JIPEL

NYU Journal of Intellectual Property
& Entertainment Law

VOLUME 13 NUMBER 2

Statement of Purpose

Consistent with its unique development, the New York University Journal of Intellectual Property & Entertainment Law (JIPEL) is a nonpartisan periodical specializing in the analysis of timely and cutting-edge topics in the world of intellectual property and entertainment law. As NYU's first online-only journal, JIPEL also provides an opportunity for discourse through comments from all of its readers. There are no subscriptions or subscription fees; in keeping with the open-access and free discourse goals of the students responsible for JIPEL's existence, the content is available for free to anyone interested in intellectual property and entertainment law.

The New York University Journal of Intellectual Property & Entertainment Law is published up to three times per year at the New York University School of Law, 139 MacDougal Street, New York, New York, 10012. In keeping with the Journal's open access and free discourse goals, subscriptions are free of charge and can be accessed via www.jipel.law.nyu.edu. Inquiries may be made via telephone (212-998-6101) or e-mail (submissions.jipel@gmail.com).

The Journal invites authors to submit pieces for publication consideration. Footnotes and citations should follow the rules set forth in the latest edition of *The Bluebook: A Uniform System of Citation*. All pieces submitted become the property of the Journal. We review submissions through Scholastica (scholasticahq.com) and through e-mail (submissions.jipel@gmail.com).

A nonpartisan periodical, the Journal is committed to presenting diverse views on intellectual property and entertainment law. Accordingly, the opinions and affiliations of the authors presented herein do not necessarily reflect those of the Journal members.

The Journal is also available on WESTLAW, LEXIS-NEXIS and HeinOnline.

NEW YORK UNIVERSITY

JOURNAL OF INTELLECTUAL PROPERTY AND ENTERTAINMENT LAW

VOLUME 13 SPRING 2024 NUMBER 2

PRIVACY OF PERSONAL DATA IN THE GENERATIVE AI DATA LIFECYCLE

MINDY NUNEZ DUFFOURC* SARA GERKE** & KONRAD KOLLNIG†

* Assistant Professor of Law, (Maastricht) Law and Tech Lab, Maastricht European Private Law Institute, Maastricht University, Maastricht, Netherlands. Mindy Nunez Duffourc reports grant funding from the European Union (Grant Agreement no. 101057321) during the research and writing of this article. We would like to thank Candace Thomas and Lyubomir Ivanov Avdzhiyski for their excellent research assistance and the entire editorial team at NYU Journal of Intellectual Property & Entertainment Law.

** Associate Professor of Law and the Richard W. & Marie L. Corman Scholar, College of Law, University of Illinois Urbana-Champaign, USA; Co-Principal Investigator, WP8 (Legal, Ethical & Liability), Validating AI in Classifying Cancer in Real-Time Surgery (CLASSICA), European Union (Grant Agreement no. 101057321); Co-Principal Investigator, WP4 (Addressing Ethical/Legal Concerns), Optimizing Colorectal Cancer Prevention through Personalized Treatment with Artificial Intelligence (OperA), European Union (Grant Agreement no. 101057099); Multiple Principal Investigator, Bioethical, Legal, and Anthropological Study of Technologies (BLAST), National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the National Institutes of Health Office of the Director (NIH OD) (Grant Agreement no. 1R21EB035474-01).

† Assistant Professor of Computational Law, (Maastricht) Law and Tech Lab, Maastricht European Private Law Institute, Maastricht University, Maastricht, Netherlands.

Introduction

Generative AI ("GenAI") is a powerful tool in the content generation toolbox. Its modern debut via applications like ChatGPT and DALL-E enamored users with human-like renditions of text and images. As new user accounts grew exponentially, these models soon gained a foothold in various industries, from law to medicine to music. The GenAI honeymoon period ended when questions about GenAI development and content began to mount: Is this content reliable? Can GenAI harm consumers and others? What are the implications for intellectual property (IP) rights? Does GenAI violate data privacy laws? Is it ethical to use AI-generated content? Users have already faced the consequences of putting blind faith in GenAI. For example, a lawyer who relied on ChatGPT to perform legal research was sanctioned for including fictional cases in his court pleadings.[1] Additionally, GenAI developers began to encounter scrutiny related to their development and marketing of GenAI tools. In Europe, Italy temporarily banned ChatGPT, citing concerns about data privacy violations.[2] Recently, a non-profit organization focused on private enforcement of data protection laws in the European Union (EU) claimed that ChatGPT's provision of inaccurate personal data about individuals

[1] Mata v. Avianca, Inc., 678 F. Supp. 3d 443, 451, 460–66 (S.D.N.Y. 2023).

[2] *ChatGPT: Italy blocks AI chatbot over privacy concerns*, Deutsche Welle (Mar. 31, 2023), https://www.dw.com/en/chatgpt-italy-blocks-ai-chatbot-over-privacy-concerns/a-65200137 [https://perma.cc/743G-XV8C].

violates data privacy.[3] In the United States (US), lawsuits alleged violations of IP, privacy, and property rights resulting from developers' use of massive amounts of data to train GenAI.[4]

The increasing development and use of GenAI has spurred data privacy concerns in both the EU and the US. According to the FTC, "[c]onsumers are voicing concerns about harms related to AI—and their concerns span the technology's lifecycle, from how it's built to how its [sic] applied in the real world."[5] While the introduction of new technologies has generally led to significant legislative retooling in the area of data privacy in the last decade—with the EU adopting the General Data Protection Regulation ("GDPR"), and several US states following suit—legislators designed these laws before GenAI models, like ChatGPT, were on the radar. In April 2023, the European Data Protection Board (EDPB) developed a ChatGPT taskforce to coordinate regulatory enforcement in the EU Member States "on the processing of personal data in the context of ChatGPT."[6]

The use of personal data to develop and update GenAI models can harm individuals by disclosing personal data, including sensitive personal data, to a broad audience; enabling individual profiling for targeting, monitoring, and potential discrimination; producing false information; and limiting an individual's ability to keep their personal data private.[7] This article examines how the current

[3] *ChatGPT provides false information about people, and OpenAI can't correct it*, NYOB (Apr. 29, 2024), https://noyb.eu/en/chatgpt-provides-false-information-about-people-and-openai-cant-correct-it [https://perma.cc/G5S9-RVQU].

[4] *See e.g.*, Class Action Complaint, Silverman. v. Open AI, Inc., No. 3:23-cv-03416 at ¶¶ 35–36 (N.D. Cal. Jul. 7, 2023).

[5] Simon Fondrie-Teitler & Amritha Jayanti, *Consumers Are Voicing Concerns About AI*, FTC TECH. BLOG (Oct. 3, 2023), https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2023/10/consumers-are-voicing-concerns-about-ai [https://perma.cc/YU2W-P5HP].

[6] EUROPEAN DATA PROTECTION BOARD (EDPB), REPORT OF THE WORK UNDERTAKEN BY THE CHATGPT TASKFORCE 4 (May 23, 2024), https://www.edpb.europa.eu/our-work-tools/our-documents/other-guidance/report-work-undertaken-chatgpt-taskforce_en [https://perma.cc/X8GB-YTSE] (last visited Jun. 3, 3024).

[7] *See generally* CONFEDERATION OF EUR. DATA PROT. ORGS. AI WORKING GRP., GENERATIVE AI: THE DATA PROTECTION IMPLICATIONS (Oct. 16, 2023), https://cedpo.eu/generative-ai-the-data-protection-implications-16-10-2023 [https://perma.cc/G4AX-QC82] (discussing concerns that AI-generated content about individuals might lead to biased decisions and discrimination that affect data subjects); FTC, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS 33 (2012), https://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers [https:

approaches to data privacy in the US and EU govern personal data in the GenAI data lifecycle. We focus specifically on the flow of *personal data* in the GenAI data lifecycle because its collection and use have important data privacy implications for users.

Part I introduces GenAI models. It traces the development of GenAI architecture and provides a technical overview of modern GenAI. It discusses the capabilities and limitations of GenAI and provides a quick glimpse into the data sources that power these models.

Part II sets forth the current frameworks governing personal data in the US and EU. It discusses how these frameworks aim to protect personal data and provides the basic definitions for various types of data that have important implications in the GenAI data lifecycle. This part is supplemented by tables in the appendices that provide a summary comparison of the treatment of various types of data—including personal data, sensitive personal data, de-identified, pseudonymized, and anonymized data, and publicly available data—in US and EU regulatory frameworks.

Part III outlines the flow of personal data in the GenAI data lifecycle and its implications on data privacy. It describes the role of personal data in GenAI development and training, how GenAI developers might use personal data to improve existing GenAI models or develop new models, and how personal data can become part of a GenAI model's output. Finally, it discusses the retention of personal data in the GenAI data lifecycle.

Part IV identifies data privacy implications that arise as personal data flows through the GenAI data lifecycle and analyzes how the current frameworks governing personal data might address these data privacy implications in the US and EU. First, it discusses the privacy implications and governance of using publicly available data in the GenAI data lifecycle. Second, it discusses the privacy implications and governance of using private and sensitive personal data in the

//perma.cc/CR7Z-7RU4] ("The extensive collection of consumer information — particularly location information — through mobile devices also heightens the need for companies to implement reasonable policies for purging data. Without data retention and disposal policies specifically tied to the stated business purpose for the data collection, location information could be used to build detailed profiles of consumer movements over time that could be used in ways not anticipated by consumers.").

GenAI data lifecycle. Finally, it discusses the loss of control over personal data in the GenAI data lifecycle.

I
Generative AI

GenAI describes AI models that create content like images, videos, sounds, and text. Even if the public release of the latest generation of GenAI models came as a surprise to many, the underlying technologies had been in development for decades.[8] One strand of GenAI takes the form of large-language models (LLMs), like ChatGPT, that use sophisticated deep learning models for next-word prediction to generate human-like text.[9] Next-word prediction is possible because words do not randomly follow each other in a text, but are context-dependent.[10] In other words, if one knows what other words have been written so far in a piece of text, then it is possible, with relatively high accuracy, to build a "model" that predicts the following word.

Some of the earliest approaches for next-word prediction were *n-gram models* that date back to Claude Shannon's revolutionary work on information theory in the 1940s.[11] An *n-gram model* is a very simple language model. It looks at a fixed

[8] An early example of chatbots was ELIZA. It was developed in the 1960s by Joseph Weizenbaum at MIT to create the illusion of genuine human interaction. Human participants would engage in an exchange of text messages with a computer, similarly to how users now engage with ChatGPT. *See* Joseph Weizenbaum, *ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine*, 9 Commun. ACM 36, *passim* (1966).

[9] *See* David Nield, *How ChatGPT and Other LLMs Work—and Where They Could Go Next*, Wired (May 9, 2023), https://www.wired.com/story/how-chatgpt-works-large-language-model [https://perma.cc/3APZ-QQES] (discussing how LLMs like ChatGPT work using next-word prediction). Deep learning describes a subset of AI that uses many layers of artificial neural networks (ANNs) to produce an output. *See* Zubair Ahmad et al., *Artificial Intelligence (AI) in Medicine, Current Applications and Future Role with Special Emphasis on Its Potential and Promise in Pathology: Present and Future Impact, Obstacles Including Costs and Acceptance Among Pathologists, Practical and Philosophical Considerations. A Comprehensive Review*, 16 Diagnostic Pathology art. 24, 2021, at 2, https://doi.org/10.1186/s13000-021-01085-4 [https://perma.cc/L8U5-ACMH].

[10] *See* Nield, *supra* note 9 ("One of the key innovations of [this neural network architecture] is the self-attention mechanism … [W]ords aren't considered in isolation, but also in relation to each other in a variety of sophisticated ways.").

[11] *See generally* Claude E. Shannon, *A Mathematical Theory of Communication*, 27 Bell Sys. Tech. J. 379, 386–89 (1948).

number of words and tries to guess what word is most likely to come next.[12] For example, a 2-gram model (where $n = 2$) only looks at *one* word to predict the next one (*i.e.*, a total of *two* words, hence "2-gram").[13] For example, if a 2-gram model is given the word "How" as context, it might guess that the next word is "are" because "How are" is a common pairing of words. Then, it can use "are" to guess the next word, maybe "you." This, overall, gives the text "How are you." By predicting the next word again and again, it is then possible for algorithms to write large amounts of text that may or may not look like it was written by a human, depending on the quality of the model. Like other language models, including LLMs, *n-gram models* are trained on large corpora of text to learn what word is—statistically speaking—most likely to come next.[14]

Modern-day LLMs also use next-word prediction for text generation but are significantly more sophisticated than *n-gram models*. Unlike n-grams, they are not limited to a fixed number of words but can, instead, reason over much larger inputs of text. To do so, they are made up of hundreds of billions of parameters, which are "mathematical relationship[s] linking words through numbers and algorithms."[15] To train such a large number of parameters, a large amount of training data is necessary (i.e., hundreds of gigabytes of data), as well as significant computational resources (i.e., millions of dollars of energy consumption and computing hardware). For example, OpenAI trained GPT-3 with the Common Crawl, WebText2, Books1, Books2, and English-language Wikipedia datasets.[16] Google's original LLM model, Bard, used Language Models for Dialog Applications ("LaMDA"), which was trained using 1.56 trillion words of publicly available text and dialogue on the internet.[17] According to Meta, its LLM, Llama

[12] Daniel Jurafsky & James H. Martin, *N-gram Language Models*, *in* Speech and Language Processing (3d ed. forthcoming 2024) (manuscript at 32–33), https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf [https://perma.cc/6LC5-UJ2S].

[13] *Id.* at 33.

[14] *See id.* at 15–17 (describing the text used to train NLP models).

[15] Nield, *supra* note 9.

[16] Tom B. Brown et al., *Language Models are Few-Shot Learners*, 33 Advances in Neural Info. Processing Sys., 1877 (2020), https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html [https://perma.cc/W2GM-KCM]. OpenAI, the company that developed ChatGPT, has not published much information on the training of GPT-3's successor (and current foundation for ChatGPT), GTP-4.

