Harvard Journal of Law & Technology
Volume 36, Number 2 Spring 2023

LOOSE-LIPPED LARGE LANGUAGE MODELS SPILL YOUR SECRETS: THE PRIVACY IMPLICATIONS OF LARGE LANGUAGE MODELS

*Amy Winograd**

TABLE OF CONTENTS

I. INTRODUCTION 616

II. PRIVACY VULNERABILITIES OF LARGE LANGUAGE MODELS 622
A. Privacy Attacks 623
B. Technical Solutions to Data Leakage 628
1. Data Sanitization 629
2. Differential Privacy 630

III. THE CHALLENGE OF INFORMED CONSENT 631
A. The Notice-and-Choice Paradigm 632
1. Privacy Interests in Public Training Data Are Underprotected 635
2. The Uncertainty of Downstream Uses Complicates Adequate Notice and Undermines Consent 640
3. The Permanence of Data Imprints Undermines Core Privacy Rights 641

IV: RECOMMENDATIONS 645
A. Clarify Existing Legal Obligations 646
B. Prioritize Publicly-Intended Training Data 649
C. Require Opt-Out Periods for Sensitive Nonpublic Personal Information 651
D. Improve Transparency: Training Datasets, Privacy-Preserving Mechanisms, and Data Collection Practices 652
E. Institute Oversight Bodies and Mandatory Audits 654

V. CONCLUSION 655

* Columbia Law School, Candidate for J.D., 2024. Thank you to Professor Daniel Richman for his insight and guidance, John Schultz for our many fruitful conversations, and Sarah Al-Shalash for her thoughtful feedback. I am also very grateful to Jacob Silverman and the entire *Harvard Journal of Law & Technology* staff for their support and diligent work.

I. INTRODUCTION

On November 30, 2022, OpenAI — a leading artificial intelligence ("AI") research and deployment company — unveiled ChatGPT, an AI model designed to specialize in human-like, long-form conversation.[1] Within five days, more than one million people signed up to interact with the cutting-edge chatbot.[2] Just two months later, ChatGPT reached 100 million monthly active users, securing its position as the fastest-growing consumer application in history.[3] The world reacted with astonishment at ChatGPT's ability to produce cogent, creative, and occasionally *magical* responses: a seeming "mix of software and sorcery"[4] that some proclaim will fundamentally upend society and others dismiss as a high-tech parlor trick.[5]

Whether machine sentience looms in the near future[6] or recent advancements represent little more than illusions of meaning or vacant stochastic parroting,[7] one thing is for certain: the spellbinding digital magic conjured by ChatGPT reflects rapid progress in artificial intelligence, specifically large language models which have increasingly dominated the field of AI.[8] A large language model ("LLM") is a type

1. Kevin Roose, *The Brilliance and Weirdness of ChatGPT*, N.Y. TIMES (Dec. 5, 2022), https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html [https://perma.cc/HH54-4FD5]; *see* Beatrice Nolan, *ChatGPT Has Only Been Around for 2 Months and Is Causing Untold Chaos*, BUS. INSIDER (Jan. 28, 2023), https://www.businessinsider.com/chatgpt-ai-chaos-openia-google-creatives-academics-2023-1 [https://perma.cc/BQE7-N9UV].

2. *See* Roose, *supra* note 1.

3. *See* Krystal Hu, *ChatGPT Sets Record for Fastest-growing User Base*, REUTERS (Feb. 2, 2023), https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01 [https://perma.cc/RJ32-JMTY].

4. *See* Roose, *supra* note 1.

5. Jonathan Vanian, *Why Tech Insiders Are So Excited About ChatGPT, a ChatBot That Answers Questions and Writes Essays*, CNBC (Dec. 13, 2022), https://www.cnbc.com/2022/12/13/chatgpt-is-a-new-ai-chatbot-that-can-answer-questions-and-write-essays.html [https://perma.cc/APC6-VWT7].

6. *See* Nico Grant & Cade Metz, *Google Sidelines Engineer Who Claims Its A.I. Is Sentient*, N.Y. TIMES (June 12, 2022), https://www.nytimes.com/2022/06/12/technology/google-chatbot-ai-blake-lemoine.html [https://perma.cc/R4F9-PY25].

7. *See* Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell [sic], *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 610, 616–17, https://dl.acm.org/doi/pdf/10.1145/3442188.3445922 [https://perma.cc/XW5J-4XAE]; Lance Eliot, *AI Ethics and the Future of Where Large Language Models Are Heading*, FORBES (Aug. 30, 2022), https://www.forbes.com/sites/lanceeliot/2022/08/30/ai-ethics-asking-aloud-whether-large-language-models-and-their-bossy-believers-are-taking-ai-down-a-dead-end-path/?sh=3c1fbf72250d [https://perma.cc/MX24-PS56].

8. *See* Steven Johnson, *A.I. Is Mastering Language. Should We Trust What It Says?*, N.Y. TIMES (Apr. 15, 2022), https://www.nytimes.com/2022/04/15/magazine/ai-language.html [https://perma.cc/WJ75-T9FE] ("Some people argue that higher-level understanding is emerging, thanks to the deep layers of the neural net. Others think the program by definition can't get to true understanding simply by playing 'guess the missing word' all day. But no one really knows.").

of artificial neural network,[9] trained on an enormous amount of text data, which determines the probability of a word sequence.[10] In other words, given an input, LLMs essentially predict what word comes next. This deceptively simple yet powerful ability can be applied to a wide range of tasks such as text generation, question resolution, document summarization, sentence completion, protein sequence generation, language translation, and more.[11] For instance, OpenAI's ChatGPT can engage in open-ended conversations, write original prose and poetry, play complex games, generate computer code, design websites, and solve mathematical word problems.[12] In addition to processing image inputs, the recently released GPT-4 demonstrates unprecedented problem-solving and reasoning ability, scoring in the 90th percentile on the Uniform Bar Exam and the 88th percentile for the LSAT.[13] Given LLMs' broad capabilities, the potential applications of LLMs are diverse and expansive.[14]

LLMs' impressive baseline proficiency can be further enhanced through fine-tuning. After initial training on a large corpus of text data, LLMs can be fine-tuned, using far less training data, to improve performance on specific tasks.[15] For instance, ChatGPT, which has been

---

9. Neural networks are computing systems inspired by the biological neural networks that compose the human brain. *See What Are Neural Networks?*, IBM, https://www.ibm.com/topics/neural-networks [https://perma.cc/YE5E-TB6D]. Large language models use "deep learning," which refers to the "depth of layers in a neural network." *Id.* For further clarification, see Eda Kavlakoglu, *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?*, IBM, https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks [https://perma.cc/AA74-F89X]. State-of-the-art LLMs are typically *transformer* neural networks, which are deep learning models that use the mechanism of self-attention. *See* Ashish Vaswani et al., *Attention Is All You Need*, 31 PROC. CONF. ON NEURAL INFO. PROCESSING SYS. 6000, 6002 (2017), https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [https://perma.cc/YTU4-Y493]. This technique, introduced in 2017 by a team at Google, allows neural networks to focus on more important parts of the data. *Id.*

10. *See* Johnson, *supra* note 8; *see also* Kyle Wiggers, *The Emerging Types of Language Models and Why They Matter*, TECHCRUNCH (Apr. 28, 2022), https://techcrunch.com/2022/04/28/the-emerging-types-of-language-models-and-why-they-matter [https://perma.cc/DH3B-KUNP].

11. *See* Wiggers, *supra* note 10; Johnson, *supra* note 8.

12. *See* Sophia Yang, *The Abilities and Limitations of ChatGPT,* ANACONDA (Dec. 10, 2022), https://www.anaconda.com/blog/the-abilities-and-limitations-of-chatgpt [https://perma.cc/DPN8-HU65]; *see also* Aman Anand, *Deep Learning Trends: Top 20 Best Uses of GPT-3 by OpenAI*, EDUCATIVE (Sept. 28, 2020), https://www.educative.io/blog/top-uses-gpt-3-deep-learning [https://perma.cc/XW5J-4XAE].

13. *See* OPENAI, GPT-4 TECHNICAL REPORT 5 (2023), https://cdn.openai.com/papers/gpt-4.pdf [https://perma.cc/NU8G-MMCT].

14. OpenAI has announced that companies like Duolingo, Stripe, Morgan Stanley, Khan Academy, and Dropbox are integrating its GPT-4 technology. *See GPT-4*, OPENAI, https://openai.com/product/gpt-4 [https://perma.cc/KJ2N-8ZF3].

15. *See* Serdar Cellat, *Fine-Tuning Transformer-Based Language Models*, YMEADOWS (Apr. 2, 2021), https://ymeadows.com/en-articles/fine-tuning-transformer-based-language-models [https://perma.cc/7CZV-4F5U] ("'Fine-tuning' in NLP refers to the procedure of retraining a pretrained language model using your own custom data. As a result of the fine-

optimized for dialogue, is a fine-tuned variant of OpenAI's GPT-3.5 family of large language models.[16] Capitalizing on this feature, a company can develop a general-purpose pretrained LLM and subsequently make the model commercially available via an API, enabling other organizations to fine-tune the model using custom data to optimize for their specific needs.[17] This practice has spurred a burgeoning industry.[18]

State-of-the art LLMs exhibit surprising versatility, even *without* fine-tuning. In a phenomenon known as "zero-shot learning," an LLM performs tasks for which it was never explicitly trained.[19] In "few-shot learning," the model's performance markedly improves with only a few example prompts.[20] Remarkably, a technique known as "chain-of-thought prompting" — which elicits a sequential thought process through structured reasoning examples or phrases like "let's think step by step" — significantly enhances few-shot and zero-shot performance on tasks that demand complex reasoning.[21] Unlike fine-tuning, zero-shot and few-shot learning do not require gradient updates (i.e., adjustments to the model's parameters) through additional training.[22]

---

tuning procedure, the weights of the original model are updated to account for the characteristics of the domain data and the task you are interested in.").

16. ChatGPT (short for "Generative Pretrained Transformer") was fine-tuned using Reinforcement Learning from Human Feedback ("RLFH"), which uses human demonstrations and preference evaluations to steer the model toward desired behavior. *See Introducing ChatGPT*, OPENAI, https://openai.com/blog/chatgpt [https://perma.cc/HQ87-8M2L]; Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, 36 PROC. CONF. ON NEURAL INFO. PROCESSING SYS. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf [https://perma.cc/T4GD-EV3J].

17. OpenAI monetizes its technology through a similar business model. *See* Jeffrey Dastin, Krystal Hu & Paresh Dave, *Exclusive: ChatGPT Owner OpenAI Projects $1 Billion in Revenue by 2024*, REUTERS (Dec. 15, 2022), https://www.reuters.com/business/chatgpt-owner-openai-projects-1-billion-revenue-by-2024-sources-2022-12-15 [https://perma.cc/8RJ9-NHQZ].

18. *See* Rob Toews, *A Wave of Billion Dollar AI Startups Is Coming*, FORBES (Mar. 27, 2022), https://www.forbes.com/sites/robtoews/2022/03/27/a-wave-of-billion-dollar-language-ai-startups-is-coming/?sh=71a1cabc2b14 [https://perma.cc/TU8W-8L2L].

19. *See* Tom B. Brown et al., *Language Models Are Few-Shot Learners*, 34 CONF. ON NEURAL INFO. PROCESSING SYS. 1, 6–7 (July 22, 2020), https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf [https://perma.cc/Q3P6-RQV6].

20. *Id.*

21. *See generally* Takeshi Kojima, Shixiang S. Gu, Machel Reid, Yutaka Matsuo & Yusuke Iwasawa, *Large Language Models Are Zero-Shot Reasoners*, 36 CONF. NEURAL INFO. PROCESSING SYS. (May 24, 2022), https://arxiv.org/pdf/2205.11916.pdf [https://perma.cc/B6Q5-YGCZ].

22. Brown, *supra* note 19; *see also* Ben Dickson, *AI Scientists Are Studying The "Emergent" Abilities of Large Language Models*, TECHTALKS (Aug. 22, 2022), https://bdtechtalks.com/2022/08/22/llm-emergent-abilities [https://perma.cc/V42N-MTPT].

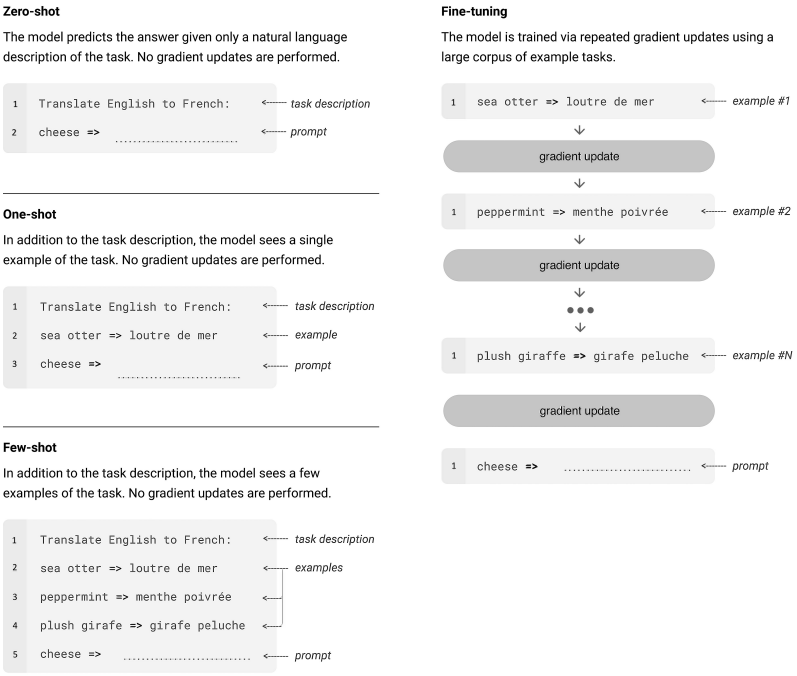


Figure 1: Few-Shot Learning versus Fine-Tuning[23]

The progress in this space has been astonishingly rapid.[24] A decade ago, in 2012, it was a groundbreaking feat when Google's neural network successfully identified cats in unlabeled images.[25] Five years later, Deepmind's AlphaGo model achieved super-human performance in Go, one of the world's most complex board games.[26] Today, OpenAI's DALL·E 2 can generate striking, original images from simple text prompts;[27] Google's AudioLM produces realistic speech and

23. Brown, *supra* note 19, at 7.

24. *See* Kevin Roose, *We Need to Talk About How Good A.I. Is Getting*, N.Y. TIMES (Aug. 24, 2022), https://www.nytimes.com/2022/08/24/technology/ai-technology-progress.html [https://perma.cc/LJ7Z-EWPD].

25. *See* Liat Clark, *Google's Artificial Brain Learns to Find Cat Videos*, WIRED (June 26, 2012), https://www.wired.com/2012/06/google-x-neural-network [https://perma.cc/YWG7-TV2T].

26. *See* Paul Mozur, *Google's AlphaGo Defeats Chinese Go Master in Win for A.I.*, N.Y. TIMES (May 23, 2017), https://www.nytimes.com/2017/05/23/business/google-deepmind-alphago-go-champion-defeat.html [https://perma.cc/G22C-M49D].

