D. Reed Freeman Jr. ARENTFOX SCHIFF LLP

Generative Artificial Intelligence, Data Minimization, and Today's Gold Rush

This article discusses the principle of data minimization in the context of commercial applications of generative artificial intelligence (GenAI) technology and tools.

IN THE UNITED STATES, THE PRINCIPLE OF DATA minimization is embedded firmly within the Federal Trade Commission (FTC) Act, through FTC enforcement activities, and in the host of state-level privacy laws and rules that have proliferated in recent years.

The explosive emergence in recent months of commercial applications of GenAI technology and tools, their requirements to train on very large data sets, and the need to continue to develop user-generated data supplied in GenAI prompts (prompt data) present challenges in applying this principle.

Now is the time to take stock of your data-minimization strategies to ensure that your technology and tools based on GenAI are resilient, can withstand regulatory scrutiny, and can position your organization to compete effectively in a market estimated to experience a compound annual growth rate of over 35% through 2030—more than 10 times higher than the rate of the U.S. economy.[1]

Data Minimization Laws

In general, the data-minimization principle holds that controllers should only collect and process the personal information they need to accomplish a disclosed purpose or a contextually compatible purpose, should only transfer such data consistent with those purposes, and should only maintain personal information as long as is necessary for those purposes.

The FTC's enforcement posture has changed dramatically over the past 11 years. As far back as 2012, the FTC advocated reasonable collection limitation.[2] Now, according to the FTC, using an interface to steer consumers to an option to provide more information than the context makes necessary may be considered a dark pattern, in violation of Section 5.[3]

Focusing more narrowly on AI and machine learning in a recent case, all three sitting commissioners stated that "machine learning is no excuse to break the law. Claims from businesses that data must be indefinitely retained to improve algorithms do not override legal bans on indefinite retention of data. The data you use to improve your algorithms must be lawfully collected and lawfully retained." In a clear warning shot far beyond the contours of the case at hand, the FTC continued, "companies would do well to heed this lesson."[4]

The FTC's Commercial Surveillance Advanced Notice of Proposed Rulemaking makes clear that the FTC is considering codifying data minimization into federal law.[5] In the meantime, the FTC has already brought a number of enforcement actions focused on data minimization. These cases allege that companies violated laws enforced by the FTC when they:

- Collected more personal information than they disclose or need for the purposes for which it was collected[6]
- Used[7] or shared[8] personal information for incompatible purposes
- Retained the information in violation of their own representations, or beyond the period for which the data is required for the purposes for which it was collected[9]

U.S. Laws

The California Privacy Protection Act, as amended by the California Privacy Rights Act, was the first comprehensive privacy law in the United States to reduce the data-minimization principle to codified law. Collection of personal information must be proportionate to the purpose for which it was collected or reasonably necessary for another purpose, provided that purpose is compatible with the context of collection.[10] New laws taking effect this year in Colorado,[11] Connecticut,[12] Virginia,[13] and laws passed this legislative cycle that take effect in 2024 and beyond in Indiana,[14] Iowa,[15] Tennessee,[16] Montana,[17] and Texas[18] all share common principles. In short, it is now black-letter law in the United States that personal information can only be collected for disclosed and contextually relevant purposes.

Contracts

One risk associated with licensing GenAI technology is that it may have been trained on data sets including personal information or sensitive personal information—or both. Companies can limit their risk in this regard by focusing their attention on the representations, warranties, limitations of liability, and indemnity provisions. In the GenAI context, these terms are not yet standard. The market is still

[1] *Compare* Grand View Research, *Generative AI Market Size To Reach $109.37 Billion By 2030* (Sept. 2023) *with* Congressional Budget Office, *The Economic Outlook for 2023 to 2033 in 16 Charts* (Feb. 21, 2023).

[2] *See* Fed. Trade Comm., *Protecting Consumer Privacy in an Era of Rapid Change* (March 2012).