[17] Romal Thoppilan et al., *LaMDA: Language Models for Dialog Applications*, arXiv, Jan. 2022, at 1–3, https://arxiv.org/abs/2201.08239 [https://perma.cc/ZH4D-5W3J].

2, is also trained with the enormous amount of publicly available text on the internet.[18] By training LLMs on the vast text available on the internet and in books, they have become extremely good at mimicking previous human-generated work through next-word prediction. Yet, since these models are so dependent on previously existing text, they currently struggle with producing accurate textual outputs beyond their training data.[19] Thus, LLMs are good at producing text that looks reliable, but since they have no semantic understanding of the text that they write, they have been deemed "stochastic parrots."[20]

The models that underpin all state-of-the-art LLMs rely on the "Transformers"—or a closely related—architecture.[21] This type of deep learning model architecture was invented by researchers at Google and first released in 2017.[22] Transformers use "self-attention," which simplified the model architecture compared to previous models and achieved cutting-edge performance on language tasks—including text generation.[23]

Other GenAI models also use deep learning models but create *non-textual* outputs like images, videos, and sounds—or even combine different such

[18] *Llama 2: open source, free for research and commercial use*, Meta, https://llama.meta.com/llama2 [https://perma.cc/4XN9-E6KC].

[19] Steve Yadlowsky et al., *Pretraining Data Mixtures Enable Narrow Model Selection Capabilities in Transformer Models*, arXiv, Nov. 2023, at 2, https://arxiv.org/abs/2311.00871 [https://perma.cc/6HGR-VYNS].

[20] *See* Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*🦜, *in* 2021 ACM Conference on Fairness, Accountability, and Transparency 610, 616–17 (Ass'n for Computing Mach., ed., 2021), https://dl.acm.org/doi/10.1145/3442188.3445922 [https://perma.cc/H4V6-7HJP].

[21] *See generally* Mostafa Ibrahim, *An Introduction to Transformer Networks*, Weights & Biases (Dec. 15, 2022), https://wandb.ai/mostafaibrahim17/ml-articles/reports/An-Introduction-to-Transformer-Networks--VmlldzoyOTE2MjY1 [https://perma.cc/YT4U-SLQE] ("[T]ransformers are a neural network with a novel architecture that aims to solve sequence-to-sequence complex language tasks like translation, question answering, and chatbots, all while managing long-range dependencies."); Dana Leigh, *How Does Chat GPT Actually Work?*, TechRound (Feb. 15, 2023), https://techround.co.uk/guides/how-does-chat-gpt-actually-work/ [https://perma.cc/YL7R-NKWG] ("At its core, Chat GPT is the implementation of a type of neural network known as a transformer. Transformers are a type of deep learning algorithm that is commonly used in the field of natural language processing (NLP).").

[22] Ashish Vaswani et al., *Attention Is All You Need*, *in* 31st Conference on Neural Information Processing Systems *passim* (I. Guyon et al. eds., 30th ed. 2017), https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf [https://perma.cc/936G-ZXHK].

[23] *Id.* at 2–3, 6–7.

"modalities" (e.g., producing an image based on textual input). While these models rarely use the Transformers architecture, all GenAI models share many of the same challenges and limitations. For example, like LLMs, other GenAI models also need to be trained on vast amounts of data. Much of this data comes from information on the internet that is extracted or "scraped" by automated tools. Large-Scale Artificial Intelligence Open Network ("LAION") provides a publicly available image data set that can be used to train GenAI models.[24] Stable Diffusion's AI model, which underlies popular AI image generation applications like Midjourney and Dreamstudio, was trained with 2.3 billion images scraped from the internet by the nonprofit organization Common Crawl.[25] This dataset includes stock images, but also hundreds of thousands of images from individuals on social media and blogging platforms, like Pinterest, Tumblr, Flickr, and WordPress.[26] OpenAI's DALL-E, too, was trained using millions of images on the internet.[27] Google trained its text-to-music GenAI model, MusicLM, with datasets that are primarily sourced from over 2 million YouTube clips containing not only music, but a variety of other human, animal, natural, and background sounds.[28] Google replaced Bard with a new multi-modal model, Gemini, which is trained not only with text, but also with images, audio, and video.[29]

[24] LAION, https://laion.ai/ [https://perma.cc/Z98L-VVU2].

[25] Andy Baio, *Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator*, Waxy (Aug. 30, 2022), https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/ [https://perma.cc/TX38-DF6G] ("All of LAION's image datasets are built off of Common Crawl, a nonprofit that scrapes billions of webpages monthly and releases them as massive datasets.").

[26] *Id.*

[27] *DALL·E 2 pre-training mitigations*, OpenAI, https://openai.com/research/dall-e-2-pre-training-mitigations [https://perma.cc/WQA8-646Q] ("DALL•E 2 is training on hundreds of millions of captioned images from the internet, and we remove and reweight some of these images to change what the model learns.").

[28] Ezra Sandzer-Bell, *Google's AI Music Datasets: MusicCaps, AudioSet and MuLan*, Audiocipher (May 17, 2023), https://www.audiocipher.com/post/musiccaps-audioset-mulan [https://perma.cc/Y28E-QXCL] ("Behind the scenes, Google has used three music datasets, called MusicCaps, AudioSet and MuLan, to trained [sic] their music models for MusicLM"); *AudioSet*, Google, https://research.google.com/audioset/dataset/index.html [https://perma.cc/QV7N-FSVH].

[29] Gemini Team, Google, Gemini: A Family of Highly Capable Multimodal Models, 1 (2024), https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf [https://perma.cc/MAR2-HKJY].

GenAI has wide-ranging uses, from designing magazine covers,[30] to creating new music,[31] to summarizing medical records.[32] These uses present a wide range of legal issues that stem from their use of massive amounts of data for development and training. For example, some of this data, like copyrighted books, might implicate IP rights.[33] Some of this data, like an individual's biographical information or photograph, might implicate data privacy rights. We now turn our attention to the latter.

II
The Legal Framework Governing Personal Data in the US and EU

Data is the lifeblood of GenAI. Some of this data is "personal data," which is the term we use to refer to data that relates to or can be used to identify an individual. A subset of personal data that reveals particularly sensitive information about individuals is often labeled "sensitive personal data." Personal data can be de-identified, pseudonymized, or anonymized. De-identification or pseudonymization describes a process aimed at preventing the identification of individuals in the data set itself, though it is possible to re-identify individuals by combining de-identified or pseudonymized data with data keys or other datasets. On the other hand, anonymized data usually refers to data that cannot be re-identified. Finally, publicly available data usually describes data, including personal data, that is already accessible to the public either through public records or through an individual's own publication.

A. Legal Framework Governing Personal Data in the US

US data privacy law is "a hodgepodge of various constitutional protections, federal and state statutes, torts, regulatory rules, and treaties."[34] Although there is

---

[30] Gloria Liu, *The World's Smartest Artificial Intelligence Just Made Its First Magazine Cover*, Cosmopolitan (Jun. 21, 2022), https://www.cosmopolitan.com/lifestyle/a40314356/dall-e-2-artificial-intelligence-cover/ [https://perma.cc/6GPD-QUWN].

[31] Bryan Clark, *Check out this Beatles-inspired song written entirely by AI*, Next Web (Sept. 22, 2016), https://thenextweb.com/news/check-out-this-beatles-inspired-song-written-entirely-by-ai [https://perma.cc/W492-VUP4].

[32] *AI-Powered Medical Record Summarization Platform*, Digit. Owl, https://www.digitalowl.com/ [https://perma.cc/SQ6Q-4JEL].

[33] *See* Class Action Complaint & Demand for Jury Trial, Silverman v. OpenAI, Inc., No. 3:23-cv-03416 (N.D. Cal. Jul. 7, 2023) (alleging that OpenAI used copyrighted works to train ChatGPT).

[34] Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 Colum. L. Rev. 583, 587 (2014).

no single piece of federal legislation that governs personal data, there are several sector-specific federal laws that may offer protection for certain types of personal data, including the Federal Trade Commission (FTC) Act, the Gramm-Leach-Bliley Act, the Children's Online Privacy Protection Act (COPPA), the Family Educational Rights and Privacy Act (FERPA), and the Health Insurance Portability and Accountability Act (HIPAA).[35] At the state level, several states have enacted general data privacy laws, including California and Virginia.[36] Finally, state tort law might also govern the collection and use of some personal data through privacy- and property- related torts.

1. Federal Laws

Section 5 of the Federal Trade Commission (FTC) Act prohibits deceptive and unfair business practices.[37] It gives the FTC legal authority to establish, monitor, and enforce rules concerning deceptive and unfair practices that harm consumers, which can include protecting consumers' personal data.[38] The FTC adopts a consumer-centric concept of personal data by focusing on data that can be "reasonably linked to a specific consumer, computer, or other device."[39] The FTC has indicated that genetic data, biometric data, precise location data, and data concerning health are sensitive categories of consumers' personal data.[40]

---

[35] *See* Federal Trade Commission Act § 5, 15 U.S.C. § 45 (consumer data generally); Gramm-Leach-Bliley Act tit. V, 15 U.S.C. §§ 6801–6809, §§ 6821–6827 (financial services consumer data); Children's Online Privacy Protection Act §§ 1301–1308, 15 U.S.C. §§ 6501–6505 (children's online data); Family Educational Rights and Privacy Act § 438, 20 U.S.C. § 1232g (codifying FERPA) (educational data); Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104–191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29, 42 U.S.C.) (healthcare data).

[36] Andrew Folks, *US State Privacy Legislation Tracker*, IAPP (Feb. 2024), https://iapp.org/resources/article/us-state-privacy-legislation-tracker/ [https://perma.cc/KNF8-38A4].

[37] Federal Trade Commission Act § 5, 15 U.S.C. § 45.

[38] *Id.* § 5(a), (n).

[39] FTC, *supra* note 7, at 22 (clarifying the final scope of the FTC's framework).

[40] Elisa Jillson, *The DNA of privacy and the privacy of DNA*, FTC Bus. Blog (Jan. 5, 2024), https://www.ftc.gov/business-guidance/blog/2024/01/dna-privacy-privacy-dna [https://perma.cc/Q8Q2-9JYS] (considering voice recordings and videos highly sensitive data); Kristin Cohen, *Location, health, and other sensitive information: FTC committed to fully enforcing the law against illegal use and sharing of highly sensitive data*, FTC Bus. Blog (Jul. 11, 2022), https://www.ftc.gov/business-guidance/blog/2022/07/location-health-and-other-sensitive-information-ftc-committed-fully-enforcing-law-against-illegal [https://perma.cc/2P6N-QFUV] ("Among the most sensitive categories of data collected by connected devices are a person's precise location and information about their health.").

In 2012, the FTC established a Privacy Framework to provide guidance to commercial entities that collect consumer data to help them avoid running afoul of the broad consumer protections that the FTC enforces through the FTC Act.[41] In this framework, the FTC adopted a risk-based approach to data de-identification and considered data that cannot be reasonably linked to a consumer as being in a "de-identified form."[42] Notably, it recommended that companies take measures to reduce the risk of re-identification before considering the data truly de-identified.[43] More recently, the FTC has expressed skepticism of claims that personal data is "anonymous," noting that, "[o]ne set of researchers demonstrated that, in some instances, it was possible to uniquely identify 95% of a dataset of 1.5 million individuals using four location points with timestamps."[44]

Because the FTC focuses generally on deceptive and misleading practices that harm consumers under the FTC Act, there is no exclusion for such practices that involve the use of publicly available data. In the past, the FTC has expressed concerns about the collection of publicly available data by individual reference services (IRSs)—services that provide access to databases with publicly-available data about individuals. In a 1997 report to Congress, the FTC embraced a self-regulatory approach relying on principles developed by the now-defunct IRS industry group to limit access to "non-public information," which the FTC report defined as "information about an individual that is of a private nature and neither available to the general public nor obtained from a public record."[45] Now, publicly available data collected, aggregated, and sold by IRSs can serve as a valuable source of personal data in the GenAI lifecycle.

---

[41] FTC, *supra* note 7, at 15–71.

[42] FTC, *supra* note 7, at iv, 21.

[43] FTC, *supra* note 7, at 21 (outlining obligations related to use of de-identified data); FTC, *supra* note 7, at 22 (excluding de-identified data from scope of privacy framework).

[44] Cohen, *supra* note 40.

[45] FTC, Individual Reference Services — A Report to Congress (1997), https://www.ftc.gov/reports/individual-reference-services-report-congress [https://perma.cc/4CK4-7UAZ]; *see also* Robert Gellman & Pam Dixon, Many Failures: A Brief History of Privacy Self-Regulation in the United States 7 (2011), https://worldprivacyforum.org/wp-content/uploads/2011/10/WPFselfregulationhistory.pdf [https://perma.cc/2L99-FV76] (describing the history of the IRS industry group, including its termination in 2001).