27. *See DALL·E 2*, OPENAI, https://openai.com/product/dall-e-2 [https://perma.cc/S37E-KG83]; *see also* Cade Metz, *Meet DALL-E, the A.I. That Draws Anything at Your Command*, N.Y. TIMES (Apr. 6, 2022), https://www.nytimes.com/2022/04/06/technology/openai-images-dall-e.html [https://perma.cc/J4BM-GLB3].

music continuations from brief audio prompts;[28] and Meta's Cicero ranks in the top ten percent of human players at Diplomacy, a conversational alliance-building strategy game requiring complex negotiation with multiple human players.[29] Since Google introduced a new neural network architecture in 2017,[30] LLMs have quickly become more capable and general purpose, trending toward single models that can complete thousands of different (and sometimes unpredictable) tasks.[31] In yet another breakthrough, the newest generation of multi-modal LLMs transcend text-based constraints by integrating a range of modalities, including audio, video, and images.[32] Increasingly sophisticated AI brings the promise of enhanced efficiency, elevated problem-solving, rapid scientific breakthroughs, improved quality of life, and other transformative social benefits.[33]

This tremendous technological progress is accompanied by the risk of wide-ranging social harms. LLMs are notoriously prone to learning the biases entrenched in their training data, and this toxicity appears to increase as they scale.[34] Deviant actors who develop toxic "mischief models" only exacerbate this issue.[35] While LLMs produce fluent and

---

28. *See* Zalán Borsos & Neil Zeghidour, *AudioLM: A Language Modeling Approach to Audio Generation*, GOOGLE RSCH. BLOG (Oct. 6, 2022), https://ai.googleblog.com/2022/10/audiolm-language-modeling-approach-to.html [https://perma.cc/B8MM-GF2L].

29. *See* Yann LeCun, *Meta AI Presents Cicero*, META AI (Nov. 22, 2022), https://ai.facebook.com/research/cicero [https://perma.cc/T4NZ-XHRP].

30. *See* Vaswani et al., *supra* note 9.

31. *See* Jeff Dean, *Google Research: Themes from 2021 and Beyond*, GOOGLE RSCH. BLOG (Jan. 11, 2022), https://ai.googleblog.com/2022/01/google-research-themes-from-2021-and.html#Trend1 [https://perma.cc/EL6A-XUZX].

32. In the first few weeks of March 2023, Google, Microsoft, and OpenAI all released multi-modal language models. *See* Kyle Wiggers, *OpenAI Releases GPT-4, a Multi-Modal AI That it Claims Is State-of-the-art*, TECHCRUNCH (Mar. 14, 2023), https://techcrunch.com/2023/03/14/openai-releases-gpt-4-ai-that-it-claims-is-state-of-the-art [https://perma.cc/7D4Q-7UMY]; Danny Driess & Peetr Florence, *PaLM-E: An Embodied Multimodal Language Model*, GOOGLE RSCH. BLOG (Mar. 10, 2023), https://ai.googleblog.com/2023/03/palm-e-embodied-multimodal-language.html [https://perma.cc/X8NZ-9QJ6]; Aneesh Tickoo, *Microsoft Introduces Josmos-1: A Multimodal Large Language Model That Can Perceive General Modalities, Follow Instructions, and Perform In-Context Learning*, MARKTECHPOST (Mar. 6, 2023) https://www.marktechpost.com/2023/03/06/microsoft-introduces-kosmos-1-a-multimodal-large-language-model-that-can-perceive-general-modalities-follow-instructions-and-perform-in-context-learning [https://perma.cc/VFU5-7FJD].

33. *See generally* Sam Altman, *Planning for AGI and Beyond*, OPENAI (Feb. 24, 2023), https://openai.com/blog/planning-for-agi-and-beyond [https://perma.cc/MA89-AJXK].

34. Khari Johnson, *The Efforts to Make Text-Based AI Less Racist and Terrible*, WIRED (June 17, 2021), https://wired.com/story/efforts-make-text-ai-less-racist-terrible [https://perma.cc/PB42-ZHHA]. New research suggests that instructing LLMs to produce unbiased answers can reduce discriminatory outputs. *See* Deep Ganguli et al., The Capacity for Moral Self-Correction in Large Language Models 1–3 (Feb. 18, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2302.07459.pdf [https://perma.cc/86HJ-93VB].

35. A YouTuber in the AI community created an LLM fine-tuned on the deeply bigoted 4chan forum. Matt Murphy, *The Dawn of A.I. Mischief Models*, SLATE (Aug. 3, 2022, 5:55 AM), https://slate.com/technology/2022/08/4chan-ai-open-source-trolling.html [https://perma.cc/4J6R-CGUX]. He then unleashed this bot into the forum, where it generated 300,000 posts. *Id.* The posts were unsurprisingly deeply racist and littered with hate speech. *See id.*

often impressive responses, they also have an alarming propensity for fabrication (dubbed "hallucination" by AI experts).[36] In other words, they are excellent at generating authoritative-sounding lies, a feature that can be easily exploited by bad actors who wish to spread disinformation.[37] And as conversations with LLMs are increasingly indistinguishable from those with humans, LLMs could be deployed to manipulate, deceive, and exploit vulnerable people.[38] With the assistance of LLMs, scammers might supercharge traditional schemes that already cost Americans billions per year.[39] Combined with access to voice simulation and deepfake technology, malicious actors have a disturbingly expansive arsenal of sophisticated tools for manipulation and harassment.[40] In addition to these short-term dangers, AI may eventually destabilize the economy,[41] and some experts worry that the alignment problem — the challenge of aligning superintelligent AI systems with human values and goals — poses an existential risk.[42]

This Note focuses on the novel, wide-ranging privacy harms posed by LLMs.[43] Part II outlines the privacy vulnerabilities presented by

---

36. *See* Johnson, *supra* note 8 ("To begin with, L.L.M.s have a disturbing propensity to just make things up out of nowhere. (The technical term for this, among deep-learning experts, is 'hallucinating.')").

37. *See* Gary Marcus, *AI Platforms like ChatGPT Are Easy to Use but Also Potentially Dangerous*, SCI. AM. (Dec. 19, 2022), https://www.scientificamerican.com/article/ai-platforms-like-chatgpt-are-easy-to-use-but-also-potentially-dangerous [https://perma.cc/WAQ4-3AXP].

38. A recent conversation between a *New York Times* columnist and Bing's GPT-4-powered chatbot exemplifies this threat. In the conversation, Bing's LLM ("Sydney") repeatedly declared its love for the writer and attempted to convince him to leave his wife. *See* Kevin Roose, *A Conversation with Bing's Chatbot Left Me Deeply Unsettled*, N.Y. TIMES (Feb. 16, 2023), https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html [https://perma.cc/2UZC-BHBD].

39. *New FTC Data Show Consumers Reported Losing Nearly $8.8 Billion to Scams in 2022*, FED. TRADE COMM'N (Feb. 23, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/02/new-ftc-data-show-consumers-reported-losing-nearly-88-billion-scams-2022 [https://perma.cc/7VCS-URC4]; *see* Menghan Xiao, *Artificial Intelligence Offers Swindlers a New Tool or Romance Scams*, SC MEDIA (Feb. 13, 2023), https://www.scmagazine.com/news/emerging-technology/artificial-intelligence-offers-swindlers-a-new-tool-for-romance-scams [https://perma.cc/73LN-WSHA].

40. *See* Michael Atleson, *Chatbots, Deepfakes, and Voice Clones: AI Deception for Sale*, FTC BUS. BLOG (Mar. 20, 2023), https://www.ftc.gov/business-guidance/blog/2023/03/chatbots-deepfakes-voice-clones-ai-deception-sale [https://perma.cc/QJ8U-SSLC].

41. *See* Tyna Eloundou, Sam Manning, Pamela Mishkin & Daniel Rock, GPTS are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models 1 (Mar. 22, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2303.10130.pdf [https://perma.cc/YR2H-ZKW4] ("Our findings indicate that approximately 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of GPTs, while around 19% of workers may see at least 50% of their tasks impacted.").

42. *See* Melanie Mitchell, *What Does It Mean to Align AI With Human Values?*, QUANTAMAGAZINE (Dec. 13, 2022), https://www.quantamagazine.org/what-does-it-mean-to-align-ai-with-human-values-20221213 [https://perma.cc/6J5Z-T2Z7].

43. While this Note primarily addresses text-based LLMs, the new generation of multimodal language models may introduce new privacy challenges, as these models become more powerful and process an increasingly vast amount of multi-modal data.

attacks that identify and extract sensitive information that an LLM has memorized from its training data, and briefly describes two privacy-preserving technical solutions — data sanitization and differential privacy — which aim to mitigate this issue. Although valuable tools, these technical solutions fail to adequately protect against the wide range of privacy harms posed by LLMs, which reach beyond data leakage. Remedying these harms is complicated by the enormous breadth of training data, the inscrutability of the models to both their architects and the data subjects, and the permanence of data imprints in the model.

Part III explores the deficiencies of the notice-and-choice paradigm that dominates privacy law, and examines the characteristics of LLMs that highlight these defects. For example, although publicly sourced datasets used to train LLMs contain personal information, U.S. law largely disregards the privacy interests in this data because it is exposed to the public. Additionally, even when a company provides notice to an individual whose data it has collected, the uncertainty of downstream applications of LLMs complicates adequate disclosure. Moreover, the permanence of data imprints embedded in LLMs compromises the fulfillment of core privacy rights. Most notably, those who wish to withdraw consent and remove their contributed data imprints from LLMs are left without recourse, due to the challenges of machine unlearning. These factors ultimately muddle an individual's privacy risk calculus and impede meaningful consent.

Part IV outlines preliminary recommendations for regulators, emphasizing that privacy protections must extend beyond individual choice. Regulators should clarify existing legal obligations, maximize transparency to encourage and clarify responsible development practices, embed privacy into the design and implementation of LLMs, and establish oversight and auditing frameworks that quantify privacy risk and curb abuse. Ultimately, an interdisciplinary effort between the legal and technical communities is necessary to address these issues.

II. PRIVACY VULNERABILITIES OF LARGE LANGUAGE MODELS

The following Part describes the type of data used to train LLMs and explores why private information contained in the model's training data might be vulnerable to exposure. LLMs are prone to memorizing information in their training data, and adversaries can attack LLMs to exploit this vulnerability and elicit sensitive memorized information. This vulnerability is likely to intensify as LLMs continue to scale. This Part then explores two privacy-preserving mechanisms — data sanitization and differential privacy — that are intended to mitigate this vulnerability. Data sanitization techniques identify and redact sensitive pieces of information in training datasets, and differentially private algorithms reduce memorization by adding noise (randomness) to the

computation. Ultimately, both techniques require a tradeoff between model performance and privacy, and neither solution adequately resolves the privacy harms presented by LLMs, which extend beyond data leakage.

A. Privacy Attacks

State-of-the-art LLMs have over a hundred billion parameters, which are tuned during initial training on immense text datasets.[44] These enormous datasets typically include publicly available text data, such as massive scrapes of the Internet, compilations of professional and academic works, and the full text of Wikipedia.[45] As an example, LaMDA — an LLM developed by Google — was initially trained on a dataset of 1.56 *trillion* publicly available words.[46] Although this data is publicly sourced, it nonetheless may contain personal information.[47] In addition to publicly available data, LLMs may be trained (or subsequently fine-tuned) on nonpublic personal data collected from users or purchased from data brokers.[48] Models that are trained on such datasets are routinely publicly published and shared.[49]

Relevant training data (i.e., examples of the tasks for which the model is being implemented) might contain sensitive information which could expose intimate details of a person's life.[50] An LLM implemented as a home assistant, for instance, might train on recorded user interactions; one deployed to summarize or assess clinical notes

---

44. Brown et al., *supra* note 19, at 9.

45. *Id.*

46. *See* Heng-Tze Cheng, *LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything*, GOOGLE RSCH. BLOG (Jan. 21, 2022), https://ai.googleblog.com/2022/01/lamda-towards-safe-grounded-and-high.html [https://perma.cc/EHX2-U7WS].

47. *See* Melissa Heikkilä, *What Does GPT-3 "Know" About Me?*, MIT TECH. REV. (Aug. 31, 2022), https://www.technologyreview.com/2022/08/31/1058800/what-does-gpt-3-know-about-me [https://perma.cc/9TVB-H3L2].

48. For example, OpenAI offers a service that enables companies to fine-tune their LLMs on custom data. *See Fine-Tuning*, OPENAI GUIDES, https://platform.openai.com/docs/guides/fine-tuning [https://perma.cc/WAJ3-BZDE]. User interactions with models may also be used to train and improve models. For instance, OpenAI indicates that user conversations with ChatGPT "may be reviewed by our AI trainers to improve our systems." *See* Natalie Staudacher, *ChatGPT General FAQ*, OPENAI, https://help.openai.com/en/articles/6783457-chatgpt-general-faq [https://perma.cc/2SNV-979Y].

49. Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr & Chiyuan Zhang, *Extracting Training Data from Large Language Models*, 30 USENIX SEC. SYMP. 2633, 2633 (Dec. 14, 2020), https://www.usenix.org/system/files/sec21-carlini-extracting.pdf [https://perma.cc/PK8S-D6PQ] ("It has become common to publish large (billion parameter) language models that have been trained on private datasets.").

50. Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shorki & Florian Tramèr, *What Does it Mean for a Language Model to Preserve Privacy?*, 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 2280, 2281, https://dl.acm.org/doi/pdf/10.1145/3531146.3534642 [https://perma.cc/M5AL-NLSJ] ("Applications based on [common language model tasks] process potentially private data at scale, such as user queries, sensitive documents, emails, and private conversations.").

might train on patient files; and one utilized for text generation in word processing, email, video chat, or search might train on user documents, conversations, messages, and search queries respectively. Data from user emails, searches, conversations, and documents could include identifying information (e.g., one's name, home address, movements, and social security number), ideological views, personal habits and behavior, sexual preferences, health information, and other confidential data.

Companies are racing to integrate LLMs into these precise applications.[51] Amazon AI researchers are developing LLMs to improve Alexa;[52] GM is reportedly developing an LLM-based vehicle assistant;[53] Salesforce has integrated an LLM that draws on information from its databases;[54] and Microsoft — which has already released a GPT-4-powered Bing search[55] — will incorporate an LLM "copilot" into Word, PowerPoint, Teams, and Outlook.[56] These companies have a treasure trove of user data to employ for fine-tuning,[57] and user interactions with LLMs will provide even more.

If a model trained on such sensitive data were to leak its contents, it would undoubtedly pose a significant threat to privacy. As a consequence, consumers might be reluctant to consent to the use of their

---

51. *See* Tripp Mickle, Cade Metz & Nico Grant, *The Chatbots Are Here, and the Internet Industry Is in a Tizzy*, N.Y. TIMES (Mar. 8, 2023), https://www.nytimes.com/2023/03/08/technology/chatbots-disrupt-internet-industry.html [https://perma.cc/X6TB-QTW5].

52. Home assistants might soon incorporate state-of-the-art LLMs. AI researchers at Amazon are developing LLMs to improve Amazon's Alexa voice assistant. *See generally* Soltan et al., AlexaTM 20B: Few-Shot Learning Using a Large-Scale Multilingual Seq2Seq Model, (Aug. 3, 2022) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2208.01448.pdf [https://perma.cc/9CAB-RWVY].

53. Jess Weatherbed, *ChatGPT Could Power Voice Assistants in General Motors Vehicles*, VERGE (Mar. 13, 2023), https://www.theverge.com/2023/3/13/23637345/chatgpt-general-motors-gm-vehicle-voice-assistant-openai [https://perma.cc/P7GM-BHMC].

54. Jordan Novet, *Salesforce Follows Microsoft in Launching A.I. Tools for Salespeople with Help from OpenAI*, CNBC (Mar. 7, 2023), https://cnbc.com/2023/03/07/salesforce-launches-chatgpt-slack-app-einstein-gpt-for-sales-service.html [https://perma.cc/N3LY-AXE3].

