Response to the RFI on Executive Branch Agency Handling of Commercially Available Information Containing Personally Identifiable Information

Rachel Cummings[1], Shlomi Hod[2], Palak Jain[2], Gabriel Kaptchuk[3], Tamalika Mukherjee[1], Priyanka Nanayakkara[4], and Jayshree Sarathy[5]

[1]Columbia University, {rac2239,tm3391}@columbia.edu
[2]Boston University, {shlomi,palakj}@bu.edu
[3]University of Maryland, College Park, kaptchuk@umd.edu
[4]Harvard University, priyankan@g.harvard.edu
[5]Northeastern University, j.sarathy@neu.edu

December 16, 2024

1 Introduction

On October 16th, 2024, the Office of Management and Budget (OMB) issued a Request for Information (RFI) titled *Executive Branch Agency Handling of Commercially Available Information Containing Personally Identifiable Information*. The RFI is part of the OMB's implementation of the Biden Administration Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." While the Executive Order the OMB is responding to concerns artificial intelligence (AI), the primary questions raised in the RFI center on how to handle Personally Identifiable Information (PII) within Commercially Available Information (CAI). Among other things, OMB is asking for input on how best to protect people's privacy and sensitive data, what risks OMB should be aware of, and what resources, frameworks, or tools should be used in the process.

In this response we focus on two main ideas. **First, when thinking about the risks associated with using CAI—and particularly the use of CAI within an "AI-enabled" world—it is important to think about PII in broader terms than currently assumed.** Simplistic models of PII (e.g., categorizing characteristics as "identifying" versus "non-identifying") are incompatible with current scientific understanding of data privacy [1–3]. Using these out-dated models puts data subjects represented in CAI at risk. **Second, there are technical and procedural paradigms for risk mitigation that governments can use in order to minimize the risks of using CAI.** These include existing practices that should be considered a baseline requirement (including access control and re-evaluation of risks due to new threats) and more advanced privacy enhancing technologies, such as differential privacy and secure multiparty computation.

We structure our response thematically around these ideas, and identify how these ideas address the questions enumerated in the RFI.

1.1 About the authors.

We are a multi-institution group of academic researchers who specialize in cybersecurity and privacy enhancing systems. Our expertise spans designing privacy enhancing technologies, systems, and workflows [4–21], deploying privacy enhancing systems [22,23], identifying vulnerabilities in deployed systems [24,25], and analyzing the human and social factors associated with privacy enhancing technologies [26–32]. Our group has deep expertise with differential privacy in particular, a formal privacy notion that has been receiving significant attention from governments [23,26–28,33–35]. We have previously responded to other governmental requests for information, focusing on the promise and applicability of privacy enhancing technologies [36–38], described policy challenges of modernizing privacy protections [39, 40], written policy guidance on the use of privacy enhancing technologies [41], and trained policymakers in responsible computing [42, 43].

Given our qualifications, we focus this response specifically on *technical* aspects of the listed questions. We emphasize that the choice to focus on technical aspects is *not* because we believe these are the only important considerations around government use of CAI. Rather, we consider protections (both technical and procedural) that can minimize the risks associated with using CAI *after the choice to purchase CAI has already been made.* However, there are numerous legal and ethical factors that must be considered within the choice to purchase and use CAI; we leave these for other experts with relevant expertise. Importantly the technical and procedural protections that we discuss below should not be understood as sufficiently addressing these legal and ethical factors alone. There may be times in which there are legal or ethical barriers to using CAI that no amount of technical protection can overcome. Thus, an interdisciplinary approach integrating diverse areas of expertise is *required.*

2 Rethinking "PII" (Q1 and Q4)

OMB Circular A-130 [44] defines PII as "information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual." In practice, however, the focus is often placed narrowly on direct identifiers like SSNs (which uniquely identify individuals) and "quasi-identifiers" like zip codes and birth dates (which aid identification by significantly narrowing down the possible individuals) [45,46]. Contrary to the assertion that "[other attributes] usually would not provide for direct or indirect identification of an individual. . . " [47], this limited focus significantly understates the identification risk posed by a dataset.

All attributes should be treated as quasi-identifying. Recent attacks show that seemingly innocuous information can facilitate re-identification of individuals:

- A landmark 2008 study published in the IEEE Symposium on Security and Privacy [3] showed that an attacker could re-identify individuals in a dataset containing "anonymous" movie ratings of 500,000 Netflix subscribers. Although the dataset contained no PII (as typically understood and, e.g., enshrined in HIPPA's safe harbor rules), attackers were still able to learn individuals' political affiliations and other personal characteristics using only approximate knowledge of when they watched a handful of movies. This attack leveraged publicly-available movie ratings from IMDb.com, and succeeded even though some of the attacker's information was incorrect.

- Even data handled by de-identification experts in accordance with strict privacy regulations still contains identifying information. A study published in USENIX Security Symposium in 2022 showed that an attacker could re-identify individuals in EdX data published by MIT

and Harvard using information available publicly on Linkedin, despite the data being "de-identified" by statistical experts in compliance with FERPA [48].