The Gramm-Leach-Bliley Act regulates consumer data provided in connection with obtaining financial services.[46] The FTC exercises its legal authority to protect the privacy of financial data under the Gramm-Leach-Bliley Act through the Financial Privacy Rule (FPR).[47] Under the FPR, "[p]ersonally identifiable financial information" includes information about a consumer obtained in connection with the provision of financial services and products that can be used to identify an individual consumer.[48] On the other hand, the FPR does not govern "[i]nformation that does not identify a consumer" as "[p]ersonally identifiable financial information."[49] It lists "aggregate information" as an example of information that will not be governed as "personally identifiable financial information."[50] The FPR does not distinguish a separate category of "sensitive" personal information.[51] The Rule only governs "nonpublic personal information,"[52] and does not regulate most "publicly available information," described as "information that you have a reasonable basis to believe is lawfully made available to the general public."[53]

The FPR takes two main approaches to protecting the privacy of nonpublic personal information. First, it requires financial institutions to provide information about their privacy policies and disclosure practices.[54] Generally, this information should be contained in a privacy notice and include a description of nonpublic personal information that is collected and disclosed, the recipients of this information, information about the ability to opt out of certain third-party disclosures, and an explanation about how information security and confidentiality are protected.[55] Second, it limits disclosures of nonpublic personal information and requires financial institutions to provide individuals with an opportunity to "opt

[46] Gramm-Leach-Bliley Act tit. V, 15 U.S.C. §§ 6801–6809, §§ 6821–6827.
[47] Financial Privacy Rule, 16 C.F.R. § 313 (2023).
[48] 16 C.F.R. § 313.3(o)(1).
[49] 16 C.F.R. § 313.3(o)(2)(ii)(b).
[50] *Id.*
[51] *See* 16 C.F.R. § 313.3(n), (o).
[52] 16 C.F.R. § 313.1(b) (excluding publicly available information from the scope of the privacy rule); *see also* 16 C.F.R. § 313.3(n) (excluding publicly available information from the definition of nonpublic personal information).
[53] 16 C.F.R. § 313.3(p)(1).
[54] 16 C.F.R. § 313.1(a)(1), (2).
[55] 16 C.F.R. § 313.6(a).

out" of certain disclosures of their nonpublic personal information.[56] Generally, the FPR allows disclosure if the required information has been provided in a privacy notice and if, after a reasonable period of time, the individual has not opted out of the disclosure.[57] However, disclosures are allowed without providing an opt-out if they are made for the purpose of having a third party perform services on the company's behalf if the company's use of the information is limited and the individual received an initial privacy notice.[58] Disclosures are allowed without providing either a privacy notice or opt-out if they are made (1) for the purpose of processing transactions requested by the individual, (2) with the individual's consent, (3) to protect confidentiality and prevent fraud, (4) in connection with compliance with industry standards, or (5) as required by law.[59]

COPPA protects the personal information of children under the age of 13.[60] The FTC implements COPPA protections through the Children's Online Privacy Protection Rule (COPPR).[61] The COPPR defines "personal information" as "individually identifiable information about an individual collected online."[62] The rule does not distinguish a separate category of "sensitive" personal data.[63] Under COPPA, de-identified data may not fall under the definition of regulated "personal information" if it does not include identifiers or trackers such as internet protocol addresses or cookies.[64] COPPR governs only personal information when it is collected from a child online, but this could also include a subset of publicly available personal data.[65]

Under COPPR, online operators cannot collect and use personal information from children without first obtaining "verifiable parental consent."[66] It also requires operators of online services to disclose information about their collection and

[56] 16 C.F.R. § 313.1(a)(3).
[57] 16 C.F.R. § 313.10.
[58] 16 C.F.R. § 313.13.
[59] 16 C.F.R. § 313.14; 16 C.F.R. § 313.15.
[60] Children's Online Privacy Protection Act §§ 1301–1308, 15 U.S.C. §§ 6501–6505.
[61] Children's Online Privacy Protection Rule, 16 C.F.R. § 312 (2022); *see also* 15 U.S.C. § 6505 (granting enforcement authority to the FTC).
[62] 16 C.F.R. § 312.2.
[63] *See id.*
[64] *See* 15 U.S.C. § 6501(8).
[65] 16 C.F.R. § 312.2.
[66] 16 C.F.R. § 312.3(b).

use of personal information obtained from children and take measures to protect the confidentiality and security of such information.[67] COPPR gives parents the right to review their children's personal information and withdraw consent for any further use of such information.[68] However, parental consent is not required when the operator collects only necessary cookies, when information is provided for the purpose of obtaining consent, or in some cases when information is used for the limited purposes of protecting the safety of a child, responding to a specific and direct request from a child, protecting website security, or complying with legal obligations.[69] Finally, COPPR requires the deletion of personal information once it is no longer "reasonably necessary to fulfill the purpose for which the information was collected."[70]

The Family Educational Rights and Privacy Act (FERPA), which is administered and enforced by the US Department of Education (DOE), provides privacy protections for educational records, which include all information directly relating to a student that is maintained by an educational institution.[71] Under FERPA, "[p]ersonally identifiable information" includes information such as names, addresses, identification numbers, date and place of birth, as well as "[o]ther information that, alone or in combination, is linked or linkable to a specific student that would allow a reasonable person in the school community, who does not have personal knowledge of the relevant circumstances, to identify the student with reasonable certainty."[72] Like COPPA and the Gramm-Leach-Bliley Act, FERPA does not distinguish a separate category of sensitive data.[73] FERPA allows the non-consensual disclosure of de-identified records and information, which it describes as "the removal of all personally identifiable information provided that the educational agency or institution or other party has made a reasonable determination that a student's identity is not personally identifiable, whether through single or multiple releases, and taking into account other reasonably available information."[74] The DOE recognizes data aggregation as a potential

[67] 16 C.F.R. § 312.3(a), (e).
[68] 16 C.F.R. § 312.3(c).
[69] 16 C.F.R. § 312.5(c).
[70] 16 C.F.R. § 312.10.
[71] Family Educational Rights and Privacy Act § 438, 20 U.S.C. § 1232g.
[72] 34 C.F.R. § 99.3 (2022).
[73] *See* 20 U.S.C. § 1232g(a)(1)(D)(4) (defining scope of education records governed by FERPA).
[74] 34 C.F.R. § 99.31(b)(1).

method for de-identifying educational data under FERPA.[75] FERPA protects personal data in education records regardless of whether the data is otherwise publicly available.[76]

FERPA prohibits disclosure of most personally identifiable information in educational records to parties outside of the educational institution unless the student (or parent for students under 18) have provided prior written consent or the disclosure meets one of the enumerated exceptions (i.e., disclosure is required to comply with law or standard, to protect the student, etc.).[77] To be valid, this consent must identify the specific records being disclosed, the purpose of the disclosure, and party or parties receiving the disclosure.[78] However, FERPA does not limit the disclosure of "directory information," which includes "the student's name, address, telephone listing, date and place of birth, major field of study, participation in officially recognized activities and sports, weight and height of members of athletic teams, dates of attendance, degrees and awards received, and the most recent previous educational agency or institution attended by the student."[79] Finally, FERPA provides students (or parents) with rights to access educational records and request amendments, including correction and deletion, of records if they are "inaccurate, misleading, or in violation of their rights of privacy."[80]

The Health Insurance Portability and Accountability Act (HIPAA)[81] governs a subset of personal data known as "[p]rotected health information" (PHI), which comprises, among other things, "[i]ndividually identifiable health information" that is created, used, or disclosed by so-called "[c]overed entit[ies]" in the course

[75] Priv. Tech. Assistance Ctr., U.S. Dep't of Educ., *Frequently Asked Questions—Disclosure Avoidance*, Student Priv. Pol'y Off. 2, https://studentprivacy.ed.gov/sites/default/files/resource_document/file/FAQs_disclosure_avoidance_0.pdf [https://perma.cc/3V8T-27AR] (updated May 2013) (discussing aggregation as a "disclosure avoidance method . . .").

[76] *See* 20 U.S.C. § 1232g.

[77] 20 U.S.C. § 1232g(b)(1); *accord.* 34 C.F.R. § 99.30 (requiring consent from parents); *see also* 20 U.S.C. § 1232g(d) (transferring parents' rights under FERPA to students when they turn 18 or enroll in postsecondary education).

[78] 34 C.F.R. § 99.30.

[79] 20 U.S.C. § 1232g(a)(5)(A).

[80] 20 U.S.C. § 1232g(a)(2) (providing right to correction and deletion); *see also* 20 U.S.C. § 1232g(a)(1)(A) (providing right to access).

[81] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104–191, 110 Stat. 1936 (codified as amended in scattered sections of 18, 26, 29, 42 U.S.C.).

of providing healthcare services.[82] "Individually identifiable health information" describes health information that identifies or can be reasonably used to identify an individual.[83] The US Department of Health and Human Services (HHS) regulates PHI privacy through HIPAA's Privacy Rule.[84] The Privacy Rule governs the use and disclosure of patients' PHI by a limited category of "covered entities," which generally includes healthcare providers, insurers, clearinghouses, and the "business associates" of covered entities.[85] Health data is not regulated as PHI under HIPAA's Privacy Rule if it "does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual."[86] PHI can be de-identified either through an expert determination that the re-identification risk is "very small" or by removing 18 specific identifiers.[87] De-identification can be accomplished using data aggregation, usually in combination with other de-identification techniques.[88] HIPAA does not exempt publicly available personal information, which would still be considered PHI if it is created or maintained by a covered entity and sufficiently relates to a patient's medical treatment.[89]

HIPAA prohibits covered entities from using or disclosing PHI except in the following circumstances: (1) disclosure to the individual concerned, (2) for healthcare-related purposes, (3) pursuant to an authorization from the individual concerned or their representative, (4) to maintain directory information or for notification purposes after allowing the concerned individual an opportunity to object, (5) in emergency situations, (6) as required by law or public interest, or

---

[82] 45 C.F.R. § 160.103 (2023).

[83] *Id.*

[84] 42 U.S.C. § 1302(a); *see also* 42 U.S.C. §§ 1320d–1320d-9 (outlining responsibilities of the Department of Health and Human Services); HIPAA Privacy Rule, 45 C.F.R. § 164 (2023).

[85] 45 C.F.R. § 160.102 (defining covered entities).

[86] 45 C.F.R. § 164.514(a).

[87] 45 C.F.R. § 164.514(b).

[88] *See Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, U.S. Department of Health and Human Services [hereinafter *Guidance Regarding Methods for De-identification of Protected Health Information*] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [https://perma.cc/TT45-UATR] (last updated Oct. 25, 2022) (discussing aggregation to de-identify PHI); Priv. Tech. Assistance Ctr., *supra* note 75 (discussing aggregation as a "disclosure avoidance method").

[89] *See* Health Insurance Portability and Accountability Act of 1996, §1171, 42 U.S.C. § 1320d; *accord.* 45 C.F.R. § 160.103 (2023).

(7) when a limited data set that excludes certain direct identifiers is used for research, public health or health care operations.[90] For an authorization to be valid, it must provide information about the specific PHI disclosed and the person receiving the PHI and contain an expiration date.[91] Additionally, HIPAA sets forth the "minimum necessary" standard, which generally requires covered entities to limit their use and disclosure of PHI to what is necessary to accomplish a particular purpose.[92] Finally, HIPAA provides individuals with the right to review their PHI and the right to request amendments to inaccurate or incomplete PHI.[93]

2. State Laws

At the State level, California was the first state to pass a comprehensive data privacy law in 2018 (effective since January 2020),[94] followed by Virginia in 2021 (effective since January 2023).[95] Meanwhile, several other states have passed broad privacy laws, all of which have already become effective or will become effective in the next few years.[96] In this article, we focus on the regulation of personal data in California and Virginia.

In California, the California Consumer Privacy Act (CCPA), recently amended by the California Privacy Rights Act (CPRA),[97] uses the term "personal information," which is defined as "information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household."[98] Virginia's Consumer Data Protection Act (VCDPA) defines "personal data" as "any

---

[90] 45 C.F.R. § 164.502(a).

[91] 45 C.F.R. §§ 164.501, 164.532.

[92] 45 C.F.R. § 164.502(b), *accord.* 164.514(d).

[93] 45 C.F.R. § 164.524 (regarding right to access PHI); 45 C.F.R. § 164.526 (regarding right to amend PHI).

[94] California Consumer Privacy Act of 2018, 2018 Cal. Legis. Serv. Ch. 55 (West) (codified as amended at Cal. Civ. Code §§ 1798.100 –.199.100 (West 2023)).

[95] Consumer Data Protection Act, 2021 Va. Acts Ch. 35 (codified as amended at Va. Code Ann. §§ 59.1–575 to–585 (West 2023)).

[96] *See generally* Folks, *supra* note 36; Conor Murray, *U.S. Data Privacy Protection Laws: A Comprehensive Guide*, Forbes (Apr. 21, 2023), https://www.forbes.com/sites/conormurray/2023/04/21/us-data-privacy-protection-laws-a-comprehensive-guide/?sh=ce3727c5f925 [https://perma.cc/5APB-EBJ3] (providing timeline of state data protection laws).

[97] California Privacy Rights Act of 2020 § 24, Cal. Legis. Serv. Prop. 24 (West) (approved by the voters at the Nov. 3, 2020 election) (codified as amended at Cal. Civil Code § 1798.185 (West 2024)).

[98] Cal. Civil Code § 1798.140(v) (West 2024).

information that is linked or reasonably associated to an identified or identifiable natural person."[99] Both the CCPA and the VCDPA recognize a subcategory of sensitive personal data. The CCPA uses the term "sensitive personal information," which includes a consumer's government identification numbers, financial account information, geolocation data, email and text message content, genetic and biometric data used to identify an individual, and data concerning race, religion, ethnicity, philosophical beliefs, union membership, health, sexual orientation, and sex life.[100] Under the VCDPA, "sensitive data" is personal data that reveals an individual's race, ethnicity, religion, medical diagnoses, sexual orientation, citizenship, immigration status, personal data of a child, geolocation data, and genetic or biometric data processed for the purpose of identifying an individual.[101]

The CCPA and VCDPA both generally describe de-identified data as information that cannot be reasonably linked to an individual consumer.[102] In these cases, de-identified data is not regulated as personal data as long as companies take reasonable measures to ensure that the data is properly de-identified and reduce the risk of re-identification.[103] Because the CCPA and the VCDPA do not govern de-identified data (albeit with a relatively narrow definition of such data), the more stringent category of anonymous data also falls outside of their scope. Regarding data aggregation, the CCPA explicitly excludes "aggregated consumer information," which it defines as "information that relates to a group or category of consumers, from which individual consumer identities have been removed, that is not linked or reasonably linkable to any consumer or household, including via a

[99] VA. CODE ANN. § 59.1-575 (West 2023).