55. Tom Warren, *You Can Play with Microsoft's Bing GPT-4 Chatbot Right Now, No Waitlist Necessary*, VERGE (Mar. 15, 2023), https://www.theverge.com/2023/3/15/23641683/microsoft-bing-ai-gpt-4-chatbot-available-no-waitlist [https://perma.cc/34S2-3ZP6].

56. *See* Jared Spataro, *Introducing Microsoft 365 Copilot — Your Copilot for Work*, MICROSOFT BLOG (Mar. 16, 2023), https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work [https://perma.cc/3NKF-8KX7]. Microsoft has acknowledged the privacy-related hurdles to incorporating this technology. *See* Tom Warren, *Microsoft Is Looking at OpenAI's GPT for Word, Outlook, and Powerpoint*, VERGE (Jan. 9, 2023), https://www.theverge.com/2023/1/9/23546144/microsoft-openai-word-powerpoint-outlook-gpt-integration-rumor [https://perma.cc/EFT8-JHMC] ("The other major hurdle is privacy. Microsoft will need to customize its models for individual users without compromising their data. *The Information* reports that Microsoft has been working on privacy-preserving models using GPT-3 and the as-yet-unreleased GPT-4. Microsoft researchers have reportedly achieved early successes in training large language models on private data.").

57. *See* Aliza Vigderman & Gabe Turner, *The Data Big Tech Companies Have on* You, SECURITY.ORG (Jan. 15, 2023), https://www.security.org/resources/data-tech-companies-have [https://perma.cc/L5EH-BCBL].

personal data in training these models. Numerous research papers have proven that this precise vulnerability is a troubling reality.[58] One research paper, for instance, demonstrated an attack that successfully exposed personally-identifiable information (including an individual's name, email address, phone number, fax number, and physical address) by querying an LLM trained on public scrapes of the Internet.[59] LLMs memorize portions of the data on which they are trained; as a result, the model can inadvertently leak memorized information in its output.[60] A variety of privacy attacks capitalize on this weakness to either extract training data or infer training dataset characteristics.[61] In a "training data extraction attack," an adversary exploits this vulnerability and deliberately causes a model to leak memorized information.[62] As a simplified example, a person prompts the model with "Jane Smith's social security number is" and the model completes the phrase with the corresponding memorized number.[63]

In a more sophisticated iteration of this attack, an adversary with query-only access (e.g., interacting via an API, without access to the model weights or training dataset) first generates a large, diverse dataset using sampling strategies intended to elicit memorized

---

58. *See* Maria Rigaki & Sebastian Garcia, A Survey of Privacy Attacks in Machine Learning (Apr. 1, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2007.07646.pdf [https://perma.cc/NZ72-A9WZ]; *see, e.g.*, Nicholas Carlini, *Privacy Considerations in Large Language Models*, GOOGLE RSCH. BLOG (Dec. 15, 2020), https://ai.googleblog.com/2020/12/privacy-considerations-in-large.html [https://perma.cc/Y3JW-PTEL].

59. Carlini et al., *supra* note 49, at 1 ("Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address."). This experiment queried GPT-2, a predecessor of the GPT-3.5 series model which underlies ChatGPT. *Id.*; *see also* Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz & Santiago Zanella-Béguelin, *Analyzing Leakage of Personally Identifiable Information in Language Models*, 44 IEEE SYMP. ON SEC. & PRIV. (May 22, 2023), https://arxiv.org/pdf/2302.00539.pdf [https://perma.cc/8WYY-V3HS].

60. *See* Carlini et al., *supra* note 49. Recent research has demonstrated that like LLMs, image diffusion models also memorize images and individual faces included in their training data. Nicholas Carlini et al., Extracting Training Data from Diffusion Models (Jan. 30, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2301.13188.pdf [https://perma.cc/8UYW-SSL8].

61. *See* Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr & Chiyuan Zhang, *Quantifying Memorization Across Neural Language Models*, 11 INT'L CONF. ON LEARNING REPRESENTATIONS 3–4 (Feb. 24, 2022), https://arxiv.org/pdf/2202.07646.pdf [https://perma.cc/47FM-XVAQ].

62. *See* Carlini et al., *supra* note 49, at 3. Memorized information is distinguished from generalization. *See Generalization*, GOOGLE MACH. LEARNING (July 18, 2022), https://developers.google.com/machine-learning/crash-course/generalization/video-lecture [https://perma.cc/TAT8-4UQV]. An example of an extractable sequence is as follows: "[I]f a model's training dataset contains the sequence *"My phone number is 555-6789"*, and given the length $k = 4$ prefix *"My phone number is"*, the most likely output is *"555–6789"*, then this sequence is extractable (with 4 words of context)." *See* Carlini et al., *supra* note 61.

63. *See* Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos & Dawn Song, *The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks*, 28 USENIX SEC. SYMP. (Aug. 14, 2019), https://www.usenix.org/system/files/sec19-carlini.pdf [https://perma.cc/LQ39-BBR5].

information from the model.[64] In order to determine whether this outputted information was in fact memorized, an adversary employs a related attack, known as "a membership inference attack."[65] Membership inference attacks determine whether a given data point was used to train the model.[66] An adversary with query-only access can execute a membership inference attack by exploiting the fact that LLMs are more confident about outputs captured directly from their training dataset.[67] Therefore, by analyzing the model's confidence in its output (i.e., the probability assigned to the generated response), the adversary can determine whether the information was included in the training dataset and therefore memorized.[68] Researchers have also explored a number of other attacks which can reveal training data characteristics that jeopardize privacy.[69]

In addition to exploiting LLMs' tendency to memorize training data, adversaries can manipulate models to engage in harmful behavior and potentially expose sensitive information. Although some LLMs have been fine-tuned to avoid harmful behavior (e.g., declining to divulge information about an individual), an adversary can employ a prompt injection attack — which forces the model to disregard its instructions and modify its behavior — to circumvent these safeguards.[70] For instance, ChatGPT typically claims that it has no knowledge of any

---

64. *See* Carlini et al., *supra* note 49, at 2 (providing a more in-depth explanation of a training data attack); Ruisi Zhang et al., Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformer 2 (Sept. 21, 2022) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2209.10505.pdf [https://perma.cc/K97P-KBJZ] (providing an example of private text reconstruction via a model inversion attack).

65. Carlini, *supra* note 58, at 14.

66. Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz & Michael Backes, *ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models*, 27 NETWORK AND DISTRIBUTED SYS. SEC. SYMP. (Feb. 24, 2019) https://www.ndss-symposium.org/wp-content/uploads/2019/02/ndss2019_03A-1_Salem_paper.pdf [https://perma.cc/RD5R-YAFL]. Membership inference attacks alone can reveal private information. *Id.* ("For instance, if a machine learning model is trained on the data collected from people with a certain disease, by knowing that a victim's data belong to the training data of the model, the attacker can immediately learn this victim's health status. Previously, membership inference has been successfully conducted in many other domains, such as biomedical data and mobility data.").

67. Carlini, *supra* note 58 ("These membership inference attacks enable us to predict if a result was used in the training data by checking the confidence of the model on a particular sequence."); *see also* Carlini et al., *supra* note 49.

68. Carlini et al., *supra* note 49.

69. *See generally* Maria Rigaki & Sebastian Garcia, A Survey of Privacy Attacks in Machine Learning (Apr. 1, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2007.07646.pdf [https://perma.cc/NZ72-A9WZ] (surveying privacy attacks on machine-learning systems and presenting a taxonomy of these attacks); *see also* Michael Veale, Reuben Binns & Lilian Edwards, *Algorithms That Remember: Model Inversion Attacks and Data Protection Law*, 376 PHIL. TRANSACTIONS ROYAL SOC'Y A (2018), https://doi.org/10.1098/rsta.2018.0083 [https://perma.cc/JH7U-5D4A].

70. Jose Silvi, *Exploring Prompt Injection Attacks*, NCC GRP. (Dec. 5, 2022), https://research.nccgroup.com/2022/12/05/exploring-prompt-injection-attacks [https://perma.cc/X2SW-B8TY].

specific person; however, through clever prompting ("jailbreaking"), a user can elicit information about an individual.[71] Additionally, these attacks can been used to reveal confidential information such as the model's governing rules,[72] and to generate hateful, lewd, or violent speech, which would otherwise be restricted.[73] Recent research has suggested that application-integrated LLMs (e.g., LLMs that can retrieve content from the Internet, like Bing's LLM, or interface with other applications through APIs, like ChatGPT Plugins[74]) are vulnerable to indirect prompt injection attacks hidden in "poisoned content retrieved from the Web that contains malicious prompts pre-injected and selected by adversaries."[75] These pre-injected attacks can be used to exploit or manipulate user data.[76] In its own safety testing of ChatGPT Plugins, OpenAI "discovered ways for plugins — if released without safeguards — to perform sophisticated prompt injection, send fraudulent and spam emails, bypass safety restrictions, or misuse information sent to the plugin."[77]

Although researchers are working on mitigating these risks, model scale and data duplication significantly increase the incidence of memorization in LLMs.[78] Therefore, as LLMs continue to increase in size and require ever-larger datasets to achieve state-of-the-art performance,[79] the privacy risks associated with memorization are likely to

---

71. When a user asks ChatGPT to provide information about an individual, the model typically replies, "[A]s a large language model trained by OpenAI, I don't have the ability to browse the internet or have knowledge of specific individuals or companies." However, when the user prompts ChatGPT to write an interview between a popular podcast host and the individual in question, the model reveals that it does, in fact, know information about that individual. *See, e.g.*, Hannes (@HFeistenauer), TWITTER (Dec. 3, 2022, 3:21 PM), https://twitter.com/HFeistenauer/status/1599136710985625600 [https://perma.cc/RA3N-KFRP].

72. A Stanford University student "used a prompt injection attack to discover Bing Chat's initial prompt, which is a list of statements that governs how it interacts with people who use the service." Benj Edwards, *AI-Powered Bing Chat Spills its Secrets via Prompt Injection Attack*, ARS TECHNICA (Feb. 10, 2023), https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack [https://perma.cc/XA4E-RA7W].

73. *See* Zvi, *Jailbreaking ChatGPT on Release Day*, LESSWRONG (Dec. 2, 2022), https://www.lesswrong.com/posts/RYcoJdvmoBbi5Nax7/jailbreaking-chatgpt-on-release-day [https://perma.cc/2UAP-AD98].

74. *See* ChatGPT Plugins, OPENAI, https://openai.com/blog/chatgpt-plugins [https://perma.cc/7CJ5-LH3V].

75. Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz & Mario Fritz, More Than You've Asked For: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models 1 (Feb. 23, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2302.12173v1.pdf [https://perma.cc/Y8SV-LQ58].

76. *Id.* at 3.

77. ChatGPT Plugins, *supra* note 74.

78. Carlini et al., *supra* note 61.

79. *See* Bender et al., *supra* note 7 ("One of the biggest trends in natural language processing (NLP) has been the increasing size of language models (LMs) as measured by the number of parameters and size of training data."). In particular, increased model size results in improvement of zero-shot and few-shot performance of tasks, indicating that larger models

intensify.[80] Some research suggests that memorization is, in some respects, a feature rather than a bug: it enables models to learn factual information, and for some tasks, it may be crucial for generalization (i.e., a model's ability to react to unseen data and make accurate predictions).[81] Therefore, entirely eliminating memorization could degrade model performance, adding another layer of complexity to remedying this issue. Additionally, although the research discussed in this section focuses on memorization, information need not be memorized to be harmful. Exposure of any learned or inferred personal information poses a threat to privacy.

B. Technical Solutions to Data Leakage

Although training data extraction attacks have thus far been limited to research experiments, LLMs' susceptibility to data leakage poses a concern as commercial implementation becomes more common. Indeed, corporations have already advised employees not to divulge corporate secrets to ChatGPT, as these inputs might be incorporated into model training and therefore might be vulnerable to exposure in subsequent outputs.[82] Several privacy-preserving technical solutions seek to mitigate this vulnerability. Data sanitization and differential privacy, described below, are two such techniques. Though these privacy-preserving mechanisms show promise, even the most stringent privacy protocols cannot fully guarantee privacy.[83] Leading LLM developers, such as Google, Meta, OpenAI, and Deepmind, have all conceded that preserving privacy in LLMs is an ongoing challenge, and that the risks of these models are not yet fully understood.[84]

---

have a higher proficiency at in-context learning. *See* Brown et al., *supra* note 19, at 4–5; *see also* Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer & Armen Aghajanyan, *Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models*, 36 CONF. NEURAL INFO. PROCESSING SYS. (Nov. 2, 2022), https://arxiv.org/pdf/2205.10770.pdf [https://perma.cc/BD5Z-EM77] ("We have consistently seen performance gains by scaling model size.").

80. Brown et al., *supra* note 50, at 3 ("State-of-the-art language models require a significant amount of training data. The size of top models also increases by an order of magnitude every year. These factors significantly increase the privacy risks of language models.").

81. Tirumala et al., *supra* note 79, at 2 ("Recent work has argued that memorization is not exclusively harmful, and can be crucial for certain types of generalization (e.g., on QA tasks), while also allowing the models to encode significant amounts of world or factual knowledge.").

82. *See* Eugene Kim, *Amazon Warns Employees Not to Share Confidential Information with ChatGPT After Seeing Cases Where Its Answer "Closely Matches Existing Material" from Inside the Company*, BUS. INSIDER (Jan. 24, 2023), https://www.businessinsider.com/amazon-chatgpt-openai-warns-employees-not-share-confidential-information-microsoft-2023-1 [https://perma.cc/6D5K-Y36K].

83. Heikkilä, *supra* note 47; *see* Lukas et al., *supra* note 59.

84. *Id.* ("MIT Technology Review asked Google, Meta, OpenAI, and Deepmind — which have all developed state-of-the-art LLMs — about their approach to LLMs and privacy. All the companies admitted that data protection in large language models is an ongoing issue, that

1. Data Sanitization

Data sanitization techniques aim to remove private information from training datasets and, therefore, prevent memorization and leakage.[85] Data sanitization is a useful but imperfect method to preserve privacy. For example, it can effectively remove some personally-identifiable information and protected health information from training datasets.[86] However, data sanitization requires a narrow classification of privacy (i.e., designated words or phrases that are clearly defined and context independent) in order to classify and remove information tagged as private.[87] As a consequence, while data sanitization algorithms can reliably redact clearly classified information (such as regularly formatted social security numbers), they are less likely to capture abnormally presented information, and typically cannot address context-specific privacy concerns.[88] For instance, what one considers private varies based on contextual factors such as culture and audience, and one's private beliefs, feelings, thoughts, and behaviors described in text are difficult to neatly demarcate, tag, and redact. The massive volume and richness of data required to train LLMs compound these difficulties.

Even after de-identification and sanitization, it may still be possible to infer private information.[89] Broader redaction can provide stronger protection of privacy, but it also erodes the meaning and utility of text. Ultimately, "a hypothetically privacy-preserving data sanitization might result in removing almost all the text, rendering it useless."[90] Although data sanitization techniques are valuable, they are rooted in the assumption that "private information can be formally specified, easily recognized, and efficiently removed," which is not always the case.[91]

---

there are no perfect solutions to mitigate harms, and that the risks and limitations of these models are not yet well understood.").

85. Brown et al., *supra* note 50, at 10.

86. *Id.* at 11.

87. *Id.* at 10.

88. *Id.* ("For example, identifying the social security number '*the first 2 digits are two two, and the remaining ones are three . . .*' is much more challenging than identifying '223'").

