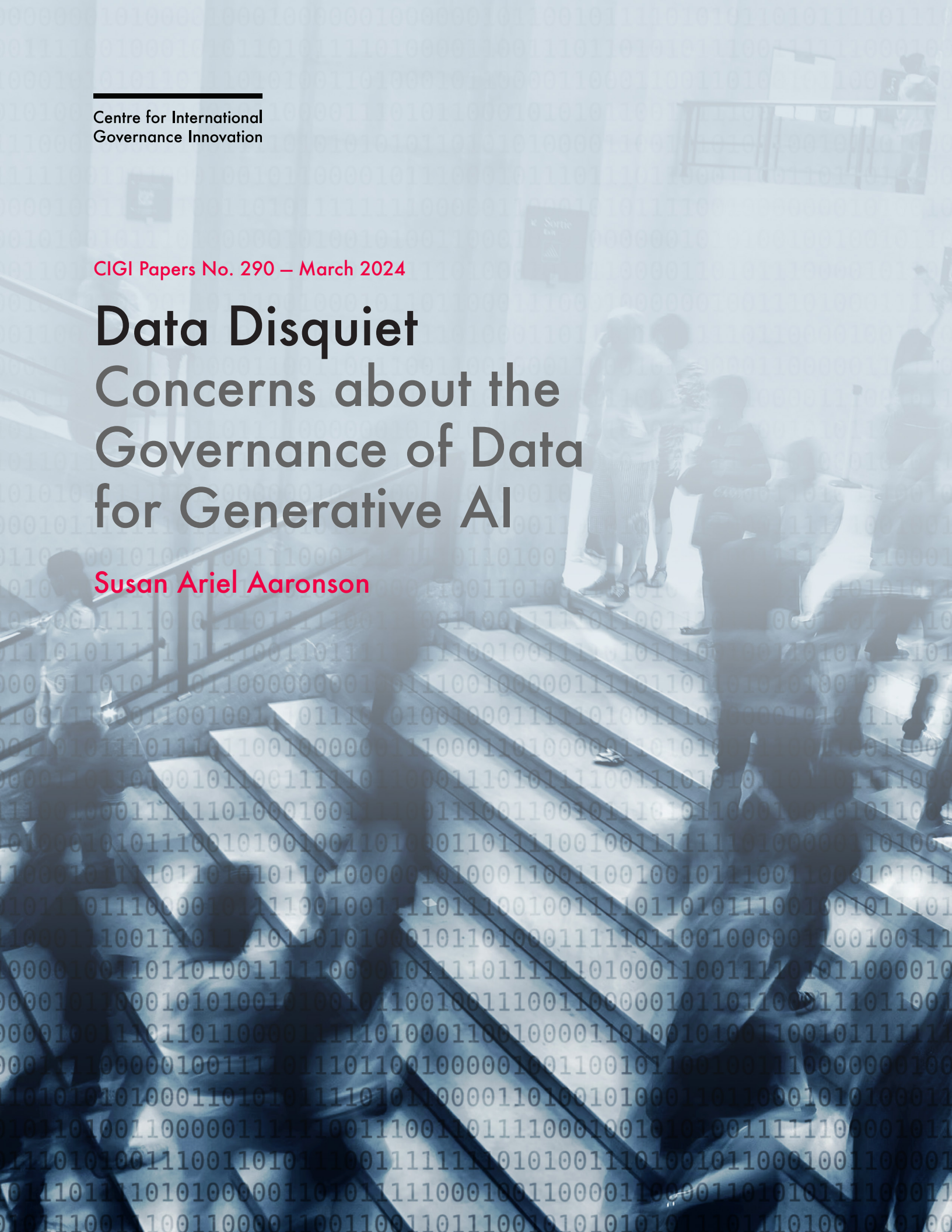

Centre for International
Governance Innovation

CIGI Papers No. 290 – March 2024

Data Disquiet
Concerns about the Governance of Data for Generative AI

Susan Ariel Aaronson

Centre for International
Governance Innovation

CIGI Papers No. 290 — March 2024

Data Disquiet
Concerns about the Governance of Data for Generative AI

Susan Ariel Aaronson

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

Credits

Managing Director of Digital Economy (until February 2024) Robert Fay
Director, Program Management Dianna English
Program Manager Jenny Thiel
Publications Editor Susan Bubak
Senior Publications Editor Jennifer Goyder
Graphic Designer Abhilasha Dewan

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Table of Contents

About the Author

Susan Ariel Aaronson is a CIGI senior fellow, research professor of international affairs at George Washington University (GWU) and co-principal investigator with the NSF-NIST Institute for Trustworthy AI in Law & Society, where she leads research on data and AI governance. She was also named GWU Public Interest Technology Scholar.

Susan directs the Digital Trade and Data Governance Hub at GWU. The Hub was founded in 2019 and educates policy makers, the press and the public about data governance and data-driven change through conferences, webinars, study groups, primers and scholarly papers. It is the only organization in the world that maps the governance of public, proprietary and personal data at the domestic and international levels. The Hub's research has been funded by foundations such as Ford and Minderoo.

Susan directs projects on defining AI protectionism; how governments may incentivize more accurate, complete and representative data sets; and how open-source AI builds trust. She regularly writes op-eds for *Barron's* and has been a commentor on economics for NPR's *Marketplace*, *All Things Considered* and *Morning Edition*, and for NBC, CNN, the BBC and PBS.

Previously, Susan was a guest scholar in economics at the Brookings Institution (1995–1999) and a research fellow at the World Trade Institute (2008–2012). Susan was also the Carvalho Fellow at the Government Accountability Project and held the Minerva Chair at the National War College. She has served on the business and human rights advisory board at Amnesty International and the advisory board of Human Rights under Pressure, a joint German and Israeli initiative on human rights.

In her spare time, Susan enjoys triathlons and ballet.

Acronyms and Abbreviations

AI	artificial intelligence
FTC	Federal Trade Commission
GDPR	General Data Protection Regulation
GPT	generative pre-trained transformer
IP	intellectual property
LLM	large language model
NIST	National Institute of Standards and Technology
NSF	National Science Foundation
OECD	Organisation for Economic Co-operation and Development
WTO	World Trade Organization

Executive Summary

The world's people increasingly rely on large language model (LLM) chatbots such as ChatGPT or Copilot to receive and organize information. But these chatbots often make mistakes or provide made-up or false information (hallucinations). They hallucinate because they are built on problematic data sets or incorrect assumptions made by the model, creating disquiet among users, developers and policy makers.

The author argues that policy makers have responded to this challenge in a piecemeal fashion. The paper[1] uses qualitative methods to examine these issues in several countries. While some policy makers are responsive to some concerns, these same policy makers have not developed a systemic approach — one that reflects the complexity of LLMs as well as the complicated nature and magnitude of the data that underpins these systems.

The paper begins by describing what the author means by a systemic approach, then turns to the history and economics of LLMs, which provide insights into why it is so hard to govern these LLMs. Next, the author discusses some of the challenges in data governance related to LLMs, and what some governments are doing to address these concerns. The author concludes that if policy makers want to effectively address the data underpinning LLMs, they need to incentivize greater transparency and accountability regarding data-set development.

Introduction: What Hath Generative Artificial Intelligence Wrought?

Generative artificial intelligence (AI) is a technology rife with challenges for policy makers. At times, generative AI chatbots make mistakes or invent facts. In February 2024, Air Canada learned this lesson. In 2022, a customer used Air Canada's chatbot to understand the company's bereavement flight policies. The customer booked a flight and took a screenshot of the advice provided by the company's chatbot: "If you need to travel immediately or have already travelled and would like to submit your ticket for a reduced bereavement rate, kindly do so within 90 days of the date your ticket was issued by completing our Ticket Refund Application form." The customer followed that advice, but the company refused his request for a lower rate. After the customer went to court, a judge required Air Canada to give a partial refund to the grieving passenger, arguing that the company was responsible for the chatbot's mistake (Belanger 2024).[2]

Liability is not the only problem; policy makers must find ways to incentivize accuracy, transparency and trust in these systems. This is why: growing numbers of people are turning to chatbots such as OpenAI's ChatGPT[3] and Google's Bard to find and create new forms of information. Yet because many of these systems are proprietary, their algorithms, models and data sources are not transparent. Outsiders cannot utilize scientific methods to reproduce the LLMs that underpin generative AI and, in so doing, build trust in these systems. Moreover, the world knows very little about the sources of that data (data provenance) and whether such data sets are accurate, complete and representative. Finally, only a few companies have the staff; computing power; computer and data science expertise; and the large data sets necessary to build, explain, expand and improve the models that underpin the technology. As a result, generative AI could be controlled by a few giant data

1 This material is based on work supported, in part, by the NIST-National Science Foundation (NSF) Institute for Trustworthy AI in Law and Society, which is supported by the National Science Foundation under award no. 2229885. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

2 *Moffatt v Air Canada*, 2024 BCCRT 149 (CanLII), online: <www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html>/.

3 GPT stands for "generative pre-trained transformer," which is a program that can write like a human.

companies that control the use and reuse of much of the world's data (Staff in the Bureau of Competition & Office of Technology 2023). To effectively address these challenges, policy makers must view both AI and the data that underpins it as a system.

Generative AI chatbots are an LLM application that uses language as both an input and output (hereafter LLM chatbots). The author only examines herein LLMs that can be applied to create conversational chatbots such as Bard. Such LLMs are designed to predict the most likely next word and to output text that will satisfy the goals of a human, whether by following an instruction or retrieving important information (Wolfe 2023c).

At first, ChatGPT and its like inspired awe because they could perform tasks that previously only humans could do, such as coding, translating languages or writing poetry (James 2023). Moreover, they seemed human-like as they interacted with users. But they also inspired lawsuits, bans and public concern (Southern 2023; The Fashion Law 2024). These LLM chatbots are fallible — they frequently communicate incomplete, outdated, inaccurate or distorted information, as well as lies and disinformation (Pelk 2016; Sirimanne 2023; O'Brien 2023; Thorbecke 2023). AI developers admit that they do not know yet how to fix this problem (hallucinations).[4] Researchers attribute such hallucinations to problems in the underlying data sets and assumptions made by the models (Dziri et al. 2022; Khan and Hanna 2023).

Some observers argue that, over time, reliance on such chatbots could undermine open science, reduce access to information, jeopardize shared facts about the world, reduce trust in institutions, and threaten the financial stability of credible information sources such as book publishers or scholarly journals.[5] Not surprisingly, the public is divided about reliance on these LLM chatbots (Thomson-DeVeaux and Yee 2023; Madiega 2023; Bowman 2023; Aaronson 2023).

LLMs are generally constructed from two main pools of data (pre-filtered data sets). The first pool is comprised of data sets created, collected or acquired by the model developers. This pool of data can be considered proprietary because it is owned and controlled by the LLM developer. It may include many different types of data from many different sources, as well as computer-generated (synthetic) data created to augment or replace real data to improve AI models, protect sensitive data and mitigate bias.[6]

The second pool consists of scraped data. When researchers scrape the Web, they use a bot to copy code off the internet, which they can then use for innovation, business or research purposes. Some of these can be open source, such as the Pile, an "open source language modelling data set that consists of 22 smaller, high-quality datasets combined."[7] But, in general, there is very little information about the data sets created from web scraping. *The Washington Post* analyzed one of Google's LLM data sets and reported that the top sites for that data set were: "patents.google.com No. 1, which contains text from patents issued around the world; wikipedia.org No. 2, the free online encyclopedia; and scribd.com No. 3, a subscription-only digital library" (Schaul, Chen and Tiku 2023; Congressional Research Service 2023).[8] Scraped data sets can also include data illegally obtained from data subjects or intellectual property (IP) holders without permission or informed consent, as well as data scraped from open-access websites such as Wikipedia and Reddit. Although these open-access sites have no paywall, LLM developers often utilize such data without direct consent, compensation or attribution.