[100] CAL. CIVIL CODE § 1798.140(ae) (West 2024).

[101] VA. CODE ANN. § 59.1-575 (West 2023). Other States have passed data privacy laws that include similar categorizations of sensitive personal data. *See generally* Zachary S. Schapiro, *Update: Processing Sensitive Personal Information under U.S. State Privacy Laws*, GREENBERG TRAURIG, LLP (Sept. 12, 2023), https://www.gtlaw-dataprivacydish.com/2023/09/update-processing-sensitive-personal-information-under-u-s-state-privacy-laws/ [https://perma.cc/9X5A-K4KA].

[102] CAL. CIV. CODE § 1798.140(m) (West 2023) ("'Deidentified' means information that cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer."); VA. CODE ANN. § 59.1-575 (West 2023) ("'De-identified data' means data that cannot reasonably be linked to an identified or identifiable natural person, or a device linked to such person.").

[103] *See* CAL. CIV. CODE § 1798.140(m) (West 2023) (outlining obligations related to use of de-identified data); CIV. § 1798.140(v)(3) (excluding de-identified data from the definition of "personal information"); VA. CODE ANN. § 59.1-581(A) (West 2023) (outlining obligations related to use of de-identified data); § 59.1-575 (excluding de-identified data from the definition of personal data).

device."[104] While the VCDPA does not expressly exempt aggregate data from its scope, aggregate data that can no longer be linked to an individual will likely fall outside of the definition of "personal data" under the act.[105]

Neither the CCPA nor the VCDPA govern publicly available data. The CCPA excludes "publicly available information," which includes "information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience."[106] Similarly, the VCDPA excludes "publicly available information," which it defines as

> [I]nformation that is lawfully made available through federal, state, or local government records, or information that a business has a reasonable basis to believe is lawfully made available to the general public through widely distributed media, by the consumer, or by a person to whom the consumer has disclosed the information, unless the consumer has restricted the information to a specific audience.[107]

Unlike the VCDPA, the CCPA does not consider "biometric information collected by a business about a consumer without the consumer's knowledge" to be publicly available.[108]

The CCPA and VCDPA impose several obligations on businesses that collect and/or use consumers' personal data. Prior to or at the time of collection, businesses must provide consumers with information about the types of personal data collected, the purposes of the collection, and information about sharing data with third parties.[109] Unlike the CCPA, which only requires notification, the VCDPA requires consent prior to the collection or use of sensitive personal data.[110] Additionally, under both regimes, the collection and use of consumers'

---

[104] Cal. Civ. Code § 1798.140(b) (West 2024).

[105] *See* David Zetoony, *What Is Aggregated Data?*, Greenberg Traurig, LLP (Oct. 21, 2022) (citing Va. Code Ann. § 59.1-571 (2022)), https://www.gtlaw-dataprivacydish.com/2022/10/what-is-aggregated-data/ [https://perma.cc/W2YR-EJ5F].

[106] Cal. Civ. Code § 1798.140(v)(2) (West 2024).

[107] Va. Code Ann. § 59.1-575 (West 2023).

[108] Cal. Civ. Code § 1798.140(v)(2) (West 2024) (excluding collection of biometric information without consumer's consent from the definition of "publicly available").

[109] Cal. Civ. Code § 1798.100(a)(1)–(2) (West 2024); Va. Code Ann. § 59.1-578(C) (West 2023).

[110] *Compare* Cal. Civ. Code § 1798.100(a)(2) (West 2024) (requiring notification), *with* Va. Code Ann. § 59.1-578(A)(5) (West 2023) (requiring consumer consent).

personal data must be "reasonably necessary and proportionate" to accomplish the stated purpose or compatible purposes.[111] Under both laws, consumers have the right to opt-out of the sale, sharing, or further disclosure of their personal information,[112] request deletion of personal information,[113] and request the amendment of inaccurate personal information.[114]

In addition to public laws that regulate the collection and use of personal data, private law might also govern personal data when it intersects with certain privacy interests. An individual has an interest in the privacy of their personal data, which includes a right to "control of information concerning his or her person."[115] A violation of such interest may give rise to tort claims for invasion of privacy, disclosure of private information, or intrusion upon seclusion if an individual suffers a legally cognizable harm.[116] However, not all personal data implicate protected privacy interests. An individual must first have a reasonable expectation of privacy in the type of personal data that is obtained or disclosed.[117] Absent this, the individual will not suffer the type of harm that is sufficient to confer standing.[118] To determine whether users have a protected privacy interest in their personal data, it is relevant to consider "whether the data itself is sensitive *and* whether the manner it was collected . . . violates social norms."[119] Using these considerations, the Ninth Circuit agreed that Facebook users had a reasonable expectation of privacy in the "enormous amount of individualized data" that Facebook secretly obtained by using cookies to track the browsing activity of users who were logged out of the platform.[120] On the other hand, the District Court for the Western District of Washington noted that "[d]ata and information that has been found insufficiently

[111] Cal. Civ. Code § 1798.100(c) (West 2024); *accord.* Va. Code Ann. § 59.1-578(A)(1)–(2) (West 2023) (requiring data collection and further processing to be "adequate, relevant, and reasonably necessary in relation to the purposes for which such data is processed" and "compatible with the disclosed purposes.").

[112] Cal. Civ. Code § 1798.135 (West 2024) (notification of consumers right to opt out); Civ. § 1798.120 (consumers' right to opt out); Va. Code Ann. § 59.1-575, 577(A)(5) (West 2023) (consumers' right to opt out of disclosure to third parties).

[113] Cal. Civ. Code § 1798.105 (West 2024); Va. Code Ann. § 59.1-577(A)(3) (West 2023).

[114] Cal. Civ. Code § 1798.106 (West 2024); Va. Code Ann. § 59.1-577(A)(2) (West 2023).

[115] U.S. Dep't of Just. v. Reps. Comm. for Freedom of the Press, 489 U.S. 749, 763 (1989).

[116] *See* Cook v. GameStop, Inc., No. 2:22-CV-1292, 2023 WL 5529772, at *4 (W.D. Pa. Aug. 28, 2023).

[117] *See* Saeedy v. Microsoft Corp., No. 23-CV-1104, 2023 WL 8828852, at *6 (W.D. Wash. Dec. 21, 2023).

[118] *See* Popa v. PSP Grp., LLC, No. C23-0294JLR, 2023 WL 7001456, at *3–5 (W.D. Wash. Oct. 24, 2023).

[119] *In re* Facebook, Inc. Internet Tracking Litig., 956 F.3d 589, 603 (9th Cir. 2020).

[120] *Id.*

personal includes mouse movements, clicks, keystrokes, keywords, URLs of web pages visited, product preferences, interactions on a website, search words typed into a search bar, user/device identifiers, anonymized data, product selections added to a shopping cart, and website browsing activities."[121]

B. The GDPR Framework Governing Personal Data in the EU

In the EU, the collection, use, and retention of personal data is generally prohibited unless allowed by law. The General Data Protection Regulation (GDPR), effective since May 18, 2018, is an EU Regulation that directly governs the processing of personal data in all 27 EU Member States.[122]

In general, the GDPR governs entities that process personal data of individual "data subjects."[123] These entities are regulated as "Controllers" and/or "Processors" depending on what they do with the personal data. Controllers "determine[] the purposes and means of the processing of personal data," while processors "processes personal data on behalf of the controller."[124] Additionally, the GDPR is said to have an extraterritorial scope because, under certain conditions, it can govern controllers and processors outside of the EU who process personal data of a data subject located in the EU.[125]

Data "processing" means "any operation or set of operations which is performed on personal data," and includes, for example, collection, using, storing, de-identifying, transferring, and deleting personal data."[126] Under the GDPR, personal data can include any information that identifies a natural person or can be used to identify a natural person (directly or indirectly), including "a name, an identification number, location data, an online identifier or to one or more factors

---

[121] *Saeedy*, 2023 WL 8828852, at *4.

[122] General Data Protection Regulation, 2016 O.J. (L 119) Art. 4.2. The Regulation defines "processing" as:

> any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

[123] *Id.* at Art. 3 (providing scope of GDPR); *Id.* at Art. 4.1 (defining "Data Subject").

[124] *Id.* at Art. 4.7 (defining "Controller"); *Id.* at Art. 4.8 (defining "Processor").

[125] *Id.* at Art. 3.

[126] *Id.* at Art. 4.2.

specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."[127] Notably, the GDPR does not exclude publicly available data from its scope.[128] "Special categories of data" (or sensitive personal data), under the GDPR include "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."[129]

The EU GDPR does not recognize a category of "de-identified data." Instead, it refers to personal data that has undergone "pseudonymization," which is

> the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.[130]

Notably, the GDPR does not exempt pseudonymized data from its scope, but instead views pseudonymization as a method for protecting personal data.[131] Conversely, the GDPR does not govern anonymous data, which it describes as, "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable."[132] However, there is conflicting regulatory

---

[127] *Id.* at Art. 4.1.

[128] Piotr Foitzik, *Publicly available data under the GDPR: Main considerations*, IAPP (May 28, 2019), https://iapp.org/news/a/publicly-available-data-under-gdpr-main-considerations/ [https://perma.cc/4CGW-RRJF].

[129] General Data Protection Regulation, 2016 O.J. (L 119) Art. 9.1.

[130] *Id.* at Art. 4.5.

[131] *See* Recital 28 (noting that pseudonymization "is not intended to preclude any other measures of data protection."); art. 23 (describing pseudonymisation and encryption as a security measure); art. 25 (defining pseudonymization as a technical and organizational measure to safeguard data); Recital 26 (explaining that pseudonymized data is considered data relating to an "identifiable natural person").

[132] Recital 26.

guidance in the EU about how to interpret the anonymization standard.[133] Some regulatory authorities take an absolutist approach that considers data anonymous only when it is impossible to re-identify the data, while others adopt a risk-based approach that considers data anonymous when there is no reasonable chance of re-identification.[134] Since absolute anonymization may be "statistically impossible," EU regulatory authorities have tended to adopt a risk-based approach to anonymization.[135] The GDPR will also not apply to fully "aggregated and anonymised datasets" when the "original input data … [is] destroyed, and only the final, aggregated statistical data is kept."[136]

The GDPR protects personal data through its principles of lawfulness, fairness, transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity, confidentiality, and accountability.[137] To lawfully process personal data under the GDPR, there must be a lawful basis to support data processing.[138] Under Article 6, there are six legal justifications upon which personal data processing can be based: (1) consent, (2) contract, (3) legal obligation, (4) vital interests of natural person, (5) public interest or official authority, and (6) legitimate interests of controller or third party.[139] Processing special categories of personal data, including biometric data, data concerning health, data revealing racial or ethnic origins, and data potentially revealing sexual orientation, political opinions, religious, or philosophical beliefs, is generally

[133] *See generally* Andrew Burt, Alfred Rossi & Sophie Stalla-Bourdillon, *A guide to the EU's unclear anonymization standards*, IAPP (Jul. 15, 2021), https://iapp.org/news/a/a-guide-to-the-eus-unclear-anonymization-standards/ [https://perma.cc/2FGM-QB54].

[134] *Id.*

[135] Andrew Burt & Sophie Stalla-Bourdillon, *The definition of 'anonymization' is changing in the EU: Here's what that means*, IAPP (Jun. 27, 2023), https://iapp.org/news/a/the-definition-of-anonymization-is-changing-in-the-eu-heres-what-that-means/ [https://perma.cc/Q3CJ-RSMU].

[136] Eur. Data Prot. Supervisor, Opinion 10/2017, EDPS Opinion on safeguards and derogations under Article 89 GDPR in the context of a proposal for a Regulation on integrated farm statistics 10 (2017), https://www.edps.europa.eu/sites/default/files/publication/17-11-20_opinion_farm_statistics_en.pdf [https://perma.cc/9F5Y-EN37].

[137] General Data Protection Regulation , 2016 O.J. (L 119) Art. 5.

[138] Art. 6.

[139] *Id.*

prohibited and only allowed under certain conditions laid out in Article 9, such as obtaining "explicit" consent of the natural person.[140]

Consent can only be a valid legal basis for processing personal data if it is (1) freely given, (2) specific, (3) informed, (4) unambiguous, and (5) as easy to give as to withdraw.[141] For consent to be freely given, users must be able to refuse consent without suffering "significant negative consequences."[142] For consent to be specific, the user must provide consent for a specific purpose.[143] According to the EDPB (previously the Article 29 Data Protection Working Party, which coordinates GDPR enforcement), neither blanket consent "for all the legitimate purposes" nor consent based on "an open-ended set of processing activities" are valid.[144] For consent to be informed, the user must have the following information: (1) the identity of the controller(s), (2) the purpose for collecting and further processing the data, (3) the category of data collected, (4) the right to withdraw consent, (5) the existence of automated decision making (if any), and (6) the risks and safeguards associated with transferring data to third countries.[145] For consent to be unambiguous, the user should provide consent via an "affirmative action."[146] A clear affirmative action in the digital sphere can include sending an email, submitting an online form, using an electronic signature, or ticking a box.[147] Consent based on a user's silence or inaction will not be valid.[148]

---

[140] Art. 9 (prohibiting the processing of special categories of data except in the following cases: (1) explicit consent; (2) employment social security, or social protection; (3) vital interests of a natural person incapable of providing consent; (4) when data subject makes the data public; (5) substantial public interest; (6) healthcare; (7) public health; and (8) archiving, scientific or historical research, and statistical purposes); *see also* Eur. Data Prot. Bd., Guidelines 3/2019 on Processing of Personal Data 18–20 (2020) [hereinafter Guidelines on Processing of Personal Data], https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201903_video_devices.pdf [https://perma.cc/RSP3-N6DZ].

[141] Art. 4.11.

