epic.org

Electronic Privacy Information Center
1519 New Hampshire Avenue NW
Washington, DC 20036, USA

+1 202 483 1140
+1 202 483 1248
@EPICPrivacy
https://epic.org

COMMENTS OF THE ELECTRONIC PRIVACY INFORMATION CENTER

to the

White House Office of Science and Technology Policy

RFI on Advancing Privacy-Enhancing Technologies

87 Fed Reg. 35,250

July 8, 2022

The Electronic Privacy Information Center (EPIC) submits these comments in response to White House Office of Science and Technology Policy's (OSTP) June 9, 2022 Request for Information on Advancing Privacy-Enhancing Technologies.[1] OSTP is interested in privacy preserving technologies for data sharing and analytics including "secure multiparty computation, homomorphic encryption, zero-knowledge proofs, federated learning, secure enclaves, differential privacy, and synthetic data generation tools."[2] Through this RFI, OSTP is seeking guidance on adopting "a national strategy on privacy-preserving data sharing and analysis," including "Federal laws, regulations, authorities, research priorities, and other mechanisms across the Federal Government that could be used, modified, or introduced to accelerate the development and adoption of PETs."[3]

EPIC is a public interest research center in Washington, D.C. EPIC was established in 1994 to focus public attention on emerging privacy and related human rights issues and to protect privacy,

[1] 87 Fed. Reg. 35,250, *available at* https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies.
[2] *Id*. at 35,251.
[3] *Id*.

the First Amendment, and constitutional values. EPIC has a longstanding interest in federal efforts to develop privacy-enhancing technologies and regularly comments on proposed federal planning efforts at the intersection of technology and privacy.[4] EPIC has also repeatedly intervened to ensure proper privacy protections, including differential privacy, are used on the Census.[5]

EPIC urges OSTP to (1) prioritize the adoption of differential privacy in its national research plan and (2) direct federal agencies to shift and increase funding toward the development of privacy-enhancing technologies.

I. The federal government should invest more in differential privacy instead of relying on deidentification techniques that do not work.

Traditional techniques for deidentifying and anonymizing datasets are ineffective and do not account for the ease with which information in multiple datasets can be combined to reidentify individuals. Differential privacy is the intentional injection of controlled amounts of statistical noise into data products to provide a mathematical guarantee of privacy while preserving the ability to use

[4] *See, e.g.*, Comments of EPIC, Public and Private Sector Uses of Biometric Technologies, Office of Sci. & Tech. Policy (Jan. 15, 2022), https://epic.org/documents/epic-comments-to-ostp-on-public-and-private-sector-uses-of-biometric-technologies/; Comments of EPIC, Artificial Intelligence Risk Management Framework, Nat'l Instit. of Standards & Tech. (Aug. 18, 2021), https://epic.org/documents/regarding-the-artificial-intelligence-risk-management-framework/; Comments of EPIC, Request for Information (RFI) on an Implementation Plan for a National Artificial Intelligence Research Resource, Office of Sci. & Tech. Policy & Nat'l Sci. Found. (Oct. 1, 2021), https://epic.org/wp-content/uploads/2021/10/EPIC-Comment-NAIRR-Oct2021.pdf; Comments of EPIC, Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, Comptroller of the Currency et al., (July 1, 2021), https://archive.epic.org/apa/comments/EPIC-Financial-Agencies-AI-July2021.pdf; Comments of EPIC, Solicitation of Written Comments by the National Security Commission on Artificial Intelligence, 85 Fed. Reg. 32,055 (Sep. 30, 2020), https://epic.org/wp-content/uploads/apa/comments/EPIC-comments-to-NSCAI-093020.pdf; Comments of EPIC, Request for Comments on a Draft Memorandum to the Heads of Executive Departments and Agencies, "Guidance for Regulation of Artificial Intelligence Applications," (Mar. 13, 2020), https://epic.org/apa/comments/EPIC-OMB-AI-MAR2020.pdf.