This paper examines how policy makers in some countries responded to the rise of LLM chatbots as a means to receive and create information. As people started paying attention to how these LLMs are designed and developed, they became more aware of the data sets that underpin these models, leading to disquiet over how data is governed. Individuals, content creators, IP rights holders and data subjects provide much of the input for these data sets. In many countries, these same people provide taxpayer funds for research to improve these systems. Their personal and professional data fuels these AI systems. However, many of the

4 Even some of the chatbots' biggest boosters were honest about their flaws. Sam Altman (2023), CEO of OpenAI, tweeted on March 14, 2023, that GPT-4 "is more creative than previous models, it hallucinates significantly less, and it is less biased…[but] it still seems more impressive on first use than it does after you spend more time with it." Also see O'Brien (2023); Heikkilä (2023).

5 See Birhane et al. (2023); Whang (2023); Huang and Siddarth (2023); Fabre (2023); Knight (2023b); Belanger (2023).

6 See https://research.ibm.com/blog/what-is-synthetic-data.

7 See https://pile.eleuther.ai/.

8 The sites include GitHub, Kaggle (www.kaggle.com/) and Data.world (https://data.world/).

Box 1: How Do LLMs Work?

An LLM algorithm scans enormous volumes of text to learn which words and sentences frequently appear near one another and in what context. LLMs can be adapted to perform a wide range of tasks across different domains. Developers take and combine various data sets, then remove redundant, missing or low-quality data through a filtering process (Dermawan 2023). The data is then fed into machine-learning software known as a transformer, which is a type of neural network (Organisation for Economic Co-operation and Development [OECD] 2018; Knight 2023a). The LLM learns the patterns in that training data and eventually becomes proficient at predicting the letters and words that should follow a piece of text. In this way, these LLMs are less human-like than parrot-like (Bender et al. 2021; Nicholas and Bhatia 2023).

entities developing these systems provide little information about how they constructed, filtered and organized their underlying data sets (Khan and Hanna 2023; Huang and Siddarth 2023).[9]

The author argues that policy makers have responded to this challenge in a piecemeal fashion:

- They have focused on addressing data by type (such as making personal data protection understandable), but they have not thought systemically about the mix of data that underpins generative AI systems.
- They have not addressed the legality of web scraping internationally, given that the internet is a shared global resource (Surman 2016; Bhatia 2022). To do so effectively, policy makers need to address web scraping across borders, which in turn means they need to address the free flow of data — an issue currently governed by bilateral and regional trade agreements.
- They have not focused sufficiently on the importance of establishing data provenance and transparency as a means of ascertaining if the data sets underpinning LLMs are accurate, complete and representative.

To tell this story, the author focuses on four issues:

- how web scraping may affect individuals and firms that hold copyrights;
- how web scraping may affect individuals and groups who are supposed to be protected under privacy and personal data protection laws;
- how web scraping revealed the lack of protections for content creators and content providers on open-access websites; and
- how there are no clear and universal rules to ensure the accuracy, completeness and representativeness of the data sets underpinning LLM chatbots.

The author uses qualitative methods to examine these issues. The paper discusses only those governments that adopted specific steps (actions, policies, new regulations and more) to address web scraping, LLMs or generative AI. The author acknowledges that these examples do not comprise a representative sample of governments based on income, LLM expertise and geographic diversity. However, these examples do illuminate that while some policy makers are responsive to some concerns, these same policy makers have not developed a systemic approach — one that reflects the complexity of LLMs as well as the complicated nature and magnitude of the data that underpins these systems (see Box 1).

The paper begins by describing what the author means by a systemic approach, then turns to the history and economics of LLMs, which provides insights into why it is so hard to govern these LLMs. Next, the author discusses some of the challenges in data governance related to LLMs,

9 The author notes that researchers at these firms draft scholarly papers on their models but provide few specifics on the data sets. See, for example, Radford et al. (2018, 20n–23n).

Box 2: Key Words

Data provenance: Entails providing information on the origin of the data underlying a model and any changes or modifications the data set has undergone, and details supporting the confidence or validity of the data. The concept of provenance provides a chain of custody for data, which can help developers build and sustain trust in a data set.

Generative AI: Consists of AI models that emulate the structure and characteristics of input data to generate derived synthetic content.

Hallucinations: Incorrect or misleading results that AI models generate because they are built on incomplete, inaccurate or unrepresentative data sets and/or incorrect assumptions made by the model.

LLMs: Underpin generative AI to create natural language text. These models are trained on vast amounts of textual data scraped broadly from the internet or from specific focused data sets.

Model weight: Refers to a numerical parameter within an AI model that helps determine the model's outputs in response.

Synthetic data: Generated on a computer to augment or replace real data to improve AI models, protect sensitive data and mitigate bias.

Sources: www.nnlm.gov/guides/data-glossary/data-provenance; https://csrc.nist.gov/glossary/term/data_provenance; https://cloud.google.com/learn/what-is-artificial-intelligence; The White House (2023a); US General Services Administration (2023).

and what some governments are doing to address these concerns. The author then argues that if policy makers want to effectively address the data underpinning LLMs, they need to incentivize greater transparency and accountability regarding data-set development. Finally, the author suggests how policy makers might address this dilemna.

Why Is a Systemic Approach to the Data Underpinning LLM Chatbots Important?

As Box 2 illustrates, generative AI systems are complex — they are trained on large pools of various types of data. That data is also part of a complex system. Hence, policy makers should adopt an approach to data governance that reflects this complexity and can adapt as these systems evolve over time.

While there are many definitions of data governance (World Bank 2021),[10] herein the author uses that of the OECD: "Data governance refers to diverse arrangements, including technical, policy, regulatory or institutional provisions, that affect data and their cycle (creation, collection, storage, use, protection, access, sharing and deletion) across policy domains and organisational and national borders."[11] In so doing, policy makers must find ways to maximize the benefits of data access and sharing, while addressing related risks and challenges.[12]

But data is different from other goods and services produced by humans. Data is multidimensional. Researchers in the public and private sectors can reuse troves of data indefinitely without that data

10 See, for example, https://coe.gsa.gov/coe/ai-guide-for-government/data-governance-management/.

11 See www.oecd.org/digital/data-governance/.

12 Ibid.

losing its value. Individuals can use the same data to create new products or research complex problems. Moreover, data can simultaneously be a commercial asset and a public good. When raw data is organized, it becomes information — information that society uses to grow economies, hold governments to account, and solve wicked problems that transcend borders and generations. So, how societies govern various types of data has direct effects on democracy, economic progress and social stability (Aaronson 2018). Given these complexities, data governance requires adaptability — as information systems change, so too must data governance.

As the author will describe later, LLM chatbots rely on many different sources of data. Moreover, data and algorithm production, deployment and use are distributed among a wide range of actors from many different countries and sectors of society who together produce the system's outcomes and functionality. Thus, today, LLMs are not only part of the internet ecosystem, but are also a complex system of data. LLMs are at bottom a global product built on a global supply chain with numerous interdependencies among those who supply data, those who control data, and those who are data subjects or content creators (Cobbe, Veale and Singh 2023).

The US National Academy of Sciences notes that the only way to govern such complex systems is to create a governance ecosystem that cuts across sectors and disciplinary silos. Government officials should also consistently solicit and address the concerns of many stakeholders (Marchant and Wallach 2015). But, generally, these officials govern data by type (such as personal data, IP, public data and so forth) and not by use or purpose. Moreover, policy makers are in the early stages of linking data governance to AI governance.

The History and Economics of LLM Data Sets

AI language models are not new, and neither are LLM chatbots. The earliest LLMs were created in the early 1980s and were used as components in systems for automatic speech recognition, document classification and other tasks.[13] As with other approaches to AI, LLM developers experienced periods of boom and bust. However, recent advances in computing power and speed, combined with the ability to accumulate, analyze and store massive data sets, have made more advanced LLMs possible. Due to these advances, LLMs are transforming education, productivity and business (OECD 2023). Not surprisingly, policy makers in many countries want to ensure that they create an enabling environment that nourishes LLM innovation while protecting people from harm.

The earliest LLMs were generally open source (Wolfe 2023a). The Open Source Initiative defines "open source" as a development method for software that harnesses the power of distributed peer review and transparency of process. Open-source approaches can facilitate an environment of collaboration and idea sharing. When developers make their algorithms and underlying data sets (and other criteria) publicly available, many people can contribute to the development, improvement and customization of these models (ibid.).[14]

But open-source models have costs and benefits. Openness can lead to greater accountability, as analysts can gain a better understanding of how the LLM was developed, how it operates and how it can be improved. By being open, these LLMs may inspire greater dialogue and innovation (Castelvecchi 2023). But openness can be risky, as bad actors could insert incorrect code or malware that hopefully other researchers will correct and point out because it is open.

In contrast, developers of closed-source LLMs do not reveal specific details of their architecture, training data and algorithms to the public

13 See Zhou et al. (2023); Bender et al. (2021); https://onlim.com/en/the-history-of-chatbots/.

14 See https://opensource.org/about/.

(Bommasani, Liang and Lee 2023; Digital Public Goods Alliance and UNICEF 2023). Developers of these models may require others to obtain licences or subscriptions for their use. These LLM developers argue that their models will be more secure because they are protected and proprietary.

LLM developers provide various degrees of transparency — some providing more, others less (Barr 2023).[15] Hence, openness of LLMs is more like a continuum than a dialectic.

Open-source models are easier to govern because policy makers and the broader public can see and test the model and its underlying data sets (Digital Public Goods Alliance and UNICEF 2023; Aaronson 2023). Consequently, some governments are trying to encourage open-source LLMs. The governments of France[16] and Taiwan (Schneier 2024), for example, have tried to promote open-source LLMs to ensure that technological development and access to data remain open and global. They hope that their support for open source will reduce the concentration of LLM behemoths and reduce the entry costs for other competitors (Pai 2023; Stokel-Walker and Van Noorden 2023). In 2021, the French government gathered researchers from 60 countries and more than 250 institutions to create a very large multilingual neural network language model and a very large multilingual text data set, on a French supercomputer near Paris. BLOOM is open to everyone, but one must sign documentation that commits developers to not use the model for malicious or inappropriate ends, such as generating fake news (Gibney 2022).

Despite this momentum for open source, the producers of LLMs are, in general, a small number of extremely large data giants that are very concerned about their proprietary data — their algorithms, underlying data sets, model weights and so forth. Only some 20 firms possess the cloud infrastructure, computing power, access to capital and vast troves of data to develop and deploy tools to create LLMs (Staff in the Bureau of Competition & Office of Technology 2023). These firms are also concentrated in a few advanced developed countries — in North America, Asia and Europe. As a result, a few companies with expertise in generative AI could hold outsized influence over a significant swath of economic activity (Staff in the Bureau of Competition & Office of Technology 2023; Hacker, Engel and Mauer 2023; Khan 2023). These companies may not be motivated or encouraged to ensure that their data sets are broadly representative of the people and data of the world.

Moreover, many of the firms producing LLMs have, over time, become less forthcoming about their data. For example, the first paper published by OpenAI in 2018 describes the training data in general terms. It notes, "We use the BooksCorpus dataset for training the language model. It contains over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance" (Radford et al. 2018, 4–5).[17] The AI developers also used an alternative data set: the 1B Word Benchmark. OpenAI's most recent scholarly paper on GPT-4 was even less specific. It notes that the company used "both publicly available data (such as internet data) and data licensed from third-party providers.... Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" (OpenAI 2023, 2).

Meta is only slightly more specific. In its paper describing the first iteration of its model LLaMA 1, Meta notes, "Our training dataset is a mixture of several sources...that cover a diverse set of domains. For the most part, we reuse data sources that have been leveraged to train other LLMs, with the restriction of only using data that is publicly available, and compatible with open sourcing" (Touvron et al. 2023a, 2). Meta also states that 67 percent of its data set comes from the CommonCrawl; 15 percent from the C4 data set, a filtered data set; 4.5 percent each from GitHub and Wikipedia; and smaller amounts from other data sets in the public domain (ibid.). In its more recent model, LLaMA 2, Meta provides the model code, model weights, user guides, licences, acceptable use and model card but not a full description of the data set. The accompanying paper says that the model is trained on "a new mix of data from publicly available sources, which does not include data from Meta's products or services....We made an effort to remove data from certain sites known to contain a

15 They described it as open source, but it is not fully open. See Meta (2023); Touvron et al. (2023b).

16 In June, French President Emmanuel Macron announced new funding for an open "digital common" for French-made generative AI projects. See Chatterjee and Volpicelli (2023).