[3] Fed. Trade Comm., *Bringing Dark Patterns to Light* (Sept. 2022).

[4] Statement of Commissioner Alvaro M. Bedoya Joined by Chair Lina M. Khan and Commissioner Rebecca Kelly Slaughter in United States v. Amazon.com, Inc. (May 31, 2023).

[5] Fed. Reg. 51,273 (Aug. 22, 2022).

[6] United States v. Edmodo, LLC, 3:23cv2495 (ND Cal. May 22, 2023).

[7] In the Matter of Support King, LLC, C-4756 (Fed. Trade Comm. Dec 20, 2021).

[8] In the Matter of Goldenshores Technologies, LLC, and Erik M. Geidl, 132 3087, Fed. Trade Comm. (April 9, 2014). *See also* United States v. Easy Healthcare Corp., 1:23-cv-3107 (ND Ill May 17, 2023), In the Matter of Flo Health, Inc., C-474 (Fed. Trade Comm. June 17, 2021).

[9] In the Matter of Everalbum, Inc., 192 3172, Fed Trade Comm. (May 5, 2022).

[10] Cal. Civ. Code § 1798.100(c). ("A business' collection, use, retention, and sharing of a consumer's personal information shall be reasonably necessary and proportionate to achieve the purposes for which the personal information was collected or processed, or for another disclosed purpose that is compatible with the context in which the personal information was collected, and not further processed in a manner that is incompatible with those purposes.")

[11] Colo. Rev. Statutes § 6-1-1304(4)(a)-(b).

[12] Connecticut Act Concerning Personal Data Privacy and Online Monitoring § 10(f), 2022 Ct. SB 6.

[13] Virginia Code Ann. §59.1-578.

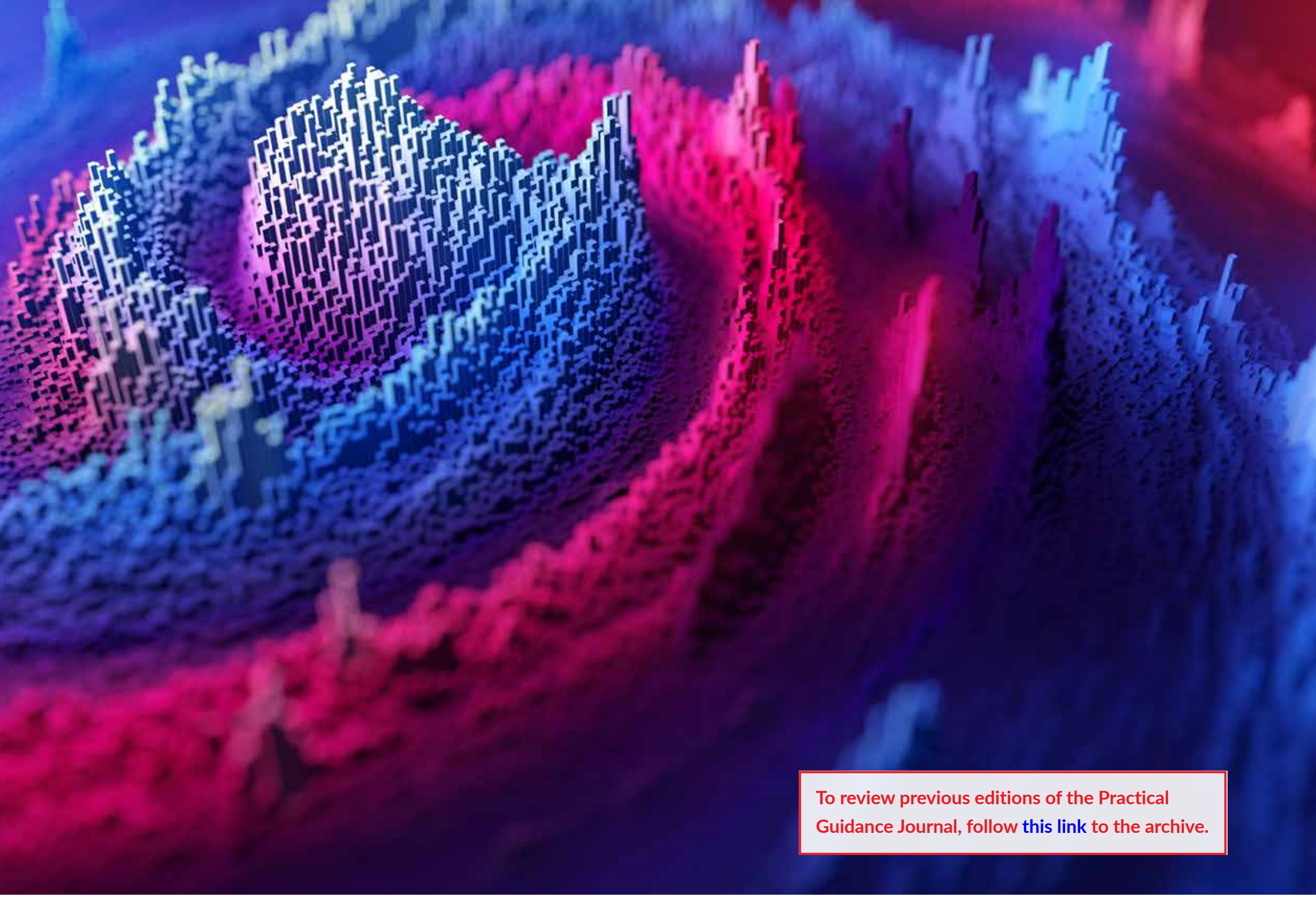
[14] Indiana Consumer Data Protection Act, Ch. 4, § 1; Ind. Code Ann. § 24-15-4-1 (Effective Jan. 1, 2026).