[5] *See* Br. EPIC as Amicus Curiae, *Alabama v. Dep't of Commerce*, 546 F. Supp. 3d 1057 (M.D. Ala. 2021), https://epic.org/wp-content/uploads/amicus/census/2020/Alabama-v-Commerce-21-cv-211-EPIC-Amicus-Brief.pdf; EPIC, *EPIC v. Commerce (Census Privacy)* (2019), https://epic.org/documents/epic-v-commerce-census-privacy/; Br. EPIC as Amicus Curie, *Dep't of Commerce v. New York*, 139 S. Ct. 2551, (2019), https://epic.org/wp-content/uploads/amicus/census/2020/Commerce-v-NY-EPIC-Amicus.pdf.

the resulting data. Differential privacy is a more robust means of protecting individual privacy that should be prioritized in OSTP's national strategy.

a. Traditional deidentification and anonymization techniques do not work.

It has been clear since the early 2000s that basic database deidentification techniques which rely on removing personally identifiable information like names and addresses are insufficient to stop reidentification attacks achieved by combining information from multiple databases. Pioneering work by Latanya Sweeney demonstrated that "87% of the U.S. Population are uniquely identified by {date of birth, gender, ZIP}."[6] Sweeney showed that re-identification was broadly possible with diverse and supposedly anonymized datasets including survey data,[7] pharmacy data,[8] data from clinical trials,[9] public health registries,[10] and partial Social Security Numbers.[11]

Subsequent research has confirmed that supposedly deidentified datasets are at best weakly anonymized and are subject to increasingly easy reidentification.[12] For example, Netflix data stripped of names can be used to reidentify more than 80 percent of users.[13] The work of Cynthia Dwork has demonstrated how broadly applicable mathematical reidentification techniques are.[14] Newer mathematical methods for reidentification rely on large, aggregate databases, allowing the

[6] Latanya Sweeney, *Simple Demographics Often Identify People Uniquely*, Carnegie Mellon Data Privacy Lab (2000), https://dataprivacylab.org/projects/identifiability/paper1.pdf.
[7] Latanya Sweeney, *Re-identification of De-identified Survey Data*, Carnegie Mellon Data Privacy Lab (2000).
[8] Latanya Sweeney, *Patient Identifiability in Pharmaceutical Marketing Data*, Carnegie Mellon Data Privacy Lab (2011), https://dataprivacylab.org/projects/identifiability/pharma1.pdf.
[9] Latanya Sweeney, *Identifiability of De-identified Clinical Trial Data*, Carnegie Mellon Data Privacy Laboratory (2009).
[10] Latanya Sweeney, *Iterative Profiler*, Carnegie Mellon Data Privacy Lab (1997).
[11] *See* Data Privacy Lab, SOS Social Security Number Watch, Harvard University, https://dataprivacylab.org/dataprivacy/projects/ssnwatch/index.html.
[12] Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701 (2010), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006.
[13] *Id.*
[14] Dwork, C., A. Smith, T. Steinke, & J. Ullman, *Exposed! A Survey of Attacks on Private Data*, 4 Annual Review of Statistics and Its Application 61–84 (2017), https://privacytools.seas.harvard.edu/files/privacytools/files/pdf_02.pdf.

identification of sensitive features of individuals in datasets and allowing a bad actor to determine whether a particular individual is included.[15] The most up-to-date reidentification models can identify virtually all Americans (99.87 percent) in any dataset containing 15 demographic attributes.[16] In short, traditional deidentification techniques do not work and do not reliably preserve individual privacy.

b. Differential privacy allows for responsible use of datasets containing personal information while reducing the risk of reidentification.

Although traditional deidentification and anonymization techniques are fundamentally ineffective, there is a viable alternative for preserving the privacy of individuals contained in datasets while still preserving the research value of the data: differential privacy. "'Differential privacy' describes a promise, made by a data holder, or curator, to a data subject: 'You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available.'"[17] Differential privacy is achieved through the controlled injection of statistical noise into a study or analysis of a dataset, providing a mathematical guarantee of privacy while preserving the research value of the information.[18] Applying differentially private algorithms to datasets allows researchers to perform one or multiple analyses on the data while minimizing the risk that such analysis will reveal the identity of any person in the dataset. Differential privacy has proven valuable in a broad range of applications, including statistical research, machine learning, emoji suggestions on Apple devices,

[15] *Id*.
[16] Luc Rocher, Julien Hendrickx, & Yves-Alexandre de Montjoye, *Estimating the success of re-identifications in incomplete datasets using generative models*, 10 Nature Commc'ns 3,069 (2019), https://www.nature.com/articles/s41467-019-10933-3.
[17] Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy* 5 (2014).
[18] Daniel L. Oberski & Frauke Kreuter, *Differential Privacy and Social Science: An Urgent Puzzle*, Harv. Data Sci. Rev. (Jan. 31, 2020).