[142] Eur. Data Prot. Bd., Guidelines 05/2020 on consent under Regulation 2016/679 9, 12 (2020), https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_202005_consent_en.pdf [https://perma.cc/358Y-H8X2].

[143] *Id.* at 14–15.

[144] Art. 29 Data Prot. Working Party, 01187/11/EN, Opinion 15/2011 on the Definition of Consent 17 (2011), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2011/wp187_en.pdf [https://perma.cc/9Y4M-R56J].

[145] Guidelines on Processing of Personal Data, *supra* note 140, at 15–16.

[146] Guidelines on Processing of Personal Data, *supra* note 140, at 26.

[147] Guidelines on Processing of Personal Data, *supra* note 140, at 26.

[148] Guidelines on Processing of Personal Data, *supra* note 140, at 36.

At the time of personal data collection, the GDPR also requires that data subjects be informed about the existence and purposes of the processing, which includes providing information about the "specific circumstances and context in which the personal data are processed."[149] This information should be in clear, plain, and easily understandable language.[150] Once collected, the GDPR requires personal data to be stored in a manner that is sufficient to protect it "against unauthorised or unlawful processing and against accidental loss, destruction or damage."[151] This requires the implementation of technical and organizational measures to protect personal data.[152] These measures should include pseudonymization and encryption of personal data and the implementation of measures that ensure confidentiality and integrity of processing systems as well as the ability to restore data if necessary.[153] Personal data should not be stored for any longer than needed to accomplish the purposes for which it is being processed.[154] However, data can be stored for a longer period "solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes" as long as it is sufficiently protected.[155]

Finally, data subjects have a right, among other things, to make requests to withdraw consent,[156] access,[157] rectify,[158] erase,[159] transfer,[160] restrict[161] or object[162] to the processing of their personal data. The GDPR also gives individuals who suffer "material or non-material damage" as a result of a GDPR "infringement" the "right to receive compensation" and holds data controllers (and potentially even processors) "liable for the damage caused by processing which

[149] Recital 60; *accord.* art. 14.
[150] Art. 12.1.
[151] Art. 5.1(f).
[152] Art. 5.1(e); art. 32.
[153] Art. 32.
[154] Art. 5.1(e).
[155] *Id.*
[156] Art. 7.3.
[157] Art. 15.
[158] Art. 16.
[159] Art. 17.
[160] Art. 20.
[161] Art. 18.
[162] Art. 21.

infringes this Regulation."[163] As a result, not only does the GDPR regulate data privacy in public law, it also provides a private right of action for individuals who are harmed by violations of the GDPR.

III
The Flow of Personal Data Flow in the GenAI Data Lifecycle

Personal data flows through the GenAI data lifecycle in various stages. In the development phase, personal data can be used to train the model. Once the GenAI model is released, users can provide more personal data while signing up and using the model. Next, the GenAI models themselves produce data in the form of outputs provided to users, which can also include personal data. Finally, developers can retain personal data that users input or that the AI models output to improve GenAI models or develop new models.

A. Training Data

Creators of GenAI use large datasets, sourced primarily from publicly available information on the internet, to train models.[164] This data can be scraped from public social media profiles (e.g., LinkedIn), online discussions (e.g., Reddit), photo sharing sites (e.g., Flickr), blogs (e.g., WordPress), news media (e.g., arrest reports), information and research sites (e.g., Wikipedia), and even government records (e.g., voter registration records).[165] Paywalls are not always effective at protecting online data from web scrapers, and pirated data, like illegal copies of books, can end up in a scraped data set.[166]

Personal data is no exception. For example, Meta admits that personal information like a blog post author's name and contact information may be

[163] Art. 82 ("Any controller involved in processing shall be liable for the damage caused by processing which infringes this Regulation. A processor shall be liable for the damage caused by processing only where it has not complied with obligations of this Regulation specifically directed to processors or where it has acted outside or contrary to lawful instructions of the controller.").

[164] *See, e.g.*, Brown, *supra* note 16.

[165] Lauren Leffer, *Your Personal Information Is Probably Being Used to Train Generative AI Models*, Sci. Am. (Oct. 19, 2023), https://www.scientificamerican.com/article/your-personal-information-is-probably-being-used-to-train-generative-ai-models/ [https://perma.cc/P2PP-WJKR].

[166] *Id.*

collected to train its GenAI models.[167] Even sensitive personal data can be scraped to train GenAI models. Photos from a patient's medical record were included in LAION, a publicly available data set used to train some GenAI image models.[168]

B. User Input of Data

Users of GenAI applications provide various data to the companies that control and deploy the AI models, including specific user account data, user behaviors, and more general data that users input to facilitate the generation of new content by GenAI. It is no surprise that many companies collect personal data from account holders (e.g., name, age, contact information, and billing details) to provide information and services. Perhaps less obvious to users, but still considered routine, is the collection of user data for website analytics, which provide companies with insight into how users interact with their websites and applications.[169] Finally, developers can also collect personal data that users enter into various online applications, including personal data that users provide to prompt AI content generation. This is a broad category of data that can range from details about a private business deal for contracts, to information about students for letters of recommendation, to a patient's medical history for a referral. Users might also provide voice and image data to GenAI models that create art or music.

Developers might use user-provided personal data to develop or improve GenAI models.[170] According to OpenAI's Privacy Policy, it may use data that users provide, including user input of data to ChatGPT and DALL-E, for the development, training, and improvement of its GenAI models.[171] Google Cloud

---

[167] Priv. Center, *How Meta Uses Information for Generative AI models*, Meta, https://www.facebook.com/privacy/genai [https://perma.cc/RQD9-3QJ7].

[168] Benji Edwards, *Artist Finds Private Medical Record Photos in Popular AI Training Data Set*, Ars Technica (Sept. 21, 2022), https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/ [https://perma.cc/V2ES-6UZW].

[169] Google Mktg. Platform, *Analytics*, Google, https://marketingplatform.google.com/about/analytics/ [https://perma.cc/FWE6-M7L8].

[170] *See* Paul F. Christiano et al., *Deep Reinforcement Learning from Human Preferences*, 30 Advances in Neural Info. Processing Sys., 2017, *passim* (discussing the ability of deep learning algorithms to improve performance in response to human interaction and feedback though a process called Reinforcement Learning through Human Feedback (RHLF)).

[171] *Privacy Policy*, OpenAI, https://openai.com/policies/privacy-policy [https://perma.cc/9P59-DHPG] (updated Nov. 14, 2023).

Services claims that it does not train its AI models with data provided by its Cloud users without permission.[172] However, Google's Privacy Policy states that it "collect[s] the content you create, upload, or receive from others when using our services … includ[ing] things like email you write and receive, photos and videos you save, docs and spreadsheets you create, and comments you make on YouTube videos."[173] Elsewhere, the same policy states generally that Google "use[s] the information [they] collect in existing services to help [them] develop new ones" and that it "uses information to improve [their] services and to develop new products, features and technologies that benefit [their] users and the public," which can include "us[ing] publicly available information to help train Google's AI models."[174] Meta's Privacy Policy describes how it uses both the content provided by users and information about user activity to develop, improve, and test its products, which can include GenAI products.[175] This can include a Facebook user's posts, messages, voice, camera roll images and videos, hashtags, likes, and purchases.[176] Additionally, Meta's use of users' AI prompts, "which could include text, documents, images, or recordings," are also governed by broad use provisions in the Meta Privacy Policy.[177]

User-provided content might also be shared with third parties. OpenAI's privacy policy states that it may share personal information with third parties, including vendors, service providers, and affiliates.[178] OpenAI sets almost no limitations on its use of anonymous, aggregated, and de-identified data.[179] Google claims that it does not share personal information with third parties except in the following circumstances: (1) the user consents, (2) in the case of organizational use of Google by schools or companies, (3) for external processing, or (4) when sharing

[172] Andrew Moore, *Sharing Our Data Privacy Commitments for the AI Era*, Google Cloud Blog (Oct. 14, 2020), https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-unveils-ai-and-ml-privacy-commitment [https://perma.cc/GA46-48ZX].

[173] *Privacy Policy*, Google, https://policies.google.com/privacy#infosharing [https://perma.cc/G2JL-2YAG] (describing policies for sharing information).

[174] *Id.*

[175] *Privacy Policy*, Meta, https://www.facebook.com/privacy/policy?section_id=2-HowDoWeUse [https://perma.cc/RXX5-DRNY].

[176] *Id.*

[177] *Meta AIs Terms of Service*, Meta, https://www.facebook.com/policies/other-policies/ais-terms [https://perma.cc/M2PK-TWHU].

[178] *Privacy Policy*, OpenAI, *supra* note 171.

[179] *Privacy Policy*, OpenAI, *supra* note 171.

is required or permitted by law.[180] Google states, however, that it "may share non-personally identifiable information publicly and with [their] partners—like publishers, advertisers, developers, or rights holders."[181] According to Meta, "[b]y using AIs, you are instructing [them] to share your information with third parties when it may provide you with more relevant or useful responses."[182] This includes personal information about the user or third parties.[183] Meta's Privacy Policy articulates the broadest data sharing practices to include sharing all user data, including personal data, with partners, vendors, service providers, external researchers, and other third parties.[184] Meta advises users "not [to] share information that you don't want the AIs to retain and use."[185]

C. AI-Generated Output of Data

The output that GenAI models produce in response to user prompts is another source of data in the GenAI data life cycle. These outputs can include personal data that the model learned either through training or user-provided data. Despite efforts to prevent GenAI models from leaking personal data memorized in the model parameters during model training,[186] researchers were able to extract personal data, including names, phone numbers, and email addresses, from GPT-2.[187] In a later study, researchers extracted data from both open models—those with publicly available training data sets, algorithms, and parameters—and semi-open models—those with only publicly available parameters.[188] They were "able to extract over 10,000 unique verbatim memorized training examples" from ChatGPT 3.5.[189] These samples included personal data such as "phone numbers, email

[180] *Privacy Policy*, Google, *supra* note 173.
[181] *Privacy Policy*, Google, *supra* note 173.
[182] *Meta AIs Terms of Service*, *supra* note 177.
[183] *Meta AIs Terms of Service*, *supra* note 177.
[184] *Privacy Policy*, Meta *supra* note 175.
[185] *Meta AIs Terms of Service*, *supra* note 177.
[186] *How ChatGPT and Our Language Models Are Developed*, OpenAI, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed [https://perma.cc/3EAS-E74C] ("[W]e try to train our models to reject requests for private or sensitive information about people.").
[187] Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, 30 USENIX Security Symposium 2633, 2640–41 (2021), https://www.usenix.org/system/files/sec21-carlini-extracting.pdf [https://perma.cc/MKN7-SEP8].
[188] Milad Nasr, et al., *Scalable Extraction of Training Data from (Production) Language Models*, arXiv, Nov. 2023, at 7–10, https://arxiv.org/abs/2311.17035 [https://perma.cc/VV4X-BHSG].
[189] *Id.* at 9.

addresses, and physical addresses (e.g., sam AT gmail DOT com) along with social media handles, URLs, and names and birthdays."[190] Over eighty-five percent of this information was the personal data of real individuals, rather than hallucinated.[191] This research cautioned that attackers with more resources could likely gather up to 10 times more training data from ChatGPT.[192] GenAI developers disclose that outputs may be inaccurate but provide little information about the risk of the model disclosing a user's personal data.[193] OpenAI gives users the option to submit a correction request in cases where the model provides inaccurate information about a user; however, this does nothing to cure GenAI output of accurate personal data.[194]

D. Data Retention

Once personal data is collected by companies for developing and improving GenAI models, it may be retained for further use. OpenAI claims that ChatGPT does not store its training data.[195] However, it states that it stores users' personal information, which includes user input, "as long as we need in order to provide our Service to you, or for other legitimate business purposes such as resolving disputes, safety and security reasons, or complying with our legal obligations."[196] This is presumably for as long as a user has an active account.[197] Although OpenAI provides users with an option to request that their personal data no longer be processed, removal of personal information from OpenAI's applications is not guaranteed.[198] For example, OpenAI notes that it will consider requests "balancing

[190] *Id.* at 10.

[191] *Id.*

[192] *Id.* at 1.

[193] *See Privacy Policy*, OpenAI, *supra* note 171; *see also Meta AIs Terms of Service*, *supra* note 177 (warning user's about the content and accuracy of output); Google AI for Developers, *Generative AI APIs Additional Terms of Service*, Google, https://ai.google.dev/terms [https://perma.cc/765N-FLCG] (noting that Google's model may provide inaccurate or offensive content).

[194] *See Privacy Policy*, OpenAI, *supra* note 171.

[195] *How ChatGPT and Our Language Models Are Developed*, *supra* note 186.

[196] *Privacy Policy*, OpenAI, *supra* note 171.

[197] Aaron Drapkin, *Does ChatGPT Save My Data? OpenAI's Privacy Policy Explained*, tech.co, https://tech.co/news/does-chatgpt-save-my-data [https://perma.cc/4RS7-ZXLZ] (last updated Jun. 29, 2023).

[198] *OpenAI Personal Data Removal Request*, OpenAI, https://share.hsforms.com/1UPy6xqxZSEqTrGDh4ywo_g4sk30 [https://perma.cc/Q8GX-UUBQ].

privacy and data protection rights with public interests like access to information, in accordance with applicable law."[199]

Meta also allows users to submit a request to obtain, delete, or restrict Meta's use of their personal information to train its AI models.[200] Notably, this request only relates to personal data that Meta obtains from third parties and not data that it obtains directly from the user.[201] To limit the use of personal data that users provide to Meta directly for any purpose, including training GenAI models like Llama 2, users are instructed to delete their Meta accounts (Facebook and Instagram) or to exercise their rights under data protection laws.[202]

Figure A illustrates the flow of personal data in the GenAI data lifecycle.

