Centre for Information Policy Leadership
HUNTON

Privacy-Enhancing and Privacy-Preserving Technologies in AI:

Enabling Data Use and Operationalizing Privacy by Design and Default

March 2025

Table of Contents

1. Introduction

As AI continues to evolve, it also continues to fuel an ongoing discussion on how to protect personal data and the rights of individuals in the context of AI training and deployment.[1] AI technologies, generative AI, traditional machine-learning and new agentic AI, rely on vast and diverse data to train and fine-tune the underlying algorithms. Those data sets can include personal or even sensitive information. The collection and processing of this data can raise privacy and cybersecurity concerns and compliance challenges.

At the same time, governments and corporate boards are realizing the power and beneficial role of data in driving digital transformation, leading to increasing demand – and sometimes mandates – for the sharing of data. This creates additional challenges for companies and public sector organizations, who are reluctant to share their key asset and want to preserve commercial interests and intellectual property rights as well as avoid privacy and security pitfalls.

Privacy-enhancing or privacy-preserving technologies (PETs and PPTs)[2] provide opportunities to protect privacy and cybersecurity in the development and deployment of AI while enabling broader beneficial sharing and use of data across different organizations and sectors to boost further AI adoption. As such, PETs also serve as key business enablers, allowing companies and public sector organizations to access, share and use data that would otherwise be unavailable. In addition to safeguarding privacy, PETs also help protect confidential information, trade secrets, commercial interests, and ensure regulatory compliance. As AI adoption grows, the need for solutions that safeguard privacy and enable responsible data access and sharing has become more urgent, with organizations like the Organisation for Economic Co-operation and Development (OECD) also actively working on this subject.[3]

This family of technologies can be used to train AI models while safeguarding privacy in a range of different ways.[4] For example, federated learning provides the possibility to train an AI model without exposing personal data to the party training the model.[5] Similarly, homomorphic encryption enables secure cross-border data sharing, allowing organizations in different countries to collaboratively train AI models without revealing sensitive data. PETs may also help to de-identify or anonymize data used to train AI models. PETs hold immense potential to help operationalize privacy by design and by default when developing AI systems[6].

While PETs can play a significant role in reducing data, privacy and security risks, they are not a panacea and may not work in all the situations where the trade-offs between utility and protection need to be made. Rather, they should be seen as one of many tools to help mitigate risks and address commercial and legal challenges. Furthermore, PETs are not a one-size-fits-all solution. As this paper will demonstrate, different PETs are most effective when applied at various stages of the AI life cycle. In many circumstances, to achieve the best outcomes, PETs should be used in combination.

In December 2023, CIPL published its white paper entitled "Privacy-Enhancing and Privacy-Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age".[7] It provides insights into the different types of PETs available, demonstrates their application through case studies, and explores how PETs support data-protection principles and innovation. It also explores how organizations are approaching PETs, potential challenges to their use and what solutions could look like. Below are the key recommendations identified by CIPL in this previous paper to address the challenges preventing widespread PET adoption:

- **Regulators and lawmakers should provide greater legal certainty by issuing regulatory guidance on and incentivizing PETs**. Organizations need legal certainty to be able to operate, invest and grow. Official regulatory guidance addressing PETs in the context of specific legal obligations or concepts (such as anonymization) will certainly drive adoption, as will mitigations in enforcement and "safe harbors" from liability for organizations that implement PETs in good faith. By supporting such initiatives, regulators and policymakers will also incentivize greater private sector investment in fundamental and application-specific research to advance these technologies.
- **Increase education and awareness about PETs**. To achieve widespread adoption, PET developers and providers need to show tangible evidence of the value of PETs and how such technologies can facilitate responsible data use. Case studies of deployments are especially useful for this purpose. Equally, businesses must understand the limitations of PETs and the conditions that determine which PET or combination of PETs is most suitable for a given use case. Individuals whose data is being processed via PETs also need a better understanding of the technology and the protection measures put in place. This will foster further trust and digital confidence.
- **Develop industry standards for PETs**. The lack of industry standards for many PETs is an obstacle to their wider adoption. While standards do exist for some PETs (such as homomorphic encryption), other PETs (like differential privacy) are at an earlier stage of development. Industry standards would help facilitate interoperability among PETs across jurisdictions. Common frameworks would establish compatibility and consistency, enabling different PETs to communicate and work together. Standards would also help codify best practices, thereby ensuring a level of sophistication and technical reliability to foster trust in the technologies.
- **Recognize PETs as a demonstrable element of accountability**. PETs complement robust data and privacy management programs that are grounded in principles of organizational accountability, such as CIPL's Accountability Framework. By helping to mitigate risk and avoid harm, PETs support compliance efforts and demonstrate effective accountability. Organizations developing, deploying and investing in PETs are able to demonstrate their commitment to protecting privacy, while at the same time enabling beneficial uses of data in a systematic, sustainable and an accountable way.