As demonstrated by both of the real-world attacks, the risk of re-identification can increase significantly with access to even small amounts of outside information. Additionally, very high dimensional data generally poses a higher re-identification risk, as individuals are more unique as more information about them is captured. This is especially concerning when multiple datasets are processed through the same system—a standard practice in AI model training. For example, an AI model might be trained with multiple, high dimensional datasets with the goal of predicting features of individuals. Whenever an algorithm or a data scientist has access to both these datasets, there is an increased risk of linkage. Beyond even the risk of linkage between datasets, the opacity of modern AI models renders rigorous auditing for such linkages infeasible.

2.1 What does this mean for CAI? (Q1a)

The previous discussion on PII and the potential to re-identify individuals from seemingly innocuous data has important implications for the government's use of CAI. In particular, agencies must incorporate these insights to move beyond the binary classification of CAI as being with or without PII or quasi-identifiers, and must instead take a nuanced view of the privacy risks from using CAI. We again recommend *treating all attributes of CAI as quasi-identifying* in risk assessment procedures.

As illustrated in both the Netflix dataset and EdX dataset examples above, allegedly de-identified datasets can be combined with other data sources to reveal identities or protected information about the individuals represented in the dataset. The use of CAI in combination with other data sources—as is likely to occur when training an AI model—carries the risk of combining multiple sources to data to inadvertently enable linkage attacks. Agencies should consider the full spectrum of potential identification risks in advance of using CAI, especially when CAI will be combined with other data sources.

Additionally, since CAI comes from external organizations, the federal agencies using these datasets may lack important visibility into the data collection process. This means that government users of CAI may not be aware of other datasets—such as those publicly available, commercially available, or even held by the CAI provider—that can be combined with the data to perform linkage attacks. There may also be errors or missing values in the data, or incorrect documentation describing the dataset that cannot be verified by government users of CAI. As illustrated with the Netflix dataset, these types of errors do not necessarily prevent re-identification attacks from being successful, and may even be additional sources of information leakage. For example, if user IDs are generated sequentially by birthdate instead of randomly, then an attacker could use this attribute as a quasi-identifier that encodes partial information about birthdate. Finally, if individuals are not aware of their data being repurposed for a secondary use as CAI, any data breaches or attacks could seriously erode public trust, both in the federal government and in the seller supplying CAI.

3 Rethinking Data: Privacy Risks of AI (Q1 and Q9)

AI models are typically created using statistical and/or machine learning techniques. These models are trained to perform tasks based on *training datasets*—collections of input-output data pairs, which often contain data about individuals. At their core, AI models encode task performance through *weights*: collections of numerical values describing patterns the model has "learned" from its training dataset. In simpler models, like logistic regression, these weights can be interpreted

as measures of each input feature's *importance* to the output (i.e., prediction). However, for more complex models, such as neural networks, there is no clear interpretation for these numerical values.

Largely speaking, the rise of powerful, modern machine learning techniques poses two privacy risk for government use of CAI: (1) AI models can be effective tools in conducting the types of linkage attack discussed in Section 2; and (2) when CAI is used to train AI models, information about the individuals in the training set can be extracted. Given that we have already discussed these attacks above, we focus on (2) below.

Disclosure via training AI on CAI. Although AI models store information in a format drastically different from their training data (which may include records, text, or images), research over the past decade has definitively shown that model weights encode detailed information about the training data itself at a granular level. In other words, they do not just represent abstract rules on how to perform the task at hand. This phenomenon, known as *memorization* [49], presents significant concerns, as information about individuals in the training data can be disclosed inadvertently. More troubling still, in some cases, malicious actors can extract training data points directly from model weights (i.e., weights access), or even through normal model usage by analyzing outputs for given inputs (i.e., query access).

This threat very clearly translates to Large Language Models (LLMs) like GPT, Llama, and Claude, which have all been shown to be vulnerable to the same type of data extraction. These models have been documented to reproduce text verbatim from their training data, including PII (as normatively understood) [50, 51], during normal operation. Similarly, researchers have demonstrated methods to extract training images from computer vision models [52, 53], including sensitive data such as biometric photos and other PII [54].

A related privacy threat, known as membership inference [55], enables determining whether specific data points were used in a model's training dataset. This capability poses serious privacy risks, particularly when dataset membership could reveal sensitive information (e.g., being part of a dataset of HIV-positive individuals). Researchers have also shown that such membership inference attacks are possible both through direct access to model weights and through careful analysis of model outputs [55, 56]. When working with CAI, this means that it may be possible to determine if a particular individual's data was included in a CAI dataset. There may be times in which revealing this piece of information is problematic (e.g., if the dataset was restricted to individuals who purchased particular products or had a particular medical condition). Even if it were possible for malicious actors to *buy* this information by purchasing the CAI dataset on their own, extracting this information from a deployed model can significantly lower the bar for access.

This body of evidence suggests that AI models should be categorized, to a certain extent, as "data" rather than treated as a distinct category of data-processed products, despite their different format. Consequently, any reduced privacy or security standards for AI models compared to "regular" data—including for complex models such as LLMs and when PII is included in the training data—should face rigorous evaluation and require thorough justification.