17 See Rastogi (2023).

high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations" (Touvron et al. 2023b, 4, 5). Moreover, the firm notes that during the supervised fine-tuning process, it set aside "millions of examples from third-party datasets and using fewer but higher-quality examples from our own vendor-based annotation efforts, our results notably improved" (ibid., 9). The authors did not describe the millions of examples that Meta kept or filtered out, nor did they describe the "higher-quality examples." So, despite being relatively open, Meta has also provided vague and incomplete detail about its data sets.

Clearly, LLMs require extremely large data sets of various types of data. So, the firms specializing in LLM chatbots have an incentive to get control over as much data as possible when innovation is data-driven (Martens 2018). As Iain M. Cockburn, Rebecca Henderson and Scott Stern (2018) noted, if there are increasing returns to scale or scope in data acquisition, it is possible that early or aggressive entrants into a particular application area may be able to create a substantial and long-lasting competitive advantage over potential rivals merely through the control over data. Over time, the companies with more and better data will be better able to improve the quality of algorithms through learning by doing. These companies will thus be well positioned to control ever more of the market for LLMs and their applications (Whang 2023; Hagiu and Wright 2023). Moreover, many of the most powerful models are only accessible via paid application programming interfaces[18] and trained using large amounts of proprietary data (OpenAI et al. 2023), thus limiting the research community from accessing or reproducing such models (Wolfe 2023b). For example, OpenAI's terms of service for its chatbot state that users cannot "attempt to or assist anyone to reverse engineer, decompile or discover the source code or underlying components of our Services, including our models, algorithms, or systems (except to the extent this restriction is prohibited by applicable law)."[19] If these companies continue to thwart outsiders' knowledge and testing of their models, it could have implications for scientific replicability and the basic human right of access to information (Cockburn, Henderson and Stern 2018; Aaronson 2023). But it could also incentivize developers to rethink how they obtain data, or to find ways to train LLMs on smaller or synthetic data sets (Whang 2023). However, because synthetic data sets are often proprietary, large developers of LLMs are unlikely to encourage data sharing or reuse of their synthetic data. Global society could be the big loser, as data sharing is important to economic, social and scientific progress.

LLM chatbots are becoming *where* individuals go to get and analyze information (Perri 2023; Stokel-Walker and Van Noorden 2022).[20] For example, ChatGPT was first released in November 2022. By March 2023, the chatbot had 170 million users, becoming one of the fastest-growing applications the world has ever seen (Tarnoff 2023; Duarte 2024).[21] Recognizing the technology's potential, other entities rushed out their own LLM chatbots, such as Facebook's LLaMA, Baidu's ERNIE, Anthropic's Claude and Dubai's Falcon (Grant and Weise 2023; Hacker, Engel and Mauer 2023).

Some of the data giants want to use these chatbots both to improve and, ultimately, replace browsers (which provide ranked links to sites) such as Bing or Google Chrome (Abbas 2023). Some have integrated chatbots with search engines to obtain more up-to-date information.[22] For example, Google combined its Gemini (formerly Bard) chatbot and various Google apps, making it easier to do two tasks simultaneously — for example, search for travel information and book flights (Pinsky 2023). But others have abandoned search engines for a more interactive approach. For example, users provide prompts to Perplexity AI, which in turn asks the user specific questions, so that it can then fetch the information it perceives that the user wants.[23]

LLM chatbots are also changing who creates and distributes information. For example, LLM chatbots can already create most types of written, image-based, video, audio and coded content. In the future, our news and culture may be machine generated (McKinsey 2023). LLM chatbots are

18 See https://docs.anthropic.com/claude/docs/guide-to-anthropics-prompt-engineering-resources.

19 See https://openai.com/policies/terms-of-use.

20 Statista has statistics on ChatGPT-related mobile app downloads worldwide between May and December 2023; see www.statista.com/statistics/1386342/chat-gpt-app-downloads/.

21 See https://openai.com/gpt-4.

22 See, for example, https://copilot.microsoft.com.

23 See https://blog.perplexity.ai/faq/what-is-copilot.

altering who provides information. They free individuals to concentrate on higher-value tasks. Moreover, they can help facilitate knowledge sharing and empower knowledge workers (Alavi and Westerman 2023). However, these LLMs may augment skills, but they could also be deskilling (Alexander 2023; boyd 2023). As a result, unionized workers are demanding and winning some protections from generative AI in new union contracts, such as those of the Screen Actors Guild and Screen Writers Guild (Niedzwiadek 2023).

Finally, these LLM chatbots are also having a major impact on where and how students receive and judge information. Educators can use LLM chatbots to create class outlines, generate ideas for classroom activities and update curricula. These chatbots can also provide more personalized learning and greater time and ability to meet specific student needs. According to Teach For America (2023), they may also unlock "the potential for greater student agency, creativity, and higher order thinking."

Despite the potential magnitude of these changes, governments have responded in an ad hoc manner. The next section describes their actions.

The Data Governance Challenges

How Web Scraping May Affect Individuals and Firms that Hold Copyright

On August 24, 2023, the Office of the Australian Information Commissioner and 11 of its international data protection and privacy counterparts released a joint statement on web scraping (data collected by a bot from a wide range of websites). The 12 signatories warned that "data protection authorities are seeing increasing incidents involving data scraping, particularly from social media and other websites that host publicly accessible data" (Office of the Australian Information Commissioner 2023). They stressed that operators of websites that host publicly accessible personal data have obligations to protect personal information on their platforms from unlawful data scraping (ibid.).

Researchers, governments and companies have scraped the Web for years. In 1993, Matthew Gray created the first web crawler, the World Wide Web Wanderer, to chart the Web's growth (Roth 2022).[24] Today, researchers rely on bots that search and scrape the Web to index web content, or gauge political sentiment to sustain and improve the internet (Web Scraper 2021; Nagel 2023).[25] AI developers may scrape the Web themselves, or rely on existing web scrapes to quickly create a large and diverse data set.[26] Web scraping is legal in most countries, although some types of web scraping may violate consumer protection, personal data protection or privacy laws.[27] However, web scraping can lead to unanticipated side effects. For example, developers who rely on scraped data may struggle to identify falsified or manipulated data in large data sets (United Nations Educational, Scientific and Cultural Organization 2023, 42). Some critics assert that by building their data sets with scraped material, including from sites open to all, these firms capture much of the value of the digital commons and gain ever greater control over the reuse of such data. Moreover, because their data sets may include inaccurate, false or incomplete information, these LLMs may pollute the shared digital and information commons — the collected open-access, open-source infrastructure and data underpinning the World Wide Web (Huang and Siddarth 2023; Jones and Steinhardt 2022). Mozilla recently published a study noting the dangers of relying on the Common Crawl for trustworthy AI. Author Stefan Baack noted that the crawl's mission does not align with the needs of trustworthy AI developers. He also pointed out that because so many important domains such as Facebook and *The New York Times* ban

24 See https://en.wikipedia.org/wiki/World_Wide_Web_Wanderer.

25 See www.geektime.com/the-history-of-web-scraping-and-what-the-future-holds/.

26 See https://huggingface.co/datasets/EleutherAI/pile.

27 For example, the *Computer Fraud and Abuse Act* (18 USC § 1030) imposes liability when a person "intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains…information from any protected computer." But some court cases have held that this prohibition does not apply to public websites — meaning that scraping publicly accessible data from the internet does not violate US law (Congressional Research Service 2023b). In contrast, Canadian courts have found violations of copyright and personal data protection laws. See Lifshitz (2019); Whittaker (2021).

the crawl from their pages, no one should view it as representative of "the Web" (Baack 2024).

Moreover, some studies show that web scraping could eventually blow up the utility of generative AI (Chiang 2023). Ilia Shumailov et al. (2023) found that using AI-generated text to train another AI invariably "causes irreversible defects." The authors note that over time, the original content distribution disappears, leading to the collapse of the model. Hence, AI developers have some incentives to find different ways to obtain a large sample of various types of data. But as of February 2024, many of these firms still rely on web-scraped data to underpin their LLMs (Mims 2024; Baack 2024).

Officials in some countries have tried to provide regulatory certainty to those who create data sets, including those who rely on web scraping. They recognize that researchers in the public, private and civil society sectors create data sets for a wide variety of reasons, and those creators deserve some form of legal protections (R. Morrison 2023; Huang and Siddarth 2023).

For example, the EU database directive establishes exclusive ownership rights for "databases," subject to some exceptions. Entities can gain a copyright for databases if that data set is original and constitutes the author's own intellectual creation. The directive also provides for another right to protection, as long as there has been "substantial investment in obtaining, verifying or presenting the contents" (Martens 2018, 17). Copyright holders can prevent others from conducting text and data mining, when doing so breaches their copyrights (Dermawan 2023).

EU law protects the collection of data sets, but it does not address its constituent elements (for example, the various types of data included). These elements may or may not be protected separately from any protection afforded to the database. Moreover, any software that is used in the making or operation of a database is specifically excluded from protection as a database. Even though the 2019 EU copyright directive provides an exception from copyright for text and data mining, this provision does not appear to have fully resolved the issue. Thus, some want the upcoming EU Artificial Intelligence Act (EU AI Act) to include language that clarifies if copyrighted content can be included in LLMs and the conditions under which royalties must be paid (Bania 2023; Marcus 2023; Margoni and Kretschmer 2022).

The UK government sought to exempt text and data mining from copyright protection. However, a committee in the UK Parliament warned in August 2023 that this approach risks reducing arts and cultural production to mere "inputs" in AI development, so the government is currently reconsidering the proposal (Culture, Media and Sport Committee 2023; Dickens 2023).

In 2021, Singapore created an exception in its copyright law for computational data analysis, which applies to text and data mining, data analytics and machine learning. The exception applies for both commercial and non-commercial databases, and policy makers anticipate that the exception will encourage basic and applied innovation (Norton Rose Fulbright 2021).

Meanwhile, Japan has revamped its approach to copyright to facilitate AI development and to encourage the development of databases based on copyrighted material. Its 2018 copyright law asserts that entities can conduct text and data mining without permission from the relevant rights holders "if the exploitation is aimed at neither enjoying nor causing another person to enjoy the work unless such exploitation unreasonably prejudices the interests of the copyright holder" (Dermawan 2023, 11). It is based on a presumption that there is no need for copyright protection if the exploitation of the work was not designed to prevent another person from enjoying a copyrighted work of art, movies or novels (Dermawan 2023; Ueno 2021). While this regulatory change was not specific to generative AI, Japanese government officials stated in May 2023 that they would not enforce some forms of copyright in the hopes of encouraging their use for generative AI.[28]

In contrast, US federal law says nothing explicit about web scraping as a means of creating a data set. US courts have upheld the right to scrape as a form of fair use, if the scraped data is not used to cause harm to society, a firm or an individual (Dilmegani 2024; Whittaker 2022).[29] Fair use is a legal doctrine that promotes freedom of expression by permitting the unlicensed use of copyright-

28 In Japan, copyrights are automatically generated when content is created, so not enforcing copyright made it easier to use older content. See Nishino (2022); Wan (2023); Technomancers.ai (2023).