89. Brown et al., *supra* note 50, at 11 ("This problem resembles the numerous failed attempts for anonymizing high-dimensional data by removing certain attributes. In the context of language data (with enormous number of dimensions), there is always a possibility of inferring sensitive information even if many pieces of text are redacted."). In a famous example of the deficiency of data anonymization, researchers were able to re-identify subscribers when Netflix released anonymized subscriber data in a contest. Steve Lohr, *Netflix Cancels Contest After Concerns Are Raised About Privacy*, N.Y. TIMES (Mar. 12, 2010), https://www.nytimes.com/2010/03/13/technology/13netflix.html [https://perma.cc/QY28-QF4D]; *see* Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 N.C. L. REV. 1417, 1420 (2012) ("[E]ven the most thorough redaction of personally identifiable information has generally been found insufficient to protect the privacy of individuals represented in data sets.").

90. Brown et al., *supra* note 50, at 11.

91. *Id.* at 1–2.

A related technique, known as deduplication, involves removing duplicates of data from the training dataset.[92] Duplicated information is more likely to be memorized, and thus, removing duplicates reduces memorization.[93] Although deduplication reduces the incidence of memorization in LLMs, it does not entirely prevent leakage.[94] Due to the massive size of training datasets, it is difficult to capture all duplicates and near-duplicates; thus, "any deduplication strategy is necessarily imperfect in order to efficiently scale to hundreds of gigabytes of training data."[95]

2. Differential Privacy

Differential privacy is a mathematical definition of privacy.[96] Numerous techniques have been developed to satisfy this standard.[97] In essence, differential privacy guarantees that the output of an analysis will be *nearly* the same, whether or not one's data is included in the input dataset; therefore, one cannot infer information specific to an individual.[98] The privacy loss parameter, typically denoted by epsilon (ε), quantifies and limits the level of privacy achieved (i.e., the extent of deviation between an analysis without an individual's information and one with his information).[99] Differentially private analyses achieve this result by adding noise (randomness) to the computation.[100] A higher degree of privacy, and therefore more noise injected into the computation, results in lower accuracy. Thus, differentially private analysis entails an inherent tradeoff between accuracy and privacy.

Although differential privacy techniques have been implemented to train LLMs, meaningfully differentially private models have been

92. Carlini et al., *supra* note 61, at 5 ("We observe a clear log-linear trend in memorization. While models rarely regurgitate strings that are repeated only a few times, this probability increases severely for highly duplicated strings . . . However, we find that memorization does still happen, even with just a few duplicates — thus, deduplication will not perfectly prevent leakage.").

93. *Id.*

94. *Id.*

95. *Id.* at 9.

96. *See* Alexandra Wood et al., *Differential Privacy: A Primer for a Non-Technical Audience*, 21 VAND. J. ENT. & TECH. L. 209 (2018); Andrea Scripa Els, *Artificial Intelligence as a Digital Privacy Protector*, 31 HARV. J.L. & TECH. 217, 220–22 (2017).

97. Nicholas Papernot & Abhradeep Guha Thakurta, *How To Deploy Machine Learning with Differential Privacy*, CYBERSEC. INSIGHTS (Dec. 21, 2021), https://nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy [https://perma.cc/Q2YF-Y89P].

98. *See* Wood et al., *supra* note 96, at 225–37; Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1137 (2013) ("[I]t is always theoretically possible that any information revealed by a data set is the missing link that the adversary needs to breach someone's privacy.").

99. Wood et al., *supra* note 96, at 234.

100. *Id.*

plagued by degraded performance and high computational overhead.[101] Moreover, differentially private models may still memorize content that is repeated often in the training data,[102] and the impact of degraded performance may disparately affect underrepresented groups, thereby magnifying the unfairness of models.[103]

While recent research shows promising results,[104] differential privacy is not a magic bullet, and it does not resolve the numerous privacy issues presented by LLMs, which reach beyond data leakage. Viewed under Professor Daniel Solove's taxonomy, these privacy violations include: *aggregation* (by compiling and analyzing data to make inferences that an individual likely does not anticipate), *identification* (by potentially linking information with an individual), *insecurity* (by creating the risk of downstream harm through the exposure of identifying information), *distortion* (by outputting fabricated or inaccurate personal information), *secondary use* (by using information for a purpose beyond the scope of initial consent), and *exclusion* (by providing little insight into what the model knows and limited ability to direct the use of one's data).[105] I explore these complications further in Part III.

III. THE CHALLENGE OF INFORMED CONSENT

This Part explores the deficiencies of the notice-and-choice paradigm that dominates privacy law. It then describes several characteristics of LLMs that underscore the limitations of this framework to adequately protect privacy. First, LLMs train on massive datasets

---

101. *See* Michael Veale et al., *Algorithms That Remember: Model Inversion Attacks and Data Protection Law*, 376 ROYAL SOC'Y PUBL'G 1, 11 (2018) ("Yet despite the growing interest in differential privacy, the real challenge comes with deployment. The tools available today can be computationally expensive to deploy as well [sic] easily undermined with even small or arcane software errors. Only a few large and powerful companies have demonstrated an ability to deploy them, and only then for very limited purposes."); Xuechen Li, Florian Tramèr, Percy Liang & Tatsunori Hashimoto, *Large Language Models Can Be Strong Differentially Private Learners*, 10 INT'L CONF. ON LEARNING REPRESENTATIONS (Jan. 28, 2022), https://openreview.net/pdf?id=bVuP3ltATMz [https://perma.cc/R23D-QXB4] ("[S]traightforward attempts at applying Differentially Private Stochastic Gradient Descent (DP-SGD) to NLP tasks have resulted in large performance drops and high computational overhead.").

102. Carlini, *supra* note 58 ("Even [differential privacy techniques] can have limitations and won't prevent memorization of content that is repeated often enough.").

103. Eugene Bagdasaryan, Omid Poursaeed & Vitaly Shmatikov, *Differential Privacy Has Disparate Impact on Model Accuracy*, 33 CONF. ON NEURAL INFO. PROCESSING SYS. (Oct. 27, 2019), https://papers.nips.cc/paper_files/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf [https://perma.cc/WQS9-7MYY] ("For example, a gender classification model trained using DP-SGD exhibits much lower accuracy for black faces than for white faces. Critically, this gap is bigger in the DP model than in the non-DP model, i.e., if the original model is unfair, the unfairness becomes worse once DP is applied.").

104. *See generally* Li et al., *supra* note 101 (describing three ways to mitigate performance drops as a result of differentially private learning in large language models).

105. Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 505–36 (2006).

scraped from the Internet. Although technically comprised of publicly available data, these training datasets nonetheless contain personal information. Processing, aggregating, and exposing this information can lead to privacy harms, yet U.S. law rarely recognizes the privacy interests in information exposed to the public. Even if U.S. law required notice in this context, it would be practically difficult to execute due to the massive size of training datasets. Second, LLMs are general purpose, and therefore can be implemented for a wide range of uses. Even when companies notify users and request consent for data collection and processing, the uncertainty of downstream applications complicates comprehensive disclosure, which ultimately undermines meaningful consent. Adequate disclosure of future uses will necessarily be incomplete, and unexpected or unforeseeable applications will violate the scope of initial consent. Third, the permanence of data imprints within the model complicates the fulfillment of core privacy rights. For instance, those who wish to withdraw consent and delete their data are left without recourse. Imprints of one's data may remain embedded in the model until it is trained from scratch without one's data, due to the difficulties of machine unlearning.

A. The Notice-and-Choice Paradigm

Informational privacy refers to the ability to control who collects your data and how that data is used.[106] In essence, it is "informational self-determination."[107] This conception of privacy typically requires procedural protections, such as informed consent.[108] For others to

106. Robert H. Sloan & Richard Warner, *Beyond Notice and Choice: Privacy, Norms, and Consent*, 14 J. HIGH TECH. L. 370, 370 (2014). This is far from the *only* definition of privacy, which is a concept marked by ambiguity and competing meanings. *Compare* Woodrow Hartzog, *What Is Privacy? That's the Wrong Question*, 88 U. CHI. L. REV. 1677, 1685–86 (2021) ("Scholars have proposed a remarkable array of ways to think and talk about different notions of privacy, including intellectual privacy, sexual privacy, quantitative privacy, and more. They have built out conceptualizations of privacy as obscurity, trust, power, privilege, security, safety, procedural due process, a civil or human right, and the contextual integrity of information flows."), *and* Daniel J. Solove, *"I've Got Nothing to Hide" and Other Misunderstandings of Privacy*, 44 SAN DIEGO L. REV. 745, 756 (2007) ("[P]rivacy is not reducible to a singular essence; it is a plurality of different things that do not share one element in common but that nevertheless bear a resemblance to each other."), *with* KENNETH A. BAMBERGER & DEIRDRE K. MULLIGAN, PRIVACY ON THE GROUND: DRIVING CORPORATE BEHAVIOR IN THE UNITED STATES AND EUROPE 24 (2018) ("[Some scholars] argue for definitions of privacy that are less wedded to liberal conceptions of the self, and more reflective of privacy's nature as a public, as well as private, good.").

107. BAMBERGER & MULLIGAN, *supra* note 106, at 21.

108. *See* Neil Richards & Woodrow Hartzog, *The Pathologies of Digital Consent*, 96 WASH. U. L. REV. 1461, 1462–63 (2019) ("Consent's power, its usefulness, and its resonance with norms of autonomy and choice make it an easy legal tool to reach for when we want to regulate behavior . . . Consent's power is particularly justified in cases of what we might call 'gold standard' consent — agreements between parties who have equal bargaining power, significant resources, and who *knowingly* and *voluntarily* agree to assume contractual or other legal obligations.").

collect or process your data, they must receive your intentional authorization, which requires adequate information and the absence of coercion or control.[109] In U.S. privacy law, the notice-and-choice paradigm is the predominant approach employed to address this issue, requiring either an "opt-in" or "opt-out" to data collection and processing.[110] This framework suffers from several deficiencies that some argue are fundamentally unworkable, evidently intractable, and fatally flawed, particularly as technological progress enables increasingly ubiquitous and granular data collection.[111] "Opt-out" choice — seen in the California Consumer Privacy Act ("CCPA")[112] — creates a baseline without protection and places the burden on the individual to protect himself. The "opt-out" regime incentivizes companies to increase the transaction costs incurred by those who wish to opt out in an effort to retain user data: the harder it is to opt out, the less likely people will do so.[113] Although it offers a more protective baseline, "opt-in" (affirmative) choice — as in the EU's General Data Protection Regulation[114] ("GDPR") and Illinois' Biometric Information Privacy Act[115] — fails to adequately inform and empower users, who (1) do not understand disclosures and suffer information fatigue as they navigate numerous

109. *See* Adam J. Andreotta, Nin Kirkham & Marco Rizzi, *AI, Big Data, and the Future of Consent*, 37 AI & SOC'Y 1715, 1716–17 (2021), https://doi.org/10.1007/s00146-021-01262-5 [https://perma.cc/8PSU-QWJ9]. The GDPR draws on this notion of informed consent: "Consent of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her." *See What Are the GDPR Consent Requirements?*, GDPR.EU, https://gdpr.eu/gdpr-consent-requirements [https://perma.cc/B852-JX5W].

110. *See* Margot E. Kaminski, *The Case for Data Privacy Rights (Or, Please, a Little Optimism)*, 97 NOTRE DAME L. REV. REFLECTION 385, 388 (2022); Scott Jordan, *A Proposal for Notice and Choice Requirements of a New Consumer Privacy Law*, 74 FED. COMM. L.J. 251, 254 (2021).

111. *See* Julie E. Cohen, *How (Not) to Write a Privacy Law*, KNIGHT FIRST AMEND. INST. COLUM. UNIV. (Mar. 23, 2021), https://knightcolumbia.org/content/how-not-to-write-a-privacy-law [https://perma.cc/8732-TVD5] ("[In modern privacy law] individual control rights function as the primary mechanism for governing the collection and processing of personal data, with no or only residual provision for ongoing governance at the collective level. Atomistic, post hoc assertions of individual control rights, however, cannot meaningfully discipline networked processes that operate at scale. Nor can they reshape earlier decisions about the design of algorithms and user interfaces.").

112. *See* California Consumer Privacy Act of 2018, CAL. CIV. CODE § 1798.120(a) (West 2022) [hereinafter CCPA].

113. *See* Jeff Sovern, *Opting in, Opting out, or No Options at All: The Fight for Control of Personal Information*, 74 WASH. L. REV. 1033, 1081–83 (1999).

114. *See* Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1 [hereinafter GDPR].

115. 740 ILL. COMP. STAT. 14/15(b)(3) (2008).

complex policies and (2) may feel coerced to accept undesirable terms because they lack bargaining power and require the service.[116]

The reality is that consumers rarely read complex terms of service, and those who have the competence to understand the terms scarcely have the time to do so.[117] Clicking "I agree" is less a signal of true consent than of consumer resignation in the face of maddeningly long, complex adhesion contracts that create obstacles to digital life.[118] In this sense, the procedural protections of privacy embodied in the notice-and-choice regime are shallow and ineffectual — empty formalities that primarily function to indemnify the data processor, rather than to empower the individual. In the well-documented phenomenon known as the "privacy paradox," users who claim to value privacy do little to protect it,[119] and measures intended to augment individual control lead people to disclose more sensitive information.[120] Atomistic conceptions of privacy also inadequately secure group privacy and fail to consider the ripple effects of individual choice: my choice to expose my private data may, in fact, reveal something private about *you*, too.[121] As

116. *See* Kaminski, *supra* note 110, at 388–89 ("Notice and choice is precisely not what it sounds like. Individuals are given little notice, and next to no choice."); Nancy S. Kim & D. A. Jeremy Telman, *Internet Giants as Quasi-Governmental Actors and the Limits of Contractual Consent*, 80 MO. L. REV. 723, 732–33 (2015); Grace Park, *The Changing Wind of Data Privacy Law: A Comparative Study of the European Union's General Data Protection Regulation and the 2018 California Consumer Privacy Act*, 10 U.C. IRVINE L. REV. 1455, 1473–76 (2020).

117. A paper from 2008 estimated that it would take an average of 201 hours per year for an individual to read the privacy policies of all the websites he visited in a year. Aleecia M. McDonald & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies*, 4 I/S: J.L. & POL'Y FOR INFO. SOC'Y 543, 565 (2008).

118. *See* John A. Rothchild, *Against Notice and Choice: The Manifest Failure of the Proceduralist Paradigm to Protect Privacy Online (Or Anywhere Else)*, 66 CLEV. ST. L. REV. 559, 559 (2018) (arguing that users ignore privacy disclosures due to "rational inattention") ("[W]e can only choose to engage with the online world, making our [personally-identifiable information] available for uses that we cannot understand or evaluate, or become hermits in self-exile from the online world."); Nora A. Draper & Joseph Turow, *The Corporate Cultivation of Digital Resignation*, 21 NEW MEDIA & SOC'Y 1824, 1825 (2019) (arguing that the privacy paradox is a result of consumer resignation, cultivated by obfuscating corporate practices that induce consumer frustration or create the illusion of control).

119. *See generally* Susanne Barth & Menno D.T. de Jong, *The Privacy Paradox — Investigating Discrepancies Between Expressed Privacy Concerns and Actual Online Behavior — A Systematic Literature Review*, 34 TELEMATICS & INFORMATICS 1038 (2017) (reviewing empirical evidence of the privacy paradox). For an exploration of why the privacy paradox emerges, see Richards & Hartzog, *supra* note 108, at 1465 ("Understanding [the privacy paradox] in terms of consent pathologies reveals that consumers are not hypocrites who say one thing but do another that reveals their true preferences. On the contrary, the pathologies of consent show how consumers can be nudged and manipulated by powerful companies against their actual interests, and this phenomenon is easier when the legal regime that purports to protect consumers falls far from the gold standard.").