4 Recommendations

We have described above how OMB must rethink their definition of privacy risks of using CAI, and how AI exacerbates these privacy risks in several ways. To address and mitigate these risks, we recommend (1) more robust governance mechanisms and (2) the use of technical tools such as privacy-enhancing technologies (PETs).

4.1 Governance Mechanisms

Access control (Q2). The first step toward mitigating the risks associated with collecting and using data about individuals—and CAI in particular as discussed in Section 2.1—is access control. Access control is a data management practice which can be thought of as a data minimization practice: only the individuals who *need* to use the data to complete a specific task should have the technical means to access it. This fundamental cybersecurity practice is widespread throughout industry, often taking the form of *role-based access control*, where access to data and computational resources is restricted to individuals within the organizations who play a particular role (data scientist, database engineer, etc.). Employing access control mitigates two types of concrete risks:

1. *Data exfiltration during a limited breach:* A common way for an organization's infrastructure to be breached is through *compromised credentials*. For example, an adversary could learn an employee's authentication credentials (e.g., learning the employee's password through phishing attacks), gaining access to the employee's accounts. Access control mechanisms can limit the impacts of such a breach, as the set of parties who have access in the first place is minimized. Thus, the compromised credentials may be insufficient to access critical systems or sensitive datasets.

2. *Unintentional misuse of data:* Limiting access to datasets means that employees must ask for permission *before* using a dataset. This practice prevents well-meaning employees from inadvertently using a dataset in ways that violate institutional policy, violate data-use agreements, or are otherwise inappropriate. Importantly, this type of misuse may be completely innocent. For example, an employee may simply want to be more effective in completing their task by using additional data and is not aware of (or fails to consider) the reasons that such a use-case is not be allowed.

3. *Scope creep:* Minimizing access also minimizes the risk that datasets are reused beyond the purpose for which they were initially procured. For example, a dataset might have been purchased in order to inform a resource allocation process within an agency. If that data is later used to inform individuals' access to benefits (based on information that the government would not have had without access to the CAI), there is a significant change in the *way* the data is being used. The former operates on population-level aggregates, whereas the latter might "punish" individuals because their data happened to be included in the CAI—and someone who is "similar" to them might not be punished in this way. Thus, risks of unfair treatment resulting from these two use-cases are meaningfully different, and repurposing a dataset in this way may lead to significant backlash from the public.

Preventing each of these risks is important, but the second and third risks are especially important when considering government use of CAI. Data exfiltration as part of a data breach may open the agency up to liability issues. For example, the company from which the dataset was purchased may require that the data is not made public. A malicious actor gaining access to the data may constitute a violation of the agreement between the government and the company, resulting in legal and reputational repercussions. That said, the CAI was *available* to malicious actors before a potential breach, albeit for a price. On the other hand, government *misuse* of datasets and scope creep can put the government in a morally compromised position that should be avoided where possible.

New risk evaluations for new contexts (Q9 and Q10) Agencies should evaluate CAI anew each time it is proposed for use in a new context and purpose. As mentioned above, access control

is an important partial solution, but it is not sufficient. This is because privacy harms of using CAI can occur even via actions taken by parties who have been given legitimate access to the data. Thus, it is important for agencies to re-evaluate the risks of using a dataset containing CAI every time its context of use evolves — whether that means it is proposed for a different use case, deployed over a long time horizon, has been linked with other datasets, has been shown to have negative impacts on individuals, and so on.

Reasoning about privacy risks, compliance with ethical standards, and trust boundaries cannot be a "one and done" event, but rather must be a continual process of evaluating and responding to new information and use-cases. Given the ways in which the impacts of PII and AI on privacy are contextual [57], it is not feasible to understand the risks of data in a vacuum. Risk evaluations should be grounded in use cases, and each new use case should trigger a fresh evaluation of the risks of employing CAI within that data and decision pipeline. Otherwise, scope creep and repurposing of the data for new use cases undermine the trustworthiness of government use of CAI.

4.2 Transparency (Q7a)

We strongly recommend creating *public inventories* describing CAI datasets an organization is holding for multiple reasons. At heart, making these lists public facilitate increased trust with the public and provide an opportunity for members of the public to provide *meaningful and timely* feedback to government flagging improper use of CAI datasets. When more people and organizations have an opportunity to provide comment and feedback on the use of CAI, there is a better chance that problems will be identified. Below are some concrete reasons:

- *Risks are cumulative:* As discussed in Section 2, the risks of using data increases as the amount of data—and the dimensionality of that data—increases. Specifically, even if datasets have no direct PII, large amounts of indirect PII can be just as problematic. Critically, this risk is not just about a *single* dataset, but rather about the contents of *all* the datasets held and used by an organization. Specifically, it might be that the risk of holding and using dataset A and dataset B are minimal on their own, but making use of both A and B together poses significant risks. Creating lists of all the datasets that are in use, thus, is strictly *necessary* when it comes to evaluating risk. This is true both if these lists are made public and when they are just internal to an agency.