29 See *Computer Fraud and Abuse Act*, *supra* note 26.

protected works in certain circumstances.[30] Despite the import of the generative AI sector, Congress has not yet taken steps to provide regulatory certainty regarding the creation of databases for AI. Databases are generally protected by copyright law as compilations. Under the Copyright Act, a compilation is defined as a "collection and assembling of preexisting materials or of data that are selected in such a way that the resulting work as a whole constitutes an original work of authorship."[31] The Copyright Act specifically states that the copyright in a compilation extends only to the compilation itself, and not to the underlying materials or data.[32]

LLM developers' reliance on web scraping has inspired both litigation and policy maker actions. As of November 30, 2023, Microsoft, OpenAI and Google are facing several lawsuits for misuse of copyrighted data in US courts (Gordon-Levitt 2023; De Vynck 2023). A November 2023 court filing argues that the defendants "have built a business valued into the tens of billions of dollars by taking the combined works of humanity without permission. Rather than pay for intellectual property, they pretend as if the laws protecting copyright do not exist. Yet the United States Constitution itself protects the fundamental principle that creators deserve compensation for their works."[33]

Meanwhile, public and private entities that have been crawled are taking steps to gain greater control over their data. News sites such as *The Guardian* and BBC News as well as public websites such as Reddit have moved to block web crawlers from accessing their sites to create LLM data sets (David 2023a, 2023b). To prevent further actions, the major AI chatbot firms have been trying to negotiate licensing deals in which they compensate the media (but not the journalists) for their stories. As an example, the Associated Press is exploring using LLMs as part of a partnership with OpenAI, which is paying to use part of the former's text archive to improve its AI systems (Di Stefano 2023). On December 27, 2023, *The New York Times* sued OpenAI, contending that the company violated its copyrighted articles and is using this information to directly compete with the *Times* and other trusted information sources (Grynbaum and Mac 2023).[34] In response to such cases, a senior Google official claimed that under the fair use provisions of the US approach to copyright, firms can use public information to create new beneficial uses. However, it is unclear if such web scraping is truly a case of fair use, or if *The New York Times* or other relatively open websites provide "public information" (Dean 2023).

Some companies are worried that their employees might leak proprietary data when they use generative AI chatbots (Campbell 2019; Sherry 2023; Rossi 2016; Appel, Neelbauer and Schweidel 2023; Bania 2023). In response, major AI developers such as Google and OpenAI provided instructions on how to block their web crawlers using "robots.txt." The robots.txt file tells search engine crawlers which URLs the crawler can access on a particular site.[35] The owners and designers of most websites want to be crawled by search engines because they want to be seen, which means they must rank highly in searches. But these sites do not want their data or analysis to be freely crawled and taken by OpenAI and other generative AI chatbots (Milmo 2023). In 2023, researchers at Originality.AI found that 306 of the top 1,000 sites on the Web blocked GPTBot, but only 85 blocked Google-Extended and 28 blocked anthropic-ai. The author concluded that companies are learning that they cannot keep up with crawling by AI firms; the bot cannot save them from the theft of IP (Pierce 2024).

On August 31, 2023, the US Copyright Office (which is part of the Library of Congress) announced it would study and seek public comment on the copyright law and policy issues raised by generative AI (US Copyright Office 2023).[36] In October 2023, the White House also stated in the Executive Order

30 See www.copyright.gov/fair-use/.

31 See *Copyrights*, 17 USC § 101; www.bitlaw.com/copyright/database.html.

32 See www.uspto.gov/learning-and-resources/ip-policy/database-protection-and-access-issues-recommendations; www.bitlaw.com/copyright/database.html.

33 *Julian Sancton, on behalf of himself and all others similarly situated, v OpenAI and Microsoft Corporation*, USDC, SDNY at 1.

34 *The New York Times Company v Microsoft Corporation and OpenAI*, (SD NY), online: <https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf>.

35 See https://developers.google.com/search/docs/crawling-indexing/robots/intro.

36 The call noted, "The NOI seeks factual information and views on a number of copyright issues raised by recent advances in generative AI. These issues include the use of copyrighted works to train AI models, the appropriate levels of transparency and disclosure with respect to the use of copyrighted works, the legal status of AI-generated outputs, and the appropriate treatment of AI-generated outputs that mimic personal attributes of human artists."

on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (Executive Order on AI) that it would ask the director of the US Copyright Office to issue recommendations to the president on potential executive actions relating to copyright and AI. The White House (2023a) also called on various departments to develop a plan to mitigate AI-based IP rights theft.

Meanwhile, on April 16, 2023, an independent regulatory agency, the US Federal Trade Commission (FTC), warned that "Generative AI tools that produce output based on copyrighted or otherwise protected material may, nonetheless, raise issues of consumer deception or unfairness. That's especially true if companies offering the tools don't come clean about the extent to which outputs may reflect the use of such material.... When offering a generative AI product, you may need to tell customers whether and the extent to which the training data includes copyrighted or otherwise protected material" (Atleson 2023).

The problem of inadequate governance at the intersection of scraping and copyright stems from the failure of LLM developers to document data provenance and to ensure that they have legal rights to use and reuse the data they collect. A widely cited 2021 paper, "Datasheets for Data Sets," recommended that every AI data set should be accompanied by a "data sheet" that documents its motivation, composition, collection process, recommended uses and so on (Gebru et al. 2021).

Policy makers are starting to recommend and, in some instances, require such documentation of data sets. For example, the National Institute of Standards and Technology's (NIST's) Artificial Intelligence Risk Management Framework suggests that designers and deployers build data sheets for data sets by documenting the AI system's data provenance, including sources, origins, transformations, augmentations, labels, dependencies, constraints and metadata (NIST 2023a). AI actors should also state the motivation for creating the data set and provide a means of ensuring that the data collected is adequate, relevant and not excessive in relation to the intended purpose (NIST 2023b). However, because the framework is a set of recommendations for best practice, firms could ignore it. In Canada, meanwhile, the proposed Artificial Intelligence and Data Act requires businesses conducting regulated activities to be held accountable for ensuring that employees implement measures to address risks associated with high-impact AI systems. But it is relatively vague regarding their responsibilities for data, as noted above. In a companion document to the act, the Canadian government says that firms must document the data and algorithms these firms utilize and assess and address potential bias in data sets. But it does not delineate how.[37] In the absence of clear legislation, the government worked with citizens to devise a voluntary code for generative AI. It states that "organizations will publish information on systems and ensure that AI systems and AI-generated content can be identified" (Government of Canada 2023). But the "how" was left vague. The EU AI Act (discussed later) also states that AI firms should provide documentation on the provenance of their data and requires such documentation for high-risk variants of AI (European Council 2024).

China has done more than other countries to link data governance to its governance of AI (O'Shaughnessy and Sheehan 2023). China finalized its generative AI regulations in August 2023, which apply to both domestic and overseas providers that use generative AI technology within China's territory. The rules apply to developers that provide generative AI to the public, but not to those that are not consumer facing. The regulations provide very specific directives for data governance. Generative AI service providers must:[38]

- → use data and foundation models from lawful (legitimate) sources;
- → not infringe others' legally owned IP;
- → obtain personal data with consent or under situations prescribed by the law or administrative measures;
- → take effective measures to increase the quality of training data, its truthfulness, accuracy, objectivity and diversity;
- → obtain consent from individuals whose personal information was processed;
- → take effective measures to improve the training data quality, authenticity, accuracy, objectivity and diversity;
- → ensure that LLM training activities are conducted in compliance with China's Cybersecurity Law,

37 See https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document.

38 This regulation is the latest addition to AI regulations in China after the Algorithm Provisions in 2021 and the Deep Synthesis Provisions in 2022.

Data Security Law and Personal Information Protection Law;

- not illegally retain input information and usage records, which can be used to identify a user; and
- not illegally provide users' input information and usage records to others (Gamvros, Yau and Chong 2023; Cooley LLP 2023).

However, thus far, no nation has adopted mandates that require LLMs to delineate data provenance.

How Web Scraping May Affect Individuals and Groups Who Are Supposed to Be Protected under Privacy and Personal Data Protection Laws

Most data protection laws around the world permit the collection and processing of personal data under specific conditions, such as when the individual's consent is given or as required by law.[39] Yet many people cannot meaningfully provide consent for the use of their data in LLMs. Many people are not aware that their data — including their tweets, Facebook posts, searches and other information created for one specific purpose — could be utilized for another purpose as part of the data set used to train an LLM (Romero 2023).[40] In the interest of transparency, a growing number of firms are admitting that they use personal data they collect to train variants of AI. For example, Google recently altered its privacy policies,[41] admitting it will use publicly available information to help train its AI models and build products and features such as Google Translate, Bard and Cloud AI capabilities (Germain 2023; Tiku and De Vynck 2023). However, most LLM developers do not inform data subjects that they use their personal data for several reasons. First, because they often rely on scraped data, they do not have direct access to users. Second, because they did not create these data sets or directly collect such data, it is difficult to find and notify individuals whose data they used (Argento 2023). Moreover, it would be extremely difficult for an individual or group of individuals to prove that an LLM used their data (R. Morrison 2023).

Policy makers in some countries have taken steps to protect their citizens' personal data. In March 2023, the Italian Data Protection Authority, the Garante, initially banned ChatGPT because Italian officials assumed that the company was violating Europe's General Data Protection Regulation (GDPR). The Garante listed measures that it said OpenAI must implement to have the suspension order lifted by the end of April — including adding age-gating to prevent minors from accessing the service and amending the legal basis claimed for processing local users' data. It lifted the ban after OpenAI announced a set of privacy controls (Lomas 2023a, 2023b). In June 2023, the French data protection body, the National Commission on Informatics and Liberty, developed an action plan focused on generative AI, LLMs and derived applications (especially chatbots). The action plan aims to:

- understand the functioning of AI systems and their impact on people;
- enable and guide the development of privacy-friendly AI;
- federate and support innovative players in the AI ecosystem in France and in Europe; and
- audit and control AI systems and protect people from harm.

But the plan said little about determining the provenance of the various types of data underpinning LLMs.[42]

These steps at the national level are not assuaging concerns that web scraping violates the GDPR. A Polish security researcher filed a complaint with the Polish data protection authority, alleging that ChatGPT's violation of privacy was systemic. The complaint accuses OpenAI of acting in an "untrustworthy, dishonest, and perhaps unconscientious manner" by failing to be able to comprehensively detail how it processed people's data (Lomas 2023c).

39 See https://ico.org.uk/for-organisations/sme-web-hub/your-beginner-s-guide-to-data-protection/.

40 For a real-world example, see Data Protection Commission of Ireland, *In the matter of the General Data Protection Regulation Data Protection Commission Reference: IN-21-4-2, In the matter of Meta Platforms Ireland Ltd. (Formerly Facebook Ireland Ltd.), Decision of the Data Protection Commission made pursuant to Section 111 of the Data Protection Act 2018 and Article 60 of the General Data Protection Regulation*, s G.5 at 94; also see Future of Privacy Forum (2018).

41 See https://policies.google.com/privacy#whycollect.

42 See www.cnil.fr/en/artificial-intelligence-action-plan-cnil.

Many nations are seeking public input on how to address this problem. For example, in April 2023, the US Department of Health and Human Services sought public comment on whether it should allow patients access to electronic health records and, in particular, the personally identifiable information that firms utilize for predictive modelling, such as those designed to identify future cancer patients.[43] Singapore's Personal Data Protection Commission, meanwhile, initiated a public consultation on proposed guidelines concerning the use of personal data in AI recommendation and decision systems. The guidelines seek to clarify the application of the 2012 Personal Data Protection Act to organizations using personal data in the development and deployment of AI systems.[44]

Some nations are probing the business practices of companies creating LLMs. The FTC is investigating whether OpenAI offered or made available products or services "incorporating, using, or relying on Large Language Models engaged in unfair or deceptive privacy or data security practices or engaged in unfair or deceptive privacy or data security practices relating to risks of harms to consumers, including reputational harm," in violation of US laws (Zakrzewski 2023).[45] US President Joe Biden also decided to use his bully pulpit, getting public commitments from the seven largest developers[46] of generative AI to "commit to publicly reporting their AI systems' capabilities, limitations, and areas of appropriate and inappropriate use. This report will cover both security risks and societal risks, such as the effects on fairness and bias" (The White House 2023c). The AI giants also agreed to develop robust mechanisms, including provenance and/or watermarking systems for audio or visual content created by any of their publicly available systems introduced after the watermarking system is developed. However, it is too early to tell if these commitments will include public reporting on how these firms collected, reviewed and utilized data for their LLMs along the lines of the NIST's risk management framework (The White House 2023a). The UK Communications and Digital Committee of the House of Lords is examining "what needs to happen over the next 1–3 years to ensure the UK can respond to the opportunities and risks posed by large language models. This will include evaluating the work of Government and regulators, examining how well this addresses current and future technological capabilities, and reviewing the implications of approaches taken elsewhere in the world."[47]

As noted in the previous section, China has adopted very clear rules regarding the use of personal data for AI. Some analysts believe China's requirements are simultaneously too vague and onerous and will require further clarification (Arcesati and Brussee 2023). Others argue that the requirements are too demanding and impractical (Toner et al. 2023). Nonetheless, as Matt Sheehan of the Carnegie Endowment for International Peace noted, "Governments around the world…can draw lessons from China's experience. A vertical and iterative approach to regulation requires constant tending and updating. But by accumulating experience and creating reusable regulatory tools, that process can be faster and more sophisticated" (ibid.).