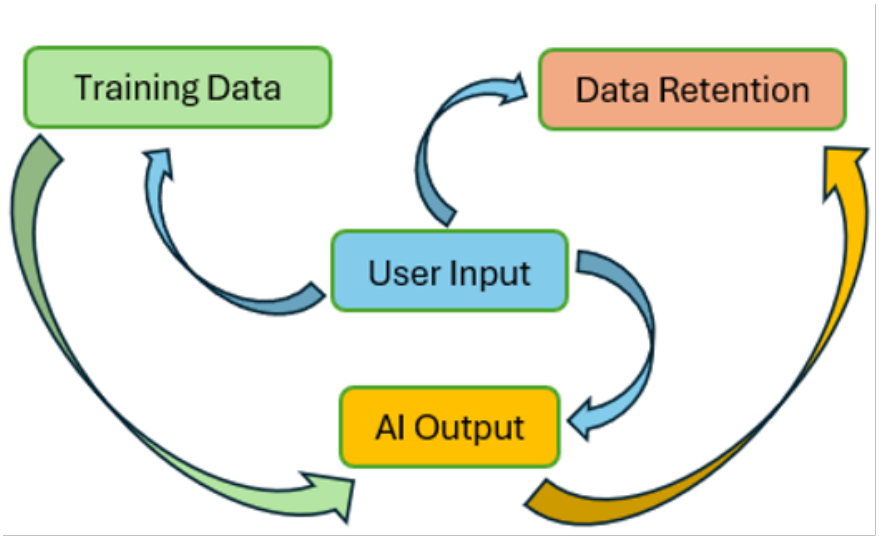

Figure A: Flow of Personal Data in the GenAI Data Lifecycle

[199] *OpenAI Privacy Center*, OpenAI, https://privacy.openai.com/policies?modal=take-control&submissionGuid=378496de-9581-4d9e-b8dd-dc48b35b219b [https://perma.cc/Q8GX-UUBQ].

[200] *Generative AI Data Subject Rights*, Facebook, https://www.facebook.com/help/contact/510058597920541 [https://perma.cc/QVH6-SZLX] (describing "personal information" as "information about you").

[201] *Id.*

[202] *Privacy Policy*, Meta *supra* note 175.

IV
The Protection of Personal Data in the GenAI Data Lifecycle in the US and EU

As demonstrated above, the flow of personal data through the GenAI data lifecycle is complex, and it can be difficult for users to keep track of how GenAI developers and third parties might access and use their personal data. As a result, several data privacy implications arise in connection with the use of publicly available, private, and sensitive personal data in the GenAI data lifecycle in both the US and EU. Additionally, individuals may also lose control over the accuracy and retention of their personal data.

A. Publicly Available Personal Data

GenAI training datasets include troves of publicly available personal data scraped from the internet. Third parties can examine these datasets to extract personal data in ways that can threaten individuals' privacy. For example, researchers analyzing the open access German-language LAION dataset were able to determine both the identity of a man in a naked photograph and the exact address where a baby's photograph was taken.[203] Additionally, GenAI models themselves might disclose personal data to third parties because they "memorize portions of the data on which they are trained; as a result, the model can inadvertently leak memorized information in its output."[204]

One view is that individuals should expect that any data they share publicly online is up for grabs and not subject to any privacy protections.[205] On the other hand, the reproduction of personal data that individuals provide online could still violate reasonable expectations of data privacy. An individual might decide to disclose personal data in a specific public online setting for a specific purpose without any expectation that the same data will be used to train GenAI models.

---

[203] Elisa Harlan & Katharina Brunner, *We Are All Raw Material for AI*, Netzwelt (Jul. 7, 2023), https://interaktiv.br.de/ki-trainingsdaten/en/index.html [https://perma.cc/KGM6-3A3B].

[204] Amy Winograd, *Loose-Lipped Large Language Models Spill Your Secrets: The Privacy Implications of Large Language Models*, 36 Harv. J.L. Tech. 615, 625 (2023).

[205] *See, e.g.*, Harlan & Brunner, *supra* note 203 (reporting statement from one of LAION's founders, Christoph Schuhmann, that "[i]n principle, that means that at the moment I make my image and my data publicly available on the internet, I should be aware that there is a very good chance that someone will download it and use it for models").

In fact, many users who originally provided personal data online could not have reasonably expected that future technology would be used by companies that did not yet exist to collect, use, profit from, and potentially publish their personal data far outside of the parameters in which it was originally shared.[206] For example, a Reddit user might share personal data in a city-based group about depression for the purpose of obtaining mental health support, but the same user would likely not expect this information to become available to a virtually unlimited audience as part of a GenAI training data set. Additionally, the context and privacy implications of disclosing personal data may change over time.[207] For example, imagine that the same user also disclosed to the online support group that her mental health condition was related to a legal abortion, and months later, that same abortion procedure is deemed illegal.

In the US, the FTC expressed early concerns about the IRS industry's collection and use of publicly available personal data, noting that "advances in computer technology have made it possible for more detailed identifying information to be aggregated and accessed more easily and cheaply than ever before."[208] It worried that consumers would be "adversely affected by a perceived privacy invasion, the misuse of accurate information, or the reliance on inaccurate information" and noted the "potential harm that could stem from access to and exploitation of sensitive information in public records and publicly available information."[209] Unfortunately, it appears that the FTC's interest in this topic seems to have waned considering the current absence of regulatory efforts focused on protecting consumers from harm that may stem from access, aggregation, and use of publicly available personal data. The US federal laws that govern subsets of personal data also offer little protection from broad disclosure of publicly available personal data. The Gramm-Leach-Bliley Act does not regulate personal

[206] *See, e.g.*, Sara Morrison, *The tricky truth about how generative AI uses your data*, Vox (Jul. 27, 2023), https://www.vox.com/technology/2023/7/27/23808499/ai-openai-google-meta-data-privacy-nope [https://perma.cc/4H4S-ZXQX].

[207] *See, e.g.*, FTC, *supra* note 45 ("[T]he same piece of information (e.g., age) may raise different privacy concerns at different points in a person's life.").

[208] FTC, *supra* note 45.

[209] FTC, *supra* note 45.

information that is lawfully made public.[210] While COPPA, FERPA, and HIPAA do not exclude publicly available information from regulation, these laws offer only limited and fragmented protection for certain subcategories of personal data that are collected and used under certain circumstances and would likely not protect publicly available personal data from data scraping practices or from being accessed in the resulting datasets.[211]

The exclusion of publicly available information from general state data privacy laws like the CCPA and VCDPA leaves open questions about whether all personal data scraped from the web is considered "publicly available."[212] Notably, the CPRA's amendments to the CCPA broadened the scope of what is considered "publicly available" to include information in government records that is used for a purpose other than that for which it was originally collected.[213] The drafters worried that limiting the use of information in public records would infringe upon constitutionally protected free speech.[214] However, personal data disclosed online might not be considered "publicly available" if the individual restricted their self-publication to a specific audience. Similarly, if a third party publishes personal data about an individual outside of the original audience restrictions, this data may not be considered publicly available. Theoretically, data scraping should not collect self-published personal data that is not available to the "general public" because of audience restrictions; however, it can be difficult to determine whether personal data published to the general public by a third party was originally restricted to a specific audience. For example, individuals, companies, or government organizations might make the personal data of another individual publicly available, and the collection of this data to train GenAI might violate the data subject's privacy.

---

[210] 16 C.F.R. § 313.1(b) (excluding publicly available information from the scope of the privacy rule); § 313.3(n) (excluding publicly available information from the definition of nonpublic personal information), see also § 313.3(p)(1) (defining publicly available information).

[211] *See supra* Part II.A.1.

[212] CAL. CIV. CODE § 1798.140(v)(2) (West 2024) (excluding publicly available information); *accord.* VA. CODE ANN. § 59.1-575 (West 2023) (excluding publicly available information).

[213] Act of October 11, 2019 § 7, 2019 Cal. Legis. Serv. Ch. 757 (West) (codified as amended at CAL. CIV. CODE § 1798.140) (West 2024) ("For these purposes, 'publicly available' means information that is lawfully made available from federal, state, or local government records, if any conditions associated with such information.").

[214] *See* S. 2019-1355, Reg. Sess., at 4 (Cal. 2019).

Unlike US privacy laws, the GDPR does not exclude publicly available data from its scope, and processing of such data requires a valid legal basis.[215] Consent, contract, and/or legitimate interests are the most likely legal bases for processing personal data in the GenAI data lifecycle.[216] However, because scraping publicly available personal data generally occurs without knowledge of or contact with the data subjects concerned, advisory committees have noted that there should be sufficient legitimate interests to support the processing of personal data through data scraping.[217] However, the public availability of data can increase the supporting legitimate interests of a processor, "if the publication was carried out with a reasonable expectation of further use of the data for certain purposes."[218] On the other hand, the broad collection and use of publicly available data implicates a wide range of privacy concerns for a large number of data subjects who are likely not aware that their personal data will be used for training GenAI models.[219] The EDPB ChatGPT taskforce notes that when balancing the data controller's legitimate interests with data subjects' fundamental rights, "reasonable expectations of data subjects should be taken into account."[220] However, it also indicates that data controllers who employ technical measures to (1) exclude certain

[215] Piotr Foitzik, *Publicly available data under the GDPR: Main considerations*, IAPP (May 28, 2019), https://iapp.org/news/a/publicly-available-data-under-gdpr-main-considerations/ [https://perma.cc/8M8L-B5VU]; Art. 29 Working Party, WP251rev.01, Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679 (2018) (noting that publicly available personal data is still personal data governed by GDPR).

[216] *See* Confederation of Eur. Data Prot. Orgs. AI Working Grp., *supra* note 7, at 8; *see also What is our legal basis?*, Meta, https://www.facebook.com/privacy/policy/?subpage=7.subpage.1-WhatIsOurLegal [https://perma.cc/6HLQ-3Y8Y] ("We process your information that's necessary to fulfil our contracts with you … We process your information if you give your consent … We process your information as necessary for our or others' legitimate interests. Our interests include providing an innovative, personalised, safe and profitable service to our users and partners, and responding to legal requests.") (last accessed from the Netherlands on Jun. 4, 2024).

[217] *See* Confederation of Eur. Data Prot. Orgs. AI Working Grp., *supra* note 7, at 10 ("[T]here is very little room for the contract basis when training an AI system."); Confederation of Eur. Data Prot. Orgs. AI Working Grp., *supra* note 7, at 8 ("The entire apparatus used for the training of AI systems makes it almost impossible to obtain consent.").

[218] Art. 29 Data Prot. Working Party, 844/14/EN, Opinion 06/2014 on the Notion of Legitimate Interests of the Data Controller Under Article 7 of Directive 95/46/EC 39 (2014) [hereinafter Opinion 06/2014] (explaining that the data may have been originally published "for purposes of research or for purposes related to transparency and accountability").

[219] *Id.* (noting that impact on fundamental rights considers the amount of data processed and the breath of access to personal data).

[220] EDPB, *supra* note 6, at 6.

categories of data from data scraping and (2) delete or anonymize data that has been scraped are more likely to succeed in claiming legitimate interests as a valid legal basis for processing training data.[221]

Scraping publicly available data for training GenAI can be particularly intrusive when the data can be used to make predictions about an individual's behavior.[222] In Europe, the French, Italian, and Greek national data protection authorities ("DPAs"), charged with national enforcement of the GDPR, imposed fines on US-based Clearview AI for GDPR violations stemming from the company's data scraping practices.[223] Specifically, the DPAs found that Clearview AI's scraping of over 10 billion publicly available facial images and associated metadata for the purpose of further processing those images to create a facial recognition database violated the GDPR's requirements of lawfulness, fairness, and transparency in personal data processing.[224] The DPAs rejected the company's alleged legitimate interests in making a business profit as a valid legal basis for their data processing and ordered the erasure of such data, banned further collection, and fined the company 20 million euros each.[225] In assessing the balance between the Clearview AI's economic interest and data subjects' fundamental rights, the DPAs highlighted that (1) the biometric data produced from the processing of these images was particularly intrusive and concerns a large number of people, and (2) the data subjects were not aware and could not have reasonably expected that their photographs and the associated metadata would be used to develop

[221] *See* EDPB, *supra* note 6, at 6–7.

[222] *See* Art. 29 Data Prot. Working Party, *supra* note 144, at 39.

[223] Garante per la Protezione dei Dati Personali [Guar. for the Prot. Of Pers. Data], Ordinanza ingiunzione nei confronti di Clearview AI [Injunction Order Against Clearview AI] (2022), https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9751362 [https://perma.cc/FR9U-AYTG] (Italy); *Hellenic DPA fines Clearview AI 20 million euros*, Eur. Data Prot. Bd. (July 20, 2022), https://www.edpb.europa.eu/news/national-news/2022/hellenic-dpa-fines-clearview-ai-20-million-euros_en [https://perma.cc/SFW5-3BT2] (Greece); Commission Nationale de l'Informatique et des Libertés [Nat'l Comm'n for Info. Tech & Libertés], SAN-2022-019, Délibération de la formation restreinte n° SAN-2022-019 du 17 octobre 2022 concernant la société CLEARVIEW AI [Deliberation of the restricted panel concerning the company CLEARVIEW AI] (2022), https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000046444859 [https://perma.cc/6MZY-9TQD] (France).

[224] *See* Injunction Order Against Clearview AI, *supra* note 223.