- *What variables are in use:* It is well understood that not all types of data carry the same privacy risks—both from a normative and a legal perspective. Thus, if the government wants the public to feel trust in a particular government use of CAI, it is important to know what types of CAI (and the variables contained therein) are being used. For example, citizens may feel very differently about a government algorithm that determines loan eligibility if it was trained using card purchase history vs. internet browsing history. Given that these feelings may be hard to predict or may change over time, the government should be committed to continually giving the public access to this information.

- *Minimize costs:* On a more basic level, governments should be responsible for minimizing the costs associated with providing services. Maintaining lists of CAI datasets helps support this goal is two ways: (1) government agencies can *re-use* previously purchased datasets to accomplish multiple task, and having a list of the datasets currently in use provides a convenient way to support this re-use—although care is required when managing this re-use to prevent *scope creep*, as discussed in Section 4.1 above; (2) there may be many different CAI datasets that can support the same analyses and (roughly) the same the same level of effectiveness.

Providing government watchdogs with visibility into the government's purchase history can help reduce the risk of wasteful spending. As with (1), there are caveats: minimizing costs at the expense of data quality can lead to worse outcomes, undermining the motivation for purchasing the CAI in the first place.

- *Identifying use of incorrect or low-quality datasets:* Datasets often contain incorrect information, sometimes in systematic ways. If and when CAI datasets are identified as containing incorrect information, it is important that governments modify systems in response. Making a public list of the CAI datasets used by agencies means that members of the public or civil society can ensure that the government updates these systems in a timely fashion.

- *Identifying use of non-ethical/illegal datasets:* Datasets can be sourced in unethical ways or contain illegal information. For example, datasets can be determined to be the result of data-breaches or be collected using deceptive practices. Alternatively, some datasets contain information that is outright illegal (e.g., the existence of child sexual abuse imagery (CSAM) in the widely used LION dataset [58]). When CAI datasets fall into these categories, it is even more important that there are avenues by which the public and civil society can ensure that government use of these CAI datasets stops.

- *Identifying potential harm:* In the case that a consumer of government systems believes they have experienced some harm or discrimination based on the use of CAI, it is important that they know which datasets are in use. Specifically, if they know that their information is contained in a particular CAI dataset, *knowing* that specific CAI was in use can support them as they seek to rectify the harm or seek recourse. On the other hand, if they are able to determine that none of the CAI datasets contain information about them, they may be able to conclude that any harm they have experienced was not a result of CAI use.

Taken collectively, these concerns are a clear argument that agencies should maintain public lists of the CAI datasets they use.

4.3 Privacy Enhancing Technologies (Q2)

So far, we have explored *processes* that are critical in mitigating the risks associated with using CAI within the government. These are management rules and social guidelines that provide structure necessary for supporting responsible decision making. Process, however, is rarely enough, as guidance can be ignored, individuals can fail to follow procedure, or institutional priorities can shift in a way that supersedes prior consensus. Procedural requirements paired together with technical mechanisms designed to mitigate risk will result in more robust risk mitigation than using either approach alone.[1]

The authors of this response are technologists, first and foremost. As such, we feel that it is important to highlight the existing and emerging technologies that promise to further mitigate risk. These largely fit under the umbrella of *privacy enhancing technologies* (PETs). PETs are a set of techniques deriving from computer science and statistics that attempt to carefully control the ways in which data processing systems leak information. For example, secure multiparty computation

[1]For example, the effectiveness of security protocols can be further enhanced by procedural requirements of transparency. According to Kerckhoff's principle [59], the security of a protocol should not depend on the secrecy of its algorithm or mechanism. Sharing information about the technical security protocols that are used achieves two things—(1) any potential security bugs can be identified by experts who can publicly access this information, and (2) information sharing reduces the risk of privacy theater (i.e., the appearance of privacy protection without actual protection) [60].

(MPC) facilitates different organizations to jointly compute functions of the data held by each organization without needing to actually *share* their data with one another; there have been several successful feasibility tests of using MPC within government [61, 62]. Another emerging PET is differential privacy, which carefully controls the risk incurred by data subjects when aggregate statistics are published or machine learning models are created using their data; differential privacy was recently used by the US Census Bureau within the 2020 decennial census data products in order to fulfill the Census's confidentiality requirements [63].

PETs are best understood as (technical) mechanisms for enforcing *purpose limitation* on data—that is, a way to render the social guidance on how data should be used into automatically enforced barriers. The most popular method of purpose limitation is access control, which offers a binary (technical) notion of limitation—individuals with access to a dataset are technically able to perform arbitrary computation over that data (although there may be rules or regulations that this individual is expected to follow with respect to this data, there is no technical enforcement of these limitations once access is granted), and those without access are unable to do so. PETs extend the ability to purpose limit data to non-binary notions. For example, using MPC can facilitate the computation of *certain types of computational tasks* without risking that other types of computation are conducted later. Similarly, differential privacy enforces that data releases or models only leak a limited amount of information about individuals represented in the data. By enforcing purpose limitation at a technical level—in addition to a procedural level—we are able to minimize the risk of the data being misused, both accidental misuse and intentional misuse by malicious actors. Moreover, purpose limitation provides longitudinal risk mitigation. That is, PETs can help prevent against the following types of social or institutional risks:

1. *Improperly delegating trust*: While the data scientists themselves might be trustworthy enough to have complete access to the data, giving them unlimited access to the data means that they can independently further delegate that trust. For example, they might be approached by other data scientists within that organization who want access to the same data set. The initial data scientist may not be trained or equipped to fully verify the intentions of the second data scientist, and therefore might think that this further delegation of trust is harmless. Then, the second data scientist may make use of the data in a way that is harmful. Moreover, the burden of vetting this delegation is placed on the data scientist—or at least the burden of redirecting the request back to the initial holders of the data.