While governments are acting at the national level (Tene 2023), policy makers globally have not responded to concerns about web scraping by providing international certainty. When AI developers scrape the Web or rely on previous web scraping, they are taking data from many countries. Some of that data may flow from one country to the country where that data is used to train the model.

Some bilateral and regional trade agreements have binding rules governing cross-border data flows. More than 90 nations are working at the World Trade Organization (WTO) to set rules governing such data flows (Aaronson and Struett 2020). Such rules would not clarify if web scraping per se is legal among entities in different nations, but they would delineate when nations can breach the rules to prevent cross-border data flows (for example, to protect privacy). A nation could argue that its citizens' personal data is inadequately protected and possibly challenge such practices. However, these negotiations do not discuss web scraping, generative AI, or ways to ensure that data sets are as accurate, complete and representative as possible. In the author's view, the WTO may not be

43 *Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing*, 88 Fed Reg 23746 (2023).

44 See www.pdpc.gov.sg/Guidelines-and-Consultation/2023/07/Public-Consultation-for-the-Proposed-Advisory-Guidelines-on-Use-of-Personal-Data-in-AI-Recommendation-and-Decision-Systems.

45 See copy of the FTC's order at www.washingtonpost.com/documents/67a7081c-c770-4f05-a39e-9d02117e50e8.pdf.

46 Amazon, Anthropic, Google, Inflection, Meta, Microsoft and OpenAI.

47 See https://committees.parliament.uk/call-for-evidence/3183/.

the best venue to discuss these topics, yet it is the only international organization that has a rules-based system addressing data (Aaronson 2018). In the future, policy makers will need to find common ground on these topics with their international counterparts.

How Web Scraping Revealed the Lack of Protections for Content Creators and Content Providers on Open-Access Websites

Much of the data underpinning today's LLMs comes from widely used open-access platforms and websites such as Wikipedia, X (formerly Twitter), Facebook, Stack Overflow[48] and Reddit. These sites are open to all who sign up to use them, and these users provide comments, conversations, real-time reactions and other information for free (Schaul, Chen and Tiku 2023).

However, many open-access websites delineate in their terms of service that outsiders should not scrape their sites. Facebook provides a good example (although it does allow researchers access to some of its data) (Octoparse 2022). Clearly, individuals ignore and frequently breach these terms of service (Schaul, Chen and Tiku 2023).

After ChatGPT and other chatbots gained widespread use, some of the managers of these sites recognized that they needed to think differently about their data and its value to others. Reddit provides a good example.

In June 2023, Reddit's management decided to start charging third-party developers for access to its data. Company officials made that decision because they wanted to be compensated when others (whether researchers or other businesses) scrape Reddit's webpages to create new analysis or services such as LLM chatbots (Goswami 2023). On June 12, the moderators of thousands of Reddit forums, called "subreddits," collectively began to protest this decision, which cut off their access to applications they used to perform their (unpaid) duties. Many of the moderators opposed Reddit's decision to begin charging for access to the site's data. They also felt that management was ignoring their unappreciated and unpaid contributions.[49]

The moderators at Reddit were not alone in their concern that their contributions to Reddit were undervalued and ignored. Contributors to Wikipedia argued that these chatbots were cannibalizing their site (Gertner 2023). Elon Musk, CEO of X, announced he was going to limit how many tweets users can view daily. But he pulled back due to user protests (Nolan 2023; Arcesati and Brussee 2023). Stack Overflow's CEO Prashanth Chandrasekar explained that "allowing AI models to train on the data developers have created over the years, but not sharing the data and learnings from those models with the public in return, would lead to a tragedy of the commons....Unless we all continue contributing knowledge back to a shared, public platform, we risk a world in which knowledge is centralized inside the black box of AI models that require users to pay in order to access their services" (Diaz 2023).

The web scraping of open-access sites raised several issues: Should LLM developers compensate these sites for the data they scrape? Should content creators and moderators on these sites be compensated too and, if so, how? And, finally, should this data be controlled by a few big companies that reap the benefits of shared efforts to expand knowledge? The author could find no country thus far addressing the first two issues. However, policy makers in some countries are investigating whether a few companies could control and define information through their LLMs. The FTC announced it was investigating OpenAI's use of data (Zakrewski 2023). Competition authorities in Sweden and several other countries are investigating whether these AI companies should control the reuse of that data and whether they control too much of the world's data through network effects (AI Now Institute 2023; msmash 2023; Ikeda 2023; Pandey 2023; Holmes 2020). The FTC is also investigatting whether it is legal for companies such as Reddit to sell user-generated data to companies, which then use such data to train AI. Such actions raise significant privacy, copyright and fairness concerns (Dave 2024).

48 Stack Overflow is a programming forum that offers a collaborative environment to its users, who are mostly developers. It is a popular place for programmers to ask about coding problems and programming language and works as a learning resource for its more than 20 million users.

49 Reddit is a US-based news aggregation, content rating and discussion website. Registered users submit content to the site such as links, text posts, images and videos, which are then voted up or down by other members. Reddit is manufactured by its members, who do tasks such as moderate content; see www.redditinc.com/policies/user-agreement; www.redditinc.com/. On the protest, see S. Morrison (2023).

Policy makers' failure to address these issues could have significant effects on humankind. Over time, these content creators could hoard their data or not participate in open websites. If content creators decide to do so, this could result in less access to information as well as less data for everyone to use.

Thus far, there is little evidence that policy makers are worried about this possibility, which has implications for access to information, a basic human right (United Nations Development Programme 2004). Nor do they yet seem worried about whether it is appropriate for LLMs to explain crucial global information such as scientific research. As noted above, LLMs generate predictions of the "statistically likely continuations of word sequences." They lack capacity for scientific reasoning and cannot capture the uncertainties, limitations and nuances of research that are obvious to the human scientist. These LLMs also generate non-existent and false content. Scientists may become reluctant to share their data for peer review and replication if they feel it will be misrepresented. Policy makers should weigh these potential scenarios (Bender et al. 2021; Birhane et al. 2023).

How the Debate Over Open- and Closed-Source LLMs Revealed the Lack of Clear and Universal Rules to Ensure the Quality and Validity of Data Sets

The NIST has warned that many LLMs depend on large-scale data sets, which can lead to data quality and validity concerns: "The difficulty of finding the 'right' data may lead AI actors to select datasets based more on accessibility and availability than on suitability....Such decisions could contribute to an environment where the data used in processes is not fully representative of the populations or phenomena that are being modeled, introducing downstream risks" — in short, problems of quality and validity (NIST 2023b, 80).

By relying on data scraped from the web, LLMs are likely producing incomplete and inaccurate outputs. Scraped data, in essence, provides a snapshot of the internet in time, but it is likely an incomplete, incorrect, outdated picture (Kim et al. 2003; Rossi 2016; Riley 2023). Unfortunately, by relying on web scraping plus proprietary data as their data foundation, LLMs may be relying on a model that, by definition, produces biased and incomplete data.

One can only scrape the World Wide Web that exists, not the Web we wish to see. The Web is dominated by content from and about people who are online, and those people live mainly in Europe, North America and Asia. Throughout Europe, the Commonwealth of Independent States and the Americas, between 80 and 90 percent of the population uses the internet, approaching universal use (defined for practical purposes as an internet penetration rate of at least 95 percent). Approximately two-thirds of the population in the Arab states and Asia-Pacific countries (70 percent and 64 percent, respectively) use the internet, in line with the global average, while the average for Africa is just 40 percent of the population.[50] However, in 2022, the International Telecommunication Union reported that 34 percent of the world's population has never used the internet. Most of these people live in rural areas in the developing world. These people are not visible in most web scraping.[51]

The author is not aware of efforts in developing countries to ensure that their contributions to knowledge and culture are included in web searches.[52] Officials in African nations have expressed concerns that their workers are involved in data labelling — and, in that way, they help train LLMs. African policy makers are also concerned about their citizens' data being used without informed consent (Kannan 2022; Birhane 2020). But these officials have not yet made an issue of incomplete and inaccurate data from web scraping.

One option is to require information on both data provenance and data accuracy. The EU AI Act was approved March 13, 2024. The act delineates how the European Union will regulate AI risk, particularly that of high-risk foundation models, and it describes how AI developers should build more accurate and trustworthy

50 See www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-internet-use/.

51 The author is grateful to Angie Raymond, Indiana University, for making this point. See www.itu.int/itu-d/reports/statistics/2022/11/24/ff22-internet-use-in-urban-and-rural-areas/.

52 The African Union has unveiled the Artificial Intelligence Continental Strategy for Africa, which is intended to facilitate the participation of stakeholders, initiate capacity-building efforts, and fortify regulatory frameworks for AI technology and data management.

data sets.[53] The law highlights high-impact foundation models, a particular type of AI:

> High-impact capabilities in general purpose AI models means capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models....According to the state of the art at the time of entry into force of this Regulation, the cumulative amount of compute used for the training of the general purpose AI model measured in floating point operations (FLOPs) is one of the relevant approximations for model capabilities. The amount of compute used for training cumulates the compute used across the activities and methods that are intended to enhance the capabilities of the model prior to deployment, such as pre-training, synthetic data generation and fine-tuning. Therefore, an initial threshold of FLOPs should be set, which, if met by a general-purpose AI model, leads to a presumption that the model is a general-purpose AI model with systemic risks. This threshold should be adjusted over time to reflect technological and industrial changes, such as algorithmic improvements or increased hardware efficiency, and should be supplemented with benchmarks and indicators for model capability.[54]

Firms providing high-impact foundation models are required to enable traceability of their systems, to verify compliance and develop technical documentation of how they built their models. Developers of these systems must be transparent about their design before these systems are placed on the market. Outsiders should be able to oversee their functioning and ensure they are used as intended.[55]

Meanwhile, Canada's Directive on Automated Decision-Making governs a wide range of AI systems procured by the Canadian government. The directive requires that the data be relevant, accurate, up to date and traceable; protected and accessed appropriately; and lawfully collected, used, retained and disposed.[56] However, the directive says nothing about data provenance or transparency.

Conclusion

In his executive order on AI, President Biden stressed that "AI reflects the principles of the people who build it, the people who use it, and the data upon which it is built" (The White House 2023a). However, the world's people need to know more about how data is used to create LLM chatbots. They will also need to govern data differently if they want to ensure that current and future AI systems are accurate, complete and representative of the world, as well as robust, equitable and safe (Bender et al. 2021, 2022; Bommasani, Liang and Lee 2023; Bommasani et al. 2023).

This paper examined how policy makers in some countries responded to the rise of LLM chatbots as a venue to receive and create information. These LLM chatbots are becoming a key venue where people obtain and create information.

As people started to pay attention to the design and development of LLMs, they became more aware of enforcement problems and governance gaps, leading to disquiet over how data is governed. Policy makers have responded to this challenge in a piecemeal fashion:

- They have focused on addressing data by type (such as making personal data protection

53 *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))*, online: <www.europarl.europa.eu/doceo/document/TA-9-2024-03-13_EN.html#sdocta2>. One can download the text dated February 2 (European Council 2024). Recital 44 of the law notes that "datasets for training, validation and testing, including the labels, should be relevant, sufficiently representative, and to the best extent possible free of errors and complete in view of the intended purpose of the system." It should also include "transparency about the original purpose of the data collection[.] The datasets should also have the appropriate statistical properties, including as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used, with specific attention to the mitigation of possible biases in the datasets, that are likely to affect the health and safety of persons, negatively impact fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations ('feedback loops')."

54 *AI Act, supra* note 52, recital 60n.

55 *Ibid*, arts 6–12.

56 See www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592.

understandable), but they have not thought systemically about the mix of data that underpins generative AI systems, or about whether data and information governance needs to change in light of this new venue to receive and create information.