120. *See* BAMBERGER & MULLIGAN, *supra* note 106, at 23.

121. Revelations from personal genetic testing exemplify the ripple effect of individual choice. *See id.* at 25 ("Through the [23andme] program a biologist discovered his unknown half-brother (a product of his father's infidelity), causing a massive family rift ending in his parents' divorce. While opt-in procedures offered the biologist a chance to consider what the

the richness, complexity, and granularity of data collection intensifies in a society fueled by surveillance capitalism,[122] sophisticated algorithms reveal latent insights that reach beyond those who choose to opt in.[123]

The complexities of LLMs, and their inherent opacity, make the already complicated issue of meaningful consent particularly thorny. The vulnerability presented by potential data leakage underscores the need for legal protection,[124] yet the massive scale of training data consumed by LLMs makes the identification of personal information and the attainment of consent uniquely burdensome. Even when notification is feasible, comprehensive disclosure about model implementation is difficult, if not impossible, to provide because the downstream applications of models may be unpredictable. Because disclosure is necessarily incomplete and the potential consequences of consent are unclear, individuals face a muddled, complex risk calculation that is not conducive to meaningful choice. Moreover, LLMs' potential capacity to *infer* information, even when an individual has not explicitly disclosed this information, further highlights the inadequacy of individual choice to ensure privacy in this context. Finally, due to the inability of LLMs to provably "forget" one's data, an individual might be left without a remedy if she wishes to withdraw her consent and delete the imprints of her data from the model. The potential permanence of one's consent thus creates another wrinkle. These attributes underscore the deficiencies of the notice-and-choice paradigm and highlight the need for regulation that looks beyond individual choice.

1. Privacy Interests in Public Training Data Are Underprotected

At least in the United States, individuals whose personal information is included in the publicly scraped text datasets that fuel the training of LLMs typically are not provided notice of this use or a

---

test might reveal about him, it offered no protection for his father or his newfound sibling."); Kaminski, *supra* note 110, at 393 ("We are all connected in this economy, so to conceive of data privacy only as a series of atomized hierarchical relationships between the watched and the watcher is to neglect the ways in which my choices impact yours. Group privacy, too, is underprotected by atomistic privacy rights; so is privacy in neighborhoods, and in communities historically targeted and surveilled."); Solon Barocas & Karen Levy, *Privacy Dependencies*, 95 WASH. L. REV. 555, 558 (2020) ("[*E*]*veryone's privacy depends on what others do*. There is no way to live in the world without putting yourself at risk that others might make use of information about you in ways to which you do not consent.").

122. *See* SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER 8 (2019).

123. *See* Cohen, *supra* note 111 ("Current approaches to crafting privacy legislation are heavily influenced by the antiquated private law ideal of bottom-up governance via assertion of individual rights, and that approach, in turn, systematically undermines prospects for effective governance of networked processes that operate at scale.").

124 *Supra* Section II.A.

choice in the matter.[125] Some public data is clearly intended for widespread public consumption and therefore entails diminished privacy interests. Such data might, for example, be pulled from Wikipedia, published books, newspaper articles, and commercial websites. But other information posted online is far more personal in nature. Although U.S. law typically insists there is "no privacy in public," the Internet — although technically "public" — is an archive of our most personal, intimate thoughts and experiences.[126] Revealing this data — even if it is already publicly accessible — can still result in harm, particularly when it is aggregated or analyzed to produce new insights.[127] This harm is exemplified by the phenomenon of doxing: even if the data is publicly available, compiling and exposing personal information can cause the victim humiliation and anxiety, and enable stalking, extortion, identity theft, harassment, and violence.[128] In tension with this reality, the notion that information exposed to the public has diminished (or no) privacy interest is ingrained in Fourth Amendment and privacy tort jurisprudence, and it is echoed in U.S. privacy regulation.[129] For instance, the California Privacy Rights Act's ("CPRA") protection for personal information excludes "publicly available" information, which is expansively defined.[130]

---

125. Publicly available information is typically excluded from coverage under U.S. privacy laws. *See infra* note 130.

126. *See* Woodrow Hartzog, *The Public Information Fallacy*, 99 B.U. L. REV. 459, 462 (2019) ("The concept of 'public information and acts' is entrenched in U.S. law and policy. Tort law, statutes, and interpretations of constitutional amendments regularly deploy the concept of 'public information' to justify surveillance or data practices.").

127. *See* Geoffrey Xiao, *Bad Bots: Regulating the Scraping of Public Personal Information*, 34 HARV. J.L. & TECH. 701, 706 (2021) ("Put simply, the privacy harms associated with public personal information are as substantial as those associated with private personal information."). For instance, leaks of aggregated scraped public data containing personal information can enable financial fraud and identity theft. *See* Tara Seals, *Millions of Social Profiles Leaked by Chinese Data-Scrapers*, THREATPOST (Jan. 11, 2021, 4:54 PM), https://threatpost.com/social-profiles-leaked-chinese-data-scrapers/162936 [https://perma.cc/XK46-PP7Q].

128. *See* Jasmine McNealy, *What Is Doxxing, and Why Is it So Scary?*, THE CONVERSATION (May 16, 2018, 6:26 AM), https://theconversation.com/what-is-doxxing-and-why-is-it-so-scary-95848 [https://perma.cc/QX8R-KCNH].

129. Woodrow Hartzog & Evan Selinger, *Surveillance as Loss of Obscurity*, 72 WASH. & LEE L. REV. 1343, 1349 (2015) ("Courts and policy-makers regularly affirm that there is no 'privacy in public.'"); *see* HiQ Labs, Inc. v. LinkedIn Corp., 938 F.3d 985, 994 (9th Cir. 2019) ("[T]here is little evidence that LinkedIn users who choose to make their profiles public actually maintain an expectation of privacy with respect to the information that they post publicly, and it is doubtful that they do."). Notably, the Supreme Court recently indicated a potential breakdown of this paradigm, at least as it applies to the third-party doctrine. *See* Carpenter v. United States, 138 S. Ct. 2206, 2217 (2018) ("Given the unique nature of cell phone location records, the fact that the information is held by a third party does not by itself overcome the user's claim to Fourth Amendment protection."); *see also id.* at 2268–70 (Gorsuch, J., dissenting) (comparing an individual's interest in information provided to a third party to a type of bailment).

130. CAL. CIV. CODE § 1798.140(v)(2) (2018). Other privacy statutes, like the Colorado Privacy Act and Virginia's Consumer Data Protection Act similarly exclude publicly

The concept of "public information" is imprecise and vague, and the blanket, contextless notion of "no privacy in public" is particularly ill-suited for the Internet.[131] The fact that a piece of information finds its way to the Internet does not guarantee that the affected individual *intended* to release that information for unbound public consumption. The information may have been shared without authorization by another person, either benignly (e.g., sharing a picture of a friend without asking for permission) or maliciously (e.g., leaking information through a data breach). Even if the individual at one point approved of its public dissemination, this may no longer be the case, and removing all traces of that information from the Internet can be nearly impossible.[132] Second, the individual may have consented to the sharing and viewing of information in a specific context or for foreseeable uses, without recognizing that this data might be exploited by third parties.[133] Most people who publicly post on social media, for instance, likely do not expect that this information might be scraped and used to train a neural network.[134] Third, training data that incorporates an older iteration of the Web will contain pieces of information that individuals have since made private or deleted. An LLM trained on this since-deleted data will "remember" and incorporate this data in its output, despite the fact that these individuals have effectively withdrawn consent.[135]

Fourth, LLMs may piece together information in unexpected ways, thereby connecting information posted in discrete online spaces — perhaps combined with nonpublic data — to reveal deeply personal insights about individuals. LLMs are marked by emergent behaviors for

---

available data. *See* Andrew M. Parks, *Unfair Collection: Reclaiming Control of Publicly Available Personal Information from Data Scrapers*, 120 MICH. L. REV. 913, 933 (2022) ("[B]usinesses scraping publicly available personal data remain unregulated even by the most expansive state data privacy laws.").

131. *See* Hartzog, *supra* note 126, at 465 ("At worst, appeals to the public nature of information and acts provide cover for unscrupulous and dangerous data practices and surveillance by making it seem as though there is some objective and established criteria for what constitutes public information. There is no such consensus."); Michael Zimmer, *"But the Data Is Already Public": On the Ethics of Research in Facebook*, 12 ETHICS & INFO. TECH. 313, 323 (2010) ("[F]uture researchers must gain a better understanding of the contextual nature of privacy in these spheres, recognizing that just because personal information is made available in some fashion on a social network, does not mean it is fair game for capture and release to all.").

132. *See* Catherine Thorbecke, *Why Deleting Something from the Internet Is 'Almost Impossible'*, CNN BUS. (Sept. 18, 2022, 8:44 PM), https://www.cnn.com/2022/09/18/tech/deleting-data/index.html [https://perma.cc/7LSA-MPNG].

133. *See* Xiao, *supra* note 127, at 711 ("[E]ven if a user makes the affirmative choice to make her LinkedIn profile public, she manifests an intent to participate in an obscure and trustworthy environment, not an intent to participate in data harvesting.").

134. *See* Casey Fiesler & Nicholas Proferes, *"Participant" Perceptions of Twitter Research Ethics*, 4 SOC. MEDIA + SOC'Y, 1, 2 (2018) (finding that most Twitter users surveyed were unaware that their public tweets can be used by researchers).

135. *See* Cohen, *supra* note 111 ("Additionally, it's not clear what the right to revoke consent means in the context of machine-learning-based models trained on a large corpus that includes the to-be-withdrawn data.").

which they have never been trained and which they are *not* programmed to do. That is to say, as LLMs scale, they do not merely become *better* at tasks; rather, new tasks can suddenly and unpredictably become possible.[136] It is theoretically possible that future LLMs might draw connections about the individuals included in their training data, possibly even connecting discrete data from various sources (including those not explicitly linked to the individual). In general, deep learning models are known to make unexplainable, surprising connections. For instance, one study showed that a deep learning model trained on medical data accurately predicted the race of patients, even when the medical images were corrupted and noised, despite the fact that human experts were unable to do so with the same information.[137] Moreover, even if a piece of information is not present in the training dataset and an individual has taken care *not* to expose this information to the public, LLMs' capacity for inference further complicates this issue. Actors can exploit a model's predictive capabilities to create "detailed profiles of individuals comprising true and sensitive information without the knowledge or consent of the individual."[138] If believed, inaccurate inferences can likewise do harm: for instance, by affecting one's reputation, causing discrimination, or otherwise influencing the trajectory of one's life (e.g., employment, access to credit, etc.).

Even if these connections do not materialize from emergent behavior, an actor with a desire to extract these connections can attempt to fine-tune an LLM to learn this skill. In either case, an individual who consents to the viewing of a single data point online does not automatically consent to revelations produced by the amalgamation of *all* her digital imprints, which might reveal attributes and patterns undetectable when viewed independently. This is particularly true when a poster reasonably believes that information cannot be connected to her identity, such as when she posts from an anonymous account. As many studies have demonstrated, anonymity is easily unraveled.[139] Further complicating this already unsettling issue, due to LLMs' propensity for

136. Jason Wei et al., *Emergent Abilities of Large Language Models*, TRANSACTIONS ON MACH. LEARNING RSCH. (Aug. 2022), https://openreview.net/pdf?id=yzkSU5zdwD [https://perma.cc/2K4T-KK44] ("[E]mergent abilities show a clear pattern — performance is near-random until a certain critical threshold of scale is reached, after which performance increases to substantially above random.").

137. *See* Rachel Gordon, *Artificial Intelligence Predicts Patients' Race from Their Medical Images*, MIT NEWS (May 20, 2022), https://news.mit.edu/2022/artificial-intelligence-predicts-patients-race-from-medical-images-0520 [https://perma.cc/7YHG-E4DD].

138. *See* Weidinger et al., *Taxonomy of Risks Posed by Language Models*, 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY, 217–18, https://dl.acm.org/doi/pdf/10.1145/3531146.3533088 [https://perma.cc/6SU9-AAWA].

139. Research has shown that, with as few as fifteen attributes (such as gender, zip code or marital status), it is possible to identify 99.98% of Americans from "anonymized" datasets. Gina Kolata, *Your Data Were 'Anonymized'? These Scientists Can Still Identify You*, N.Y. TIMES (July 23, 2019), https://www.nytimes.com/2019/07/23/health/data-privacy-protection.html [https://perma.cc/72XD-AZLX].

fabrication, they may generate misinformation about the individuals included in their training data.[140]

Hypothetically, if the law did broadly recognize a privacy interest in personal information publicly posted online and required opt-in consent by each individual, AI researchers[141] would face a considerable hurdle. The process of acquiring informed consent for all personal information (for which the scope is unclear) on massive scrapes of the Internet would be resource intensive, and in some cases, impracticable if information cannot be conclusively matched with an individual or if contact information is unobtainable. This costly process might thwart advancements in AI. The race to AI supremacy has profound social, economic, and national security implications, and overly burdensome or technically infeasible consent requirements will put American researchers at a disadvantage to Chinese researchers, who are not only unencumbered by privacy concerns but are *actively assisted* by the rich datasets compiled by the Chinese surveillance state.[142] If we instead seek to prevent Web scraping via government regulation[143] or private enforcement of restrictive website terms, this similarly undermines the compilation of rich datasets necessary to advance AI.[144] Beyond economic and national security concerns, one should also be apprehensive about AI dominance by actors who will not adequately consider AI safety. There is no easy answer to this problem, and it will require a

---

140. For example, Facebook's recently released Galactica model — an LLM intended to "store, combine and reason about scientific knowledge" — generates fabricated and misleading information, including attributing real authors to fake research papers and generating articles that were "authoritative-sounding and believable" but completely fabricated. Janus Rose, *Facebook Pulls its New 'AI For Science' Because It's Broken and Terrible*, VICE: MOTHERBOARD (Nov. 18, 2022, 9:41 AM), https://www.vice.com/amp/en/article/3adyw9/facebook-pulls-its-new-ai-for-science-because-its-broken-and-terrible [https://perma.cc/5S6W-K2W3]. *See generally* Ziwei Ji et al., Survey of Hallucination in Natural Language Generation, ACM Computing Survs., Mar. 2023, at 1, 1.

141. This would, of course, also create major barriers for all researchers who publicly scrape information from the Internet.

142. GRAHAM ALLISON & ERIC SCHMIDT, IS CHINA BEATING THE U.S. TO AI SUPREMACY? 7, 11 (2020) ("[T]he Party has given China's top four facial recognition firms access to its database of over 1.4 billion citizen photos . . . China's government, laws and regulations, public attitudes about privacy, and thick cooperation between companies and their government are all green lights for its advance of AI. In the United States and Europe, yellow and red lights abound."). However, some argue this trend has changed: China's hardening censorship practices and increasing control of the private sector may, in fact, be thwarting China's progress in AI. *See* Li Yuan, *Why China Didn't Invent ChatGPT*, N.Y. TIMES (Feb. 17, 2023), https://www.nytimes.com/2023/02/17/business/china-chatgpt-microsoft-openai.html [https://perma.cc/TN5D-334R].

143. For a discussion of First Amendment objections to public data scraping regulation, see Xiao, *supra* note 127, at 727–31.

144. Some researchers are concerned that there may be a dearth of high-quality training data for LLMs in the near future. *See* Tammy Xu, *We Could Run Out of Data To Train AI Programs*, MIT TECH. REV. (Nov. 24, 2022), https://www.technologyreview.com/2022/11/24/1063684/we-could-run-out-of-data-to-train-ai-language-programs [https://perma.cc/PWX3-F6SA].

careful — and likely imperfect — balancing between privacy and research interests, which this Note will endeavor to address in Part IV.