2. *Trustworthiness can change over time:* Once unlimited access to the dataset is given to a data scientist, it cannot be revoked (from a technical perspective); after all, the data scientist may have already seen the full dataset or even made a copy of it for convenience. Later, that same data scientist may find themselves in a situation where that trust is no longer warranted. For example, consider a scenario in which the data scientist is in a financial situation in which they are tempted to sell access to that data or they lose their job and want to share access to the dataset as an act of retribution. Significantly, this risk can also happen at the structural level (in addition to the individual level). For example, if access to data is given to an external organization (e.g., a contractor), the priorities of that organization might change over time.

3. *Restructuring or turnover during project lifespan:* Sometimes projects may be sufficiently complex that they take a long time to complete. During this window of time, the organization initially granted access may experience turnover, which would change the specific *people* who have access to the data (even if the organization itself stays the same). This change poses a risk, as the new people (both at the data scientist level and management level) may not be as

trustworthy as the individuals initially granted access. This could lead to uses of data that violate the initial parameters of the data use agreement.

In technical parlance, these are all examples of the entity with access to the data becoming adversarial (or being corrupted) after access to the data is granted. PETs prevent or limit these types of misuse by limiting the type of access data scientists are given to datasets.

4.4 Examples of Privacy Enhancing Technologies (Q2)

We briefly summarize some types of PETs that are currently being used (sometimes experimentally) within government, industry, and academia. We suggest that interested readers look to several existing PETs summaries that have been written recently with a public policy audience in mind, including the United Nations PET Guide [64] and guides specifically geared around the use of specific PETs like differential privacy [41, 65].

- *Differential Privacy [66]:* Differential privacy is a property that can be enforced in data processing pipelines that limits the amount of information about individuals contained in the output of the pipeline. Some examples of differential privacy use cases might include: (1) a de-identification process for inter-agency data sharing; (2) an aggregation process that releases tabular summaries of collected data, similar to the way that the US Census Bureau processed data for the 2020 decennial census [63]; (3) training a machine learning model with the data; and (4) producing synthetic data with provable privacy guarantees (e.g., [23]). Differential privacy is generally achieved by adding randomized statistical noise to computations in the data processing algorithm, which comes with an impact on accuracy of estimates. With respect to CAI datasets, applying differential privacy ensures that downstream users of the dataset (e.g., users of trained models or micro-data products) are limited in their ability to extract information about particular data subjects within the dataset.

- *Secure Multiparty Computation:* Secure multiparty computation (MPC) is a set of techniques that provides *input privacy* while facilitating computation. Namely, MPC logically allows joining multiple datasets (possibly held by different organizations or agencies) and computing over the joined datasets *without actually requiring the data to leave its owner's control*. All the parties can then get the results of this particular computation, but no other computations are possible to compute on the joined data after the fact (unless the data owners choose to later engage in a further MPC interaction). For example, the Department of Education ran a pilot study using MPC to combine two data sources (the National Postsecondary Student Aid Study group at the National Center for Education Statistics and the National Student Loan Data System) in order to produce government reports without requiring intra-departmental data sharing [61]. With respect to CAI, MPC can facilitate computation on CAI datasets, in conjunction with in-house datasets, without requiring sharing of data across agency boundaries.

- *Private set intersection/Privacy Preserving Record Linkage:* Private set intersection, often called privacy preserving record linkage within government, enables two (or more) organizations to identify individuals present in multiple datasets, without needing to reveal the datasets to one another. For example, the National Secure Data Service is piloting a study in which they use private set intersection to link health survey data to records of receiving graduate degrees in order to study the mental health of PhD holders [62]. With respect to CAI, private set intersection can help identify the subset of CAI datasets which can help

agencies reach important conclusions without requiring sharing the *entire* dataset. We note that in Section 2 we discussed linkage itself as a privacy risk and here we are discussing a privacy-preserving method for accomplishing that same goal; this reveal the ways in which system designers must be clear about they ways in which a CAI dataset is *intended* to be used before conducting system design.

5 Conclusions

We conclude by re-emphasizing the main points that we hope readers will take away from our response:

- When considering the impact of PII within any dataset—and for CAI datasets in particular—we need to go beyond traditionally recognized "direct identifiers" like name, age, birth date, etc. Rather, *all attributes should be considered PII*, especially with high dimensional data coming from many data sources. This reframing challenges some premises implicitly embedded into the Request for Information's questions.