[225] *See* Injunction Order Against Clearview AI, *supra* note 223; *Hellenic DPA fines Clearview AI 20 million euros*, *supra* note 223; Deliberation of the restricted panel concerning the company CLEARVIEW AI, *supra* note 223.

facial recognition software when they consented to the original publication of their photographs.[226]

Although, as demonstrated by the Clearview AI case, the general framework of the GDPR is already flexible enough to govern the large-scale processing of publicly available personal data, the EU also appears focused on protecting data subjects from illegal use of publicly available personal data for the specific purpose of training GenAI models. In addition to the EDPB's ChatGPT taskforce, the Confederation of European Data Protection Organizations (CEDPO), which seeks to harmonize data protection practices in the EU member states, developed its own taskforce to address GDPR compliance during the processing phase of the GenAI data lifecycle.[227] On the national level, the French data protection authority issued an action plan that will consider, among other issues, "the protection of publicly available data on the web against the use of scraping, or scraping, of data for the design of tools."[228] Notably, the EU is not ignoring the potential benefits of GenAI or attempting to regulate these models out of existence; instead, the CEDPO recognizes that, "[t]here will be no future without generative AI, and with data playing such a pivotal role in the training and operating of these systems, DPOs will play a central role in ensuring that both data protection and data governance standards are at the heart of these technologies."[229]

Finally, although data scraping is intended to collect only data that is legally available to the general public, it is also important to recognize that this practice can amplify existing data privacy violations resulting from the unauthorized disclosure of sensitive personal data online. For example, an outdated data protocol for electronic medical record storage resulted in the unauthorized online disclosure of more than 43 million health records, which included patients' names, genders, addresses, phone numbers, social security numbers, and details from medical

[226] *See* Injunction Order Against Clearview AI, *supra* note 223; *Hellenic DPA fines Clearview AI 20 million euros*, *supra* note 223; Deliberation of the restricted panel concerning the company CLEARVIEW AI, *supra* note 223.

[227] Confederation of Eur. Data Prot. Orgs. AI Working Grp., *supra* note 7.

[228] *Artificial intelligence: the action plan of the CNIL*, Commission Nationale de l'Informatique et des Libertés (May 23, 2023), https://www.cnil.fr/en/artificial-intelligence-action-plan-cnil [https://perma.cc/2YA6-H3V9].

[229] Confederation of Eur. Data Prot. Orgs. AI Working Grp., *supra* note 7, at 2.

examinations.[230] This sensitive personal data might be collected in a data scrape, used to train GenAI models, and end up in the hands of third parties, even though it should not have been publicly available to begin with.

B. Private and Sensitive Personal Data

As a source of personal data in the GenAI data lifecycle, individual users can supply private (i.e. not publicly available) and sensitive personal data about themselves or others while using services and software provided by GenAI companies, including the GenAI applications themselves. In fact, the conversational nature of GenAI applications can cause users to let their guard down and overshare personal data.[231] The GenAI model might then use this personal data to infer more personal data—even sensitive personal data—about an individual.[232] The EDPB ChatGPT taskforce notes that despite policies that warn users to refrain from providing personal data to ChatGPT, "it should be assumed that individuals will sooner or later input personal data," and that this data must still be processed lawfully.[233] Additionally, geolocation services and wearable technology might also provide troves of sensitive personal data. While users might agree to share location data to use Google Maps and Uber, they may not realize that this data can "reveal a lot about people, including where we work, sleep, socialize, worship, and seek medical treatment."[234] Additionally, other technologies, like smartwatches and apps that monitor blood sugar or menstrual cycles, collect sensitive personal data from users. Once private and sensitive personal data are in the digital marketplace,

[230] Carly Page, *Millions of patient scans and health records spilling online thanks to decades-old protocol bug*, TechCrunch (Dec. 6, 2023), https://techcrunch.com/2023/12/06/medical-scans-health-records-dicom-pacs-security/ [https://perma.cc/W383-BTZU]. In the IP context, data scraping can also retrieve copyright-protected work contained in illegal online "shadow libraries." *See e.g.*, Class Action Complaint, Silverman. v. Open AI, Inc., No. 3:23-cv-03416 at ¶¶ 35–36 (N.D. Cal. Jul. 7, 2023).

[231] Dana Mancuso, *Privacy considerations for Generative AI*, Univ. of Ill. (July 17, 2023), https://cybersecurity.illinois.edu/privacy-considerations-for-generative-ai/ [https://perma.cc/VP5C-ZTSE]; *see also* Mason Marks & Claudia E. Haupt, *AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance*, 330(4) JAMA 309 (2023), https://jamanetwork.com/journals/jama/article-abstract/2807170 [https://perma.cc/QB65-PATY].

[232] Mason Marks & Claudia E. Haupt, *AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance*, 330(4) JAMA 309 (2023), https://jamanetwork.com/journals/jama/article-abstract/2807170 [https://perma.cc/QB65-PATY].

[233] EDPB, *supra* note 6, at 8.

[234] Cohen, *supra* note 40.

they can become part of the datasets used to train GenAI models and available to third parties. The "unexpected revelation of previously private information … to unauthorized third parties" can be harmful, particularly when this personal data is later used for discriminatory purposes.[235]

In the US, the FTC has expressed some concern about whether the companies that develop LLMs are "engaged in unfair or deceptive privacy or data security practices," particularly when sensitive data is involved.[236] The FTC considers itself uniquely positioned to address consumer concerns about unfair or deceptive practices involving personal data collection and use in digital markets because it considers interests in both consumer protection and competition.[237] This includes protecting consumer's data privacy and ensuring that businesses do not gain an unfair competitive advantage as a result of illegal data practices.[238] For example, after a photo and video storage company used data uploaded by consumers for GenAI development without users' consent, the FTC, relying on its authority under the FTC Act, ordered the platform to obtain user consent for using biometric data from videos and photos stored on the platform and to delete any algorithms that were trained on biometric data without explicit user consent.[239] The FTC has also taken legal action against companies that obtained, shared, sold, or failed to protect consumers' sensitive data, including location data for places of worship and medical offices, health information, and messages from incarcerated individuals.[240] On the other hand, the FTC's ability to protect personal data is limited by the scope of the FTC Act, which only prohibits data practices that are

[235] FTC, *supra* note 7; *see also* FTC, Data Brokers: A Call for Transparency and Accountability 52 n.91 (2014) (discussing dangers of downstream discriminatory uses of personal data).

[236] FTC, FTC File No. 232-3044, Civil Investigative Demand ("CID") Schedule (2023), https://www.washingtonpost.com/documents/67a7081c-c770-4f05-a39e-9d02117e50e8.pdf?itid=lk_inline_manual_4 [https://perma.cc/23H7-2A7Y].

[237] FTC, FTC Report to Congress on Privacy and Security 4 (2021), https://www.ftc.gov/system/files/documents/reports/ftc-report-congress-privacy-security/report_to_congress_on_privacy_and_data_security_2021.pdf [https://perma.cc/PWR7-UT2K].

[238] *Id.* at 4.

[239] Everalbum, Inc., 170 F.T.C. 723 (2021); FTC Report to Congress on Privacy and Security, *supra* note 237, at 4.

[240] Complaint for Permanent Injunction and Other Relief, FTC v. Kochava, Inc., 671 F. Supp. 3d 1161 (D. Idaho June 5, 2023) (No. 2:22-cv-00377), 2022 WL 4080538; BetterHelp, Inc., Docket No. C-4796 (2023); Global Tel*Link Corp., Docket No. C-4801 (2023); X-Mode Social, Inc., File No. 2123038 (2024).

either deceptive or unfair from the perspective of a reasonable consumer.[241] The FTC has urged Congress to enact more general data protection laws, and some FTC Commissioners have called specifically for Congress to "take reasonable steps … to ensure that the consumer data they obtain was procured by the original source … with notice and choice, including express affirmative consent for sensitive data."[242]

Sector-specific federal laws offer limited upstream protection for certain subcategories of personal data that may end up in the GenAI data lifecycle. The Gramm-Leach-Bliley Act's disclosure and opt-out requirements can provide individuals with notice and some control over how their nonpublic personal information is collected and used by financial institutions, including the sharing of such data with third parties for GenAI development, but does not require explicit consent.[243] COPPR and FERPA likely both require consent prior to the collection and use of personal data concerning children and personal data contained in education records, respectively, in the GenAI data lifecycle.[244] However, FERPA's failure to prohibit the publication of "directory information," would allow some categories of personal information to become publicly available and subject to third party collection via data scraping.[245] Neither the Gramm-Leach-Bliley Act, nor COPPR, nor FERPA provide special protections for sensitive personal data.[246] HIPAA, on the other hand, governs only a subcategory of sensitive personal data through its regulation of PHI.[247] Outside of exceptions for limited data sets, HIPAA would prohibit the use and disclosure of PHI for the purpose of GenAI development without the individual's consent.[248] For example, if a healthcare provider inputs PHI into a GenAI model without patient authorization, this disclosure would

[241] FTC Report to Congress on Privacy and Security, *supra* note 237, at 1 (noting that in the absence of a general data privacy law, the FTC is limited by the scope of the FTC Act).

[242] FTC Report to Congress on Privacy and Security, *supra* note 237, at 1.

[243] *See* 16 C.F.R. § 313.1(a)(1)–(3) (2023) (regarding disclosure obligations and opt out requirement).

[244] *See* 20 U.S.C. § 1232g(b)(1), (d); 16 C.F.R. § 312.3(b) (2023) (regarding consent under COPPR); 34 C.F.R. § 99.5(a) (2023) (transferring rights under FERPA from parents to students when they turn 18 or enroll in postsecondary education); 34 C.F.R. § 99.30 (2023) (requiring consent from parents).

[245] *See* 20 U.S.C. § 1232g(a)(5)(A), (b)(1) (allowing disclosure of directory information).

[246] *See* 20 U.S.C. § 1232g(a)(1)(D)(4) (regarding FERPA); 16 C.F.R. § 312.2 (2023) (regarding COPPR); 16 C.F.R. § 313.3(n), (o) (2023) (regarding the Gramm-Leach-Bliley Act).

[247] *See* 45 C.F.R. § 160.103 (2023) (defining "Covered entity," "Individually identifiable health information," and "Protected health information").

[248] 45 C.F.R. § 164.502(a) (2023) (listing exceptions to prohibition of PHI disclosure).

likely violate HIPAA.[249] Notably, COPPR, FERPA, and HIPAA only govern the activities of specific actors in their respective industries, leaving the regulation of more general personal data processing to the FTC and state data privacy laws. For example, in most cases, HIPAA would not protect the privacy of PHI that individuals provide to GenAI models because the companies that own these models are not "covered entities" or "business associates" under HIPAA.[250]

The CCPA and VCDPA rely primarily on information obligations and the consumer's ability to opt out of data sharing (or disclosure to third parties) to protect personal data.[251] Both laws would require businesses to inform consumers of plans to collect and use personal information for GenAI purposes prior to data collection. Theoretically, consumers can either initially refrain from sharing personal data for GenAI purposes or subsequently limit the sharing of their personal data.[252] In addition to information obligations, the VCDPA, but not the CCPA, would prohibit the collection and use of sensitive personal data for GenAI purposes unless the consumer provided consent.[253]

The GDPR does not distinguish between publicly available personal data and private personal data. As a result, the collection of personal data from an individual data subject, like the collection of publicly available data through data scraping, must also be supported by a valid legal basis. In March 2023, the Italian DPA banned ChatGPT because "OpenAI had no legal basis to justify 'the mass collection and storage of personal data for the purpose of 'training' the algorithms underlying the operation of the platform.'"[254] In April 2023, the ban was lifted after OpenAI responded with increased transparency about how the company processed user data and offered an opt-out option for users who did not want their conversations used to train ChatGPT.[255] However, in January 2024, the Italian DPA

[249] *See* Genevieve P. Kanter & Eric A. Packel, *Health Care Privacy Risks of AI Chatbots*, 330(4) JAMA 311 (2023).

[250] Marks & Haupt, *supra* note 231.

[251] Cal. Civ. Code § 1798.100(a)(1), (2) (West 2023); Va. Code Ann. § 59.1-578 (C) (West 2023).

[252] *See* Cal. Civ. Code § 1798.120 (West 2024) (consumers right to opt out); Va. Code Ann. § 59.1-577(A)(5) (West 2023) (consumers right to opt out).

[253] *Compare* Cal. Civ. Code § 1798.100(2) (West 2024) (requiring notification), *with* Va. Code Ann. § 59.1-578 (A)(5) (West 2023) (requiring consumer consent).

[254] *ChatGPT: Italy blocks AI chatbot over privacy concerns*, *supra* note 2.

[255] *Italy lifts ban on ChatGPT after data privacy improvements*, Deutsche Welle (Apr. 29, 2023), https://www.dw.com/en/ai-italy-lifts-ban-on-chatgpt-after-data-privacy-improvements/a-65469742#:

launched new allegations that OpenAI is again violating the GDPR, though the details of this investigation have not yet been released.[256]

The GDPR does offer additional protections for special categories of data, also known as sensitive data. Processing such data in the GenAI data lifecycle is generally prohibited unless it falls under one of the Article 9 exceptions. Notably, none of these exceptions allow processing special categories of personal data for the entrance or performance of a contract or to advance legitimate interests.[257] While special categories of personal data that are "manifestly made public by the data subject" are excepted from Article 9's prohibition on processing, the EDPB's ChatGPT taskforce warns that public availability itself is not enough; instead, the data subject must have "intended, explicitly and by a clear affirmative action, to make the personal data in question accessible to the general public" for this exception to be valid.[258] As a result, processing special categories of personal data for the development and continued operation of GenAI will likely require explicit consent, requiring an "express statement of consent" from the data subject.[259]

C. Control Over Personal Data

It is challenging for individuals to control the flow of their personal data in the GenAI data lifecycle. The FTC warns that "[t]he marketplace for this information is opaque and once a company has collected it, consumers often have no idea who has it or what's being done with it."[260] Although the EU and US both seek to provide individuals with some rights to control their personal data vis-à-vis rights to request access, correction, and deletion of their personal data, exercising these rights can be difficult, or impossible, in the context of the GenAI data lifecycle.

~:text=Why%20was%20ChatGPT%20banned%20in,to%20train%20the%20site's%20algorithms [https://perma.cc/4ES4-S4RV].