LinkedIn's Labor Market Insights reports, Microsoft's suggested replies in Office tools, and Google's reporting of COVID-19 search trends.[19]

Perhaps the most prominent example is the U.S. Census Bureau's adoption of differential privacy for the 2020 Census disclosure avoidance system.[20] In the lead-up to the 2020 Census, the Bureau determined that existing census data products were alarmingly vulnerable to reconstruction and reidentification attacks. Specifically, the Bureau found that the sex, age, race, and ethnicity of 142 million individuals could be inferred from publicly available 2010 Census data and that 52 million census respondents could be reidentified with the added use of commercial datasets.[21] And the potential harms of these reconstruction and reidentification attacks are significant:

> Anyone could construct a linkage attack by purchasing commercial data[.] . . . Most people do not view the characteristics in the decennial census as particularly sensitive, but those who are most at risk to having their data abused (and are typically also the hardest to count) do. People who are living in housing units with more people than are permitted on the lease are nervous about listing everyone living there, unless they can be guaranteed confidentiality. Same-sex couples are nervous about marking their relationship status accurately if they feel as though they could face discrimination. Yet, the greatest risks people face often stem from how census data can be used to match more sensitive data (e.g., income, health records, etc.).[22]

In order to fulfill its Title 13 confidentiality obligations, the Bureau turned to a new disclosure avoidance system for the 2020 Census based on differential privacy—one which ensures both useful statistics and a mathematical guarantee of privacy. The Bureau's adoption of and experience with differential privacy should serve as a guide for other federal agencies engaged in the collection and analysis of large datasets containing personal information.

---

[19] Damien Desfontaines, A list of real-world uses of differential privacy, desfontaines.es (last updated Jan. 27, 2022), https://desfontain.es/privacy/real-world-differential-privacy.html.
[20] JASON, *Formal Privacy Methods for the 2020 Census*, https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/privacy-methods-2020-census
[21] Michael Hawes, U.S. Census Bureau, *Differential Privacy and the 2020 Decennial Census* 13 (Mar. 5, 2020), https://www2.census.gov/about/policies/2020-03-05-differential-privacy.pdf.
[22] danah boyd, *Balancing Data Utility and Confidentiality in the 2020 US Census* 15–16 (Apr. 27, 2020), https://datasociety.net/wp-content/uploads/2019/12/Differential-Privacy-04_27_20.pdf.

EPIC urges OSTP to look especially to the work of Cynthia Dwork,[23] one of the co-inventors of differential privacy, as a starting point for the further adoption and application of differential privacy:

- Dwork, C., and V. Feldman. "Privacy-preserving prediction." In *Conference on Learning Theory*, 1693-1702, 2018, 1693-1702.[24]
- Bun, M., C. Dwork, G. N. Rothblum, and T. Steinke. "Composable and versatile privacy via truncated cdp." Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, 2018, 74-86.[25]
- Dwork, C. "Differential privacy." International Colloquium on Automata, Languages, and Programming. ICALP, 2006.[26]

II. Agencies across the federal government should dedicate more research and development funding to privacy enhancing technologies.

Directing more funding towards privacy-enhancing technologies would create substantial benefits for individuals, industry, and the federal government. Better and more widely available PETs could help protect the privacy of individuals by reducing the harms from data breaches that are common across the federal government. And investing in PETs is worthwhile and cost-effective because PETs can enhance and preserve other research opportunities. Last year the National Security Commission on Artificial Intelligence listed funding privacy-preserving technologies as a priority in the broader rollout of AI and a necessity in democratic regimes.[27]

Privacy-enhancing technologies can provide additional protection for individuals when data breaches and hacks occur. Federal agencies maintain vast amounts of personal information in databases across the government. And data breaches are both increasingly common and increasingly severe. As an example of this trend across the federal government, a 2015 data breach at the Office of Personnel Management (OPM) exposed social security numbers and other personal data from