- The need to consider *all attributes as PII* is especially acute with the rise of modern machine learning techniques. This is because many modern machine learning training processes and their resulting models are highly opaque, and may be inadvertently taking advantage of the indirect PII contained within training datasets. This can lead to individuals who happen to have their data included in CAI datasets being treated significantly differently than other members of the public.

- Careful approaches to access control and continual risk evaluation are *strictly necessary* when dealing with CAI, both to reduce agency liability and minimize the risk of misuse of CAI datasets within government.

- Creating public lists describing the CAI datasets that organizations are using is an important step to establishing trust. This is because risk should be evaluated cumulatively over *all* the datasets that are being used and in order to allow the public to quickly identify problematic datasets in use by the government.

- Privacy enhancing technologies, including differential privacy, secure multiparty computation, and private set intersection, offer technical mechanisms to enforce *non-binary notions of purpose limitation* on CAI datasets. This can limit downstream misuse of CAI datasets, and thereby reduce privacy risks to individuals.

References

[1] Latanya Sweeney. Simple demographics often identify people uniquely, 2000.

[2] Philippe Golle. Revisiting the uniqueness of simple demographics in the us population. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, pages 77–80, 2006.

[3] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE, 2008.

[4] Arka Rai Choudhuri, Matthew Green, Abhishek Jain, Gabriel Kaptchuk, and Ian Miers. Fairness in an unfair world: Fair multiparty computation from public bulletin boards. In Bhavani M. Thuraisingham, David Evans, Tal Malkin, and Dongyan Xu, editors, *ACM CCS 2017: 24th Conference on Computer and Communications Security*, pages 719–728, Dallas, TX, USA, October 31 – November 2, 2017. ACM Press.

[5] Gabriel Kaptchuk, Matthew Green, and Ian Miers. Giving state to the stateless: Augmenting trustworthy computation with ledgers. In *ISOC Network and Distributed System Security Symposium – NDSS 2019*, San Diego, CA, USA, February 24-27, 2019. The Internet Society.

[6] Arka Rai Choudhuri, Aarushi Goel, Matthew Green, Abhishek Jain, and Gabriel Kaptchuk. Fluid MPC: Secure multiparty computation with dynamic participants. In Tal Malkin and Chris Peikert, editors, *Advances in Cryptology – CRYPTO 2021, Part II*, volume 12826 of *Lecture Notes in Computer Science*, pages 94–123, Virtual Event, August 16–20, 2021.

[7] Gabrielle Beck, Aarushi Goel, Abhishek Jain, and Gabriel Kaptchuk. Order-C secure multiparty computation for highly repetitive circuits. In Anne Canteaut and François-Xavier Standaert, editors, *Advances in Cryptology – EUROCRYPT 2021, Part II*, volume 12697 of *Lecture Notes in Computer Science*, pages 663–693, Zagreb, Croatia, October 17–21, 2021.

[8] Matthew Green, Gabriel Kaptchuk, and Gijs Van Laer. Abuse resistant law enforcement access systems. In Anne Canteaut and François-Xavier Standaert, editors, *Advances in Cryptology – EUROCRYPT 2021, Part III*, volume 12698 of *Lecture Notes in Computer Science*, pages 553–583, Zagreb, Croatia, October 17–21, 2021.

[9] Gabriel Kaptchuk, Tushar M. Jois, Matthew Green, and Aviel D. Rubin. Meteor: Cryptographically secure steganography for realistic distributions. In Giovanni Vigna and Elaine Shi, editors, *ACM CCS 2021: 28th Conference on Computer and Communications Security*, pages 1529–1548, Virtual Event, Republic of Korea, November 15–19, 2021. ACM Press.

[10] Aarushi Goel, Matthew Green, Mathias Hall-Andersen, and Gabriel Kaptchuk. Efficient set membership proofs using MPC-in-the-head. *Proceedings on Privacy Enhancing Technologies*, 2022(2):304–324, April 2022.

[11] Aarushi Goel, Matthew Green, Mathias Hall-Andersen, and Gabriel Kaptchuk. Stacking sigmas: A framework to compose Σ-protocols for disjunctions. In Orr Dunkelman and Stefan Dziembowski, editors, *Advances in Cryptology – EUROCRYPT 2022, Part II*, volume 13276 of *Lecture Notes in Computer Science*, pages 458–487, Trondheim, Norway, May 30 – June 3, 2022.

[12] Gabrielle Beck, Aarushi Goel, Aditya Hegde, Abhishek Jain, Zhengzhong Jin, and Gabriel Kaptchuk. Scalable multiparty garbling. pages 2158–2172. ACM Press, 2023.

[13] Aarushi Goel, Mathias Hall-Andersen, Gabriel Kaptchuk, and Nicholas Spooner. Speed-stacking: Fast sublinear zero-knowledge proofs for disjunctions. In *Advances in Cryptology – EUROCRYPT 2023, Part II*, Lecture Notes in Computer Science, pages 347–378, June 2023.