This paper, *PETs and PPTs in AI*, is part of the next stage of CIPL's research on PETs: an in-depth exploration of how PETs can and are being deployed to address privacy concerns specifically within AI systems. The paper describes how these technologies can help address challenges and provide new opportunities in data sourcing, model training, security, collaboration, model validation and model deployment.[8] To illustrate the application and benefits of PETs, we have collected and include real-world case studies throughout this paper.[9]

Table 1 and Figure 1 at the end of this section summarize the ways in which PETs can strengthen privacy for AI development and deployment. These visualizations, supported by the case studies in the paper, demonstrate that different PETs can be used together to enhance privacy even more effectively than when used alone, providing a multi-layered approach to protecting data in AI systems. Finally, this is an increasingly competitive and growing area of applied research in organizations, with many new applications and case studies being developed and deployed on the ground. CIPL's paper provides a snapshot in time of the current state of development and use of PETs in the blossoming AI field. We will continue to gather evidence, prepare additional case-studies, and document best practices and challenges as this field continues to grow.

CIPL's Recommendations for Boosting Adoption of and Overcoming Challenges in the Use of PETs for AI

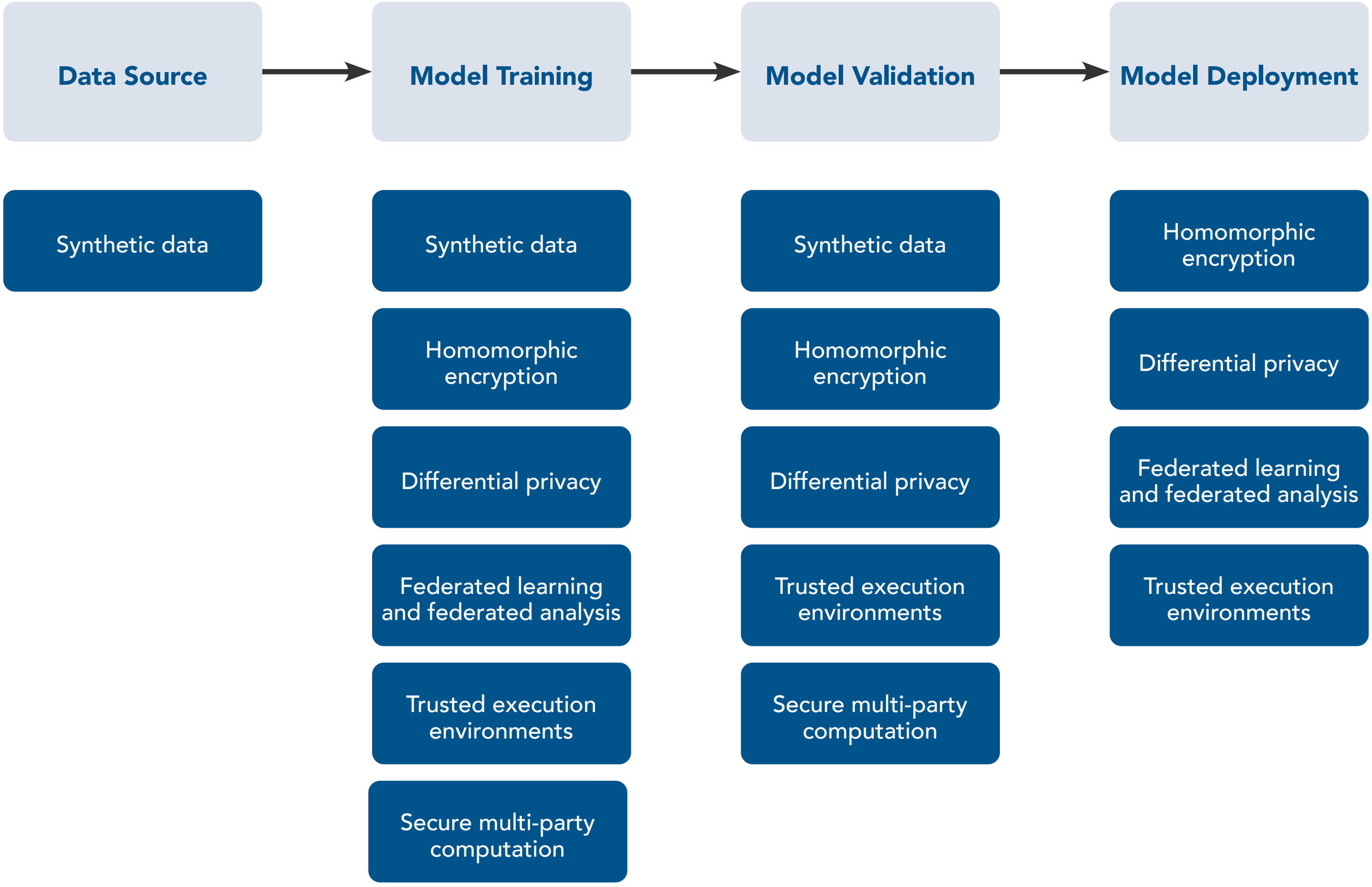
CIPL offers the following recommendations for PET deployers, regulators, and policymakers to boost development and adoption of and address challenges in the use of PETs for AI applications. These challenges are explained in more detail later in the paper:

- **Regulators should issue more clear and practical guidance to reduce regulatory uncertainty in the use of PETs in AI.** While regulators increasingly recognize the value of PETs, clearer and more practical guidance is needed to help organizations implement these technologies effectively. For instance, regulators should avoid setting unnecessarily stringent standards for determining if a particular PET achieves compliance with a specific data-protection principle, and acknowledge that PETs are not a "silver bullet" for every data governance challenge.