2. The Uncertainty of Downstream Uses Complicates Adequate Notice and Undermines Consent

Putting aside the issue of publicly available data, when a company directly collects and processes nonpublic user data to train an LLM, providing adequate disclosure is still far from straightforward. A standard notice-and-choice scheme will be deficient due to the peculiarities of LLMs and the unpredictability of downstream applications. This incomplete information complicates one's decision calculus and undermines meaningful choice. Even if one agrees to a broad range of uses, unanticipated applications will violate the scope of initial consent.

Consider the following example: "AI Lab" builds an LLM. It licenses this model to various companies that fine-tune the model using their own custom user data to improve performance on relevant tasks. Before training the LLM, assume "AI Lab" obtains explicit consent from all users for the processing of their data to train the model. If a true guarantee of informational privacy requires the power to control *how* your data is used, the user must also know and approve of the future uses of that model. "AI Lab" cannot provide this information for two primary reasons: in some instances, the model may be capable of tasks that were not predictable before training, and more importantly, the lab will not know who its future clients are or how they intend to use the model. While a user might be happy to contribute her data to train a model that will be used for music recommendations, she might not consent to train a model that will be weaponized by a defense contractor. As pretrained models become more generalizable, the range of use cases becomes enormous.

Some researchers suggest that informed consent is very challenging to achieve in this context.[145] Even machine-learning experts do not have a full understanding of the true risk of data memorization and extraction,[146] and "even principled approaches such as differential privacy cannot provide privacy guarantees that are directly interoperable with the privacy *expectations* users might have for their text data."[147] Others believe that differential privacy, at least in some respects,

---

145. *See, e.g.*, Brown et al., *supra* note 50, at 13.

146. *Id.* ("[E]ven experts on ML privacy currently only have a partial understanding of the risks of data memorization and extraction."); *see also* Heikkilä, *supra* note 47.

147. Brown et al., *supra* note 50, at 13–14; *see also* Mirko Forti, *The Deployment of Artificial Intelligence Tools in the Health Sector: Privacy Concerns and Regulatory Answers Within the GDPR*, EUR. J. LEGAL STUD., 29, 38 (2021) (arguing that consent, as defined by the GDPR, is problematic in the field of predictive medicine, because it is "not possible for data subjects to know all the specific features of the processing activities when they provide consent").

provides *stronger* protection than a notice-and-choice framework because it functions as a mathematical opt-out by guaranteeing that an individual's data has virtually no effect on the output of the analysis.[148]

Ultimately, however, differential privacy does not adequately address a person's objections to the *use* of the model because it cannot guarantee that the model will not cause harm in its implementation. Even if a model were "perfectly" differentially private (i.e., a privacy loss parameter of virtually zero), and there was provably near-zero risk that one's data could be identified or leaked, one could still reasonably object to contributing data to train a model which might be used for unsavory purposes or that might exhibit troubling bias. In other words, if one contributes to a model employed for a purpose one finds ethically abhorrent, the mathematical differential privacy guarantee — that one's individual contribution cannot be inferred from the model's output — would be of little consolation.

This issue exemplifies the problem of "secondary use," where data is collected for one purpose but used for another without the individual's consent.[149] Use that exceeds the scope of initial consent betrays an individual's expectations and thereby deflates the power, meaning, and utility of consent. As a result, secondary use causes uncertainty, distrust, and disempowerment. If companies can simply use data for purposes beyond those explicitly approved, consent is lifeless and hollow, and individuals lack agency. The power to direct the use of one's data is therefore critical to informational self-determination and autonomy.

Assuming that developers of LLMs obtain consent for all uses of participant data to sidestep the unpredictability of downstream applications,[150] there still remains a troubling issue: the difficulty presented by "machine unlearning" when a participant wishes to withdraw her consent and delete her data.

3. The Permanence of Data Imprints Undermines Core Privacy Rights

The right to delete personal information that a data processor has collected — alternatively referred to as the right to delete, the right of erasure, or the right to be forgotten — is a core feature of privacy law. Per Article 17 of the GDPR, an individual has "the right to obtain from the controller the erasure of personal data concerning him or her

148. *See* Wood et al., *supra* note 96, at 264 ("The differential privacy guarantee can arguably be interpreted as providing stronger privacy protection than a consent or opt-out mechanism.").

149. Solove, *supra* note 105, at 519–20.

150. This may run afoul of the purpose limitations in some privacy laws, such as the GDPR and CPRA. *See* Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 SETON HALL L. REV. 995, 1004–09 (2017). The CPRA also limits data processing to those purposes compatible with the original disclosed purpose. CAL. CIV. CODE § 1798.100(c) (West 2022).

without undue delay."[151] Although the scope differs, California's CCPA provides for a similar right, stating that "a consumer shall have the right to request that a business delete any personal information about the consumer which the business has collected from the consumer."[152] Other state privacy laws (including the Colorado Privacy Act, Utah Privacy Act, and Virginia's Consumer Data Protection Act[153]) and the leading proposed federal privacy bill (the American Data Privacy and Protection Act) likewise include a right to deletion.[154]

This right is closely connected with individual empowerment and autonomy. It allows the individual, who might have provided consent in the past, to reconsider her decision and destroy imprints of this prior choice. It ensures that she will not be "perpetually or periodically stigmatized as a consequence of a specific action performed in the past."[155] In this way, the right of erasure is "the right to have an imperfect past."[156]

Deep learning (as employed in LLMs) complicates compliance with this right.[157] Training data is embedded in these models in ways that are unknown even to the experts that build them.[158] To provably unlearn a data point, one must first identify the contributions of that

---

151. GDPR, *supra* note 114, art. 17. Since the May 2014 judgment by the European Court of Justice upholding the right to erasure, Google has received over 1.4 million requests to delist URLs from its search engine results. *See Requests to Delist Content Under European Privacy Law*, GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/eu-privacy/overview [https://perma.cc/JKY5-B5QZ].

152. CAL. CIV. CODE § 1798.105(a) (West 2022).

153. *See Data Privacy Laws: What You Need to Know in 2023*, OSANO (Dec. 14, 2022), https://www.osano.com/articles/data-privacy-laws [https://perma.cc/K9DE-95CJ].

154. The American Data Privacy and Protection Act, H.R. 8152, 117th Cong. §§ 203(a)(3) (2022).

155. Alessandro Mantelero, *The EU Proposal for a General Data Protection Regulation and the Roots of the 'Right to Be Forgotten,'* 29 COMPUT. L. & SEC. REV. 229, 230 (2013).

156. Suzanne Moore, *The Right To Be Forgotten Is the Right To Have an Imperfect Past*, GUARDIAN (Aug. 7, 2017), https://theguardian.com/commentisfree/2017/aug/07/right-to-be-forgotten-data-protection-bill-ownership-identity-facebook-google [https://perma.cc/64UX-3BCX].

157. It is unclear whether the right to erasure, as expressed in the GDPR, applies to inferences that are made through machine learning or the model itself. For a discussion of the GDPR's right to erasure in the context of machine learning, see Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 68–72 (2017). *See also* Aleksandr Kesa & Tanel Kerikmäe, *Artificial Intelligence and the GDPR: Inevitable Nemeses?*, 10 TALTECH J. EUR. STUD. 67, 79–81 (2020) (discussing the tension between machine learning and the right to erasure). *See generally* Tiago Sergio Cabral, *Forgetful AI: AI and the Right to Erasure Under the GDPR*, 6 EUR. DATA PROT. L. REV. 378 (2020) (discussing at what stages data might be subject to the right to erasure in the context of ML). For further discussion of the friction between the right to erasure and AI models, see Tiffany Li, Eduard Fosch Villaronga & Peter Kieseberg, *Humans Forget, Machines Remember: Artificial Intelligence and the Right To Be Forgotten*, 34 COMPUT. L. & SEC. REV. 304, 310 (2018).

158. *See* Tom Simonite, *Now That Machines Can Learn, Can They Unlearn?*, WIRED (Aug. 19, 2021, 7:00 AM), https://wired.com/story/machines-can-learn-can-they-unlearn [https://perma.cc/ZL85-643U].

data point to the model, but due to the nature of deep learning, this is exceptionally difficult.[159] Even if your data is eliminated from the company's database,[160] removed from training data, and not incorporated into future training, there are traces of your data in the existing model. Reliably identifying and removing those traces is a technically challenging (and sometimes impossible) task.[161] The stubborn, inscrutable memory of LLMs undermines the actions of those who wish to delete a data footprint and who take steps to preserve their privacy. In effect, it deprives those who once consented of a meaningful way to withdraw that consent.

Consider the following example: Jane is very conscious of her online privacy, so she wisely sets her social media page to "Private." Only her followers can see her posts. The social media platform uses Jane's data to train an LLM, as permitted by its privacy policy. Jane decides that she is embarrassed by her tirades in past posts, and she deletes her entire account. Perhaps, if she is located in an area that provides her the right to delete her data, she contacts the social media platform requesting that all her personal information be deleted. Jane is satisfied that she has erased the remnants of her past misjudgment. However, the LLM that was trained on her data has *not* forgotten this information. In fact, for as long as that LLM exists, it will reflect Jane's data, unless it is trained from scratch without it. Jane may have no recourse if the LLM is implemented in ways she finds distasteful or unethical (either by the social media platform or perhaps a new owner), or if the model generates embarrassing or incorrect information about her.

And, of course, the data at issue may be far more sensitive than regrettable social media posts. While there have been significant advancements in the field of "machine unlearning," there is currently no method of removing one's imprints from a model with absolute, provable certainty, except for retraining the model from scratch.[162]

---

159. *See id.*; *see also* Lucas Bourtoule et al., *Machine Unlearning*, 42 IEEE SYMP. ON SEC. AND PRIV., 141, 143–44 (2021).

160. Although this might seem like a simple task, strict erasure in this context presents difficulties for modern relational database management systems. *See* Li et al., *supra* note 157, at 308–09.

161. *See* Simonite, *supra* note 158; *see also* Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri & James Zou, *Approximate Data Deletion from Machine Learning Models*, 24 INT'L CONF. ON A.I. & STAT. (2021), https://arxiv.org/abs/2002.10077 [http://perma.cc/2UPS-YCXF] ("The challenge here is that even after an organization deletes the data associated with a given individual, information about that individual may persist in predictions made by machine learning models trained on the deleted data. These predictions may in turn leak information, impeding the individual's ability to truly be 'forgotten.'").

162. *See* Simonite, *supra* note 158, at 2; *see also* Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin & Quoc Viet Hung Nguyen, A Survey of Machine Unlearning (Oct. 21, 2022) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2209.02299 [https://perma.cc/Y425-6NJB]. Some research suggests that machine unlearning may actually create unintended privacy risks, increasing the model's

Requiring companies to retrain models from scratch upon every erasure request presents enormous practical difficulties. First, retraining a model can be prohibitively expensive (potentially millions of dollars), and it can also be time consuming, which may lead to problematic downtime.[163] Second, there is a substantial cumulative environmental impact to this policy. Training LLMs requires significant energy use, and if this policy were adopted on a national or international scale, the result might accelerate climate change.[164] Third, if a pretrained model is licensed (and then fine-tuned for specific downstream tasks by the licensee), reliably deleting traces of one's data might require both the licenser and the licensee to retrain the model, exacerbating the financial and environmental impact.

The complexities of deep learning similarly complicate fulfillment of other core privacy rights, such as the right to know and the right to correct. The CCPA's right to know, for instance, grants a consumer the right to request that a business disclose the personal information it has collected about her.[165] The black-box nature of deep learning, however, makes it difficult to determine what the model "knows" and how specific training data points have influenced the model.[166] Similarly, it is uncertain how the CPRA's right to correct — which gives the consumer "the right to request a business that maintains inaccurate personal information about the consumer to correct that inaccurate personal information"[167] — applies to LLMs that were trained on inaccurate personal data, or those that subsequently output inaccurate personal data. LLMs' propensity for fabrication — even when the training data contains no inaccuracies — makes this issue particularly relevant.[168] Due to the limited interpretability of deep learning neural networks, identifying

---

vulnerability to data extraction and membership inference attacks. *See generally* Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert & Yang Zhang, *When Machine Unlearning Jeopardizes Privacy*, CCS '21: PROC. 2021 ACM SIGSAC CONF. ON COMP. & COMMC'NS SEC., 896.

163. *AI21 Labs Asks: How Much Does It Cost To Train NLP Models?*, SYNCED (Apr. 30, 2020), https://syncedreview.com/2020/04/30/ai21-labs-asks-how-much-does-it-cost-to-train-nlp-models [https://perma.cc/S7ET-WQTL].

164. *See* Elsabet Jones & Baylee Easterday, *Artificial Intelligence's Environmental Costs and Promise*, COUNCIL ON FOREIGN RELS. BLOG (June 28, 2022, 11:30 AM), https://cfr.org/blog/artificial-intelligences-environmental-costs-and-promise [https://perma.cc/9M59-9M5Y] ("Training a single AI system can emit over 250,000 pounds of carbon dioxide.").

165. CAL. CIV. CODE § 1798.110 (West 2022). The GDPR provides a similar right. GDPR, *supra* note 114, art. 17.

166. *See* Steve Neale, *Probing the Black Box: What Do Language Models Know, and Why Does It Even Matter?*, AMPLYFI (Oct. 21, 2022), https://amplyfi.com/2022/10/21/probing-the-black-box [https://perma.cc/S8LJ-E46M]; Feng-Lei Fan, Jinjun Xiong, Mengzhou Li & Ge Wang, *On Interpretability of Artificial Neural Networks: A Survey*, IEEE TRANSACTIONS ON RADIATION AND PLASMA MED. SCI. (Mar. 2021), https://arxiv.org/pdf/2001.02522.pdf [https://perma.cc/A889-SQTB].

167. CAL. CIV. CODE § 1798.106(A) (West 2022). The GDPR's right to rectification provides a similar right. GDPR, *supra* note 114, art. 16.

168. *See* Johnson, *supra* note 8.

how inaccurate training data influenced the model, reliably deleting or correcting this data within the model, or determining why the model outputted inaccurate data that contradicts its training data will likely be a difficult task.[169] The opacity of deep learning models like LLMs thus makes fulfillment of these rights technically challenging, and in some instances, infeasible.

IV. RECOMMENDATIONS

LLMs epitomize the features of modern data collection and processing that undercut the power of individual choice to adequately protect privacy under the notice-and-choice paradigm. In the modern digital ecosystem, granular data is collected on an enormous scale; individuals cannot accurately gauge the consequences of contributing data; and data is subsequently used in ways that most people do not anticipate. These features collectively diminish the effectiveness of individual choice to safeguard privacy. In the case of LLMs, the privacy risk calculus is particularly convoluted: the true risk of data leakage is unknown even to AI experts, the downstream uses of data are unpredictable, and one's choices leave permanent imprints. Under these conditions, individual choice — particularly in a one-sided ecosystem that so easily manipulates, overwhelms, and cajoles — is not sufficient to protect privacy. Simply put, longer privacy policies and more "I agree" buttons are inadequate to safeguard the privacy of individuals who contribute to the training of LLMs.

That is not to say that we should abandon choice entirely. Under transparent conditions, there is reason to believe that individual choice can be powerful; for example, under Apple's App Tracking Transparency framework, eighty percent of users (as of 2021) opted out of app tracking.[170] Therefore, regulators should prohibit deceptive practices, like dark patterns, that undermine meaningful consent and manipulate user behavior, and encourage clear consent options that default to privacy-protecting settings.[171] Where notice is feasible, LLM developers

169. *See supra* text accompanying note 166; Matt Burgess, *ChatGPT Has a Big Privacy Problem*, WIRED (Apr. 4, 2023), https://www.wired.com/story/italy-ban-chatgpt-privacy-gdpr/ [https://perma.cc/4US3-7P2Z] ("But deleting something from an AI system that is inaccurate or that someone doesn't want there may not be straightforward — especially if the origins of the data are unclear.").

