meta Llama 2: Statistics on Meta AI and Microsoft's Open Source LLM*, ORIGINALITY.AI (Feb. 1, 2024), https://originality.ai/blog/meta-llama-2-statistics [perma.cc/2YQN-7UAJ].

[53] Following Elon Musk's acquisition of Twitter, the app was renamed X. For the purposes of this Article, Twitter will be used to refer to X. *See* Irina Ivanova, *Twitter is now X. Here's what that means*, CBS NEWS, https://www.cbsnews.com/news/twitter-rebrand-x-name-change-elon-musk-what-it-means [perma.cc/9YGH-7H8B] (last updated July 31, 2023, 5:18 PM).

[54] *See* Eli Tan, *What Musk and Zuckerberg entering the AI race could mean for regulation*, WASH. POST (July 14, 2023, 9:02 AM), https://www.washingtonpost.com/politics/2023/07/14/what-musk-zuckerberg-entering-ai-race-could-mean-regulation/ [perma.cc/4RXE-CN69] (describing Elon Musk announcements of his own AI venture, xAI).

[55] *See Announcing Grok*, XAI (Nov. 3, 2023), https://x.ai/blog/grok [perma.cc/X7R6-6DBY].

[56] *Id.*

[57] An AI Tutor is someone who is responsible for generating "high-quality and accurately labeled data." *See AI Tutor at xAI*, GREENHOUSE, https://boards.greenhouse.io/xai/jobs/4101903007 [perma.cc/V5HH-YX34] (last visited Dec. 18, 2023); *Grok-1 Model Card*, XAI, https://x.ai/model-card/ [perma.cc/4Q8A-HRLU] (last visited Dec. 18, 2023).

[58] *See Announcing Grok, supra* note 55.

[19] Third, Google launched Bard (now known as "Gemini"),[59] which bears a striking resemblance to ChatGPT in form and function.[60] Google's internal AI Principle guided Bard's development.[61] Bard underwent pre-training on diverse data from publicly available sources—although Google does not provide detailed information about the data.[62] However, journalists discovered language in Google's updated privacy policy indicating that Bard's training data included scraped data.[63] A Google spokesperson confirmed these findings, stating that "[o]ur privacy policy has long been transparent that Google uses publicly available information from the open web to train language models for services like Google Translate[, and the] latest update simply clarifies that newer services like Bard are also included."[64]

[59] Sissie Hsiao, *Bard becomes Gemini: Try Ultra 1.0 and a new mobile app today*, GOOGLE (Feb. 8, 2024), https://blog.google/products/gemini/bard-gemini-advanced-app/ [perma.cc/NPK4-GNM9].

[60] *See* James Manyika & Sissie Hsiao, *An overview of Bard: an early experiment with generative AI,* GOOGLE, https://ai.google/static/documents/google-about-bard.pdf [perma.cc/7CRE-DCD2] (last visited Oct. 19, 2023); *see also* Sissie Hsiao & Eli Collins, *Try Bard and share your feedback*, GOOGLE (Mar. 21, 2023), https://blog.google/technology/ai/try-bard/ [perma.cc/H2SQ-3RYU].

[61] *See id.*

[62] *See* Manyika & Hsiao, *supra* note 60.

[63] *See* Jess Weatherbed, *Google confirms it's training Bard on scraped web data, too*, THE VERGE (July 5, 2023, 11:11 AM), https://www.theverge.com/2023/7/5/23784257/google-ai-bard-privacy-policy-train-web-scraping [perma.cc/NWS7-GQJ5].

[64] *See id.*

[20] Fourth, Anthropic, founded by former OpenAI researchers,[65] developed Claude, arguably a more reliable large-scale AI system.[66] Anthropic trained Claude using three categories of data publicly available internet data, datasets licensed from third parties, and data provided by users or Anthropic's employees.[67] Anthropic acknowledges that its training data includes personal data.[68] Nevertheless, the company claims to take steps to minimize the privacy impact on individuals through the training process.[69] However, the only precaution Anthropic mentions is its abstention from scraping information from password protected pages or circumventing CAPTCHA controls.[70]

D. Generative AI's Data Scraping Problem

[21] Generative AI models require vast amounts of training data to ensure that the model "performs effectively and safely."[71] However, the data scraping practices employed by generative AI companies threaten to

[65] *See Anthropic raises $124 million to build more reliable, general AI systems*, ANTHROPIC (May 28, 2021), https://www.anthropic.com/index/anthropic-raises-124-million-to-build-more-reliable-general-ai-systems [perma.cc/62D7-KKZH] [hereinafter *Anthropic raises $124 million*]; Devin Coldewey, *Anthropic is the new AI research outfit from OpenAI's Dario Amodei, and it has $124M to burn*, TECHCRUNCH (May 28, 2021, 1:59 PM), https://techcrunch.com/2021/05/28/anthropic-is-the-new-ai-research-outfit-from-openais-dario-amodei-and-it-has-124m-to-burn [perma.cc/Z834-949B].

[66] *See Introducing Claude*, ANTHROPIC, (Mar. 14, 2023), https://www.anthropic.com/news/introducing-claude [perma.cc/HDW4-VU4D].

[67] *See How do you use personal data in model training?*, ANTHROPIC HELP CTR., https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training [perma.cc/LMJ2-XNDG] (last visited Dec. 29, 2023).

[68] *See id.*

[69] *See id.*

[70] *See id.*

[71] *See id.*

undermine public trust in technology and harm internet openness. This Section highlights existing concerns relating to generative AI models' use of scraped data.

[22] Generative AI models depend on publicly available data for training.[72] This reliance on publicly available data may be attributed to several factors. First, scraping public data may limit legal claims against the AI companies, as case law suggests that scraping publicly available data does not violate federal law.[73] Second, scraping publicly available data is relatively cheap (if not free).[74] For example, CommonCrawl and C4—some of the most popular scraping datasets—are free to access.[75] Third, the use of publicly available data is considered more transparent because it provides researchers and users with a clearer understanding of the data that went into developing the AI models.[76]

[23] By using publicly available data to train their LLMs, generative AI companies incur the direct advantages of scraping publicly available data. Conversely, the public bears the direct *disadvantages* of these companies' scraping of publicly available data. To understand the public harm, it is helpful to understand the precise definition of "publicly available data."

---

[72] Lauren Leffer, *Your Personal Information Is Probably Being Used to Train Generative AI Models,* SCI. AM. (Oct. 19, 2023), https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models [perma.cc/JTB9-FT54].

[73] *See* discussions *infra* Part III.B; CRS GENERATIVE AI REPORT, *supra* note 37, at 7.

[74] *How Much Does Web Scraping Cost?*, ZENROWS (Aug. 23, 2023), https://www.zenrows.com/blog/web-scraping-cost#web-scraping-options-and-cost [perma.cc/75LM-FS5Y].

[75] *See e.g., Get Started: Accessing the Data*, COMMON CRAWL, https://commoncrawl.org/get-started [perma.cc/42Z3-ZEX7] (last visited Mar. 17, 2024).

[76] Augustin Toma et al., *Generative AI could revolutionize health care – but not if control is ceded to big tech*, NATURE (Nov. 30, 2023), https://www.nature.com/articles/d41586-023-03803-y [https://perma.cc/C9MV-KKPY].

Courts define "publicly available data" as data "available for viewing by anyone with a web browser."[77] Webpages without a paywall, login verification, or other blocks that prevent users from readily viewing their content therefore contain "publicly available data."

[24] The scraping and use of publicly available data harms the public because companies are using individuals' public messages, posts, blogs, and online content without their consent.[78] This lack of consent presents several harms. First, companies' use of publicly available data may infringe copyrighted material.[79] The Copyright Act provides the owner of a copyrighted work the exclusive right to reproduce and distribute their work.[80] Allowing companies to reproduce and use copyrighted books, articles, and blog posts without the copyright owners' consent contradicts the purpose of copyright protection.[81] Further, several companies using publicly available data have the means to license data but instead have chosen to maximize profits.[82]

---

[77] Sharon Goldman, *Generative AI's secret sauce – data scraping – comes under attack*, VENTUREBEAT (July 6, 2023, 10:26 AM), https://venturebeat.com/ai/generative-ai-secret-sauce-data-scraping-under-attack [perma.cc/FC39-F3SA]; *see* hiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 1002, 1005 (9th Cir. 2019).

[78] *See* Meta Platforms, Inc. v. Bright Data Ltd., No. 23-cv-00077-EMC, 2024 WL 251406, at *5 (N.D. Cal. Jan. 23, 2024); Geoffrey Xiao, *Bots: Regulating the Scraping of Public Information*, 34 HARV. J. L. & TECH. 701, 710–11 (2021).

[79] *See,* e.g., Goldman, *supra* note 77.

[80] 17 U.S.C.A. § 106 (West, Westlaw through Pub. L. No. 118-41).

[81] *See* Authors Guild v. Google, Inc., 804 F.3d 202, 213 (2d. Cir. 2015) (stating that "the overall objectives of the copyright law to expand public learning while protecting the incentives of authors to create for the public good.").

[82] *Shutterstock Data Licensing and the Contributor Fund*, SHUTTERSTOCK: CONTRIBUTOR SUPPORT, https://support.submit.shutterstock.com/s/article/Shutterstock-Data-Licensing-and-the-Contributor-Fund?language=en_US [perma.cc/4QY8-9S5X] (last updated June 16, 2023).

[25] Second, companies' use of publicly available data violates the public's right to privacy. Although the US lacks a federal data privacy law, several states have enacted their own data privacy laws.[83] For example, California's Consumer Privacy Rights Act provides users with the right to know who is collecting their personal information and how it is being used.[84] Generative AI models' near indiscriminate use of scraped datasets likely contains personal information from users. For example, a ChatGPT response provided a user with the personal email of a NYT editor.[85]

[26] Finally—and most importantly—companies' use of publicly available data diminishes public trust in all AI models. Public trust may be diminished by inaccurate or problematic responses based on public training data. Publicly available data is subject to human biases and errors.[86] A Stanford study found that ChatGPT and Bard answered medical questions with racist and inaccurate responses, undoubtably due to error prone training data.[87] Public trust may also be diminished by the "black box" decision-making of AI models.[88]

---

[83] Gopal Ratnam, *Many States Have Data Privacy Laws. Where Is the Federal Law?*, GOV'T TECH. (Jan. 17, 2024), https://www.govtech.com/policy/many-states-have-data-privacy-laws-where-is-the-federal-law [https://perma.cc/V3J3-S3K5].