- **Regulators should adopt a risk-based approach to assess how PETs can meet standards for data anonymization, providing clear guidance to eliminate uncertainty.** There is uncertainty around whether various PETs meet legal standards for data anonymization. A risk-based approach to defining anonymization standards could encourage wider adoption of PETs.
- **Deployers should take steps to provide contextually appropriate transparency to customers and data subjects.** Given the complexity of PETs, deployers should ensure customers and data subjects understand how PETs function within AI models.
- **Deployers should take care to ensure that clear mechanisms exist for data subjects to exercise their rights, where applicable.** PETs may alter data in ways that affect data subject rights. Deployers must establish processes to help subjects exercise their rights.
- **Deployers must balance protecting privacy with data utility considerations.** While protecting privacy is crucial, deployers must also ensure that PETs do not impede the utility of data for AI development.
- **Policymakers and industry must work together to address the demand for large computing resources.** The use of PETs in AI, especially generative AI, can require substantial computing resources. Policymakers and industry need to work together to ensure adequate resources are available.
- **Regulators should incentivize proactive dialogue, further research, and experimentation with PETs within regulatory sandboxes.** PETs are a rapidly evolving field with great potential to enhance data privacy and security. Encouraging collaboration in regulatory sandboxes would promote ongoing dialogue and knowledge exchange between key stakeholders, helping develop adaptable regulatory frameworks that keep pace with PETs.
- **Stakeholders should adopt a holistic view of the benefits of PETs in AI.** PETs deliver value beyond addressing privacy and security concerns: They also foster trust, improve regulatory compliance, and enable data sharing while protecting sensitive information. It is crucial that stakeholders consider all these advantages when making decisions about their use.