[23] *See Cynthia Dwork*, Harvard University (2022), https://dwork.seas.harvard.edu.
[24] https://dwork.seas.harvard.edu/publications/privacy-preserving-prediction.
[25] https://dwork.seas.harvard.edu/publications/composable-and-versatile-privacy-truncated-cdp.
[26] https://dwork.seas.harvard.edu/publications/differential-privacy.
[27] Nat'l Sec. Comm'n on Artificial Intelligence, *Chapter 15: A Favorable International Technology Order* (2021), https://reports.nscai.gov/final-report/chapter-15/.**Error! Hyperlink reference not valid.**

21.5 million individuals.[28] Around the same time, OPM reported another major data breach exposing records on about 4 million federal employees.[29] Just a year before, a breach at the U.S. Postal Service led to the loss of personal information from more than 800,000 employees.[30] On August 24, 2020, a cyber-attack compromised a federal agency and documents were stolen.[31]

The greatest risks of data breaches come from the government holding large volumes of personal information that can have lasting financial and security impacts when wrongfully divulged. For example, The Federal Emergency Management Agency (FEMA) unnecessarily disclosed sensitive information from victims of the 2017 California wildfires, exposing up to 2.3 million people.[32] FEMA shared details of victims' financial institutions and personal lives, including EFT and bank transit numbers and complete addresses.[33]

While traditionally the focus on protecting federal agency databases has settled on improving cybersecurity practices, implementation of best practices has been uneven at best. In 2018 for example, the GAO found that over 700 of its cybersecurity recommendations since 2010 had not been implemented by federal agencies.[34] Privacy-enhancing technologies can complement

[28] U.S. Gov't Accountability Office, DHS Needs to Enhance Capabilities, Improve Planning, and Support Greater Adoption of Its National Cybersecurity Protection System (Jan. 2016) at 8, https://www.gao.gov/assets/680/674829.pdf.
[29] *Id*.
[30] *Id*.
[31] Cybersecurity and Infrastructure Security Agency, Federal Agency Compromised by Malicious Cyber Actor, AR20-268A, Dep't. of Homeland Sec. (Sept. 24, 2020), https://us-cert.cisa.gov/ncas/analysis-reports/ar20-268a; Duncan Riley, *DHS discloses data breach of US agency but doesn't name which was hacked*, SiliconAngle (Sept. 24, 2020), https://siliconangle.com/2020/09/24/dhs-discloses-data-breach-us-agency-doesnt-name-hacked/.
[32] Christopher Mele, *Personal Data of 2.3 Million Disaster Victims Was Released by FEMA, Report Says*, N.Y. Times (Mar. 22, 2019), https://www.nytimes.com/2019/03/22/us/fema-data-breach.html; John V. Kelly, Management Alert – FEMA Did Not Safeguard Disaster Survivors' Sensitive Personally Identifiable Information, OIG-19-32, Dep't of Homeland Sec. Off. of Inspector Gen. (Mar. 15, 2019), https://www.oig.dhs.gov/sites/default/files/assets/2019-03/OIG-19-32-Mar19.pdf.
[33] *Id*.
[34] U.S. Gov't Accountability Office, *GAO-19-105 Information Security: Agencies Need to Improve Implementation of Federal Approach to Securing Systems and Protecting Against Intrusions* (Dec. 18, 2018), https://www.gao.gov/assets/700/696105.pdf.

cybersecurity practices by making it harder to derive PII from federal databases, disincentivizing malicious hacks, and reducing the harms caused by data breaches when they do occur.

Investing in further research of PETs is a cost-effective strategy because these technologies are rapidly becoming a prerequisite to safe and ethical research. Developing better PETs and making them widely available can spur innovation across sectors by expanding the research possible for both government-supported scientists and industry. Increasing and shifting agency funding towards PETs is an investment in innovation.

III. Conclusion

EPIC applauds OSTP's development of a national strategy for advancing PETs. OSTP should focus funding on the development and adoption of differential privacy to address the shortcomings of traditional deidentification and anonymization techniques. EPIC also urges OSTP to direct federal agencies to increase funding toward PETs to protect individual privacy and promote innovation. If OSTP has any further questions, please reach out to EPIC Senior Counsel John Davisson at davisson@epic.org.

Respectfully Submitted,

John Davisson
John Davisson
EPIC Senior Counsel

Jake Wiener
Jake Wiener
EPIC Law Fellow