[84] Cal. Civ. Code § 1798.110 (West, Westlaw through Ch. 1 of 2024 Reg. Sess.).

[85] Jeremy White, *How Strangers Got My Email Address From ChatGPT's Model*, N.Y. TIMES (Dec. 22, 2023), https://www.nytimes.com/interactive/2023/12/22/technology/openai-chatgpt-privacy-exploit.html [perma.cc/L8Z6-P45C].

[86] See e.g., Jesutofunmi A. Omiye et al., *Large Language Models Propagate Race-Based Medicine,* 195 DIGIT. MED. 1 (2023), https://www.researchgate.net/publication/374885932_Large_language_models_propagate_race-based_medicine [perma.cc/U4XC-A5TC].

[87] *Id.*

[88] *See* Lou Blouin, *AI's mysterious 'black box' problem, explained*, UNIV. MICH. DEARBORN NEWS (Mar. 6, 2023), https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained [https://perma.cc/J6QK-JEW4].

III. LEGAL FRAMEWORK OF SCRAPING—PUBLIC & PRIVATE ENFORCEMENT

[27] This Part evaluates the public and private enforcement efforts aimed at limiting generative AI scraping. First, it provides a brief overview of the shift in internet actors' behavior, focusing on recent efforts taken by private parties to prevent unauthorized scraping by reducing openness. Second, it provides an overview of lawsuits brought by individuals and companies seeking to protect their data from scrapers. Third, it proceeds by breaking up the scraping cases into two periods. Fourth, it provides an overview of the US government's efforts, focusing on proposed actions by the executive and legislative branches and briefly highlighting the efforts of foreign nations. Finally, Part III concludes with an explanation of why the prior methods employed by private actors, the judiciary, and the executive branch have failed, and, consequently, why legislative action is needed to address the issue of generative AI scraping.

A. Private Regulation—Companies' Protective Measures Against Scraping

[28] Generative AI has caused companies, who for decades welcomed scraping, to revise their terms of service and close their metaphorical gates to prevent generative AI from profiting off of their data.[89] The internet has for decades been defined by its openness.[90] In *Reno v. ACLU*, the Supreme Court emphasized that the internet's public interest value lies in its

[89] Harry Guinness, *The New York Times is the latest to go to battle against AI scrapers*, POPULAR SCI., (Aug. 16, 2023, 6:00 PM), https://www.popsci.com/technology/nyt-generative-ai/ [https://perma.cc/GV2D-RYSS].

[90] *See Consumer Guide: Open Internet*, FED. COMMC'N COMM'N, https://transition.fcc.gov/cgb/consumerfacts/openinternet.pdf [perma.cc/WNV9-R4JZ] (last visited Apr. 15, 2024).

openness.[91] Consequently, companies shifting business models raises the question of whether the open internet as we know it can survive generative AI's hunger for data.

[29] Modern scraping blocks first began to emerge in the 2010s with social media platforms restricting their APIs. In 2015, Facebook, citing privacy concerns, limited its Graph API v2.0 to consenting users' personal data—notably excluding the data of users' friends.[92] Prior to this change, applications that obtained users' consent were permitted to see the consenting user's data and their friends' data—even if their friends did not consent.[93] In 2018, amidst the Cambridge Analytica data scarping scandal,

[91] *See* Reno v. ACLU, 521 U.S. 844, 886 (1997) (O'Connor, J., dissenting); DANA D. BAGWELL, AN OPEN INTERNET FOR ALL: FREE SPEECH AND NETWORK NEUTRALITY 141–42 (Melvin I. Urofsky ed., 2012) ("From the publisher's point of view, [the internet] constitutes a vast platform from which to address and hear from a worldwide audience of millions of readers, viewers, researchers, and buyers. Any person or organization with a computer connected to the Internet can 'publish' information. . . . Publishers may either make their material available to the entire pool of Internet users, or confine access to a selected group, such as those willing to pay for the privilege. 'No single organization controls any membership in the Web . . . .'").

[92] *See* Josh Constine, *Facebook is Shutting Down Its API For Giving Your Friends' Data To Apps*, TECHCRUNCH (Apr. 28, 2015, 2:06 PM), https://techcrunch.com/2015/04/28/facebook-api-shut-down/ [https://perma.cc/C8PG-2Q5J].

[93] *See* Caroline McCarthy, *Facebook: One Social Graph to Rule Them All?*, CBS NEWS (Apr. 21, 2010, 2:30 PM), https://www.cbsnews.com/news/facebook-one-social-graph-to-rule-them-all/ [https://perma.cc/DRU3-SWE6] (explaining that Mark Zuckerberg previously championed Facebook's open access. Graph API initially launched to allow applications to take advantage of users' relationships, which required seeing non-consenting users' information. "'These connections aren't just happening on Facebook, they're happening all over the Web, and today with the Open Graph we're bringing all these things together,' Zuckerberg continued.").

Facebook again limited access to its APIs.[94] At the time, Facebook's terms of service permitted outside applications to gain users' and their friends' information.[95] Despite Facebook and Instagram's restrictions, other platforms, such as Twitter and Reddit, chose to continue embracing the "openness" of the internet.[96]

[30] Twitter's format raised privacy concerns as early commenters were concerned that users' messages were published on the public website.[97] However, users adapted and came to champion Twitter's openness.[98] Openness was so central to Twitter's ethos that its decision to block former President Donald Trump from the platform led to public outcry from both

[94] *See* Christophe Olivier Schneble et al., *The Cambridge Analytica Affair and Internet-Mediated Research*, 19 EMBO REPORTS 1 (2018) (stating that Cambridge Analytica, a political consulting firm, purchased millions of Facebook users' data from thisisyourdigitallife, a quiz app); *see also* Mike Schroepfer, *An Update on Our Plans to Restrict Data Access on Facebook*, META (Apr. 4, 2018), https://about.fb.com/news/2018/04/restricting-data-access/ [https://perma.cc/LX8P-EVDN].

[95] *See* Constine, *supra* note 92 ("[P]rivacy concerns led Facebook to announce at F8 2014 that it would shut down the Friends data API in a year.").

[96] Chris Stokel-Walker, *Twitter's $42,000-per-Month API Prices Out Nearly Everyone*, WIRED (Mar. 10, 2023, 12:53 PM), https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone [perma.cc/YKT9-JBW7]; *What is the Reddit Blackout and Why Does it Matter?*, BLADE TECHS. (Sept. 1, 2023), https://www.bladetechinc.com /news/reddit-blackout-and-the-end-of-open-api [https://perma.cc/YPV5-EQUE].

[97] Michael Arrington, *Odeo Releases Twttr*, TECHCRUNCH (July 15, 2006, 10:17 PM), https://techcrunch.com/2006/07/15/is-twttr-interesting [https://perma.cc/UCR4-J2GD].

[98] Stokel-Walker, *supra* note 96.

his supporters and opponents,[99] with Republicans attempting (and failing) to pass legislation that would treat social media platforms as government actors subject to the First Amendment.[100]

[31] However, Twitter's openness fundamentally changed in 2023. In a blog post, Twitter announced that while its mission was still to promote "public conversations," it would begin to limit the reach of tweets that violated its policies.[101] The philosophy was termed "Freedom of Speech, Not Reach."[102] Later that year, Twitter implemented restrictions on tweet accessibility.[103] Prior to the restrictions, anyone could access Twitter's website and browse public tweets without limitation.[104] Elon Musk cited

[99] *See* X, *Permanent suspension of @realDonaldTrump*, X BLOG (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension [https://perma.cc/BF3G-X3XP]; *see also* Ryan Calo & Woodrow Hartzog, *Op-Ed: Banning Trump from Twitter and Facebook isn't nearly enough*, L.A. TIMES (Jan. 15, 2021, 3:30 AM), https://www.latimes.com/opinion/story/2021-01-15/facebook-twitter-extremism-donald-trump-violence [https://perma.cc/T4EV-CXKL] ("Even as many applaud Twitter and Facebook for finally 'deplatforming' this toxic president, others cower at the enormous power internet companies hold over public discourse — concerns wrapped up with deep American intuitions around enabling free speech.").

[100] *See* Mark A. Lemley, *The Contradictions of Platform Regulation*, 1 J. FREE SPEECH L. 303, 308 (2021) (citing Ending Support for Internet Censorship Act, S. 1914, 116th Cong. (2019)).

[101] X Safety, *Freedom of Speech, Not Reach: An update on our enforcement philosophy*, X BLOG (Apr. 17, 2023), https://blog.twitter.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy [https://perma.cc/RG4F-CL88].

[102] *Id.* ("Freedom of Speech, not Freedom of Reach - our enforcement philosophy which means, where appropriate, restricting the reach of Tweets that violate our policies by making the content less discoverable.").

[103] Gintaras Radauskas, *AI and data scraping: websites scramble to defend their content*, CYBERNEWS (Nov. 15, 2023, 12:53 PM), https://cybernews.com/editorial/ai-data-scraping-websites [perma.cc/WA6T-D9M5].

[104] Stokel-Walker, *supra* note 96.

generative AI as one of the main reasons for the policy change, stating that "[s]everal entities tried to scrape every tweet[.]"[105]

[32] In addition to restricting access to tweets, Musk announced a monthly API charge.[106] This new pricing structure renders the API prohibitively expensive for its primary users, which include academics and researchers.[107]

[33] Reddit also started charging for access to its API.[108] Reddit, which launched in 2005,[109] previously had a long history of fostering an open community of user-generated commentary.[110] In response to the API change, Reddit's CEO stated that there is no need "to give all of that value to some of the largest companies in the world for free[.]"[111] Stack

[105] Radauskas, *supra* note 103.

[106] *See* Vittoria Elliott, *Threads Is the Latest Move in the AI Arms Race*, WIRED (July 25, 2023, 8:00 AM), https://www.wired.com/story/threads-is-the-latest-move-in-the-ai-arms-race/ [perma.cc/XF66-NWMF].

[107] *Id.*

[108] *See* Frenkel & Thompson, *supra* note 33.

[109] *See Reddit: American social media forum website*, BRITANNICA: HIST. & SOC'Y, https://www.britannica.com/topic/Reddit [https://perma.cc/J6BK-LQE4] (last visited Mar. 18, 2024).