[14] Tushar M. Jois, Gabrielle Beck, and Gabriel Kaptchuk. Pulsar: Secure steganography for diffusion models. In Bo Luo, Xiaojing Liao, Jun Xu, Engin Kirda, and David Lie, editors, *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pages 4703–4717. ACM, 2024.

[15] Aarushi Goel, Mathias Hall-Andersen, and Gabriel Kaptchuk. Dora: A simple approach to zero-knowledge for RAM programs. In Bo Luo, Xiaojing Liao, Jun Xu, Engin Kirda, and David Lie, editors, *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS 2024, Salt Lake City, UT, USA, October 14-18, 2024*, pages 869–883. ACM, 2024.

[16] Tushar M. Jois, Gabrielle Beck, Sofia Belikovetsky, Joseph Carrigan, Alishah Chator, Logan Kostick, Maximilian Zinkus, Gabriel Kaptchuk, and Aviel D. Rubin. Socioty: Practical cryptography in smart home contexts. *Proc. Priv. Enhancing Technol.*, 2024(1):447–464, 2024.

[17] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. Visualizing privacy-utility trade-offs in differentially private data releases. *Proceedings on Privacy Enhancing Technologies*, 2022.

[18] Priyanka Nanayakkara, Hyeok Kim, Yifan Wu, Ali Sarvghad, Narges Mahyar, Gerome Miklau, and Jessica Hullman. Measure-observe-remeasure: An interactive paradigm for differentially-private exploratory analysis. *IEEE Security & Privacy*, 2024.

[19] Jeremiah Blocki, Seunghoon Lee, Tamalika Mukherjee, and Samson Zhou. Differentially private l_2-heavy hitters in the sliding window model. In *Eleventh International Conference on Learning Representations, ICLR*, 2023.

[20] Jeremiah Blocki, Elena Grigorescu, Tamalika Mukherjee, and Samson Zhou. How to Make Your Approximation Algorithm Private: A black-box differentially-private transformation for tunable approximation algorithms of functions with low sensitivity. In *Approximation, Randomization, and Combinatorial Optimization, Algorithms and Techniques, APPROX/RANDOM*, 2023.

[21] Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. Differentially private simple linear regression. *Proceedings on Privacy Enhancing Technologies*, 2022(2):184–204, 2022.

[22] Museums Moving Forward. Data study. https://museumsmovingforward.com/data-study.

[23] Shlomi Hod and Ran Canetti. Differentially private release of israel's national registry of live births. In *IEEE Symposium on Security and Privacy (S&P)*, 2025.

[24] Christina Garman, Matthew Green, Gabriel Kaptchuk, Ian Miers, and Michael Rushanan. Dancing on the lip of the volcano: Chosen ciphertext attacks on apple iMessage. In Thorsten Holz and Stefan Savage, editors, *USENIX Security 2016: 25th USENIX Security Symposium*, pages 655–672, Austin, TX, USA, August 10–12, 2016. USENIX Association.

[25] Ian Martiny, Gabriel Kaptchuk, Adam J. Aviv, Daniel S. Roche, and Eric Wustrow. Improving Signal's sealed sender. The Internet Society, 2021.

[26] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. "I need a better description": An investigation into user expectations for differential privacy. In Giovanni Vigna and Elaine Shi, editors, *ACM CCS 2021: 28th Conference on Computer and Communications Security*, pages 3037–3052, Virtual Event, Republic of Korea, November 15–19, 2021. ACM Press.

[27] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. What are the chances? Explaining the epsilon parameter in differential privacy. pages 1613–1630. USENIX Association, 2023.

[28] Mary Anne Smart, Priyanka Nanayakkara, Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. Models matter: Setting accurate privacy expectations for local and central differential privacy. *CoRR*, abs/2408.08475, 2024.

[29] Vaughn Hamilton, Gabriel Kaptchuk, Allison McDonald, and Elissa M. Redmiles. Safer digital intimacy for sex workers and beyond: A technical research agenda. *CoRR*, abs/2403.10688, 2024.

[30] Jayshree Sarathy, Sophia Song, Audrey Haque, Tania Schlatter, and Salil Vadhan. Don't look at the data! how differential privacy reconfigures the practices of data science. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[31] Patrick Song, Jayshree Sarathy, Michael Shoemate, and Salil Vadhan. " i inherently just trust that it works": Investigating mental models of open-source libraries for differential privacy. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–39, 2024.

[32] danah boyd and Jayshree Sarathy. Differential perspectives: Epistemic disconnects surrounding the US Census Bureau's use of differential privacy. *Harvard Data Science Review (Forthcoming)*, 2022.

[33] Connor Wagaman, Palak Jain, and Adam Smith. Time-aware projections: Truly node-private graph statistics under continual observation. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 237–237. IEEE Computer Society, 2024.

[34] Palak Jain, Iden Kalemaj, Sofya Raskhodnikova, Satchit Sivakumar, and Adam Smith. Counting distinct elements in the turnstile model with differential privacy under continual observation, 2023.

[35] Palak Jain, Sofya Raskhodnikova, Satchit Sivakumar, and Adam Smith. The price of differential privacy under continual observation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 14654–14678. PMLR, 23–29 Jul 2023.