- They have not addressed the legality of web scraping internationally, given that the internet is a shared global resource (Surman 2016; Bhatia 2022). To do so effectively, policy makers need to address web scraping as an international issue because when one scrapes, one is not only taking data from multiple sites but also from multiple countries. This fact is also an opportunity for developing countries to push for greater influence in the discussions about data flows at the WTO. Yet developing countries are torn — many want their data to be sovereign and under their control (Aaronson and Struett 2020).

- They have not focused sufficiently on data provenance and transparency. If users, policy makers and others could have greater insights into the data LLM developers use, we could limit hallucinations and improve these models.

LLM data sets today are large, diverse and multinational, and are thus difficult to govern (Cobbe, Veale and Singh 2023). But the world must do more to govern these LLMs for two reasons: first, because many of these systems are black boxes, whose developers provide little information about how they work; and second, because more and more people rely on LLMs for information.

Some analysts may hope that LLM developers come up with technical solutions such as synthetic data sets. But synthetic data sets are proprietary, so they are also opaque and unlikely to build trust. Policy makers will need to devise rules requiring that LLM developers hire outside auditors to vet synthetic data sets for accuracy, completeness and representativeness.

Policy makers could incentivize transparency and a more systemic approach by recognizing the complexity of these data sets and the need to go beyond data governance by type of data toward data governance by objective. Policy makers should aim to ensure that the data sets that underpin LLM chatbots are not only accurate, complete and representative but also transparent and accountable.

There are no easy policy solutions to improving these data sets. In December 2023, several members of Congress introduced the AI Foundation Model Transparency Act, which would direct the FTC, in consultation with the NIST and the Office of Science and Technology Policy, to set standards for what information high-impact foundation models must provide to the FTC and what information they must make available to the public. Information identified for increased transparency would include training data used, how the model is trained and whether user data is collected in inference (Beyer 2023). Policy makers might also consider enacting corporate governance rules based on the argument that how firms handle the data they acquire, collect, store and analyze is material to the health of the firm. Firms would be required to report quarterly on the data they acquire, collect, store and analyze and how they use it. In so doing, they would be acknowledging that the quality of their data is an important component of the quality of their LLMs. AI developers would also be required to have outsiders audit their data sets and LLMs. The developers would be required to provide outside auditors with information on the provenance of their data and how they tested for accuracy, validity and completeness as they filtered and then utilized data. Outside auditors would then verify that these firms provided complete information. Although corporate governance rules could change the culture of AI developers, some firms developing AI are government entities, privately held firms or public benefit companies, which are not covered by corporate governance rules.

Policy makers must also act internationally. So far, they have not gotten beyond the planning process. For example, in the October 2023 executive order on AI, President Biden called on the Secretary of Commerce to "to advance responsible global technical standards for AI development and use outside of military and intelligence areas....In particular, the Secretary of Commerce shall...establish a plan for global engagement on promoting and developing AI standards, with lines of effort that may include... best practices regarding data capture, processing, protection, privacy, confidentiality, handling, and analysis" (The White House 2023a).

Finally, people continue to use LLM chatbots despite inaccuracies, incomplete data, bias and hallucinations. If we want these LLM chatbots to protect personal data, content creators and

IP rights holders, users, developers and policy makers should favour LLM chatbots such as Bloom and OLMo that provide greater transparency into their underlying data.[57] If we are going to rely on chatbots to provide information about our world, we have to demand better data sets and more transparency in LLM design and development.

Works Cited

Aaronson, Susan Ariel. 2018. *Data Is Different: Why the World Needs a New Approach to Governing Cross-Border Data Flows*. CIGI Paper No. 197. Waterloo, ON: CIGI. www.cigionline.org/publications/data-different-why-world-needs-new-approach-governing-cross-border-data-flows/.

———. 2023. "How to Regulate AI? Start With the Data." *Barron's*, June 15. www.barrons.com/articles/ai-data-regulation-bfded1d4.

Aaronson, Susan and Thomas Struett. 2020. *Data Is Divisive: A History of Public Communications on E-commerce, 1998–2020*. CIGI Paper No. 247. Waterloo, ON: CIGI. www.cigionline.org/publications/data-divisive-history-public-communications-e-commerce-1998-2020/.

Abbas, Assad. 2023. "The Future of Search Engines in a World of AI and LLMs." Techopedia, October 5. www.techopedia.com/the-future-of-search-engines-in-a-world-of-ai-and-llms.

AI Now Institute. 2023. "ChatGPT And More: Large Scale AI Models Entrench Big Tech Power." AI Now Institute, April 23. https://ainowinstitute.org/publication/large-scale-ai-models.

Alavi, Maryam and George Westerman. 2023. "How Generative AI Will Transform Knowledge Work." *Harvard Business Review*, November 7. https://hbr.org/2023/11/how-generative-ai-will-transform-knowledge-work.

Alexander, Lauren. 2023. "Omidyar Network at Code 2023: How generative AI will reshape the workforce and economy." Omidyar Network, September 29. https://omidyar.com/code-2023/.

Altman, Sam. 2023. "it is more creative than previous models, it hallucinates significantly less, and it is less biased... [but] it still seems more impressive on first use than it does after you spend more time with it." (Twitter thread). Twitter, March 14, 1:02 p.m. https://twitter.com/sama/status/1635687854784540672?lang=en.

Appel, Gil, Juliana Neelbauer and David A. Schweidel. 2023. "Generative AI Has an Intellectual Property Problem." *Harvard Business Review*, April 7. https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem.

Arcesati, Rebecca and Vincent Brussee. 2023. "China's Censors Back Down on Generative AI." The Diplomat, August 7. https://thediplomat.com/2023/08/chinas-censors-back-down-on-generative-ai.

Argento, Zoe. 2023. "Data protection issues for employers to consider when using generative AI." International Association of Privacy Professionals, August 9. https://iapp.org/news/a/data-protection-issues-for-employers-to-consider-when-using-generative-ai.

Atleson, Michael. 2023. "Can't lose what you never had: Claims about digital ownership and creation in the age of generative AI." *FTC Business Blog*, August 16. www.ftc.gov/business-guidance/blog/2023/08/cant-lose-what-you-never-had-claims-about-digital-ownership-creation-age-generative-ai.

Baack, Stefan. 2024. "Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI." Mozilla Insights. February. https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/.

Bania, Konstantina. 2023. "Generative AI and the media sector: Preliminary thoughts on a legal and policy agenda." *The Platform Law Blog*, June 14. https://theplatformlaw.blog/2023/06/14/generative-ai-and-the-media-sector-preliminary-thoughts-on-a-legal-and-policy-agenda/.

Barr, Alistair. 2023. "Llama copyright drama: Meta stops disclosing what data it uses to train the company's giant AI models." Business Insider, July 18. www.businessinsider.com/meta-llama-2-data-train-ai-models-2023-7.

Belanger, Ashley. 2023. "Report: Potential NYT lawsuit could force OpenAI to wipe ChatGPT and start over." Ars Technica, August 17. https://arstechnica.com/tech-policy/2023/08/report-potential-nyt-lawsuit-could-force-openai-to-wipe-chatgpt-and-start-over.

———. 2024. "Air Canada must honor refund policy invented by airline's chatbot." Ars Technica, February 16. https://arstechnica.com/tech-policy/2024/02/air-canada-must-honor-refund-policy-invented-by-airlines-chatbot/.

57 See https://github.com/eugeneyan/open-llms. On BLOOM, see https://huggingface.co/bigscience/bloom; on OLMo, see https://allenai.org/olmo.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March): 610–23. https://doi.org/10.1145/3442188.3445922.

Beyer, Don. 2023. "Beyer, Eshoo Introduce Landmark AI Regulation Bill." Press release, December 22. https://beyer.house.gov/news/documentsingle.aspx?DocumentID=6052.

Bhatia, Karan. 2022. "Coming together to protect the global internet." *The Keyword* (blog), April 28. https://blog.google/outreach-initiatives/public-policy/coming-together-to-protect-the-global-internet.

Birhane, Abeba. 2020. "Algorithmic Colonization of Africa." *SCRIPTed* 17 (2). https://script-ed.org/article/algorithmic-colonization-of-africa/.

Birhane, Abeba, Atoosa Kasirzadeh, David Leslie and Sandra Wachter. 2023. "Science in the age of large language models." *Nature Reviews Physics* 5 (May): 277–80. https://doi.org/10.1038/s42254-023-00581-4.

Bommassani, Rishi, Percy Liang and Tony Lee. 2023. "Language Models are Changing AI: The Need for Holistic Evaluation." Center for Research on Foundation Models. https://crfm.stanford.edu/2022/11/17/helm.html.

Bommasani, Rishi, Sayash Kapoor, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Daniel Zhang, Marietje Schaake et al. 2023. "Considerations for Governing Open Foundation Models." Stanford University Human-Centered Artificial Intelligence Policy & Society Issue Brief. December. https://hai.stanford.edu/issue-brief-considerations-governing-open-foundation-models.

Bowman, Samuel R. 2023. "Eight Things to Know about Large Language Models." arXiv, April 2. https://arxiv.org/abs/2304.00612.

boyd, danah. 2023. "Deskilling on the Job." Medium, April 21. https://zephoria.medium.com/deskilling-on-the-job-bbd71a74a435.

Campbell, Fiona. 2019. "Data Scraping – Considering the Privacy Issues." *Fieldfisher* (blog), August 27. www.fieldfisher.com/en/services/privacy-security-and-information/privacy-security-and-information-law-blog/data-scraping-considering-the-privacy-issues.

Castelvecchi, Davide. 2023. "Open-source AI chatbots are booming — what does this mean for researchers?" *Nature*, June 20. www.nature.com/articles/d41586-023-01970-6.

Chatterjee, Mohar and Gian Volpicelli. 2023. "France bets big on open-source AI." Politico, August 4. www.politico.eu/article/open-source-artificial-intelligence-france-bets-big/.

Chiang, Ted. 2023. "ChatGPT Is a Blurry JPEG of the Web." *The New Yorker*, February 9. www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

Ciuriak, Dan. 2018. *The Economics of Data: Implications for the Data-driven Economy*. Waterloo, ON: CIGI. www.cigionline.org/articles/economics-data-implications-data-driven-economy/.

Cobbe, Jennifer, Michael Veale and Jatinder Singh. 2023. "Understanding accountability in algorithmic supply chains." *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (June): 1186–97. https://doi.org/10.1145/3593013.3594073.

Cockburn, Iain M., Rebecca Henderson and Scott Stern. 2018. "The Impact of Artificial Intelligence on Innovation." National Bureau of Economic Research Working Paper No. 24449. March. www.nber.org/papers/w24449.

Cooley LLP. 2023. "China Issues Measures on Generative Artificial Intelligence Services." JDSupra, August 7. www.jdsupra.com/legalnews/china-issues-measures-on-generative-3929442/.

Culture, Media and Sport Committee. 2023. "Abandon artificial intelligence copyright exemption to protect UK creative industries, MPs say." Culture, Media and Sport Committee, UK Parliament, August 30. https://committees.parliament.uk/committee/378/culture-media-and-sport-committee/news/197222/abandon-artificial-intelligence-copyright-exemption-to-protect-uk-creative-industries-mps-say/.

Congressional Research Service. 2023. "Generative Artificial Intelligence and Data Privacy: A Primer." Report No. R47569. May 23. https://crsreports.congress.gov/product/pdf/R/R47569.

Dave, Paresh. 2024. "Reddit's Sale of User Data for AI Training Draws FTC Inquiry." *Wired*, March 15. www.wired.com/story/reddits-sale-user-data-ai-training-draws-ftc-investigation/.

David, Emilia. 2023a. "The BBC is blocking OpenAI data scraping but is open to AI-powered journalism." The Verge, October 6. www.theverge.com/2023/10/6/23906645/bbc-generative-ai-news-openai.

———. 2023b. "CNN, Reuters, Australia's ABC, and other news organizations block OpenAI's web crawler." The Verge, August 25. www.theverge.com/2023/8/25/23845613/cnn-reuters-australias-abc-and-other-news-organizations-block-openais-web-crawler.