[110] *See* Elliot T. Panek, *What Is Reddit*? 4–5 (2021) ("Many Reddit users, now and throughout Reddit's history, do not engage with this commentary in any meaningful way, choosing to use Reddit only as a means of content sorting. But for millions of other users, reading and posting comments are central to the Reddit experience."); Vincent Marrazzo, *The Federalists of the Internet? What Online Platforms Can Learn From Reddit's Decentralized Content Moderation Scheme*, 2023 NEB. L. REV. 1, 5 (Jan. 19, 2023) ("In fact, online platforms are increasingly being compared to democracies, constitutions, and legal regimes. Reddit itself classifies its approach to content moderation as a form of democracy, and the company as a form of quasi-government.").

[111] *See* Radauskas, *supra* note 103.

Overflow[112] also announced that it would be using a paywall to block generative AI scraping efforts.[113] In summary, these once open companies limited data access as a direct result of generative AI's data scraping practices.

[34] Employing paywalls to prevent scraping is just one tactic used by private entities against generative AI. Companies are also responding by implementing rate limits, detections for non-human behavior, malicious entities blocks, honey potting, and firewalls.[114] However, these measures do not entirely block scraping.[115] Rather, the measures make scraping more challenging, but with the growing intelligence of scraping bots, their detectability becomes increasingly difficult.[116]

[112] Stack Overflow was launched in 2008 "to empower the world to develop technology through collective knowledge." The platform provides millions of users, whether they have accounts or not, with access to a Q&A community centered around programming and technology. *See Who we are*, STACK OVERFLOW, https://stackoverflow.co/company/careers [https://perma.cc/A8D7-QUBY] (last visited Mar. 28, 2024).

[113] *See* James Vincent, *AI is killing the old web, and the new web struggles to be born*, THE VERGE, (June 26, 2023, 11:25 AM), https://www.theverge.com/2023/6/26/23773914/ai-large-language-models-data-scraping-generation-remaking-web [https://perma.cc/4Q56-ZULZ].

[114] Lillian Pierson, *9 Fast Measures to Stop Hackers from Stealing Your Data!*, DATA MANIA, https://www.data-mania.com/blog/prevent-web-scraping-9-fast-measures-to-keep-your-data-safe [https://perma.cc/32T8-CJSJ] (last visited Mar. 28, 2024); *Firewalls and firefights*, THE ECONOMIST: BUSINESS (Aug. 10, 2013), https://www.economist.com/business/2013/08/10/firewalls-and-firefights [https://perma.cc/CF8B-TXFK].

[115] *See* Radauskas*, supra* note 103.

[116] *What is Scraping? Protection From Web Scraping & Data Scraping,* HUM. SEC., https://www.humansecurity.com/learn/topics/what-is-scraping [https://perma.cc/X8XN-2TXG] (last visited Mar. 28, 2024).

[35] Despite all of this, internet openness is not dead. Following public outcry, Stack Overflow and Reddit announced their intention to continue licensing data for free to certain individuals and companies, distinguishing themselves from Twitter's tiered approach.[117] Stack Overflow further clarified that the company's payment structure is designed to extract compensation from companies employing generative AI models for commercial purposes.[118]

B. Judicial Regulation—The Scraping Cases

[36] In 1996, the Southern District of New York defined a "web scraper" as a "software capable of automatically contacting various Web sites and extracting relevant information."[119] Following this, web scraping cases began appearing across the country. The progression of scraping cases may be categorized into two periods.

1. First Period of the Scraping Cases

[37] During the first period of scraping cases, platforms alleged that scraping violated the Computer Fraud and Abuse Act ("CFAA"), relying on the "unauthorized access" language of the CFAA to bring their claims.[120] The CFAA prohibits one who "intentionally accesses a computer without

[117] *See* Dave, *supra* note 34; *see also* Elliott, *supra* note 106 ("Musk announced that X would begin charging $42,000 a month for its API, pricing out nearly everyone that used it, particularly academics and researchers, for whom data from X was crucial for research into topics like disinformation. Later, [X/Twitter] said it would offer tiers of access priced at $125,000 and $210,000 per month.").

[118] *See* Dave, *supra* note 34.

[119] Andrew Sellars, *Twenty Years of Web Scraping and the Computer Fraud and Abuse Act*, 24 B.U. J. SCI. & TECH. L. 372, 383 (2018) (citing Shea *ex rel*. The Am. Reporter v. Reno, 930 F. Supp. 916, 929 (S.D.N.Y. 1996)).

[120] CHARLES DOYLE, CONG. RSCH. SERV., 97-1025, CYBERCRIME: AN OVERVIEW OF THE FEDERAL COMPUTER FRAUD AND ABUSE STATUTE AND RELATED FEDERAL CRIMINAL LAWS 51 (2014).

authorization or exceeds authorized access."[121] This overly broad language resulted in a circuit split, with some circuits interpreting the CFAA to cover both terms of service and employer computer policy violations, while others argued that this interpretation was overly expansive.[122] Courts acknowledged the inadequacy of the CFAA in addressing scraping, with the Ninth Circuit stating that "[t]he current broad reach of the CFAA may well have impacts on innovation, competition, and the general 'openness' of the internet . . . but it is for Congress to weigh the significance of those consequences and decide whether amendment would be prudent."[123]

[38] Finally, in *Facebook, Inc. v. Power Ventures ("Power Ventures")*, the Ninth Circuit clarified that a company's continued scraping after receiving a cease-and-desist letter constitutes "unauthorized access" and thus violates the CFAA.[124] However, in 2018, LinkedIn, following *Power Ventures*, issued a cease and desist and blocked hiQ, a competing tech startup, from accessing its platform after it discovered that the startup had been scraping its platform data.[125] hiQ subsequently sought a declaratory judgement and injunctive relief preventing LinkedIn from blocking the startup from its site.[126] The Ninth Circuit found that the public interest favored granting the preliminary injunction.[127] However, the court declined

[121] 18 U.S.C. § 1030(a)(2).

[122] *See* Samantha Hourican, *CFAA and Van Buren: A Half-Measure for a Whole-ly Ineffective Statute*, 47 SETON Hall LEGIS. J. 30, 36 (2023).

[123] Craigslist, Inc. v. 3Taps, Inc., 964 F. Supp. 2d 1178, 1187 (N.D. Cal. 2013).

[124] Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, *1068 (9th Cir. 2015) ("We therefore hold that, after receiving written notification from Facebook on December 1, 2008, Power accessed Facebook's computers 'without authorization' within the meaning of the CFAA and is liable under that statute.").

[125] *See id.*

[126] hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099, 1104 (N.D. Cal. 2017).

[127] *hiQ Labs, Inc.*, 938 F.3d, at 1005.

to issue a ruling on LinkedIn's CCFA counter claim because of the pending Supreme Court decision in *Van Buren v. United States* ("*Van Buren*").[128]

[39] In *Van Buren*, a former officer was charged with and convicted of a felony violation of CFAA for running license-plate searches in a law enforcement computer database in exchange for money.[129] The CFAA subjects anyone to criminal liability for "intentionally access[ing] a computer without authorization or exceed[ing] authorized access."[130] It defines "exceeds authorized access" as "access[ing] a computer with authorization and to use such access to obtain or alter information in the computer that the accesser is not entitled so to obtain or alter."[131] The Court found that Van Buren did not "exceed authorized access" to the database under the CFAA because he could access the data using his password.[132] An employee does not "exceed authorized access" when they obtain information for an improper purpose in violation of policies.[133] However, in footnote 8, the Court left open the possibility that the CFAA may still apply to improper purposes in violation of platforms' terms of service.[134]

[40] After the *Van Buren* decision, the Ninth Circuit applied the *Van Buren* "gates-up-or-down" inquiry to *hiQ v. LinkedIn* and found that LinkedIn's attempts to "revoke" hiQ's authorization through cease-and-

[128] hiQ Labs, Inc. v. LinkedIn Corp., 2021 U.S. Dist. LEXIS 75955, *11 (N.D. Cal. 2021).

[129] Van Buren v. United States, 141 S. Ct. 1648, 1653 (2021).

[130] 18 U.S.C. § 1030(a)(2).

[131] 18 U.S.C. § 1030(e)(6).

[132] *See Van Buren* 141 S. Ct. at 1662.

[133] *Id.*

[134] *Id.* at 1659 n.8 ("For present purposes, we need not address whether this inquiry turns only on technological (or 'code-based') limitations on access, or instead also looks to limits contained in contracts or policies.").

desist letters and anti-scraping measures did not establish liability under the CFAA.[135] It distinguished *hiQ* from *Power Ventures* by identifying the data at issue in *hiQ* as publicly available.[136] Unlike Facebook's password-protected data, LinkedIn's data was available to anyone with internet access, and consequently, hiQ did not improperly bypass any gates when scraping.[137] And with this decision, the success and use of CFAA claims in scraping cases dwindled.[138] Although a footnote in *Van Buren* left open the possibility of platforms' terms of service applying, the *hiQ* decision shut that door on policy violations constituting violations of the CFAA.[139]

2. Second Period of the Scraping Cases

[41] Even considering the CFAA setback in *hiQ*, the second period witnessed several platforms and private individuals bringing claims against unauthorized scraping. Public outcries questioning platform security and privacy also led to platforms filing lawsuits. Clearview AI provides an example of this dynamic. Clearview AI, a facial recognition app, allowed federal and state law enforcement officers to monitor citizens without their consent or knowledge.[140] The *New York Times* article goes on to state that

[135] hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1198–99 (9th Cir. 2022).

[136] *See id.* at 1184; Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1069 (9th Cir. 2016).

[137] *See Facebook, Inc.*, 844 F.3d at 1062–63; *hiQ Labs, Inc.*, 31 F.4th at 1199.

[138] *What Recent Rulings in 'hiQ v. LinkedIn' and Other Cases Say About the Legality of Data Scraping*, FARELLA BRAUN + MARTEL (Dec. 22, 2022), https://www.fbm.com/publications/what-recent-rulings-in-hiq-v-linkedin-and-other-cases-say-about-the-legality-of-data-scraping/ [https://perma.cc/4H2P-QEQ7] [hereinafter *Recent Rulings in* hiQ v. LinkedIn].

[139] *Id.*; *see hiQ Labs, Inc.*, 31 F.4th at 1199; *see Van Buren* 141 S. Ct. at 1659 n.8.