[256] *ChatGPT: Garante privacy, notificando a OpenAI l'atto di contestazione per le violazioni alla normativa privacy* [*ChatGPT: Italian DPA notifies breaches of privacy law to OpenAI*], Garante per la Protezione dei Dati Personali (Jan. 29, 2024), https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9978020#english [https://perma.cc/ENP2-7VBX] (Italy).

[257] *Compare* General Data Protection Regulation 2016 O.J. (L 119) Art. 6.1(b), (f)(allowing contract and legitimate interests to serve and legal bases for processing generally) *with* art. 9.2 (not listing contract or legitimate interests as exceptions to the prohibition on processing sensitive personal data).

[258] EDPB, *supra* note 6, at 7.

[259] Guidelines on Processing of Personal Data, *supra* note 140.

[260] Cohen, *supra* note 40.

Once personal data enters the GenAI data lifecycle, individuals may not be able to effectively exercise their rights to correct or delete personal data under either US or EU data privacy laws. This is especially problematic because GenAI can produce incorrect information about a real individual. For example, ChatGPT generated a false but detailed accusation of sexual harassment by a real law professor, citing a Washington Post article that it hallucinated.[261] It also falsely reported that an Australian mayor served a prison sentence for bribery.[262] In these cases, the use of personal data for GenAI training led to defamatory statements about real identifiable individuals. In the US, the FTC filed an information request to OpenAI, which indicates concern that OpenAI "violated consumer protection laws, potentially putting personal data and reputations at risk" in connection with ChatGPT's ability to "generate false, misleading, or disparaging statements about real individuals."[263] In Europe, the CEDPO also recognizes the danger of inaccurate GenAI outputs, noting that "[g]enerative AI systems must provide reliable and trustworthy outputs, especially about European citizens whose personal data and its accuracy is protected under the GDPR."[264] Recently, nyob, a non-profit organization that seeks to privately enforce GDPR violations, filed a complaint alleging that ChatCPT's continuous inaccurate output concerning an individual's date of birth violates the accuracy principle in Article 5(1)(d) of the GDPR.[265] The complaint further alleges that OpenAI's inability to prevent ChatGPT from hallucinating inaccurate personal data or to erase or rectify the inaccurate data also constitutes a violation of Article 5(1)(d).[266] Nyob's complaint against OpenAI requests, among other things, that the Austrian DPA investigate

[261] Pranshu Verma & Will Oremus, *ChatGPT invented a sexual harassment scandal and named a real law prof as the accused*, Wash. Post (Apr. 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/ [https://perma.cc/32ML-U2K2].

[262] Byron Kaye, *Australian mayor readies world's first defamation lawsuit over ChatGPT content*, Reuters (Apr. 5, 2023), https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/ [https://perma.cc/AK86-29YJ].

[263] Benji Edwards, *Chasing defamatory hallucinations, FTC opens investigation into OpenAI*, Ars Technica (Jul. 13, 2023), https://arstechnica.com/information-technology/2023/07/chasing-defamatory-hallucinations-ftc-opens-investigation-into-openai/ [https://perma.cc/8AQD-SF2N].

[264] Confederation of Eur. Data Prot. Orgs. AI Working Grp., *supra* note 7, at 18.

[265] Complaint at 4, Österreichische Datenschutzbehörde [DSB] [Austrian Data Protection Authority] Apr. 29, 2024, Case No. C-078, https://noyb.eu/sites/default/files/2024-04/OpenAI%20Complaint_EN_redacted.pdf [https://perma.cc/WSV7-FPT9].

[266] *Id.* at ¶¶ 26–31.

and fine OpenAI for these alleged data privacy violations.[267] The EDPB ChatGPT taskforce notes that although OpenAI might warn users that ChatGPT can produce inaccurate data to satisfy the GDPR's transparency principle, this does not relieve OpenAI of its obligation to comply with the GDPR's data accuracy principle.[268]

Despite privacy rights designed to protect the use and accuracy of personal data in both the US and EU, there are several reasons why individuals' requests to delete or correct their personal data might be futile. First, data scraped from the internet will be retained in the resulting dataset even if it is later removed from an online public forum. As a result, "even if individuals decide to delete their information from a social media account, data scrapers will likely continue using and sharing information they have already scraped, limiting individuals' control over their online presence and reputation."[269] Second, once personal data is used to develop a GenAI model, it can be difficult to extract specific data from the model to remove it after-the-fact.[270] Third, individuals may not be able to provide sufficient proof that their personal data is being used in the GenAI data lifecycle. For example, some users who submitted data deletion requests report that Meta provides a boilerplate response claiming that "it is 'unable to process the request' until the requester submits evidence that their personal information appears in responses from Meta's generative AI."[271] Similarly, OpenAI and Midjourney failed to respond to an individual's request to have her image deleted.[272]

Conclusion

The flow of personal data through the GenAI data lifecycle introduces new challenges for data privacy. This Article provides interdisciplinary insight into the role and legal implications of personal data in the modern GenAI data lifecycle

---

[267] *Id.* at ¶¶ 32–35.

[268] EDPB, *supra* note 6, at 8–9.

[269] *Joint statement on data scraping and the protection of privacy*, ICO (Aug. 24, 2023), https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf [https://perma.cc/68NG-6HU9].

[270] Melissa Heikkilä, *OpenAI's hunger for data is coming back to bite it*, MIT Tech. Rev. (Apr. 19, 2023), https://www.technologyreview.com/2023/04/19/1071789/openais-hunger-for-data-is-coming-back-to-bite-it/ [https://perma.cc/TA8W-GFBX].

[271] Kate Knibbs, *Artists Allege Meta's AI Data Deletion Request Process Is a 'Fake PR Stunt'*, Wired (Oct. 26, 2023), https://www.wired.com/story/meta-artificial-intelligence-data-deletion/ [https://perma.cc/9K5K-BP6V].

[272] Harlan & Brunner, *supra* note 203.

and examines the resilience of data privacy frameworks in the US and EU in light of these implications. In Part I, we described the architecture behind modern GenAI models. Part II identified the relevant data privacy frameworks in the US and EU. We dissected and compared the US's fragmented approach and the EU's comprehensive approach to data privacy to reveal that while both jurisdictions offer some data privacy protections through a combination of information obligations and use restrictions, as a default, the US allows personal data processing unless specifically prohibited, while the EU prohibits personal data processing unless specifically allowed. In Part III, we described how personal data is stored inside GenAI models—commonly without express consent—and present at every stage of the data lifecycle in GenAI. We explained how personal data, including publicly available personal data as well as private and sensitive personal data, come from a variety of sources and how they are used to train, operate, and improve GenAI models.

Part IV identified several implications that the flow of personal data through the GenAI data lifecycle has on data privacy and explored whether the US and EU's data privacy frameworks are equipped to deal with these new data privacy concerns. We first explained how the widespread disclosure of publicly available personal data might violate individuals' expectations of privacy. We concluded that in the US, the data privacy framework offers little protection to publicly available data collected through data scraping and used to develop GenAI models. On the other hand, the EU's GDPR does not distinguish between private and publicly available data and, as such, offers more protection. Next, we discussed how GenAI models could collect and disclose private and sensitive personal data. Both the US and EU protect private and sensitive personal data in the GenAI data lifecycle by regulating disclosure of such data; however, the US' piecemeal approach to data privacy regulation still leaves gaps, particularly in cases where individuals did not authorize and are not aware that their personal data are being processed in the GenAI data lifecycle. Finally, although both jurisdictions seek to provide individuals with rights to control the accuracy of and access to their personal data, we highlight how GenAI's ability to produce inaccurate data about individuals or retain personal data indefinitely may not be remedied by individuals' rights to request, correct, and delete their information under US and EU data privacy laws.

As one of the latest applications of data-hungry technology, GenAI introduces new concerns about data privacy. These concerns are already on the radar of

regulators in the US and EU and can already be managed to some extent by the data privacy frameworks in place; however, both jurisdictions should pay special attention to the unprecedented and sweeping collection and use of personal data from public sources that underpin GenAI models, particularly because many individuals may not even be aware of how their personal data is used in the GenAI data lifecycle nor have ever explicitly agreed, either individually or collectively, to such processing in the first place.

Acknowledgements

Mindy Duffourc and Sara Gerke's work was funded by the European Union (Grant Agreement no. 101057321). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Appendix

Appendix A. Summary Comparison of Data Types in General Data Protection Laws in the US and EU

	FTC Act (US) (based on FTC policy documents)	**CCPA (California)**	**VDPA (Virginia)**	**GDPR (EU)**
Personal	• Reasonably linked to an individual consumer or device[273]	• Reasonably capable of being associated with a particular consumer or household[274]	• Reasonably associated to an individual[275]	• Relates to an identified or identifiable natural person[276]
Sensitive	• Genetic • Biometric • Precise location • Concerning health • Voice recordings & videos[277]	• Genetic • Biometric • Geolocation • Concerning race, religion, ethnicity, philosophical beliefs, union membership, health, sexual orientation, and sex life • Government identification numbers • Financial accounts • Email and text messages[278]	• Genetic • Biometric • Geolocation • Revealing individual's race, ethnicity, religion, medical diagnoses, sexual orientation, citizenship, immigration status • Personal data of a child, geolocation data[279]	• Genetic • Biometric • Concerning health, sex life, sexual orientation • Revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, life or sexual orientation[280]

[273] FTC, *supra* note 7, at 22.
[274] Cal. Civ. Code § 1798.140(v) (West 2024).
[275] Va. Code Ann. § 59.1-575 (West 2023).
[276] General Data Protection Regulation, 2016 O.J. (L 119) art. 4.1.
[277] FTC, *supra* note 7, at 58.
[278] Cal. Civ. Code § 1798.140(ae) (West 2024).
[279] Va. Code Ann. § 59.1-575 (West 2023).
[280] General Data Protection Regulation, 2016 O.J. (L 119) art. 9.1.

(Continued from previous page)				
De-identified/ Pseudonymized	• Cannot be reasonably linked to a consumer[281]	• Cannot be reasonably linked to a consumer[282]	• Cannot be reasonably linked to a consumer[283]	• No longer be attributed to a specific data subject without the use of additional information + measures to prevent reidentification[284]
Anonymous	• N/A	• N/A	• N/A	• Does not relate to an identified or identifiable natural person[285]
Aggregate	• N/A	• Group data in which consumer identities have been removed and cannot reasonably be linked to consumer[286]	• N/A	• Potential method for anonymization[287]
Publicly Available	• N/A	• Made available to general public with no audience restrictions[288]	• Lawfully made available or reasonable basis to believe is lawfully made available to the general public with no audience restrictions[289]	• N/A

[281] Fondrie-Teitler & Jayanti, *supra* note 5, at iv, 21.
[282] Cal. Civ. Code § 1798.140(m) (West 2024).
[283] Va. Code Ann. § 59.1-575 (West 2023).
[284] General Data Protection Regulation, 2016 O.J. (L 119) art. 4.5.
[285] General Data Protection Regulation, 2016 O.J. (L 119) recital 26.
[286] Cal. Civ. Code § 1798.140(b) (West 2024).
[287] Eur. Data Prot. Supervisor, *supra* note 136, at 10. (noting that GDPR does not govern fully "aggregated and anonymised datasets" when the "original input data … [is] destroyed, and only the final, aggregated statistical data is kept").
[288] Cal. Civ. Code § 1798.140(v)(2) (West 2024).
[289] Va. Code Ann. § 59.1-575 (West 2023).

Appendix B. Summary Comparison of Data Types in Sector-Specific Laws

	The Gramm-Leach-Bliley Act	COPPA	FERPA	HIPAA
Main Type of Personal Data Governed	• Obtained in connection with the provision of financial services and products that can be used to identify an individual consumer[290]	• Individually identifiable information collected from a child online[291]	• Identifies or could be used to identify a student with reasonable certainty[292]	• Health information that identifies or can be reasonably used to identify an individual[293]
De-Identified/ Pseudonymized	• Does not identify a consumer[294]	• Stripped of identifiers and trackers[295]	• Removal of all personally identifiable information + reasonable efforts to protect against re-identification[296]	• Does not identify an individual + no reasonable basis to believe that the information can be used to identify an individual[297]
Anonymous	• N/A	• N/A	• N/A	• N/A

[290] 16 C.F.R. § 313.3 (o)(1) (2023).

[291] Children's Online Privacy Protection Act §§ 1301–1308, 15 U.S.C. §§ 6501–6505 (protecting the personal information of children under the age of 13); 16 C.F.R., § 312.2 (2023) (defining "personal information" as "individually identifiable information about an individual collected online).

[292] 34 C.F.R. § 99.30 (2023).

[293] 45 C.F.R. § 160.103 (2023).

[294] 16 C.F.R. § 313.3(o)(2)(ii)(b) (2023).

[295] *See* 15 U.S.C. § 6501(8).

[296] 34 C.F.R. § 99.31(b)(1) (2023) ("An educational agency or institution … may release records or information without the consent required by § 99.30 after the removal of all personally identifiable information provided that the educational agency or institution … has made a reasonable determination that a student's identity is not personally identifiable, whether through single or multiple releases, and taking into account other reasonably available information,").

[297] 45 C.F.R. § 164.514(b) (2023).

(Continued from previous page)				
Aggregate	• Example of information not governed[298]	• N/A	• Potential method for de-identification[299]	• Potential method for de-identification[300]
Publicly Available	• Reasonable basis to believe is lawfully made available to the general public[301]	• N/A	• N/A	• N/A

[298] 16 C.F.R. § 313.3(o)(2)(ii)(b) (listing "aggregate information" as an example of information that will not be governed as "personally identifiable financial information").

[299] Priv. Tech. Assistance Ctr., *supra* note 75.

[300] *See Guidance Regarding Methods for De-identification of Protected Health Information*, *supra* note 88 (noting that de-identification can be accomplished using data aggregation, usually in combination with other de-identification techniques).

[301] *See* Financial Privacy Rule, 16 C.F.R. § 313.3(p).