[15] Iowa SF 262 § 7(6), Iowa Code Ch. 715D (Effective Jan 1, 2025).

[16] Tenn. Code Ann. § 47-18-3304 (Effective July 1, 2025).

[17] Montana Consumer Data Privacy Act, § 7, 2023 Bill Text MT S.B. 384.

[18] Tx. Bus. and Prof. Code 11-541-101 (Effective July 1, 2024).



To review previous editions of the Practical Guidance Journal, follow this link to the archive.

developing. But savvy organizations are familiar with risk shifting. Do not let the rush-to-market period we're in now expose your organization to undue risk. Regulators have shown a willingness to seek algorithmic disgorgement—the death penalty that could cripple your GenAI rollout—for algorithms based on data improperly collected.[19] Do your best to make sure that you are building your tool on a solid foundation and that you are protected against downside risk.

What about prompt data? Consider whether this data will go to the GenAI technology developer itself, and for what purposes. Will it be used to continue the development of the tool just for your organization, or for others as well? If the toolmaker will use the data just for you, can the toolmaker be your service provider or processor just for this purpose? Appropriate data-processor or service-provider agreements under the new state laws may get your organization some control over the further use and disclosure of user prompt data, and such agreements may limit your risk to that extent. Your processor/service agreement should define the uses to which the GenAI technology developer will make of prompt data and should be parallel with the purposes you disclose at the point of collection and in your privacy policy. You should also make sure that the toolmaker is equipped to assist you in responding to consumer rights requests.

Your Disclosures: Proximate to the Prompt and Privacy Policy

Because privacy laws place an emphasis on disclosed and contextually relevant purposes, it is critical to have clear and conspicuous disclosures proximate to the prompt field. These disclosures should make clear that data submitted as a GenAI prompt will be used by your organization and (if applicable) the AI technology developer to generate content and to train the tool (and, if applicable, the underlying GenAI technology) on an ongoing basis. The company's privacy policy should also contain the same disclosures.

