Published on *The National Law Review* https://natlawreview.com

Generative Artificial Intelligence, Data Minimization, and the Gold Rush of the Early 2020s

Article By:

D. Reed Freeman Jr.

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 generative artificial intelligence (GenAI) technology and tools, and their need to train on very large data sets, and to continue to develop on user-generated data supplied in GenAI prompts (prompt data) presents some 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 US 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.

FTC Act

The FTC's enforcement posture has changed dramatically over the past 11 years. As far back as 2012, the FTC has 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; or
- 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]

US State 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.

GenAI

Two aspects of GenAI requires attention when considering data minimization. First, GenAI technology requires an extraordinary amount of training data to be useful. Much of this data is scraped from websites, and much of it contains personal or sensitive personal information. Second, both the underlying GenAI technology and commercial tools using it continue to train on prompt data, which may also contain personal or sensitive personal information.

Companies around the world are now scrambling to license commercial GenAI technology to introduce all manner of tools to their customers. By heeding these steps, organizations can meet their data minimization requirements for compliance and risk-reduction purposes and, having done so, will be poised to capture their part of the expansive new GenAI market.

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 developing. But savvy organizations are familiar with risk shifting. Don't 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're building your tool on a solid foundation and that you're 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 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. They 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.

De-Identifying Prompt Data

Because GenAI's fuel is data, and because of the expansive definition 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 excludes de-identified data[20], and contains a typical standard that organizations must meet to enjoy this protection, borrowed from Federal Trade Commission enforcement and policy work.

Section 1798.140(m) of the CCPA defines "deidentified" as:

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 NIST[22] and HHS[23] serves 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.

FOOTNOTES

[1] Compare Generative AI Market Size To Reach $109.37 Billion By 2030, Grand View Research (May 2023), available at https://www.grandviewresearch.com/press-release/global-generative-ai-market with The Economic Outlook for 2023 to 2033 in 16 Charts, Congressional Budget Office (February 2023), available at https://www.cbo.gov/publication/58880.

[2] See FTC Report, Protecting Consumer Privacy in an Era of Rapid Change, https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

[3] FTC Staff Report, Bringing Dark Patterns to Light, available at https://www.ftc.gov/system/files/ftc_gov/pdf/P214800%20Dark%20Patterns%20Report%209.14.2022%20-%20FINAL.pdf.

[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), available at https://www.ftc.gov/system/files/ftc_gov/pdf/Bedoya-Statement-on-Alexa-Joined-by-LK-and-RKS-Final-1233pm.pdf.

[5] https://www.govinfo.gov/content/pkg/FR-2022-08-22/pdf/2022-17752.pdf at p. 51284 (Q.43)

[6] United States v. Edmodo, LLC, available at https://www.ftc.gov/system/files/ftc_gov/pdf/edmodocomplaintfiled.pdf.

[7] In the Matter of Support King, LLC, available at https://www.ftc.gov/system/files/documents/cases/1923003c4756spyfonecomplaint_0.pdf.

[8] In the Matter of Goldenshores Technologies, LLC, and Erik M. Geidl, Complaint, available at https://www.ftc.gov/system/files/documents/cases/140409goldenshorescmpt.pdf, see also, e.g., United States v. Easy Healthcare Corp., available at https://www.ftc.gov/system/files/ftc_gov/pdf/2023186easyhealthcarecomplaint.pdf, In the Matter of Flo Health, Inc., available

at https://www.ftc.gov/system/files/documents/cases/192_3133_flo_health_complaint.pdf.

[9] In the Matter of Everalbum, Inc., available at https://www.ftc.gov/system/files/documents/cases/1923172_-_everalbum_complaint_final.pdf.

[10] Cal. Civ. Code § 1798.100 ("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] Co. Rev. Statutes § 6-1-1304(4)(a)-(b), available at https://leg.colorado.gov/sites/default/files/2021a_190_signed.pdf.

[12] Connecticut Act Concerning Personal Data Privacy and Online Monitoring § 10(f), available at https://www.cga.ct.gov/2022/act/pa/pdf/2022PA-00015-R00SB-00006-PA.pdf.

[13] Virginia Code Ann. §59.1-578, available at https://law.lis.virginia.gov/vacode/title59.1/chapter53/section59.1-578/.

[14] Indiana Consumer Data Protection Act, Ch. 4, § 1 available at https://iga.in.gov/legislative/2023/bills/senate/5#document-8806200c.

[15] Iowa SF 262 § 7(6), available at https://www.legis.iowa.gov/legislation/BillBook?ga=90&ba=SF%20262.

[16] Tenn. Code Ann. § 47-18-3204, available at https://www.capitol.tn.gov/Bills/113/Bill/SB0073.pdf.

[17] Montana Consumer Data Privacy Act, § 7, available at https://leg.mt.gov/bills/2023/billpdf/SB0384.pdf.

[18] Tx. Bus. and Prof. Code 11-541-101, available at https://capitol.texas.gov/tlodocs/88R/billtext/pdf/HB00004F.pdf#navpanes=0.

[19] United States v. Kurbo, Stipulated Order (March 2022), available at https://www.ftc.gov/system/files/ftc_gov/pdf/wwkurbostipulatedorder.pdf.

[20] Cal. Civ. Code § 1798.140(v)(3).

[21] Cal. Civ. Code § 1798.140(m).

[22] See NISTIR 8053, available at https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

[23] HHS, Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, available at https://www.hhs.gov/guidance/document/guidance-regarding-methods-de-identification-protected-health-information-accordance-0.

National Law Review, Volume XIII, Number 160