[140] *See* Kashmir Hill, *The Secretive Company That Might End Privacy as We Know It*, N.Y. TIMES (Nov. 2, 2021), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html [perma.cc/PN24-58GN].

"[t]he tool could identify activists at a protest or an attractive stranger on the subway, revealing not just their names but where they lived, what they did and whom they knew."[141]

[42] After *The New York Times* broke the story of Clearview AI,[142] numerous lawsuits followed.[143] Plaintiffs alleged that Clearview's facial image database consisting of over three billion images violated various state consumer protection and privacy statutes.[144] In 2022, Clearview settled a suit brought by the American Civil Liberties Union ("ACLU"), which permanently prohibited the company from providing its database to most businesses and other private entities across the US.[145]

[43] High-profile scraping cases have made the public aware of platforms' failure to prevent scraping. As a result of this awareness, individual users have begun suing to hold platforms accountable. For example, a class of Twitter users sued the platform, alleging that a defect in Twitter's API between June 2021 and January 2022 allowed cybercriminals to scrape users' personally identifiable information ("PII"), including

[141] *Id.*

[142] *Id.*

[143] *See, e.g.*, Roberson v. Clearview AI, Inc., No. 1:21-cv-00174 (N.D. Ill. 2021); Marron v. Clearview AI, Inc., No. 1:20-cv-02989 (N.D. Ill. 2020).

[144] *See* Class Action Complaint at 8, Roberson v. Clearview AI, Inc., No. 1:21-cv-00174 (E.D. Va. 2020); Class Action Complaint at 18, John et al. v. Clearview AI, Inc., No. 1:21-cv-00173 (S.D.N.Y 2020); Thornley v. Clearview AI, Inc., 984 F.3d 1241, 1242–43 (7th Cir. 2021); Renderos v. Clearview AI, Inc., 2022 Cal. Super LEXIS 70732, *1, *1–2 (Cal. Super. Ct. 2022); State v. Clearview AI, Inc., 2022 Vt. Super. LEXIS 4, at *1 (Vt. Super. Ct. 2020).

[145] Consent Order of Permanent and Time-Limited Injunctions Against Defendant, Clearview AI, Inc., Am. Civ. Liberties Union v. Clearview AI, Inc., No. 2020-CH-04353, 2022 Ill. Cir. LEXIS 2887 (Ill. Cir. Ct. 2022).

usernames, email addresses, and phone numbers.[146] The complaint asserted that this data breach exposed uses to potential harms and violated Twitter's Privacy Policy, Terms of Service, and a 2011 agreement with the FTC.[147]

[44] Moreover, the rise of generative AI has made the public more aware of platforms' use of scraped data. Plaintiffs suing companies employing generative AI are taking an everything-but-the-kitchen-sink approach to complaint drafting.[148] In *Anderson v. Stability AI* ("*Anderson*"), artists brought a class action to challenge the defendants' creation or use of Stable Diffusion, a generative AI model.[149] Although the *Anderson* plaintiffs asserted several claims, the court dismissed all claims except the copyright infringement claim.[150] Additionally, several authors filed a class action against Google alleging that Bard's training data included their copyrighted works without their authorization.[151] The plaintiffs alleged that the use of the scraped data violated plaintiffs' privacy and property rights.[152] The plaintiffs also boldly alleged that "Google has been secretly stealing

[146] *See, e.g.*, Class Action Complaint at 1–3, Gerber et al. v. Twitter, Inc., No. 3:23-cv-00186 (N.D. Cal. 2023) [hereinafter Gerber v. Twitter Complaint].

[147] *Id*. at 2–3.

[148] Complaints against OpenAI, Google, and Meta include claims such as violation of negligence, invasion of privacy, intrusion upon seclusion, larceny, conversion, unjust enrichment, copyright infringement, violation of the Digital Millennium Copyright Act ("DMCA"), and violation of California Unfair Competition Law. Class Action Complaint, J.L. v. Alphabet, No. 3:23-cv-3440 (N.D. Cal. July 11, 2023).

[149] Andersen v. Stability AI Ltd., No. 23-cv-00201-WHO, 2023 WL 7132064, at *1 (N.D. Cal. Oct. 30, 2023).

[150] *Id.* at *17.

[151] *See* Class Action Complaint at *15, *17, *27, *78, *J.L.*, No. 3:23-cv-3440.

[152] *See id.* at *26, *62.

everything ever created and shared on the internet by hundreds of millions of Americans."[153]

[45] Another class sued four defendants, including Meta, Bloomberg, Microsoft, and EleutherAI Institute,[154] alleging that EleutherAI, created and distributed the dataset "Books3," which contains information scraped from pirated eBooks.[155] The suit alleged that the remaining defendants had improperly used Books3 in their training materials.[156] Literary authors and news platforms are also suing OpenAI for scraping and using their copyrighted works.[157]

[46] In summary, public concerns about platform security and privacy have prompted litigation by both platforms and private individuals. Further, the rise of generative AI has heightened public awareness of privacy issues, leading to a surge in litigation.[158] While the outcomes of these cases are pending, the influx of litigation underscores the pressing need for regulatory measures.

[153] *See id.* at *1–2.

[154] Class Action Complaint at *1, Mike Huckabee v. Meta Platforms, Inc., No. 1:23-cv-09152 (S.D.N.Y. Oct. 17, 2023) [hereinafter Huckabee v. Meta Complaint].

[155] *See id.* at *2.

[156] *See id.* at *3.

[157] *See* Class Action Complaint at *1–2, Authors Guild v. OpenAI, No. 1:23-cv-8292 (S.D.N.Y. Sept. 19, 2023); 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 [perma.cc/VF8P-ELCX].

[158] Grynbaum & Mac, *supra* note 157.

C. Government Regulation—EU and US Efforts to Protect Consumers

[47] For a long time, the legislature has consistently lagged behind innovation, even when acknowledging the need for increased regulation of new technologies.[159] While there is no shortage of congressional interest in regulating generative AI's rampant scraping practices, Congress cannot reach a consensus.[160]

[48] This Section provides an overview of the US government's efforts to address the issues that generative AI and its unauthorized scraping practice have created. First, it reviews the executive branch's efforts to regulate how agencies and its contractors use AI. Second, it examines the legislative branch's proposed legislation addressing generative AI. Third, it details the efforts of agencies, focusing specifically on one agency, the Federal Communications Commission ("FCC"). Finally, it provides a brief overview of efforts to regulate generative AI in Europe and China.

1. US Government Regulation of AI and Unauthorized Scraping

a. Presidential Efforts to Regulate AI and Unauthorized Scraping

[49] The Biden Administration has taken a leading role in AI regulation, primarily through Voluntary Commitments, Executive Orders, and Office

[159] *See* Cecilia Kang, *OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing*, N.Y. TIMES (May 16, 2023), https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html [perma.cc/DKN2-9GAF].

[160] *See id.*

of Management and Budget ("OMB") draft guidance.[161] These efforts aim to harness the potential of AI while also managing its risks.[162]

[50] The Biden Administration's initial action in addressing AI was through the Office of Science and Technology Policy ("OSTP") and its release of the so-called "AI Bill of Rights" in October of 2022.[163] OSTP noted that the AI Bill of Rights was designed to apply to "all automated systems that have the potential to meaningfully impact individuals[] or communities[]."[164] In contrast to the AI Bill of Rights, the Voluntary Commitments are designed to be enforceable but have limited applicability.[165] The Biden Administration secured Voluntary Commitments from industry leaders (including Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI), with the overarching

[161] *See Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI*, THE WHITE HOUSE (July 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/ [perma.cc/4RG5-93CR] [hereinafter *Generative AI Voluntary Commitments*]; Press Release, The White House, OMB Releases Implementation Guidance Following President Biden's Executive Order on Artificial Intelligence (Nov. 1, 2023) (available at https://www.whitehouse.gov/omb/briefing-room/2023/11/01/omb-releases-implementation-guidance-following-president-bidens-executive-order-on-artificial-intelligence/) [perma.cc/V4G8-WKN3] [hereinafter *OMB AI Guidance Announcement*].

[162] *See OMB AI Guidance Announcement, supra* note 161.

[163] *See* Office of Science and Technology Policy, *Blueprint for an AI Bill of Rights* (2022), THE WHITE HOUSE, https://www.whitehouse.gov/ostp/ai-bill-of-rights/#applying [perma.cc/S7ST-JF6R] [hereinafter OSTP, *AI Bill of Rights*].

[164] *Id.*

[165] *Ensuring Safe Secure and Trustworthy AI* 1 (2023), THE WHITE HOUSE, https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf [perma.cc/XD49-HNPC] [hereinafter *Ensuring Safe Secure and Trustworthy AI*].

goal of ensuring safe, secure, and transparent development in generative AI technology.[166]

[51] The Voluntary Commitments make no reference to data collection or sharing.[167] Most of the Commitments concentrate on safeguarding national security, with only one referencing public concerns related to the scraping of personal information.[168] Additionally, the Commitments' omission of training data, which is generative AI's most controversial aspect, suggests that industry leaders and the executive branch failed to reach a consensus on the issue.[169] Several years prior to the Voluntary Commitments, Google made its own more robust privacy commitment for the AI era,[170] which included not using users' data without explicit consent.[171] Furthermore, OpenAI has stated that "end users' data won't be used to train our models by default," which leaves open the possibility that it intends to or is already using users' data to train its models.[172]

[52] The language within the Voluntary Commitments recognizes their limitations, describing them as "only a first step in developing and enforcing

[166] *See id.; Generative AI Voluntary Commitments supra* note 161.

[167] OpenAI, *Moving AI governance forward*, OPENAI (July 21, 2023), https://openai.com/blog/moving-ai-governance-forward [perma.cc/UBL5-8S64].

[168] Commitment seven states, "advancing privacy, protecting children, and working to proactively manage the risks of AI so that its benefits can be realized." *See id.*

[169] *See generally Generative AI Voluntary Commitments*, *supra* note 161; *see also Moving AI governance forward*, *supra* note 167.

[170] *See* Andrew Moore*, Sharing our data privacy commitments for the AI era*, GOOGLE (Oct. 14, 2020), https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-unveils-ai-and-ml-privacy-commitment [perma.cc/XMF4-6L5B].

