Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

Request for Information (RFI) on

Advancing Privacy Enhancing Technologies

MOSTLY AI Inc

MOSTLY AI Response

to the

Request for Information on

Advancing Privacy-Enhancing Technologies

from the

SCIENCE AND TECHNOLOGY POLICY OFFICE

Alexandra Ebert

Chief Trust Officer

MOSTLY AI Inc.

Type: Industry

Dr. Michael Platzer

Chief Strategy Officer & Co-Founder

MOSTLY AI Inc.

Type: Industry

Response of MOSTLY AI[1] to

OSTP RFI on Advancing Privacy-Enhancing Technologies

8th of July, 2022

Introduction

MOSTLY AI Inc. ("MOSTLY AI") welcomes the opportunity to the OSTP's RFI on advancing privacy-enhancing technologies (PETs). As the global leader in **structured synthetic data**, we are dedicated to enabling an open data ecosystem where access to high-quality, diverse, granular-level data can be democratized while privacy remains protected. Synthetic data allows to distill the insights from existing data in a fully automated manner, and makes these insights accessible by generating statistically representative, highly realistic, yet completely new data samples at scale. With that, our synthetic data platform enables already today public and private sector organizations across the world to safely innovate and collaborate on top of large-scale data assets towards building a smarter and safer future.

And it's the data that serves as the **lingua franca** of the digital era**.** It is the source for learning and exploration by machines and by humans alike. In particular, it is data at the granular level, that represents subjects and events, that can be easily understood, interpreted, and reasoned upon by people of all backgrounds. For that reason, we see that synthetic data serves a unique need among the group of emerging PETs, as it allows the involvement of much broader communities and stakeholders in the process of building and validating the algorithms that will shape our society going forward.

We would also like to use the opportunity to bring attention to a recently published report by the European Commission on synthetic data[2], that concludes:

"Synthetic data can become the unifying bridge between policy support and computational models, by unlocking the potential of data hidden in silos; thus ***becoming the key enabler of artificial intelligence****. [..] More important than focusing on how to synthesize data is what can we achieve with the new data available at scale, how to convince data owners to unleash their coveted data to the broadest audience, and how to accommodate this* ***massive new ability*** *into the policy formulation and assessment."*

[1] https://www.mostly.ai/
[2] https://publications.jrc.ec.europa.eu/repository/handle/JRC128595

Response to the RFI

1. Specific research opportunities to advance PETs

Aside from research on the technical feasibility and validity of emerging PETs, we see a need to research

1) their anticipated impact on society and economy – analyzed from a non-technical perspective

2) the interplay of the various PETs with each other - PETs are oftentimes analyzed in isolation, but they can very well complement each other

3) how to encourage or even mandate open data sharing practices for the benefits of society

2. Specific technical aspects or limitations of PETs

We consider encryption-based mechanism to provide security, but not necessarily zero-trust privacy guarantees. While these can help to provision data safely to machines to perform pre-defined computation, they fall short in enabling broad data sharing with humans.

We consider aggregation- and query-based systems to be capable of satisfying privacy requirements, but they again do not allow the broad sharing of granular-level data with humans.

We consider synthetic data to take a unique position, that comes at the cost of carefully reduced accuracy, but can then provide unrestricted access to granular-level data. To cite the aforementioned report by the European Commission on the subject:

> *"Correctly performed synthesis introduces controllable and well described distortion of the original data, which is just a small price to pay for the availability of highly granular privacy unburdened data. [..] Among the privacy-preservation technique studies analysed (differential privacy, data perturbation, homomorphic encryption, secure private computing infrastructure), data synthesis gave the best price (effort)/cost ratio."*

A current limitation of synthetic data is still the lack of commonly accepted standards and benchmarks to empirical assess the accuracy and privacy of various approaches. Not all synthetic data approaches are necessarily accurate. Not all approaches are automatically private. The involved machine learning algorithms can suffer from underfitting as well as from overfitting to the original data, if done incorrectly. First research initiatives exist[3], but more standardization is needed.

Another limitation of synthetic data exists with respect to scaling the required compute for very large data assets.

---

[3] https://www.frontiersin.org/articles/10.3389/fdata.2021.679939/full

3. **Specific sectors, applications or types of analysis that would particularly benefit from the adoption of PETs**

We consider healthcare[4] as well as the public sector to have the biggest potential for positive impact for society, if their data assets could be shared more openly in a truly privacy-preserving manner. Beyond that, we consider the financial service and the telecommunication industries to be segments that gather the largest pools of insightful behavioral data, thus posing an enormous economically opportunity, if such data can be safely utilized. But any sector, that deals with personal data at scale (retail, education, recreation, mobile services, etc.), is expected to benefit, if their insights become accessible in a privacy-safe manner.

It's important to emphasize that PETs not only allow safe data access across organizations, or across borders, but also foster data sharing within an organization. Data access is a crucial factor to accelerate innovation and thus to strengthen the competitive advantage of US entities, while safeguarding the privacy of all citizens.

4. **Specific regulations or authorities that could be used, modified or introduced to advance PETs**

PET adoption across industries can be greatly accelerated by authorities leading by example, and proactively taking them in use for protecting citizen data, while also publicly and transparently communicating about these taken initiatives.

5. **Specific laws that could be used, modified or introduced to advance PETs**

GDPR and its absolute and strict definition of anonymous data can serve as guidance for introducing a national level privacy law, which would certainly help the adoption of PETs. The more aligned these regulations are, the closer the collaboration of the western world in the digital era will be.

Aside from privacy protection, it is important to recognize the need for data to validate and assess algorithms that impact individual's lives. Ideally, upcoming laws shall consider requirements to put external parties into a position to stress-test algorithms for fairness, by being mandated to share not only model access, but also access to representative (synthetic) data samples at scale.

[4] See eg Humana's synthetic data exchange https://developers.humana.com/syntheticdata

6. **Specific mechanisms, not covered above, that could be used, modified or introduced to advance PETs**

Competitions and benchmarks are essential to advance PETs. These shall be open for submissions, both from open-source as well as from proprietary solution providers, in order to encourage broad competition and investments. However, any solution shall be openly accessible for stress testing by a broad audience, as e.g., done by MOSTLY AI with their publicly available free version of their synthetic data platform[5].

7. **Risks related to PETs adoption**

The unique strength of the US economy builds upon diversity, creativity, and open collaboration. We consider it thus as crucial to understand the non-technical implications of the various PETs in detail. Particularly with respect to whether they allow for broad, diverse communities to directly benefit from accessing data.

We see it as a potential risk, if PETs result in strengthen the already existing data monopolies of few organizations, that can afford to use PETs to exclusively gather even more data. Data is information, data is knowledge, and thus data shall be accessible to as many people as possible.

8. **Existing best practices that are helpful for PETs adoption**

We recommend a crawl, walk, run approach. We recommend to start small, and gradually and swiftly build up experience and expertise with respect to the already existing PETs, rather than investing in big initiatives with risky return in a far distant future.

[5] https://mostly.ai/