De Vynck, Gerrit. 2023. "ChatGPT maker Open AI faces a lawsuit over how it used people's data." *The Washington Post*, June 28. www.washingtonpost.com/technology/2023/06/28/openai-chatgpt-lawsuit-class-action/.

Dean, Grace. 2023. "A lawsuit claims Google has been 'secretly stealing everything ever created and shared on the internet by hundreds of millions of Americans' to train its AI." Business Insider, July 12. www.businessinsider.com/google-alphabet-bard-generative-ai-secretly-stealing-online-data-lawsuit-2023-7.

Dermawan, Artha. 2023. "Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese 'nonenjoyment' purposes?" *Journal of World Intellectual Property*, 1–25. https://doi.org/10.1111/jwip.12285.

Di Stefano, Mark. 2023. "News Corp in talks with AI firm about compensation." Financial Review, March 8. www.afr.com/companies/media-and-marketing/news-corp-in-talks-with-ai-firm-about-compensation-20230308-p5cqcp.

Diaz, Maria. 2023. "Stack Overflow joins Reddit and Twitter in charging AI companies for training data." Zdnet, April 21. www.zdnet.com/article/stack-overflow-joins-reddit-and-twitter-in-charging-ai-companies-for-training-data.

Dickens, Robert. 2023. "UK re-considers proposed exception for text and data mining." *Allen & Overy* (blog), March 2. www.allenovery.com/en-gb/global/blogs/data-hub/uk-re-considers-proposed-exception-for-text-and-data-mining.

Digital Public Goods Alliance and UNICEF. 2023. "Core Considerations for Exploring AI Systems as Digital Public Goods." Community of Practice Discussion Paper. July 27. https://policycommons.net/artifacts/4539125/ai-cop-discussion-paper/5362605/.

Dilmegani, Cem. 2024. "Is Web Scraping Legal? Ethical Web Scraping Guide in 2024." AIMultiple Research, January 5. https://research.aimultiple.com/web-scraping-ethics/.

Duarte, Fabio. 2024. "Number of ChatGPT Users (Feb 2024)." *Exploding Topics* (blog), February 2. https://explodingtopics.com/blog/chatgpt-users.

Dziri, Nouha, Sivan Milton, Mo Yu, Osmar Zaiane and Siva Reddy. 2022. "On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?" arXiv, April 17. https://arxiv.org/abs/2204.07931.

European Council. 2024. "Artificial intelligence act: Council and Parliament strike a deal on the first rules for AI in the world." Press release, February 2. www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/.

European Parliament. 2023. "EU AI Act: first regulation on artificial intelligence." December 19. www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

Fabre, Benjamin. 2023. "Generative AI Is Scraping Your Data. So, Now What?" Dark Reading, August 21. www.darkreading.com/vulnerabilities-threats/generative-ai-is-scraping-your-data-so-now-what.

Future of Privacy Forum. 2018. *The Privacy Expert's Guide To Artificial Intelligence and Machine Learning*. International Association of Privacy Professionals. October. https://iapp.org/media/pdf/resource_center/FPF_Artificial_Intelligence_Digital.pdf.

Gamvros, Anna, Edward Yau and Steven Chong. 2023. "China finalises its Generative AI Regulation." Data Protection Report, July 25. www.dataprotectionreport.com/2023/07/china-finalises-its-generative-ai-regulation/.

Gebru, Timnit, Alex Hanna, Amba Kak, Sarah Myers West, Maximilian Gahntz, Mehtab Khan and Zeerak Talat. 2023. "Five considerations to guide the regulation of 'General Purpose AI' in the EU's AI Act." AI Now Institute, April 14. https://ainowinstitute.org/wp-content/uploads/2023/04/GPAI-Policy-Brief.pdf.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford. 2021. "Datasheets for Datasets." arXiv, December 1. https://arxiv.org/abs/1803.09010.

Germain, Thomas. 2023. "Google Says It'll Scrape Everything You Post Online for AI." Gizmodo, July 3. https://gizmodo.com/google-says-itll-scrape-everything-you-post-online-for-1850601486.

Gertner, Jon. 2023. "Wikipedia's Moment of Truth." *The New York Times Magazine*, July 18. www.nytimes.com/2023/07/18/magazine/wikipedia-ai-chatgpt.html.

Gibney, Elizabeth. 2022. "Open-source language AI challenges big tech's models." *Nature*, June 22. www.nature.com/articles/d41586-022-01705-z.

Gordon-Levitt, Joseph. 2023. "If artificial intelligence uses your work, it should pay you." *The Washington Post*, July 26. www.washingtonpost.com/opinions/2023/07/26/joseph-gordon-levitt-artificial-intelligence-residuals/.

Goswami, Rohan. 2023. "Reddit will charge hefty fees to the many third-party apps that access its data." CNBC, June 1. www.cnbc.com/2023/06/01/reddit-eyeing-ipo-charge-millions-in-fees-for-third-party-api-access.html.

Government of Canada. 2023. "Minister Champagne launches voluntary code of conduct relating to advanced generative AI systems." News release, September 27. www.canada.ca/en/innovation-science-economic-development/news/2023/09/minister-champagne-launches-voluntary-code-of-conduct-relating-to-advanced-generative-ai-systems.html.

Grant, Nico and Karen Weise. 2023. "In A.I. Race, Microsoft and Google Choose Speed Over Caution." *The New York Times*, April 7. www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html.

Grynbaum, Michael M. and Ryan Mac. 2023. "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work." *The New York Times*, December 27. www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Hacker, Philipp, Andreas Engel and Marco Mauer. 2023. "Regulating ChatGPT and other Large Generative AI Models." arXiv, May 12. https://arxiv.org/abs/2302.02337.

Hagiu, Andrei and Julian Wright. 2023. "Data-enabled learning, network effects, and competitive advantage." *RAND Journal of Economics* 54 (4): 638–67. https://onlinelibrary.wiley.com/doi/abs/10.1111/1756-2171.12453.

Heikkilä, Melissa. 2023. "AI language models are rife with different political biases." *MIT Technology Review*, August 7. www.technologyreview.com/2023/08/07/1077324/ai-language-models-are-rife-with-political-biases/.

Holmes, Aaron. 2020. "Lawmakers say Facebook and Google are hoarding people's personal data and using it to grow in a 'feedback loop' of market power — with no intention to stop." Business Insider, October 14. www.businessinsider.com/facebook-google-personal-data-privacy-congress-house-antitrust-report-2020-10.

Huang, Saffron and Divya Siddarth. 2023. "Generative AI and the Digital Commons." arXiv, March 20. https://arxiv.org/abs/2303.11074.

Ikeda, Scott. 2023. "Swedish Data Protection Authority May Be Next to Take Action Against Google Analytics." *CPO Magazine*, July 12. www.cpomagazine.com/data-protection/swedish-data-protection-authority-may-be-next-to-take-action-against-google-analytics/.

James, Cordilia. 2023. "The Other A.I.: Artificial Intimacy With Your Chatbot Friend." *The Wall Street Journal*, August 6. www.wsj.com/articles/when-you-and-ai-become-bffs-ecbcda1e?mod=djemTECH.

Jones, Erik and Jacob Steinhardt. 2022. "Capturing failures of large language models via human cognitive biases." *Advances in Neural Information Processing Systems* 35: 11785–99.

Kannan, Prabha. 2022. "Neema Iyer: Digital Extractivism in Africa Mirrors Colonial Practices." Stanford University Human-Centered Artificial Intelligence, August 15. https://hai.stanford.edu/news/neema-iyer-digital-extractivism-africa-mirrors-colonial-practices.

Khan, Lina. 2023. "We Must Regulate AI: Here's How." *The New York Times*, May 3. www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html.

Khan, Mehtab and Alex Hanna. 2023. "The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability." *Ohio State Technology Law Journal* 19. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4217148.

Kim, Won, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim and Doheon Lee. 2003. "A Taxonomy of Dirty Data." *Data Mining and Knowledge Discovery* 7: 81–99. https://doi.org/10.1023/A:1021564703268.

Knight, Will. 2023a. "Generative AI Is Making Companies Even More Thirsty for Your Data." *Wired*, August 10. www.wired.com/story/fast-forward-generative-ai-companies-thirsty-for-your-data/.

———. 2023b. "Google DeepMind's CEO Says Its Next Algorithm Will Eclipse ChatGPT." *Wired*, June 26. www.wired.com/story/google-deepmind-demis-hassabis-chatgpt/.

Lifshitz, Lisa R. 2019. "Federal Court makes clear: Website scraping is illegal." *Canadian Lawyer*, May 13. www.canadianlawyermag. com/news/opinion/federal-court-makes-clear-website-scraping-is-illegal/276128.

Lomas, Natasha. 2023a. "Italy orders ChatGPT blocked citing data protection concerns." Tech Crunch, March 31. https://techcrunch.com/2023/03/31/chatgpt-blocked-italy/.

———. 2023b. "ChatGPT resumes service in Italy after adding privacy disclosures and controls." Tech Crunch, April 28. https://techcrunch.com/2023/04/28/chatgpt-resumes-in-italy/.

———. 2023c. "ChatGPT-maker OpenAI accused of string of data protection breaches in GDPR complaint filed by privacy researcher." TechCrunch, August 30. https://techcrunch.com/2023/08/30/chatgpt-maker-openai-accused-of-string-of-data-protection-breaches-in-gdpr-complaint-filed-by-privacy-researcher.

Madiega, Tambiama. 2023. "General-purpose artificial intelligence." European Parliamentary Research Service. PE 745.708. March.

Marchant, Gary E. and Wendell Wallach. 2015. "Coordinating Technology Governance." *Issues in Science and Technology* XXXI (4). https://issues.org/coordinating-technology-governance/.

Marcus, J. Scott. 2023. "Adapting the European Union AI Act to deal with generative artificial intelligence." Bruegel, July 19. www.bruegel.org/analysis/adapting-european-union-ai-act-deal-generative-artificial-intelligence.

Margoni, Thomas and Martin Kretschmer. 2022. "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology." *GRUR International* 71 (8): 685–701. https://doi.org/10.1093/grurint/ikac054.

Martens, Bertin. 2018. "The Importance of Data Access Regimes for Artificial Intelligence and Machine Learning." JRC Digital Economy Working Paper 2018-09. December. http://dx.doi.org/10.2139/ssrn.3357652.

McKinsey. 2023. *The economic potential of generative AI: The next productivity frontier*. June 14. www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-AI-the-next-productivity-frontier#/.

Meta. 2023. "Meta and Microsoft Introduce the Next Generation of Llama." Meta, July 18. https://about.fb.com/news/2023/07/llama-2/.

Milmo, Dan. 2023. "The Guardian blocks ChatGPT owner OpenAI from trawling its content." *The Guardian*, September 1. www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content.

Mims, Christopher. 2024. "It's the End of the Web as We Know It." *The Wall Street Journal*, February 16. www.wsj.com/tech/ai/its-the-end-of-the-web-as-we-know-it-2ab686a2?mod=djemTECH.

Morrison, Ryan. 2023. "How do you regulate advanced AI chatbots like ChatGPT and Bard?" Tech Monitor, February 8. https://techmonitor.ai/technology/ai-and-automation/ai-regulation-chatgpt-bard.

Morrison, Sara. 2023. "The ongoing and increasingly weird Reddit blackout, explained." Vox, June 20. www.vox.com/technology/2023/6/14/23760738/reddit-blackout-explained-subreddit-apollo-third-party-apps.

msmash. 2023. "Why YouTube could Give Google an edge in AI." Slashdot, June 15. https://news.slashdot.org/story/23/06/15/168228/why-youtube-could-give-google-an-edge-in-ai.

Nagel, Sebastian. 2023. "May/June 2023 crawl archive now available." *Common Crawl* (blog), June 21. https://commoncrawl.org/blog/may-june-2023-crawl-archive-now-available.

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

Niedzwiadek, Nick. 2023. "How labor unions are putting checks on AI." Politico, November 20. www.politico.com/newsletters/weekly-shift/2023/11/20/how-labor-unions-are-putting-checks-on-ai-00128004.

Nishino, Anna. "Japan to allow freer sharing of content with obscure copyrights." Nikkei Asia, June 3. https://asia.nikkei.com/Business/Media-Entertainment/Japan-to-allow-freer-sharing-of-content-with-obscure-copyrights.