These disclosures should also explain that the user may prevent this use by not entering any personal information into the prompt field. If possible, end users should have an opportunity to opt out of the processing of prompt data for further development of the GenAI tool and the underlying technology. But before you offer that, be sure you can honor it.

[19] United States v. Kurbo, Inc., No. 22-CV-946 (N.D. Cal. March 3, 2022).

De-identifying Prompt Data

Because GenAI's fuel is data, and because of the expansive definitions of personal information and personal data in the state privacy laws, it may not be feasible over time to sort through all of your organization's prompt data to delete all personal information before the data is used for GenAI product development. But what about de-identification? California's Consumer Privacy Act (CCPA) excludes de-identified data,[20] it and contains a typical standard that organizations must meet to enjoy this protection, borrowed from FTC enforcement and policy work.

Section 1798.140(m) of the CCPA states:

> "Deidentified" means information that cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer provided that the business that possesses the information:
>
> 1. Takes reasonable measures to ensure that the information cannot be associated with a consumer or household.
>
> 2. Publicly commits to maintain and use the information in deidentified form and not to attempt to reidentify the information, except that the business may attempt to reidentify the information solely for the purpose of determining whether its deidentification processes satisfy the requirements of this subdivision.
>
> 3. Contractually obligates any recipients of the information to comply with all provisions of this subdivision.[21]

Well-known work by the National Institute of Standards and Technology[22] and the U.S. Dept. of Health & Human Services[23] serve as tactical guideposts. The point is to do what you can to maintain the volume of data needed to develop GenAI tools while avoiding data minimization risks associated with prompt data.

Conclusion

Privacy law has long wrestled with the urge to collect and keep data for future use. What's new is that with GenAI, what was once a question of "I may want to use the data in the future" has now become "I will need to use the data in the future." Data-minimization standards do not act as a ban on the use of training data and prompt data for the development of commercial GenAI technology and tools.

In fact, done with care, you can use data-minimization standards as both a shield to avoid regulatory scrutiny and as a sword to distinguish your GenAI tools from others in an almost limitless market. L

Related Content

For an overview of the legal issues related to the acquisition, development, and exploitation of artificial intelligence (AI), see

ARTIFICIAL INTELLIGENCE KEY LEGAL ISSUES

For insight into a judge's view of the use of generative AI (GenAI), see

ARTIFICIAL INTELLIGENCE: A JUDGE'S VIEW OF GENERATIVE AI

For a sample certificate regarding the use of GenAI in federal court, see

GENERATIVE ARTIFICIAL INTELLIGENCE (AI) USE AND COMPLIANCE CERTIFICATION (FEDERAL)

For a comprehensive guide to current practical guidance on GenAI, ChatGPT, and similar tools, see

GENERATIVE ARTIFICIAL INTELLIGENCE (AI) RESOURCE KIT

For a list of key issues for performing a software and IT due diligence investigation of a seller, see

AI AND LEGAL ETHICS: WHAT LAWYERS NEED TO KNOW SOFTWARE AND INFORMATION TECHNOLOGY DUE DILIGENCE CHECKLIST

D. Reed Freeman Jr., a partner in the Washington, D.C., office of ArentFox Schiff LLP, may be contacted at reed.freeman@afslaw.com. He has represented clients in scores of FTC investigations involving privacy, data security, and advertising matters. He also defends companies in state consumer protection investigations and data breach responses. He regularly advises clients on compliance with international and domestic privacy laws and advises on compliance with advertising laws and rules.

This article was first published in Pratt's Privacy & Cybersecurity Law Report.

RESEARCH PATH: *Data Security & Privacy > Industry Compliance > Articles*

[20] Cal. Civ. Code § 1798.140(v)(3). [21] Cal. Civ. Code § 1798.140(m). [22] *See* Simson L. Garfinkel, *De-Identification of Personal Information*, NISTIR 8053 (Oct. 2015). [23] U.S. Dept. of Health & Human Services, *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* (June 8, 2020).