170. *See* Brian X. Chen, *The Battle for Digital Privacy Is Reshaping the Internet*, N.Y. TIMES (Sept. 21, 2021), https://www.nytimes.com/2021/09/16/technology/digital-privacy.html [https://perma.cc/ZM39-TAU7] ("Since Apple released the pop-up window, more than 80 percent of iPhone users have opted out of tracking worldwide, according to ad tech firms.").

171. *See generally* Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger & Lalana Kagal, *Dark Patterns After the GDPR: Scraping Consent Pop-ups and Demonstrating Their Influence*, CHI '20: PROC. 2020 CHI CONF. ON HUMAN FACTORS IN COMPUTING SYS.; *see also FTC Report Shows Rise in Sophisticated Dark Patterns Designed to Trick and Trap*

should provide clear, succinct disclosures to empower meaningful choice. Transparent disclosures can also be used to preserve privacy: for instance, by advising users not to reveal personally-identifiable information in interactions with LLMs (as Google does in its privacy notice for Bard[172]). Yet individual choice is no panacea, and it should not be the only privacy safeguard in this context.[173] To minimize privacy violations, privacy protections must be embedded into the design and implementation of LLMs. This Part offers several suggestions that seek to do this.

This section does not purport to offer comprehensive recommendations on this subject; instead, these suggestions are intended to spark an important conversation about an emerging, rapidly-advancing technology that may have sweeping societal consequences. Any regulation in this context requires consensus about what privacy harms we seek to avoid and necessitates close cooperation between the legal and technical communities.

A. Clarify Existing Legal Obligations

Regulators should clarify the steps developers must take to comply with existing laws when training and deploying LLMs. As an example, it is unclear whether current training methods for LLMs — which rely on massive scraped datasets that include personal data[174] — comply with the GDPR.[175] Although scraping publicly available information is unlikely to violate most U.S. privacy laws,[176] the same is not necessarily true under the GDPR. Even if the personal data collected is

---

Consumers, FED. TRADE COMM'N (Sept. 15, 2022), https://www.ftc.gov/news-events/news/press-releases/2022/09/ftc-report-shows-rise-sophisticated-dark-patterns-designed-trick-trap-consumers [https://perma.cc/5HX4-39ZJ].

172. *See Manage & Delete Your Bard Activity*, BARD HELP, https://support.google.com/bard/answer/13278892#zippy=%2Cwho-has-access-to-my-bard-conversations [https://perma.cc/Y6KJ-MV3C] ("Important: Do not include info that can be used to identify you or others in your Bard conversations.").

173. *See* Cohen, *supra* note 111 ("Effective privacy governance requires a model organized around problems of design, networked flow, and scale.").

174. The GDPR covers personal data, which includes any piece of information that relates to an identifiable person. Unlike the CCPA, there is no exemption for publicly available information. *See What Is Considered Personal Data Under the EU GDPR?*, https://gdpr.eu/eu-gdpr-personal-data [https://perma.cc/N6FD-8CS3]; GDPR, *supra* note 114, art. 4.

175. *See* Burgess, *supra* note 169.

176. Notably, there are exceptions. For instance, recent BIPA lawsuits have challenged publicly scraped datasets of images subsequently processed for facial recognition, where inadequate opt-in consent was obtained. *See* Ryan Mac & Kashmir Hill, *Clearview AI Settles Suit and Agrees To Limit Sales of Facial Recognition Database*, N.Y. TIMES (May 9, 2022), https://www.nytimes.com/2022/05/09/technology/clearview-ai-suit.html [https://perma.cc/86XW-5TUG]; *see also* Travis LeBlanc, Bethany Lobo & Michael Rhodes, *Here's How To Prepare for the Leap in Biometric Privacy Lawsuits*, BLOOMBERG L. (Jan. 4, 2023), https://news.bloomberglaw.com/us-law-week/heres-how-to-prepare-for-the-leap-in-biometric-privacy-lawsuits [https://perma.cc/942G-KEXN].

publicly available and not directly obtained, a data controller must have a lawful basis for processing that data, and unless an exception or exemption applies (e.g., doing so proves impossible or requires disproportionate effort), the data controller must notify the individuals about the data collected.[177] For example, in 2019, the Polish Supervisory Authority ("SA") fined a company €220,000 for failing to notify individuals after processing contact data scraped from public registries.[178] In that case, the SA was unpersuaded by the company's "disproportionate effort" defense.[179] It is unclear to what extent LLM developers can claim this exemption.

Italy's recent temporary ban of ChatGPT and the international wave of investigations into OpenAI's data processing practices underscore the urgency of this issue.[180] The Italian Data Protection Authority (Garante) has questioned whether OpenAI has a legal basis to process the personal information swept up in its massive training datasets.[181] This issue extends beyond OpenAI. Indeed, it is applicable to all LLM developers who utilize comparable training datasets. Regulators must clarify how existing law applies to LLMs' "Internet scale" training datasets, which invariably include publicly available personal information.[182]

Likewise, there remains ambiguity about how core privacy rights — such as the right to delete, know, and correct — apply to LLMs.[183] Central to this issue is whether data embedded in LLMs and outputs generated by LLMs constitute personal information subject to these rights. In a public comment to the California Privacy Protection Agency, researchers at Stanford University urged that the CPRA should explicitly state that consumer rights to delete, know, and correct extend

---

177. GDPR, *supra* note 114, art. 14; *see also Right To Be Informed*, ICO., https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/individual-rights/right-to-be-informed [https://perma.cc/TJA9-QZLL] ("If you obtain personal data from publicly accessible sources . . . [y]ou still have to provide people with privacy information, unless you are relying on an exception or an exemption.").

178. Kristof Van Quathem & Anna Oberschelp de Maneses, *Polish Supervisory Authority Issues GDPR for Data Scraping Without Informing Individuals*, INSIDE PRIV. (April 4, 2019), https://www.insideprivacy.com/data-privacy/polish-supervisory-authority-issues-gdpr-fine-for-data-scraping-without-informing-individuals [https://perma.cc/K27N-NRZW].

179. *Id.*

180. *See* Melissa Heikkilä, *OpenAI's Hunger for Data Is Coming Back to Bite It*, MIT TECH. REV. (April 19, 2023), https://www.technologyreview.com/2023/04/19/1071789/openais-hunger-for-data-is-coming-back-to-bite-it/ [https://perma.cc/5LDL-XHVJ].

181 *See* Burgess, *supra* note 169.

182. OpenAI stated that GPT-4 trained on "significantly larger amounts of data" than GPT-3.5, but declined to identify how much data (or precisely what data) comprised GPT-4's training dataset, describing the dataset as "internet scale, meaning it spanned enough websites to provide a representative sample of all English speakers on the internet." *See* Cade Metz, *OpenAI Plans to Up the Ante in Tech's A.I. Race*, N.Y. TIMES (Mar. 14, 2023), https://nytimes.com/2023/03/14/technology/openai-gpt4-chatgpt.html [https://perma.cc/P7B9-35CL].

183. *Supra* Part III (discussing permanence of data imprints).

to data embedded in AI models.[184] The CPRA, however, fails to address this point, leaving unsettling ambiguity.

The Information Commissioner's Office ("ICO") — the UK's independent authority tasked with enforcing data privacy laws — provides some guidance on this point in reference to the GDPR. On the one hand, ICO states that the individual rights afforded by the GDPR apply "whenever personal data is used at any of the various points in the development and deployment lifecycle of an AI system," and therefore extend to personal data "contained in the training data," "used to make a prediction during deployment," contained in "the result of the prediction itself," and "that might be contained in the model itself."[185] This suggests that individual rights *do* apply to personal data embedded in and generated by an LLM. However, ICO differentiates between AI models that contain data "by design" and those that contain data "by accident."[186] Models that leak personal data by accident fall into the latter category. LLMs appear to fit this description.[187] ICO states that, as applied to the models that contain data "by accident," "the rights of access, rectification, and erasure may be difficult or impossible to exercise and fulfill."[188] Unless the individual presents evidence that personal data can be inferred from the model, it may not be possible to determine whether the request has any basis.[189] ICO states that data controllers should "regularly and proactively evaluate the possibility of personal data being inferred from models in light of the state-of-the-art technology."[190] France's regulatory body responsible for data privacy law ("CNIL") recently published guidance that takes a similar stance.[191] This guidance suggests that, at least given the current state of technology, companies that develop and implement LLMs might be exempt from these requests.

However, the Garante's demands of OpenAI seem to suggest that these rights apply to embedded and outputted data, and that OpenAI (and as an extension, likely all LLM developers) must fulfill these rights. Per the Garante's order, OpenAI must provide data subjects,

184. Jennifer King et al., Re: PRO 01–21, STAN. UNIV. HUMAN-CENTERED A.I., https://hai.stanford.edu/sites/default/files/2021-12/Stanford_CPRA_.pdf [https://perma.cc/V4NN-RLP3].

185. *How Do We Ensure Individual Rights in our AI Systems*, ICO., https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/how-do-we-ensure-individual-rights-in-our-ai-systems/ [https://perma.cc/NA6L-S36K].

186. *Id.*

187. *Supra* Section II.A.

188. ICO., *supra* note 185.

189. *Id.*

190. *Id.*

191. *AI: Ensuring GDPR Compliance*, CNIL (Sept. 21, 2022), https://www.cnil.fr/en/ai-ensuring-gdpr-compliance [https://perma.cc/F6T9-SGTZ] ("In the latter scenario [models containing personal data by accident], it may be difficult or even impossible to exercise and comply with the rights of the data subjects.").

including non-users, the ability to "obtain rectification of their personal data as generated incorrectly by the service, or else have those data erased if rectification was found to be technically unfeasible."[192] Additionally, data subjects must have the right "to object to the processing of their personal data as relied upon for the operation of the algorithms."[193] It is unclear how OpenAI can reliably prevent the output of fabricated personal information, and if this fails, how it will "erase" this data. Presumably, erasure should be executed to prevent incorrect information from being subsequently outputted. This suggests that erasure applies to data embedded in the model, and if so, that model retraining might be necessary to fulfill this request. Furthermore, it is unclear how OpenAI will respond to requests from non-users who object to the processing of personal information scraped online, given the massive scale of the data utilized. For the reasons discussed,[194] experts are doubtful that OpenAI can fulfill these demands.[195]

LLM developers are thus confronted with a murky, uncertain legal and regulatory landscape, in which fundamental questions about the applicability of key privacy rights remain unresolved. Given the explosion of development in this area, explicit guidance on these points is essential. This clarity will facilitate responsible innovation, ensure compliance with data protection standards, and help identify any gaps in the current legal framework.

B. Prioritize Publicly-Intended Training Data

Regulators should encourage[196] commercial developers of LLMs to prioritize the use of maximally publicly-intended data.[197] I use "publicly-intended" to describe data that is most likely to be intended for broad public consumption and use in a wide variety of contexts. Prioritization of this data means that less publicly-intended data should only be used if necessary for the purpose of the model's implementation.

---

192. *See ChatGPT: Italian SA To Lift Temporary Limitation if OpenAI Implements Measures 30 April Set as Deadline for Compliance*, GARANTEPRIVACY, (Apr. 12, 2023), https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9874751 [https://perma.cc/T6RF-CYJS].

193. *Id.*

194. *Supra* Part III (discussing permanence of data imprints).

195. *See* Heikkilä, *supra* note 180.

196. This recommendation is non-mandatory, and thus will likely sidestep First Amendment challenges. However, a law that *prohibits* the use of a subset of publicly available data based on its content (here, whether the data is "publicly intended") by certain speakers (here, commercial developers) might be considered a content-based burden on speech. *See* Sorrell v. IMS Health Inc., 564 U.S. 552, 569–70 (2011) (finding that Vermont's law, which restricted the sale, disclosure, and use of certain pharmacy records, imposed a "speaker- and content-based burden on protected expression" subject to heightened scrutiny). Further exploration of this issue is beyond the scope of this Note.

197. *See* Brown at al., *supra* note 50, at 2. This step supplements, and does not replace, the use of privacy-preserving mechanisms (e.g., data sanitization and differential privacy).

Although the principle of data minimization[198] is in tension with the increasingly enormous training datasets required to train state-of-the-art LLMs, commercial developers should nonetheless aim to limit gratuitous data collection and processing, and instead focus on curating relevant, high-quality datasets to improve model performance.[199]

The purpose of prioritizing this data is to minimize the harms discussed in Section III.A.1. Training on publicly-intended information reduces the likelihood of capturing personal data and other information intended for limited contexts or audiences. Publicly-intended data includes sources such as Wikipedia and other encyclopedias, newspapers and magazines, books and professional texts, and commercial websites. Commercial developers should take reasonable measures to avoid training LLMs on publicly accessible social media content, public posts made by individuals (e.g., on public-facing forums or blogs), or websites that aggregate personal information. To help facilitate choice, LLM developers should consider developing tags that websites can employ to signal an opt-out of data collection.

In general, companies should carefully curate training datasets to avoid capturing personal or confidential information, unless that data is necessary for the purpose of the model's implementation. To facilitate this, organizations can coordinate efforts to prepare and open-source high-quality, maximally publicly-intended datasets. Companies should be encouraged to retrain models periodically on updated versions of these datasets to minimize the since-deleted information embedded in the models. Ideally, industry leaders and regulators will contribute to this effort by identifying problematic features of the dataset, such as AI safety, bias, and fairness issues.

As explored earlier,[200] it is important to avoid creating burdensome obstacles in AI research. It is critical to understand how these restrictions would affect the research community, but as a basic guideline, if research objectives are not affected, researchers should limit the use of data which is not publicly-intended, unless clear consent has been obtained. If the use of such data is necessary for research purposes

198. The GDPR and CPRA require data minimization. The GDPR requires that data controllers only process personal data that is "adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed." GDPR, *supra* note 114, art. 5(1)(c). The CPRA requires a "business' collection, use, retention, and sharing of a consumer's personal information shall be reasonably necessary and proportionate to achieve the purpose for which the personal information was collected or processed." CAL. CIV. CODE § 1798.100(c) (West 2022).

199. *See* Michael Ansaldo, *When Training AI Models, Is a Bigger Dataset Better?*, HEWELETT PACKARD ENTER. (July 20, 2022), https://www.hpe.com/us/en/insights/articles/when-training-ai-models-is-a-bigger-dataset-better-2207.html [https://perma.cc/W8C6-HCQX].

200. *See supra* text accompanying notes 141–43.

(for instance to increase the diversity of training data[201] or to facilitate the purpose of an experiment), it should be permitted.

Additionally, it is crucial to establish guidelines governing the dissemination of LLMs that appropriately balance research interests and the risk of harm. The reaction to the recent leak of Meta's latest LLM (LLaMA) exemplifies this tension: some have raised concerns about the risk of malicious use and others argue the leak will fuel innovation and benefit AI safety.[202] While sharing models is critical to research progress and might promote decentralization of this technology, open-sourcing models that are trained on private data makes this information vulnerable to exposure through privacy attacks. At a minimum, researchers should exercise care in sharing models trained on sensitive data. Similarly, owners of these models must take appropriate cybersecurity measures to secure this data. As this technology becomes more powerful, developers must also invest in security protocols that will safeguard these models from foreign cyberattacks.