NIST. 2023a. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. January. Washington, DC: US Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1.

———. 2023b. *AI RMF Playbook*. Washington, DC: US Department of Commerce. https://airc.nist.gov/docs/AI_RMF_Playbook.pdf.

Nolan, Beatrice. 2023. "'Can't even see my own tweets': Twitter users despair over Elon Musk's decision to limit views." Business Insider, July 3. www.businessinsider.com/elon-musk-twitter-limit-views-users-angry-2023-7.

Norton Rose Fulbright. 2021. "New Singapore Copyright Exception will propel AI revolution." *Inside Tech Law* (blog), November 15.

O'Brien, Matt. 2023. "Chatbots can make things up. Can we fix AI's hallucination problem?" *PBS News Hour*, August 1. www.pbs.org/newshour/science/chatbots-can-make-things-up-can-we-fix-ais-hallucination-problem.

Octoparse. 2022. "5 Things You Need to Know Before Scraping Data from Facebook." Medium, June 14. https://medium.com/dataseries/5-things-you-need-to-know-before-scraping-data-from-facebook-f4e84b2ab80.

OECD. 2018. *Artificial Intelligence in Society*. Paris, France: OECD. www.oecd-ilibrary.org/sites/8b303b6f-en/index.html?itemId=/content/component/8b303b6f-en.

———. 2023. "AI Language Models: Technological, Socio-Economic and Policy Considerations." OECD Digital Economy Papers, No. 352. April. Paris, France: OECD. www.oecd.org/publications/ai-language-models-13d38f92-en.htm.

Office of the Australian Information Commissioner. 2023. "Global expectations of social media platforms and other sites to safeguard against unlawful data scraping." News release, August 24. www.oaic.gov.au/newsroom/global-expectations-of-social-media-platforms-and-other-sites-to-safeguard-against-unlawful-data-scraping.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman et al. 2023. *GPT-4 Technical Report*. https://arxiv.org/abs/2303.08774.

O'Shaughnessy, Matt and Matt Sheehan. "Lessons From the World's Two Experiments in AI Governance." Carnegie Endowment for International Peace, February 14. https://carnegieendowment.org/2023/02/14/lessons-from-world-s-two-experiments-in-ai-governance-pub-89035.

Pai, Nitin. 2023. "The immediate and important task for AI policy is to govern the industry." Mint, August 14. www.nitinpai.in/2023/08/14/the-immediate-and-important-task-for-ai-policy-is-to-govern-the-industry.

Pandey, Mohit. 2023. "The Biggest Winner from Threads: LLaMA." *Analytics India Mag*, July 7. https://analyticsindiamag.com/the-biggest-winner-from-threads-llama/.

Pelk, Henk. 2016. "Will chatbots change our lives?" Medium, November 16. https://itnext.io/will-chatbots-change-our-lives-e8d515cf3320.

Perri, Lori. 2023. "Generative AI Can Democratize Access to Knowledge and Skills." Gartner, October 17. www.gartner.com/en/articles/generative-ai-can-democratize-access-to-knowledge-and-skills.

Pierce, David. 2024. "The text file that runs the internet." The Verge, February 14. www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders.

Pinsky, Yury. 2023. "Bard can now connect to your Google apps and services." *The Keyword* (blog), September 19. https://blog.google/products/bard/google-bard-new-features-update-sept-2023/.

Radford, Alec, Karthik Narasimhan, Tim Salimans and Ilya Sutskever. 2018. "Improving Language Understanding by Generative Pre-Training." https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

Rastogi, Ritvik. 2023. "Papers Explained 43: GPT." Medium, April 24. https://medium.com/dair-ai/papers-explained-43-gpt-30b6f1e6d226.

Riley, Tonya. 2023. "Open AI lawsuit reignites privacy debate over data scraping." Cyberscoop, June 30. https://cyberscoop.com/openai-lawsuit-privacy-data-scraping/.

Romero, Jessica. 2023. "Leading the Fight Against Scraping-for-Hire." Meta, January 12. https://about.fb.com/news/2023/01/leading-the-fight-against-scraping-for-hire.

Rossi, Ben. 2016. "The dangers of web scraping." Information Age, September 5. www.information-age.com/dangers-web-scraping-2478/.

Roth, Mike. 2022. "Scraping the World (Wide Web)." Medium, June 24. https://medium.com/@rollingstorms/scraping-the-world-wide-web-80fc35020d1e.

Schaul, Kevin, Szu Yu Chen and Nitasha Tiku. 2023. "Inside the secret list of websites that make AI like ChatGPT sound smart." *The Washington Post*, April 19. www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

Schneier, Bruce. 2024. "How Public AI Can Strengthen Democracy." *Schneier on Security* (blog), March 7. www.schneier.com/blog/archives/2024/03/how-public-ai-can-strengthen-democracy.html

Sherry, Ben. 2023. "Why Generative A.I. Poses Risks for Early Adopters." Inc., April 28. www.inc.com/ben-sherry/why-generative-ai-poses-risks-for-early-adopters.html.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot and Ross Anderson. 2023. "The Curse of Recursion: Training on Generated Data Makes Models Forget." arXiv, May 31. https://arxiv.org/abs/2305.17493.

Sirimanne, Shamika N. 2023. "How artificial intelligence chatbots could affect jobs." *United Nations Conference on Trade and Development* (blog), January 18. https://unctad.org/news/blog-how-artificial-intelligence-chatbots-could-affect-jobs.

Southern, Matt G. 2023. "ChatGPT Creator Faces Multiple Lawsuits Over Copyright & Privacy Violations." *Search Engine Journal*, July 3. www.searchenginejournal.com/chatgpt-creator-faces-multiple-lawsuits-over-copyright-privacy-violations/490686/.

Staff in the Bureau of Competition & Office of Technology. 2023. "Generative AI Raises Competition Concerns." *FTC Technology Blog*, June 29. www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/06/generative-ai-raises-competition-concerns.

Stokel-Walker, Chris and Richard Van Noorden. 2023. "What ChatGPT and generative AI mean for science." Nature, February 6. www.nature.com/articles/d41586-023-00340-6.

Surman, Mark. 2016. "The Internet is a Global Public Resource." *Distilled* (blog), February 8. https://blog.mozilla.org/en/mozilla/the-internet-is-a-global-public-resource/.

Tarnoff, Ben. 2023. "Weizenbaum's nightmares: how the inventor of the first chatbot turned against AI." *The Guardian*, July 25. www.theguardian.com/technology/2023/jul/25/joseph-weizenbaum-inventor-eliza-chatbot-turned-against-artificial-intelligence-ai?.

Teach For America. 2023. "The Promises and Perils of Generative AI in Education: TFA's Evolving Perspective." Teach For America, August 11. www.teachforamerica.org/one-day/ideas-and-solutions/the-promises-and-perils-of-generative-ai-in-education-tfas-evolving.

Technomancers.ai. 2023. "Japan Goes All In: Copyright Doesn't Apply To AI Training." Communications of the ACM, June 1. www.biia.com/japan-goes-all-in-copyright-doesnt-apply-to-ai-training/.

Tene, Omer. 2023. "International Regulators' Unease With AI Data-Scraping Creates Gulf With US." Goodwin, September 7. www.goodwinlaw.com/en/news-and-events/news/2023/09/announcements-technology-dpc-international-regulators-unease-with-ai-data-scraping.

The Fashion Law. 2024. "From ChatGPT to Getty v. Stability AI: A Running List of Key AI-Lawsuits." The Fashion Law, January 25. www.thefashionlaw.com/from-chatgpt-to-deepfake-creating-apps-a-running-list-of-key-ai-lawsuits/.

The White House. 2023a. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." Presidential actions, October 30. www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

———. 2023b. "Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI." Statements and releases, July 21. 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/.

———. 2023c. "Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI." Statements and releases, September 12. www.whitehouse.gov/briefing-room/statements-releases/2023/09/12/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-eight-additional-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai.

Thomson-DeVeaux, Amelia and Curtis Yee. 2023. "ChatGPT Thinks Americans Are Excited About AI. Most Are Not." FiveThirtyEight, February 24. https://fivethirtyeight.com/features/chatgpt-thinks-americans-are-excited-about-ai-most-are-not/.

Thorbecke, Catherine. 2023. "AI tools make things up a lot, and that's a huge problem." CNN, August 29. www.cnn.com/2023/08/29/tech/ai-chatbot-hallucinations/index.html.

Tiku, Nitasha and Gerrit De Vynck. 2023. "Google shared AI knowledge with the world — until ChatGPT caught up." *The Washington Post*, May 2. www.washingtonpost.com/technology/2023/05/04/google-ai-stop-sharing-research/.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière et al. 2023a. "LLaMA: Open and Efficient Foundation Language Models." Meta, February 24. https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov et al. 2023b. "Llama 2: Open Foundation and Fine-Tuned Chat Models." Meta, July 18. https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/.

Ueno, Tatsuhiro. 2021. "The Flexible Copyright Exception for 'Non-Enjoyment' Purposes — Recent Amendment in Japan and Its Implication." *GRUR International* 70 (2): 145–52. https://doi.org/10.1093/grurint/ikaa184.

United Nations Development Programme. 2004. "Right to Information: Practical Guidance Note." July. www.undp-aciac.org/publications/other/undp/governance/righttoinfo-guidance-note-04e.pdf.

United Nations Educational, Scientific and Cultural Organization. 2023. *Open data for AI: What now?* Paris, France: United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000385841?posInSet=1&queryId=9456b23d-5453-4787-8837-932bf24226c0.

US Copyright Office. 2023. "Copyright Office Issues Notice of Inquiry on Copyright and Artificial Intelligence." NewsNet Issue 1017. August 30. www.copyright.gov/newsnet/2023/1017.html.

US General Services Administration. 2023. "Security Policy for Generative Artificial Intelligence (AI) Large Language Models (LLMs)." June 9. www.gsa.gov/directives-library/security-policy-for-generative-artificial-intelligence-ai-large-language-models-llms.

Wan, Audrey. 2023. "Why Japan is lagging behind in generative AI — and how it can create its own large language models." CNBC, July 6. www.cnbc.com/2023/07/07/why-japan-is-lagging-behind-in-generative-ai-and-creation-of-llms.html.

Web Scraper. 2021. "Brief History of Web Scraping." *Web Scraper* (blog), May 14. https://webscraper.io/blog/brief-history-of-web-scraping.

Whang, Oliver. 2023. "The Race to Make A.I. Smaller (and Smarter)." *The New York Times*, May 30. www.nytimes.com/2023/05/30/ai-chatbots-language-learning-models.html.

Whittaker, Zach. 2021. "Clearview AI ruled 'illegal' by Canadian privacy authorities." TechCrunch, February 3. https://techcrunch.com/2021/02/03/clearview-ai-ruled-illegal-by-canadian-privacy-authorities.

———. 2022. "Web scraping is legal, US appeals court reaffirms." TechCrunch, April 18. https://techcrunch.com/2022/04/18/web-scraping-legal-court/.

Wolfe, Cameron R. 2023a. "The History of Open-Source LLMs: Early Days (Part One)." Substack, July 24. https://cameronrwolfe.substack.com/p/the-history-of-open-source-llms-early.

———. 2023b. "The History of Open-Source LLMs: Better Base Models (Part Two)." Substack, July 31. https://cameronrwolfe.substack.com/p/the-history-of-open-source-llms-better.

———. 2023c "Data is the Foundation of Language Models." Medium, October 29. https://towardsdatascience.com/data-is-the-foundation-of-language-models-52e9f48c07f5.

World Bank. 2021. *World Development Report 2021: Data for Better Lives*. Washington, DC: World Bank. https://wdr2021.worldbank.org/stories/governing-data/.

Zakrzewski, Cat. 2023. "FTC Investigates OpenAI over data leak and ChatGPT's inaccuracy." *The Washington Post*, July 13. www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/.

Zhou, Ce, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang et al. 2023. "A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT." arXiv, May 1. https://arxiv.org/abs/2302.09419.

.

Centre for International
Governance Innovation

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

@cigionline