[171] *Id.*

[172] OpenAI, *New ways to manage your data in ChatGPT*, OPENAI (Apr. 25, 2023), https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt [perma.cc/P79Y-AC2S].

binding obligations to ensure safety, security, and trust."[173] Further, the Commitments conclude by advocating for the enactment of new laws addressing AI risks and committing to pursue bipartisan legislation and executive action.[174]

[53] A few months after the Voluntary Commitments, President Biden signed the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence ("AI Order").[175] Before signing the AI Order, President Biden stated that "[o]ne thing is clear, to realize the promise of AI and avoid the risk, we need to govern this technology."[176] He also explained that one of the AI Order's objectives is to help ensure that AI systems "earn [] people's trust."[177]

[54] The AI Order outlines eight pivotal principles for governing the development and use of AI in executive departments and agencies.[178] Notably, section 9 of the AI Order specifically addresses protecting privacy and instructs the Director of OMB to take several steps.[179] First, to identify

[173] *See Generative AI Voluntary Commitments*, *supra* note 161 (click "secured voluntary commitments" in the first paragraph).

[174] *See id.*

[175] *See* Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023) [hereinafter *Executive Order on AI*].

[176] ABC News, *President Biden Unveils New Executive Order Regulating Artificial Intelligence*, YOUTUBE, at 6:20 (Oct. 30, 2023), https://youtu.be/fRL1kplm1H4 [perma.cc/REQ5-T64B] [hereinafter *Biden Announces AI Executive Order*].

[177] *Id.* at 8:49.

[178] *See Executive Order on AI*, *supra* note 175 (directing the federal government to pursue broad objectives, including ensuring the safety and security of AI technology; promoting innovation and competition; supporting workers; advancing equity and civil rights; protecting consumers, patients, passengers, and students; protecting privacy; advancing federal government use of AI; and strengthening American leadership abroad.).

[179] See *Executive Order on AI*, *supra* note 175.

commercially available information ("CAI")[180] containing personally identifiable information that was either procured from data brokers or processed indirectly through vendors.[181] Second, to evaluate agency standards and procedures related to the collection, processing, and maintenance of CAI with personally identifiable information.[182] Third, to implement measures that advance research and promote the use of Privacy-Enhancing Technologies ("PETs").[183] These measures, while useful, fail to extend beyond the executive branch and instead focus on national security, rather than individuals' rights.

[55] Following the release of the Executive Order on AI, OMB released a preliminary version of Implementation Guidance ("Guidance") focusing on advancing governance, innovation, and risk management for the agency's use of AI.[184] OMB's proposed Guidance built on the AI Bill of Rights[185] and the AI Risk Management Framework.[186] The objective of the Guidance was to establish AI governance structures within federal agencies which emphasized their responsibility in identifying and managing AI risks.[187]

[180] *See id.* (defining CAI as any information or data that is made available or obtainable and sold, leased, or licensed to the general public or to governmental or non-governmental entities).

[181] *See id.*

[182] *See id.*

[183] *See id.*

[184] *See OMB AI Guidance Announcement, supra* note 161.

[185] OSTP*, AI Bill of Rights*, *supra* note 163.

[186] *See OMB AI Guidance Announcement, supra* note 161.

[187] *See id.*

[56] OMB has emphasized the role of data scraping as a crucial barrier to overcome for the responsible use of AI.[188] The Guidance also emphasized that agencies should develop the necessary infrastructure and capacity to curate datasets for AI training, testing, and operation.[189] In addition to using internal data, the OMB Guidance suggested that agencies should explore using public access datasets where appropriate.[190]

[57] Regarding the quality and appropriateness of data, the OMB Guidance directed agencies to assess the quality of data used in the AI models' developments.[191] If agencies cannot access the datasets, then obtaining sufficient descriptive information from the AI or data provider is necessary.[192] Additionally, agencies must ensure that the data used for AI development, operation, and assessment is representative of the communities impacted and that the data has been reviewed for potential bias.[193] The Guidance's recommendations on data are particularly concerning, given the data privacy concerns which have been outlined in this Article.

[188] *See* Memorandum from Shalanda D. Young, Office of Mgmt. and Budget on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence, at 9 (2023) (available at https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf) [perma.cc/EJE6-E9HY] [hereinafter Memorandum for Agency Use of Artificial Intelligence].

[189] *Id.* at 9.

[190] *Id.*

[191] *Id.* at 15.

[192] *Id.*

[193] Memorandum for Agency Use of Artificial Intelligence, *supra* note 188.

b. Legislators' Efforts to Regulate AI and Unauthorized Scraping

[58] Despite growing concerns about the implications of generative AI among the public and lawmakers, experts have argued that AI-related legislation remains a distant prospect.[194] The absence of legislation does not stem from a lack of effort. Since 2019, both the House and Senate proposed over 500 provisions directly referencing AI.[195] Only 19 of these bills have been adopted into law, and this small subset does not address AI privacy and scraping concerns.[196]

[59] The first piece of legislation mentioning AI was introduced in 1979,[197] aiming to establish a national commission studying the scientific and technological implications of information technology in education.[198] The then-Chairman of the House Subcommittee on Science, Research, and Technology, George E. Brown, Jr. highlighted the importance of understanding how to "prepare to take optimal educational advantage of new and revolutionary information and communications technologies,"

---

[194] Ann-Marie Luciano & Meghan Stoppel, *NAAG Consumer Protection Conference Explores Dark Patterns, Advertising Law, Fake Reviews & AI*, JD SUPRA (Nov. 15, 2023), https://www.jdsupra.com/legalnews/naag-consumer-protection-conference-4129898/ [perma.cc/Q3DG-Y2HF].

[195] *See generally* CONGRESS.GOV, https://www.congress.gov/ [https://perma.cc/VBC6-A6QZ] (click "legislation" in the dropdown menu and search "artificial intelligence" with quotation marks around artificial intelligence) (last visited Jan 1, 2024).

[196] *See, e.g.*, National Defense Authorization Act for Fiscal Year 2024, Pub. L. No. 118-31 (2023); Countering Human Trafficking Act of 2021, Pub. L. No. 117-322 (2022); AI Training Act, Pub. L. No. 117-207 (2022); Chips and Science Act, Pub. L. No. 117-167 (2022); Infrastructure Investment and Jobs Act, Pub. L. No. 117-58 (2021); Information Technology Modernization Centers of Excellence Program Act, Pub. L. No. 116-194 (2020).

[197] H.R. 4326, 96th Cong. (1979).

[198] *Id.*

echoing sentiments expressed by the Biden administration over 40 years later.[199]

[60] While bipartisan and industry support for AI regulation exists,[200] disagreements over specifics have stalled progress, with companies expressing concerns that excessive regulation will hinder innovation.[201] Political divisions and lobbying are key obstacles to legislation.[202] For example, OpenAI's CEO, Sam Altman, actively engaged with over 100 lawmakers.[203]

[61] Despite the legislative stall, lawmakers continue to take steps to address AI. For example, lawmakers are attempting to educate themselves on AI by hosting forums, convening hearings, and requesting briefings, to understand the implications of AI on issues like jobs, the proliferation of

[199] CONG. RSCH. SERV., Serial QQQ, INFORMATION TECHNOLOGY IN EDUCATION: PERSPECTIVES AND POTENTIALS 105 (1980).

[200] Faiza Patel & Melanie Geller, *Senate AI Hearings Highlight Increased Need for Regulation*, BRENNAN CTR. JUSTICE (Oct. 13, 2023), https://www.brennancenter.org/our-work/analysis-opinion/senate-ai-hearings-highlight-increased-need-regulation [perma.cc/F5QS-96GJ].

[201] *Id.*

[202] Melissa Heikkilä, *How judges, not politicians, could dictate America's AI rules*, MIT TECH. REV. (July 17, 2023), https://www.technologyreview.com/2023/07/17/1076416/judges-lawsuits-dictate-ai-rules/ [perma.cc/6QMD-6ZRQ].

[203] Adam Satariano, *Europeans Take a Major Step Toward Regulating A.I.*, N.Y. TIMES (June 14, 2023), https://www.nytimes.com/2023/06/14/technology/europe-ai-regulation.html [perma.cc/4M6X-5572] [hereinafter Satariano, *EU Takes Major Steps*].

disinformation, and intellectual property theft.[204] Notably, during a closed-door Senate forum with CEOs of technology companies, the CEOs unanimously agreed that "government is needed to play a role in regulating AI."[205] Congress also introduced bills addressing various aspects of AI, ranging from workforce development to consumer protection.[206] However, the introduction of ChatGPT in late 2022 pushed lawmakers to begin to seriously scrutinize the potential risks associated with generative AI.[207] The post-ChatGPT proposed bills sought to protect consumers by mandating the disclosure or labeling of AI generated content[208] and further AI usage transparency measures.[209] Additionally, Senator Josh Hawley introduced a

[204] A closed-door listening session for lawmakers as they try to devise regulations for AI technologies. *See Oversight of A.I.: Legislating on Artificial Intelligence*, U.S. S. COMM. ON THE JUDICIARY (Sep. 12, 2023), https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-legislating-on-artificial-intelligence [perma.cc/QK4G-MGQA]; Anna Edgerton, et al., *Tech Leaders Discuss AI Policy in Closed-Door Senate Meeting*, BLOOMBERG, https://www.bloomberg.com/news/articles/2023-09-13/tech-leaders-head-to-capitol-hill-to-propose-their-own-ai-rules [perma.cc/2AFJ-F6XP] (last updated Sept. 13, 2023, 4:21 PM); Kang, *supra* note 159.

[205] Patel & Geller, *supra* note 200.

[206] *See*, *e.g.*, H.R. 6553, 117th Cong. (2022) ("To promote a 21st century artificial intelligence workforce."); H.R. 7296, 117th Cong. (2022) (establishing the Artificial Intelligence Hygiene Working Group); AI Training Act, Pub. L. No. 117–207, 136 Stat. 2238 (2022) (establishing an artificial intelligence training program).

[207] Whitney Downard, *Lawmakers Grapple with Legal, Educational Implication of AI*, THE74 (Oct. 31, 2023), https://www.the74million.org/article/lawmakers-grapple-with-legal-educational-implication-of-ai/ [perma.cc/LLT9-QXL3].

[208] *See*, *e.g.*, H.R. 3831, 118th Cong. (2023) ("To require generative AI to disclose that their output has been generated by AI"); S. 2765, 118th Cong. (2023) (requiring watermark for AI generated materials); S. 2691, 118th Cong. (2023) (requiring disclosures for AI generated content).