[36] Ran Canetti, Gabe Kaptchuk, Leonid Reyzin, Adam Smith, and Mayank Varia. Response to the rfi on advancing privacy-enhancing technologies. Available at https://www.nitrd.gov/rfi/2022/87-fr-35250/Canetti-Kaptchuk-Reyzin-Smith-Varia-PET-RFI-Response-2022.pdf, July 2022.

[37] Rachel Cummings, Shlomi Hod, Gabriel Kaptchuk, Priyanka Nanayakkara, Jayshree Sarathy, and Jeremy Seeman. Comment on "nist sp 800-226: Guidelines for evaluating differential privacy guarantees". Available at https://www.cs.umd.edu/~kaptchuk/publications/nist24-dp-public-comment.pdf, March 2024.

[38] Rachel Cummings. Modern solutions to modern problems: leveraging the latest in data privacy and algorithmic fairness. Available at https://www.regulations.gov/comment/FTC-2022-0053-1103, Nov 2022. Response to FTC Advance Notice of Proposed Rulemaking on Commercial Surveillance and Data Security.

[39] Rachel Cummings and Jayshree Sarathy. Centering policy and practice: Research gaps around usable differential privacy. In *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 122–135. IEEE, 2023.

[40] Jayshree Sarathy and danah boyd. Statistical imaginaries, state legitimacy: Grappling with the arrangements underpinning quantification in the US census. *Critical Sociology*, page 08969205241270898, 2024.

[41] Priyanka Nanayakkara and Jessica Hullman. What to consider when considering differential privacy for policy. *Policy Insights from the Behavioral and Brain Sciences*, 11(2):132–140, 2024.

[42] Congressial Workshop - Operationalizing Responsible AI. `https://gov.responsibly.ai/23-congress`, Aug 2023.

[43] Shlomi Hod, Karni Chagal-Feferkorn, Niva Elkin-Koren, and Avigdor Gal. Data science meets law. In *Commununications of the ACM (CACM)*, 2022.

[44] Office of Management and Budget. Managing information as a strategic resource. Circular A-130, Executive Office of the President, Washington, DC, 2016.

[45] Tore Dalenius. Finding a needle in a haystack or identifying anonymous census records. *Journal of official statistics*, 2(3):329, 1986.

[46] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[47] Erika McCallister, Tim Grance, and Karen Scarfone. Guide to protecting the confidentiality of personally identifiable information (pii). Special Publication 800-122, National Institute of Standards and Technology, Gaithersburg, MD, 2010.

[48] Aloni Cohen. Attacks on deidentification's defenses. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1469–1486, 2022.

[49] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284, 2019.

[50] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security)*, 2021.

[51] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *CoRR*, abs/2311.17035, 2023.

[52] Niv Haim, Gal Vardi, Gilad Yehudai, Ohad Shamir, and Michal Irani. Reconstructing training data from trained neural networks. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, (NeurIPS)*, 2022.

[53] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In Joseph A. Calandrino and Carmela Troncoso, editors, *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 5253–5270. USENIX Association, 2023.

[54] Francesco Pinto, Nathalie Rauschmayr, Florian Tramèr, Philip Torr, and Federico Tombari. Extracting training data from document-based VQA models. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

[55] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[56] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s):235:1–235:37, 2022.

[57] Helen Nissenbaum. Privacy in context: Technology, policy, and the integrity of social life, 2009.

[58] David Thiel. Identifying and eliminating csam in generative ml training data and models. Published by the Stanford Internet Observatory. Available at `https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf`., 2023. Accessed on 15 December 2024.

[59] La cryptographie militaire, 1883.

[60] Rohit Khare. Privacy theater: Why social networks only pretend to protect you, 2022.

[61] David Archer, Amy O'Hara, Rawane Issa, and Stephanie Straus. Sharing sensitive department of education data across organizational boundaries using secure multiparty computation, may.

[62] Lisa Mirel, Cordell Goldent, Rob Zybrick, Rui Wang, Chrystine Tadler, and Christine Cox. Linking data in a shared service environment. *International Journal of Population Data Science*, 9(5), Sep. 2024.

[63] John M Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, et al. The 2020 census disclosure avoidance system topdown algorithm. *Harvard Data Science Review*, 2, 2022.

[64] The United Nations. The pet guide: The United Nations guide on privacy-enhancing technologies for official statistics. Available at `https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf`, 2023. Accessed on 15 December 2024.

[65] Joseph P Near, David Darais, Naomi Lefkovitz, Gary Howarth, et al. Guidelines for evaluating differential privacy guarantees. *National Institute of Standards and Technology, Tech. Rep*, 2023.

[66] Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R. O'Brien, Thomas Steinke, and Salil Vadhan. Differential privacy: A primer for a non-technical audience. *Vanderbilt Journal of Entertainment & Technology Law*, 21(1):209–275, 2018. `http://www.jetlaw.org/journal-archives/volume-21/volume-21-issue-1/differential-privacy-a-primer-for-a-non-technical-audience/`.