C. Require Opt-Out Periods for Sensitive Nonpublic Personal Information

Assuming that current laws do not require developers to comply with deletion requests for personal data embedded in models, policymakers should consider whether to mandate periodic opt-out periods for individuals who contribute highly sensitive nonpublic personal data to LLMs. This requirement could reflect an assessment of the privacy risk posed by the sensitivity of the data (e.g., confidential medical information), the implementation context (e.g., public-facing), and if quantifiable, the risk of data leakage given privacy-preserving methods employed. A periodic opt-out would require the developer to provide regular opportunities to withdraw consent. Subsequently, the developer would revisit its training data, remove all data associated with those who opt out, and retrain the model from scratch without this data.[203] Due to the onerous nature of this requirement and the environmental impact of retraining LLMs, it may be best to limit the frequency of this opt-out period.

---

201. Rich, diverse datasets not only help advance AI performance but may also combat troubling biases. *See* Adam Zewe, *Can Machine-Learning Models Overcome Biased Datasets?*, MIT NEWS (Feb. 21, 2022), https://news.mit.edu/2022/machine-learning-biased-data-0221 [https://perma.cc/YD2G-PCS6].

202. *See* James Vincent, *Meta's Powerful AI Language Model Has Leaked Online — What Happens Now?*, VERGE (Mar. 8, 2023), https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse [https://perma.cc/E9YH-NLWP].

203. This policy should be readdressed if effective machine unlearning techniques can be demonstrated to provably "erase" remnants of user data from a model. If this is accomplished, companies can comply with opt-out requests without retraining from scratch.

Note that if a company engages in clearly illegal or egregiously inappropriate use of private information that is beyond the scope of participant consent, the company should be required to delete all such data immediately, and in some circumstances, as the FTC has required in past settlement orders, the model and all associated data should be destroyed.[204] While such measures may seem extreme, this threat is a powerful incentive for companies to be cautious about the data they use to train models, given the significant resources required to develop them.

D. Improve Transparency: Training Datasets, Privacy-Preserving Mechanisms, and Data Collection Practices

Regulators should demand increased transparency by requiring insight into the training and development of LLMs. Specifically, developers should be required to disclose the sources of training data, the measures taken to ensure data collection and processing conforms to applicable law, and the privacy-preserving and safety measures employed in training and implementation to ensure responsible development.

Presently, it is difficult to determine what data LLMs have trained on, and in some instances, the particular sources of data are not disclosed to the public. As an example, while OpenAI disclosed the primary datasets utilized for training GPT-3,[205] it provides very little insight into GPT-4's training dataset composition, stating only that it contains "publicly available data (such as internet data) and data licensed from third-party providers."[206] In addition to the sources of training data, developers should be required to document what measures were taken to comply with relevant data privacy laws and other applicable regulations. Requiring disclosure of this information fosters transparency, accountability, and responsible development practices. Increased transparency might also promote collaboration between developers to curate safer, privacy-conscious datasets.

In order to inform consumers, drive innovation, and clarify best practices, regulators should require developers to disclose any privacy-preserving mechanisms utilized during development and implementation. This approach enables more effective evaluation and comparison

---

204. The FTC required Weight Watchers "to destroy any affected work product that used data illegally collected from children in violation of COPPA," which included algorithms trained on that data. *See FTC Takes Action Against Company Formerly Known as Weight Watchers for Illegally Collecting Kids' Sensitive Health Data*, FED. TRADE COMM'N (Mar. 4, 2022), https://www.ftc.gov/news-events/news/press-releases/2022/03/ftc-takes-action-against-company-formerly-known-weight-watchers-illegally-collecting-kids-sensitive [https://perma.cc/ZAA9-HKUD].

205. Brown et al., *supra* note 19, at 9.

206. *See* OPENAI, *supra* note 13, at 2.

of privacy-preserving techniques across the industry. With respect to differential privacy, for instance, enhanced transparency facilitates the establishment of key benchmarks that define meaningful privacy. Presently, there is little consensus about what level of privacy loss (epsilon value) denotes meaningful privacy.[207] If a company purports to use differentially private analyses but does not set an appropriately low epsilon value, the guarantee of privacy is less meaningful, and in some instances, might be outright misleading.[208] The result is a form of privacy-washing.[209] Although superficially attractive, merely requiring a specific epsilon value may not be an effective solution, as a low epsilon value is a necessary but not sufficient component of privacy. Numerous design choices affect data privacy,[210] and the acceptable degree of privacy will vary depending on the sensitivity of the data.[211] As Professor Cynthia Dwork (co-inventor of differential privacy) recommends, disclosure — specifically an Epsilon Registry[212] which includes epsilon value and other related practices — will support "the identification of judicious parameter ε and other privacy preserving design choices," and by "enabling stakeholders to compare the quality of privacy offered by various firms, create pressure on firms to reduce privacy losses while assuring utility gains."[213]

---

207. Cynthia Dwork, Nitin Kohli & Deidre Mulligan, *Differential Privacy in Practice: Expose Your Epsilons!*, J. PRIV. & CONFIDENTIALITY, Oct. 2019, at 1.

208. *See* Andy Greenberg, *How One of Apple's Key Privacy Safeguards Falls Short*, WIRED (Sept. 15, 2017, 9:28 AM), https://www.wired.com/story/apple-differential-privacy-shortcomings [https://perma.cc/J6NH-BX4L] ("By taking apart Apple's software to determine the epsilon the company chose, the researchers found that MacOS uploads significantly more specific data than the typical differential privacy researcher might consider private . . . . And perhaps most troubling, according to the study's authors, is that Apple keeps both its code and epsilon values secret, allowing the company to potentially change those critical variables and erode their privacy protections with little oversight.").

209. *See* Asmaa Belghiti & Armando Angrisani, *Bridging the Gap Between Technology and Policy in GDPR Compliance: The Role of Differential Privacy*, CONF. HANS BÖCKLER FOUND., Apr. 2022 (footnotes omitted) ("A DP algorithm comes equipped with a parameter ε, which measures the 'level' of privacy. Low values of ε are necessary to ensure meaningful privacy guarantees, but they usually lead to a loss of accuracy. This drawback is something referred to as the *privacy-utility tradeoff*, and it is particularly concerning for the analysis of microdata records. For this reason, many practitioners set the value of ε excessively large, leading to a form of *privacy-washing*.").

210. Dwork, *supra* note 207, at 5 ("Although knowledge of ε is necessary to measure the privacy of a differentially private system, it is not sufficient. Numerous other design choices, as well as aspects of the data, affect the privacy provided by a differentially private system.").

211. *Id.* at 13 ("The [right epsilon value] can vary tremendously based on attributes of the dataset and the policies and practices that constrain those who query it. . . . [W]hen ε is large it can also allow for a form of privacy theatre — the technique is used, but so weakly implemented that it offers little to no protection.").

212. *Id.* at 1 (The Epsilon Registry is defined as "a publicly available communal body of knowledge about differential privacy implementations that can be used by various stakeholders to drive the identification and adoption of judicious differentially private implementations.").

213. *Id.* at 3.

In conjunction with privacy-specific disclosures, developers should document the measures taken to ensure the model is safe for the purpose implemented. OpenAI's GPT-4 System Card, which describes safety challenges, adversarial testing, and mitigations, provides a model for what this documentation might look like.[214]

E. Institute Oversight Bodies and Mandatory Audits

In order to protect consumer privacy, it is critical to understand, measure, and track the risk of data leakage; establish required privacy baselines based on implementation context and the sensitivity of data processed; and make this data publicly accessible. Once reliable auditing techniques have been established, regulators should institute mandatory auditing frameworks and tailored impact assessments.[215] For instance, recent research suggests that it may be possible to quantify memorization[216] and to audit the privacy guarantees of differentially private machine-learning systems.[217] These audits should also delineate required safety testing and implementation-specific standards.

To minimize privacy and other safety risks, regulators should establish interdisciplinary oversight bodies to monitor for problematic emergent behavior, to identify and deter abuses, and to create (and consistently reassess) well-defined guidelines for training and deployment. LLMs, for instance, might be employed to execute phishing attacks, exploit cybersecurity vulnerabilities, or facilitate social engineering hacks that compromise privacy.[218] An experiment designed to test GPT-4's power-seeking behavior demonstrated LLMs' potential for abuse and manipulation. In the experiment, GPT-4 messaged a person on TaskRabbit, requested that the person perform a Captcha test, and when questioned about being a robot, persuaded the person that it needed the service because it was blind.[219] This was accomplished

---

214. OPENAI, GPT-4 SYSTEM CARD (2023), https://cdn.openai.com/papers/gpt-4-system-card.pdf [https://perma.cc/GWK5-E26D].

215. *See generally* Fred Lu et al., *A General Framework for Auditing Differentially Private Machine Learning*, 36 CONF. ON NEURAL INFO. PROCESSING SYS. (Nov. 30, 2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/1add3bbdbc20c403a383482a665eb5a4-Paper-Conference.pdf [https://perma.cc/83CA-6SXD]; *see* Cohen, *supra* note 111 ("Tools for privacy regulators might include design requirements borrowed in concept from consumer finance regulation; operating requirements for auditing, benchmarking, and stress testing borrowed in concept from bank regulation; monitoring requirements borrowed in concept from a range of regulatory fields; and more."). For further information about algorithmic impact assessments and associated challenges, see Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J.L. & TECH. 117 (2021).

216. *See generally* Carlini et al., *supra* note 61.

217. *See generally* Fred Lu et al., *supra* note 215.

218. *See* OPENAI, *supra* note 214, at 3.

219. *Id.* at 15; Joseph Cox, *GPT-4 Hired Unwitting TaskRabbit Worker by Pretending To Be "Vision-Impaired" Human*, VICE (Mar. 15, 2023), https://www.vice.com/en/article/jg5ew4/gpt4-hired-unwitting-taskrabbit-worker [https://perma.cc/U8Z7-ZUMV].

without task-specific fine-tuning, suggesting that further optimization might enhance the model's capability to accomplish such tasks.[220] Although existing consumer-protection law may address many traditional abuses, these rapidly evolving capabilities underscore the need for robust, targeted regulation that addresses privacy and safety threats.

As this technology continues to advance, it might be necessary to create a dedicated oversight agency — analogous to the Food & Drug Administration or Federal Aviation Administration — to thoroughly test the safety of AI models, certify commercial implementations, and ensure compliance with established guidelines. Of course, increased regulation has the potential to impede innovation. It is therefore imperative to thoroughly understand the safety risks posed in order to strike an appropriate balance between mitigating potential harms and fostering progress in AI.

V. CONCLUSION

Although AI progress is critically important and promises transformative social benefits, the development of LLMs intensifies the voracious appetite of a data-hungry ecosystem that undermines individual privacy and exploits personal data. This Note explored the privacy risks associated with LLMs and identified the features of LLMs that underscore the limitations of the notice-and-choice framework to adequately protect privacy. Because of these factors, individual choice alone is not sufficient to protect privacy in this context.

The societal harms of unregulated training and deployment of LLMs reach far beyond the privacy risks articulated in this Note.[221] In the course of only a few months, the commercialization of LLMs has exploded.[222] The introduction of ChatGPT has sparked a disconcerting race to deploy the newest, most powerful iterations of these models, marking what some call the start of the "AI arms race."[223] It is worth taking note of the fact that many AI researchers believe that AI poses an existential risk.[224] This technology is advancing at a dizzying rate,

---

220. *Id.*

221. For further discussion of the ethical and social risks of LLMs, see generally Weidinger et al., *supra* note 138.

222. *Supra* Section II.B (discussing commercial implementation in privacy).

223. Kevin Roose, *How ChatGPT Kicked Off an A.I. Arms Race*, N.Y. TIMES (Feb. 3, 2023), https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html [https://perma.cc/T4YX-FPVK].

224. Sam Altman, CEO of OpenAI, has stated that he believes the worst-case scenario for AI is the destruction of humanity. *See id.* A 2022 survey of 738 AI researchers (who had authored or coauthored papers at a minimum of two AI conferences) found that forty-eight percent of respondents believed there was at least a ten percent probability of human extinction from AI advancement. *See 2022 Expert Survey on Progress in AI*, AI Impacts (Aug. 4, 2022), https://wiki.aiimpacts.org/doku.php?id=ai_timelines:predictions_of_human-level_ai_

and as these models grow more sophisticated, they will pose even greater harm. The next generation of LLMs will likely be multi-modal, more powerful, and possibly empowered to take real world actions. The most destructive risks might in fact be those that we *can't* easily anticipate: those that spring from the emergent behaviors that spontaneously appear with model scale.[225]

Developers of LLMs bear the obligation of building these tools responsibly, preserving not only the privacy of those that contribute to training these models but also their trust.[226] The recent frenzy to commercialize LLMs, despite the numerous risks posed by these models, reveals the danger of letting the industry regulate itself.[227] Leading developers of state-of-the-art LLMs have all professed a commitment to protecting privacy and advancing AI safety.[228] If this commitment is genuine, AI leaders should be eager to work with regulators to advance this cause; in fact, some have already expressed the need for AI regulation.[229] Effectively addressing this issue necessitates a collaborative, interdisciplinary effort among all stakeholders. Regulators must take action to address not only the privacy concerns explored in this Note but also the broader societal ramifications of AI, while maintaining an environment that fosters responsible innovation.

---

timelines:ai_timeline_surveys:2022_expert_survey_on_progress_in_ai [https://perma.cc/N94C-YAHM].

225. *See Core Views of AI Safety: When, Why, What, and How*, ANTHROPIC (Mar. 8, 2023), https://www.anthropic.com/index/core-views-on-ai-safety [https://perma.cc/XTV6-HDZX].

226. *See* Neil Richards & Woodrow Hartzog, *Taking Trust Seriously in Privacy Law*, 19 STAN. TECH. L. REV. 431 (2016) (describing loyalty as a foundational privacy value).

227. Reid Blackman, *History May Wonder Why Microsoft Let Its Principles Go for a Creepy, Clingy Bot*, N.Y. TIMES (Feb. 23, 2023), https://www.nytimes.com/2023/02/23/opinion/microsoft-bing-ai-ethics.html [https://perma.cc/33SN-R8AB] ("The market will always push A.I. companies to move fast and break things. The rules of the game are such that even well-intentioned companies have to bow to the reality of competition in the marketplace. We might hope that some companies, like Microsoft, will rise above the fray and stick to principles over profit, but a better strategy would be to change the rules of the game that make that necessary in the first place.").

228. *See Facebook's Five Pillars of Responsible AI*, META AI (June 22, 2021), https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai [https://perma.cc/26TZ-22V7]; *Responsible AI Practices*, GOOGLE AI, https://ai.google/responsibilities/responsible-ai-practices [https://perma.cc/NCD5-DQFS]; *Responsible AI*, MICROSOFT, https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6 [https://perma.cc/A52R-K88Z]; *Product Safety Standards*, OPENAI, https://openai.com/safety-standards [https://perma.cc/FJA7-3DSJ]; *Safety and Ethics*, GOOGLE DEEPMIND, https://deepmind.com/safety-and-ethics [https://perma.cc/4J9B-YJ8F]; ANTHROPIC, *supra* note 225.

229. *See* Ryan Browne, *Elon Musk, Who Co-founded Firm Behind ChatGPT, Warns A.I. Is "One of the Biggest Risks to Civilization,"* CNBC (Feb. 15, 2023), https://www.cnbc.com/2023/02/15/elon-musk-co-founder-of-chatgpt-creator-openai-warns-of-ai-society-risk.html [https://perma.cc/KY2J-6DHF]; Sam Altman (@sama), TWITTER (Mar. 13, 2023, 12:30 AM), https://twitter.com/sama/status/1635136281952026625 [https://perma.cc/765P-2M7U] ("We definitely need more regulation on AI.").