[209] *See*, *e.g.*, S. 1865, 118th Cong. (2023) (directing agencies transparent use of automated and augmented systems); S. 2708, 118th Cong. (2023) (prohibiting the use of exploitative and deceptive practices by large online operators); S. 1671, 118th Cong. (2023) (establishing a new Federal body to oversee and regulate of digital platforms).

bill to limit the scope of Section 230 of the Communications Act of 1934, enacted as part of the Communications Decency Act of 1996, which immunizes websites from claims arising out of material posted on their platform by users.[210] If Hawley's bill were passed, it would waive Section 230 immunity for claims and charges related to generative AI.[211]

c. Agencies' Efforts to Regulate AI and Unauthorized Scraping

[62] The Federal Trade Commission ("FTC") has emerged as a leading force in regulating AI.[212] In 2023, the FTC took the first steps to address generative AI's scraping problems and issued a Civil Investigative Demand ("CID") to OpenAI.[213] The FTC CID sought records to investigate whether the company engaged in unfair and deceptive privacy and data security practices or practices relating to risk of harm to consumers.[214] In response to inquiries about the investigation, representatives from OpenAI shared a Twitter thread from CEO Sam Altman, in which he said the company is "confident we follow the law."[215]

---

[210] 47 U.S.C. §230(a)(1); *Senator Hawley Introduces Legislation to Amend Section 230 Immunity for Big Tech Companies*, SENATOR JOSH HAWLEY (June 19, 2019), https://www.hawley.senate.gov/senator-hawley-introduces-legislation-amend-section-230-immunity-big-tech-companies [perma.cc/48KE-2HU7].

[211] S. 1993, 118th Cong. (2023).

[212] *FTC Authorizes Compulsory Process for AI-related Products and Services*, FED. TRADE COMM'N (Nov. 21, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/11/ftc-authorizes-compulsory-process-ai-related-products-services [perma.cc/AT6Y-LG26].

[213] *See* FED. TRADE COMM'N, FTC File No. 232-3044, CIV. INVESTIGATIVE DEMAND ("CID") SCHEDULE (2023) [hereinafter CID OPEN AI].

[214] *See id.* at 2.

[215] Heikkilä, *supra* note 202.

[63] As a leader in the generative AI field, the outcomes of the FTC's investigation into OpenAI are expected to have far-reaching implications for the industry.[216] Potential repercussions may include fines, the deletion of unlawfully acquired data, and, in the most extreme scenario, algorithmic disgorgement.[217] This relatively new enforcement measure goes beyond deletion of scraped data and requires companies to also delete any models trained on scraped data.[218]

[64] In summary, although the US executive and legislative branches are making efforts to address the issue of AI, such measures do not go far enough and do not specifically address the issue of generative AI's use of scraped data.[219] Legislation must be passed to reach larger private companies without government contracts. In Part IV, this Article proposes legal frameworks that balance societal concerns related to data scraping with private companies' interests in using scraped data.

[216] *See id.*

[217] Algorithmic disgorgement, also known as algorithmic destruction or model destruction, involves the mandated removal of computer data models or algorithms that were developed using improperly obtained data, e.g., scraped data. *See* Joshua A. Goland, *Algorithmic Disgorgement: Destruction of Artificial Intelligence Models as the FTC's Newest Enforcement Tool for Bad Data*, 29 RICH. J.L. & TECH. 1, 2 (2023); s*ee also* Heikkilä, *supra* note 202; Tonya Riley, *The FTC's biggest AI enforcement tool? Forcing companies to delete their algorithms*, CYBERSCOOP (July 5, 2023), https://cyberscoop.com/ftc-algorithm-disgorgement-ai-regulation/ [perma.cc/EHF2-JFLQ].

[218] Tiffany C. Li, *Algorithmic Destruction*, 75 SMU L. REV. 479, 498 (2022).

[219] Sabine Neschke, *Legal Challenges Against Generative AI: Key Takeaways*, BIPARTISAN POL'Y CTR. BLOG (Jan. 18, 2024), https://bipartisanpolicy.org/blog/legal-challenges-against-generative-ai-key-takeaways/# [perma.cc/NC2V-VQAC].

2. Other Governments' Efforts to Regulate AI and Unauthorized Scraping

a. EU Regulation of AI and Unauthorized Scraping

[65] In contrast to the US, the European Union ("EU") is actively enacting legislation to regulate generative AI and address unauthorized scraping.[220] For example, the EU AI Act, introduced in 2021, underwent revisions to address technological advancements, specifically in generative AI.[221] The European Parliament reached a provisional agreement with the Council on the EU AI Act, and the agreed upon text is awaiting formal adoption by both the EU Parliament and Council.[222] However, the enforcement of the EU AI Act is not anticipated to commence until 2025,[223] suggesting minimal short-term impact on established US technology companies.[224] Nevertheless, the EU AI Act is poised to serve as a global model.

---

[220] *Id.*

[221] *See* Adam Satariano, *E.U. Agrees on Landmark Artificial Intelligence Rules*, N.Y. TIMES (Dec. 8, 2023), https://www.nytimes.com/2023/12/08/technology/eu-ai-act-regulation.html [perma.cc/7LCT-RC5G] [hereinafter Satariano, *E.U. Agrees on Landmark AI Rules*].

[222] *See EU AI Act: first regulation on artificial intelligence*, EUROPEAN PARLIAMENT, https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [perma.cc/F3XD-T6TG] (last updated Dec. 19, 2023).

[223] *See* Emilia David and Jess Weatherbed, *The EU AI Act passed – here's what comes next*, THE VERGE, https://www.theverge.com/2023/12/14/24001919/eu-ai-act-foundation-models-regulation-data [perma.cc/UR7N-UDEV] (last updated Mar. 13, 2024, 8:30 AM).

[224] *See id.* ("'In the very short run, the compromise on the EU AI Act won't have much direct effect on established AI designers based in the US, because, by its terms, it probably won't take effect until 2025,' says Paul Barrett, deputy director of the NYU Stern Center for Business and Human Rights.").

[66] As in the US, commentators in Europe have expressed concerns about regulation inhibiting innovation.[225] In negotiations over the AI Act, EU member states, including France, Germany, and Italy, expressed concerns that stringent rules might diminish the EU's appeal for AI companies.[226] Ultimately, the member states reached a compromise.[227] Rather than categorizing all General Purpose AI ("GPAI") as high-risk, they adopted a two-tier system, with regulations focusing on high-risk AI systems.[228] For example, TikTok and other social media platforms use recommendation algorithms utilizing users' personal data to suggest content.[229] Because they are not deemed "high risk," these uses would go unregulated despite their potential for public harm.[230]

[67] In light of Clearview's actions, there have been increased public concerns regarding biometric data stemmed from the practice of scraping personal data from social media.[231] To address the issue of unauthorized scraping, the EU AI Act would mandate increased transparency.[232] Specifically, the Act would require the disclosure of data used in the

[225] *Id.*

[226] *EU AI Act: Germany, France and Italy reach agreement on the future of AI regulation in Europe*, EURONEWS.NEXT, https://www.euronews.com/next/2023/11/19/eu-ai-act-germany-france-and-italy-reach-agreement-on-the-future-of-ai-regulation-in-europ [perma.cc/M7E4-CC8X] (last updated Nov. 20, 2023, 10:58 AM); *See* Satariano, *EU Takes Major Steps*, *supra* note 203.

[227] David & Weatherbed, *supra* note 223.

[228] *See id.*

[229] *See* Melany Amarikwa, *Social Media Platforms' Reckoning: The Harmful Impact of TikTok's Algorithm on People of Color*, 29 RICH. J.L. & TECH. 69, 76–77 (2023).

[230] David & Weatherbed, *supra* note 223.

[231] *See* Hill, *supra* note 140; s*ee* Satariano, *EU Takes Major Steps*, *supra* note 203.

[232] *See* Satariano*, EU Takes Major Steps*, *supra* note 203.

creation of generative AI models.[233] The EU AI Act would also mandate that companies using generative AI implement safeguards against generating illegal content.[234]

[68] Apart from the EU AI Act, individual European countries have implemented measures to tackle generative AI and unauthorized scraping. For example, earlier this year, Italy temporarily banned ChatGPT due to suspected privacy violations.[235] Additionally, France's privacy regulator, the Commission Nationale de L'informatique et des Libertés, released an action plan outlining its focus on AI, including generative AI.[236] The UK's Information Commissioner's Office also issued a statement on scraping, which emphasized that the data protection laws for publicly accessible personal information imposed information safeguarding duties upon platforms and mass scraping incidents may qualify as reportable data breaches.[237]

---

[233] *See id.*

[234] *See id.*

[235] *See* Elvira Pollina, *Italy's privacy regulator looks into online data gathering to train AI*, REUTERS (Nov. 22, 2023, 3:56 PM), https://www.reuters.com/world/europe/italys-privacy-regulator-looks-into-online-data-gathering-train-ai-2023-11-22 [perma.cc/4AGH-K2VQ]; Intelligenza artificiale: Garante privacy apre un'indagine sulla raccolta di dati personali on line per addestrare gli algoritmi. l'iniziativa è volta a verificare l'adozione di misure di sicurezza da parte di siti pubblici e privati, GARANTE PER LA PROTESZIONE DEI DATAI PERSONALI (Nov. 22, 2023), https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9952078 [perma.cc/47WZ-M6JR].

[236] *See* Natasha Lomas, *France's privacy watchdog eyes protection against data scraping in AI action plan*, TECHCRUNCH (May 17, 2023, 7:35 AM), https://techcrunch.com/2023/05/17/cnil-ai-action-plan/ [perma.cc/DSW9-DFAU].

[237] *See* JOINT STATEMENT ON DATA SCRAPING AND DATA PROTECTION, INFO. COMM'R OFF. (Aug. 24, 2023), https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf [perma.cc/5VUW-M6SM] [hereinafter ICO STATEMENT ON DATA SCRAPING].

b. China's Regulation of AI and Unauthorized Scraping

[69] China's long-standing practice of internet censorship, termed the "Great Firewall,"[238] extends to generative AI.[239] Even though these AI models are not accessible in China, the People's Republic of China has nonetheless regulated them.[240] Proposed rules dictate that AI systems must align with socialist core values and prohibit the dissemination of information that undermines state power or national unity.[241] Companies must also ensure that their chatbots generate accurate content, respect intellectual property, and only use registered algorithms.[242]

[70] Alongside these regulatory efforts, China has established the National Data Bureau.[243] This agency plays a pivotal role in defining the

[238] *See* Jeremy Geltzer*, Censoring the Silk Screen: China's Precarious Balance Between State Regulation and a Global Film Market*, 6 J. INT'L MEDIA & ENT. L. 123, 151 (2016) (describing China's longstanding practice of internet censorship).

[239] *See* Charlie Campbell, *China Is Betting Big on Artificial Intelligence—Even as It Cracks Down on ChatGPT*, TIME, https://time.com/6258089/china-great-firewall-chatgpt-ai-future/ [https://perma.cc/U5EP-Z2JT] (last updated Feb. 24, 2023, 10:25 AM).

[240] *See id.*

[241] *See id.* ("That clearly won't do in China. But if China's own AI-content tools only parse data from within the confines of the Great Firewall, they should simply serve as another Party propagandist embedded onto every laptop and phone.").

[242] Chang Che, *China Says Chatbots Must Toe the Party Line*, N.Y. TIMES (Apr. 24, 2023), https://www.nytimes.com/2023/04/24/world/asia/china-chatbots-ai.html [perma.cc/K9YE-VVYN].

[243] Yan Luo et al., *China Reveals Plan to Establish a National Data Bureau*, COVINGTON (Mar. 8, 2023), https://www.nytimes.com/2023/04/24/world/asia/china-chatbots-ai.htmlhttps://www.globalpolicywatch.com/2023/03/china-reveals-plan-to-establish-a-national-data-bureau/ [perma.cc/TE8C-4SUU].

terms of data ownership, acquisition, and trade.[244] The National Data Bureau assists companies in developing datasets for training models, oversees data exported by multinational companies, enforces data collection rules, and addresses digital issues such as algorithmic manipulation, potential internet addiction among minors, and identification of data security vulnerabilities susceptible to cyberattacks.[245]

[71] In conclusion, lawmakers and private AI companies acknowledge the desperate need for clear regulation to guide the development and use of generative AI models but continue to drag their feet on meaningful action.[246] Other branches of government which are less burdened by partisan gridlock, like the Executive, are working within their powers to push AI regulation forward.[247] However, as noted by President Biden, these efforts while useful, fail to adequately address the risks AI poses.[248] Additionally, the efforts adopted in the EU, although a step in the right direction, fail to adequately address the issue of generative AI's use of scraped data.[249] The harms to intellectual property and privacy rights cannot be effectively addressed through transparency measures alone. As Senator Richard Blumenthal, one of the most active US public officials working on AI, stated in a Senate Judiciary Committee Hearing on AI, "[w]e need to learn from our experience from social media that if we let this horse get out

[244] *See* Keith Zhai, *China to Create New Top Regulator for Data Governance*, WALL ST. J. (Mar. 6, 2023), https://www.wsj.com/articles/china-to-create-new-top-regulator-for-data-governance-c9317233.

[245] *See* Keith Zhai, *China to Create New Top Regulator for Data Governance*, WALL ST. J. (Mar. 6, 2023, 4:30 AM), https://www.wsj.com/articles/china-to-create-new-top-regulator-for-data-governance-c9317233 [perma.cc/C6AV-N9ZC].

[246] *See* Kang*, supra* note 159.

[247] *See, e.g.*, EXECUTIVE ORDER ON AI, *supra* note 175.

[248] *See Biden Announces AI Executive Order*, *supra* note 176.

[249] *See* David & Weatherbed, *supra* note 223.

of the barn, [AI models] will be even more difficult to contain than social media."[250]

IV. GENERATIVE AI'S SCRAPING PRACTICES

[72] Although generative AI's unauthorized scraping and use of data presents harm, an all-out ban or overly restrictive regulation of scraping also threaten to harm the public interest. As Parts II and III of this Article outlined, the legislature's failure to act has resulted in judicial confusion, public unrest, and corporate tyranny. Part IV introduces frameworks which balance the need for scraped data by generative AI companies with preserving the public's interest in intellectual property and privacy.

[73] How to regulate AI is an open question that seemingly everyone has attempted—but failed—to answer. This Part does not attempt to solve this nebulous question. Rather it narrows the question's focus to addressing how to regulate the misuse of publicly available data by generative AI models and prevent private companies from responding with overly restrictive scraping policies. Regulation comes in three forms: pure government regulation, hybrid regulation, and private regulation.[251] This Part focuses on pure government regulation.

A. When to Regulate Generative AI's Scraping?

[74] AI models can be regulated at two stages: the development stage or the application stage. At the development stage, regulation focuses on how companies and individuals develop, design, and train their models.[252] At the application stage, regulation focuses on how companies and individuals use

[250] *See* CNBC Television, *Senate Judiciary Subcommittee Holds a Hearing on AI Legislation and Oversight*, YOUTUBE, at 7:53 (Sep. 12, 2023), https://www.youtube.com/watch?v=fF89vgPgI_w&t=1495s [perma.cc/4RPH-GPAQ] [hereinafter *Senate Judiciary Subcommittee Hearing on AI*].

[251] *See* discussions *supra* Part III; *see also* discussions infra Part IV.A.

[252] *See e.g.*, OSTP, *AI Bill of Rights*, *supra* note 163.

their models by, for example, requiring human decision-making in sensitive matters.[253]

[75] The development stage may be further broken down.[254] First, unsupervised pre-training involves training the model using large text datasets.[255] Companies either employ their own scraping bots or use public scraping datasets to develop their models.[256] For example, Google employs scraping bots and Meta uses a public scraping dataset.[257] Second, supervised pre-training refines the model using a smaller labeled dataset.[258] Finally, the training process involves using supervised finetuning and reinforcement learning with human feedback, which is a dataset noting human annotators preferred outputs.[259]

[76] Regulation addressing the misuse of publicly available data by generative AI models ought to be directed at the development stage—specifically the pre-training of AI models. Although training data may be used at later stages of the development process, the problems of using

[253] *See e.g.*, *id.*

[254] *See* Konstantinos I. Roumeliotis & Nikolaos D. Tselikas, *ChatGPT and Open-AI Models: A Preliminary Review*, 15 FUTURE INTERNET 192, 196 (2023).

[255] *Id.*

[256] Lauren Leffer, *Your Personal Information Is Probably Being Used to Train Generative AI Models*, SCI. AM. (Oct. 19, 2023), https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/ [perma.cc/R4KE-LZ2Z].

[257] *See* discussions *supra* Part II.B.1.

[258] *See* Roumeliotis & Tselikas, *supra* note 254, at 196.

[259] *See* HUGO TOUVRON ET AL., LLAMA 2: OPEN FOUNDATION AND FINE-TUNED CHAT MODELS 5 (2023) https://arxiv.org/abs/2307.09288 [https://perma.cc/V4EE-5K82].

scraped data arise at this level because it causes downstream issues that may infringe upon an individual's property or privacy rights.[260]

[77] For example, the issue of generative AI models' copying the tone or style of famous authors occurs at the training stage, when the model learns to mimic.[261] Furthermore, regulating generative AI models during the development stage ensures that developers identify issues before they become "invisible" or deeply ingrained in the model's functionality.[262]

1. Proposed Framework for Generative AI Scraping Regulation

[78] The scraping of publicly available data by generative AI models presents several issues for the public, with some of the most prominent issues relating to intellectual property and privacy. First, the intellectual property issue involves copyright infringement. Second, the privacy issue consists of generative AI's collection, use, and retention of private or personally identifiable information in violation of individual's right to privacy. This Subsection evaluates the saliency of these issues and present a framework to address the copyright infringement and privacy issues.

[260] *See id.* at 20 (stating that Meta, in its development of Llama 2, acknowledged that "pretraining data . . . shed[s] light on root causes of potential downstream issues[.]").

[261] James Vincent, *The scary truth about AI copyright is nobody knows what will happen next*, THE VERGE: A.I. (Nov. 15, 2022, 10:00 AM), https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data [perma.cc/U2KP-HC28].

[262] *See* James Bridle, *The stupidity of AI*, THE GUARDIAN (Mar. 16, 2023), https://www.theguardian.com/technology/2023/mar/16/the-stupidity-of-ai-artificial-intelligence-dall-e-chatgpt [https://perma.cc/EY7Q-MKMC].

a. Framework 1: Addressing Intellectual Property Concerns as They Relate to Copyright Infringement

[79] The intellectual property issue presents a complicated scenario. Although copyright law seeks to promote innovation by offering original authors exclusive rights, the US limits the rights it grants authors.[263] For example, while Europe provides authors with extensive moral rights, the US limits moral rights and favors economic rights.[264] The US thus seeks to strike a balance where authors receive incentives to create but also permit other authors to create and where the public to benefit from their works.

[80] A similar balancing of interests may be applied to address the issue of data scraping. Under the public domain framework, the reviewing body would evaluate the use of the copyrighted work. Did the author register the work? Has the author restricted use of the work? Is the author monetizing or seeking to monetize the work? Where an author is monetizing or seeking to monetize their publicly available copyright protected data, generative AI models may not scrape or use such data without authorization. Although the data is in the public domain, the author's monetization attempts demonstrate that they do not intend for the work to enter the commons.

[81] Conversely, where an author is not monetizing or seeking to monetize their publicly available copyright protected data or controlling its use, the works enter the public domain. Notably, to enter the commons, the author must make no attempts to monetize or control the work. The control requirement attempts to account for the situations where authors permit others to use their work but do not collect payment. Authors should not be penalized or forced to collect payment in order to protect their work.

---

[263] *What is Copyright?*, U.S. COPYRIGHT OFF., https://www.copyright.gov/what-is-copyright/ [https://perma.cc/JG4L-AQHB] (last visited Mar. 25, 2024).

[264] *See* Amy M. Adler, *Against Moral Rights*, 97 CAL. L. REV. 263, 264, 295 (2009).

[82] Generative AI models may then scrape and use data in the commons without authorization. Additionally, if the authors are later identified they may request licensing fees.[265] This framework resolves many of the authorship issues present in generative AI's scraping, as a substantial portion of written content on the internet cannot be easily traced back to a single author.

[83] Reddit provides a prime example of the difficulty of tracing author identity because Reddit users rarely if ever use their legal name.[266] Additionally, the average Reddit user does not monetize, seek to monetize, or control their works.[267] Consequently, under this framework, a generative AI company may use scraped Reddit posts to train its model.

[84] However, a generative AI model may not use Twitter data as readily. A generative AI model seeking to use Twitter data may not automatically assume that a post's author is not monetizing it. [268] Additionally, Twitter

[265] *See* Katherine M. Meeks, *Adverse Possession of Orphan Works*, 33 LOY. L.A. ENT. L. REV. 1, 17 (2013) ("If after a reasonably diligent search a party fails to find the owner of a copyrighted work, the Copyright Office would allow him to use that work without permission on the condition that he pay a licensing fee if the author eventually surfaces. In essence, the proposal would liberate parties to use orphaned works by capping potential damages.").

[266] Nicholas Proferes et al., *Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics*, SOCIAL MEDIA + SOC. 2 (May 26, 2021), https://journals.sagepub.com/doi/abs/10.1177/20563051211019004 [perma.cc/G9J7-54QX].

[267] Mathew Ingram, *Reddit moves to monetize its unruly community,* COLUM. JOURNALISM REV. (Mar. 14, 2024), https://www.cjr.org/the_media_today/reddit_ipo.php [perma.cc/6XKF-47CF].

[268] *See Monetization*, X CREATE, https://create.twitter.com/en/goals/monetization [https://perma.cc/YGJ4-MW4X] (explaining the many ways one can monetize their content on Twitter "[w]hether it's through your Tweets, videos, or Spaces, there are plenty of ways to monetize the content you produce on Twitter") (last visited Mar. 25, 2024).

users tend to use their real names, so the post is not an "orphan work."[269] In summary, this framework allows generative AI models to scrape data without the risk of incurring significant legal damages for copyright infringement and protect the rights of authors.

[85] Additionally, consider *The New York Times v. Microsoft*, where the plaintiff, *The New York Times*, alleged that Microsoft and OpenAI refused to recognize its copyrights and were using its content without permission to develop their models.[270] Under the public domain framework, *The New York Times* would succeed in a copyright infringement action against Microsoft and OpenAI. To succeed under the proposed public domain framework, a copyright holder must demonstrate that their work is identifiable, protected or controlled, and monetizable. *The New York Times* easily satisfies the identifiability requirement as nearly all of its articles include the authors' names and its website includes a copyright notice.[271] *The New York Times* also satisfies the control and monetizable element. Although visitors may view a few articles without restrictions, their access will eventually be limited unless they create an account and subscribe.[272]

[86] Even if *The New York Times* did not require visitors to pay for a subscription, the newspaper may still succeed under the public domain framework if it demonstrates that it had control over its works. The purpose

---

[269] Delip Rao et al., *Classifying Latent User Attributes in Twitter*, JOHN HOPKINS U. 37, 38 (Oct. 30, 2010) https://dl.acm.org/doi/pdf/10.1145/1871985.1871993 [perma.cc/TCD8-MYJG].

[270] Complaint at 2–3, The New York Times v. Microsoft, No., 1:23-cv-11195 (S.D.N.Y. 2023).

[271] *Copyright Notice*, N.Y. TIMES, https://help.nytimes.com/hc/en-us/articles/115014792127-Copyright-Notice [perma.cc/V26L-DX9C] (last visited Mar. 16, 2024).

[272] Rohit Supekar, *How The New York Times Uses Machine Learning To Make Its Paywall Smarter*, N.Y. TIMES OPEN (Aug. 10, 2022), https://open.nytimes.com/how-the-new-york-times-uses-machine-learning-to-make-its-paywall-smarter-e5771d5f46f8 [perma.cc/VUM5-FHY7].

of the control element is to ensure that the works at issue are of the type that the author, independent of third parties' scraping efforts, seeks to control. For example, if an individual posts a fan fiction book on a public forum and then invites the public to use, edit, and reimagine their work, they may not later claim that they controlled their work. Similarly, if an artist creates digital art and permits others to create and share derivatives, under the public domain framework, the artist may only collect reasonable licensing fees from the infringing generative AI model.

[87] In summary, the public domain framework seeks to allow generative AI companies to scrape publicly available data on the internet while also protecting individual copyright holders' interests by allowing them to retroactively seek licensing fees for the use of the work.

b. Framework 2: Addressing Privacy Concerns

[88] Another prominent issue relating to generative AI's scraping is the issue of privacy as it impacts everyone who uses the internet. The privacy issue may be viewed from two angles: generative AI development and usage. Generative AI usage presents privacy concerns, as it is unclear whether models retain and learn from the information users input into the models.[273]

[89] This Article deals with the issue of scraping private data at the development stage. The public records framework aims to address the issue of private information that an individual has themselves made public, e.g., where a person posts a personal essay to Substack detailing their life or includes personal information on their LinkedIn profile.

[273] *See*, e.g., U.S. Gov't Accountability Off., GAO-23-106782, Science & Tech Spotlight: Generative AI (2023) ("Information about how and when some generative AI systems retain and use information entered into them is sparse or unavailable to many users, which poses risks for using these tools. For example, if a user enters sensitive information into a prompt, it could be stored and misused or aggregated with other information in the future.").

[90] Under the public records framework, if an individual above the age of eighteen intends to make private information about themselves publicly accessible, then the information shall be treated as a public record. The intent element of this framework is crucial. For instance, a person might post something on a website intending it solely for their followers, but due to third-party actions or other factors, it reaches a much broader audience. In such cases, this information should not be automatically deemed public. Moreover, the intent element seeks to uphold individual agency over their data while also acknowledging individuals' sometimes limited and flawed understanding of how the internet and platforms function.

[91] At first sight, one might think that treating this information as a public record removes an individual's interest in privacy, but this is not the case. Treating information that individuals post about themselves as public records does not remove their right to privacy. Even public records subject to FOIA include privacy exemptions, including law enforcement records that "could reasonably be expected to constitute an unwarranted invasion of personal privacy."[274] The Supreme Court has held that even information that has been previously publicly disclosed may still be private.[275]

[92] This proposition is supported by the work of law professors Woodrow Hartzog and Frederic Stutzman, who have written on the concept of obscurity, which finds that public records may be considered private depending on the context.[276] Professors Hartzog and Stutzman quote Solove's claim that "there is a considerable loss of privacy by plucking

[274] Daniel J. Solove, *Access and Aggregation: Public Records, Privacy and the Constitution*, 86 MINN. L. REV. 1137, 1159 (2002).

[275] United States Dep't of Jus. v. Reporters Comm. for Freedom of Press, 489 U.S. 749, 764 (1989) ("[T]here is a vast difference between the public records that might be found after a diligent search of courthouse files, county archives, and local police stations throughout the country and a computerized summary located in a single clearinghouse of information.").

[276] Woodrow Hartzog & Frederic Stutzman, *The Case for Online Obscurity,* 101 CAL. L. REV. 1, 18 (2013).

inaccessible facts buried in some obscure [public] document and broadcasting them to the world on the evening news."[277]

[93] For example, when an individual publishes an article about their divorce on the internet, the information becomes a public record. Conversely, when an individual posts the same information on a private social media account, the information does not become a public record. The distinction between the two situations is rooted in the context in which the information was disclosed. In the first case, the individual did not restrict who can access the information. Additionally, the nature and intent of the individual must be considered. Public posts or articles aim to get the public or those outside of the author's circle engaged. It is not uncommon for obscure or small accounts to go "viral" or garner wide-spread attention. Thus, even if the user usually receives views only from close friends or specific individuals and does not generally anticipate a broader audience, there is a reasonable expectation that their content may be seen by anyone.

[94] The California Privacy Rights Act's ("CPRA") updated definition of "publicly available data" indicates that there may be growing support for a public record framework. The California Consumer Privacy Act ("CCPA") first defined publicly available data as "information that is lawfully made available from federal, state, or local government records[.]"[278] The CPRA later expanded this definition to include "information that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media; or information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience."[279]

---

[277] *Id.*

[278] *What Is "Publicly Available Information" Under the CCPA?*, TRUEVAULT, https://www.truevault.com/learn/ccpa/what-is-publicly-available-information-under-the-ccpa [perma.cc/HAV2-3QHJ] (last visited Apr. 15, 2024).

[279] *Id.*

[95] Consequently, a generative AI model scraping and using personal information from a private personal blog invades an individual's right to privacy, while the scraping of personal information from a public LinkedIn profile does not. The difference lies in contextual integrity.[280] In the blog example, the author does not intend to publicize their personal information such as their age or where they went to high school, instead the author intends to convey their story. Conversely, someone posting similar information to their LinkedIn profile intends for visitors to learn this information about them and thus the platform features it prominently.

[96] In conclusion, public concern relating to generative AI data scraping primarily revolves around concerns related to intellectual property protection, privacy, and competition. Legislative efforts relating to generative AI's scraping should prioritize intellectual property protection and privacy over competition concerns. To address these issues, I recommend implementing the public domain and public records frameworks. The public domain framework permits the use of copyrighted works if the author fails to claim or make use of such works, and the public records framework treats certain personal information made publicly available as a public record.

V. CONCLUSION

[97] In conclusion, while new technology brings the promise of enhancing our lives, it is met with understandable apprehension. As our society progresses into the digital era, the solidification of openness often leads to the erosion of traditional concepts like privacy. These shifts, however, are not entirely negative; they signify the necessary adjustments society must undergo to integrate new technologies.

[280] Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119, 119 (2004) ("Contextual integrity ties adequate protection for privacy to norms of specific contexts, demanding that information gathering and dissemination be appropriate to that context and obey the governing norms of distribution within it. . . . [T]his Article argues that public surveillance violates a right to privacy because it violates contextual integrity[.]").

[98] Nevertheless, the acceptance of new technology does not absolve society from the need for regulation and restraint. Too frequently, the rapid advancement of technology outpaces the ability of governments to respond, enabling private entities to exploit this gap and encroach upon individuals' rights. Generative AI serves as an example of this phenomenon, where companies, in the absence of clear legislation, have engaged in indiscriminate internet scraping, violating intellectual property and privacy rights.

[99] This Article proposes frameworks to guide legislative bodies in addressing the challenges posed by generative AI scraping while preserving the openness of the internet. Openness is no longer a mere feature; rather, it is a fundamental component of both the internet and modern democracy in the US. Consequently, it is imperative not to take this openness for granted. Striking a balance between technological innovation and safeguarding individual rights is crucial for a harmonious coexistence with the evolving digital landscape.