Table 1: PETs in AI Summary Table

Privacy-Enhancing Technology	How PETs can help in AI Life Cycle					
	Data Source	Model Training	Security	Collaboration	Model Validation	Model Deployment
Synthetic Data	✓	✓			✓	
Homomorphic Encryption		✓	✓	✓	✓	✓
Differential Privacy		✓	✓	✓	✓	✓
Federated Learning and Federated Analysis		✓	✓	✓		✓
Trusted Execution Environments		✓	✓	✓	✓	✓
Secure Multi-Party Computation		✓	✓	✓	✓	

Figure 1: PETs in AI Life Cycle



Data Source

Synthetic data

Model Training

Synthetic data

Homomorphic encryption

Differential privacy

Federated learning and federated analysis

Trusted execution environments

Secure multi-party computation

Model Validation

Synthetic data

Homomorphic encryption

Differential privacy

Trusted execution environments

Secure multi-party computation

Model Deployment

Homomorphic encryption

Differential privacy

Federated learning and federated analysis

Trusted execution environments

2. Applications for PETs in AI

There are a number of different use cases that find application across the PETs family.

a) Synthetic Data

Synthetic data refers to artificially generated data that resembles real data.[10] In the context of AI, synthetic data is generated by training models on real data to learn underlying patterns, distributions, and correlations, and then using these learnings to create new, artificial data mimicking the values of the original data.[11] Personal information may be replaced in part by machine-generated data, or fully removed. By carefully generating synthetic data, an alternative to real-world data can be provided that protects privacy, without losing the data value. Synthetic data can then be used for various purposes, including data analysis, machine learning, model training, testing and data sharing without the risk of exposing information. This is particularly useful where organizations seek data defined as "sensitive",[12] which can be critical for preventing bias and discriminatory outcomes from models or in medical research, for example.

Data Source

Synthetic data generation can yield large volumes of data valuable for AI model training. Consequently, synthetic data offers one method to provide training data, where there may be a shortage or limit to the data available. It can provide large volumes of data for rigorous testing where this may not have been possible before, enabling the evaluation of model performance across diverse, edge-case scenarios that would otherwise be difficult to replicate.

Model Training

By using synthetic data instead of real-world data, organizations can train AI models on data that mirrors the statistical properties of real-data sets without containing any real-world information. This enables data scientists and developers to train models with the appropriate data values, but without handling sensitive data directly, thus avoiding legal compliance concerns that may otherwise be associated with collecting and processing such data.

Model Validation

Organizations wishing to test an external vendor's AI model without using their own data, could instead choose to use synthetic data. By carefully generating synthetic data (or purchasing it from specialist vendors), organizations can ultimately utilize a data set similar to their own data, without disclosing personal or sensitive (or proprietary) information.

CASE STUDY 1: Synthetic data to train and fine-tune large-language models collaboratively[13]

InstructLab is an open-source project that allows developers to develop large-language models together. It operates by using a large-language model that has been fed a few examples of human-generated training data (from the developers) to create a large amount of synthetic data. This synthetic data is then used to fine-tune and customize the large-language model. As an open-source project, developers are able to collectively contribute to the large-language model's development.

By requiring only a few examples of human-generated data, this method allows large-language models to be improved in less time and for lower cost. It also protects privacy by allowing large-language models to be built without the need for large amounts of real-world data which could include personal data.

CASE STUDY 2: Synthetic data used in combination with web data to train small-language models[14]

In comparison to large-language models, small-language models require less data. These models may be better suited for simple tasks, such as automated customer support or summarizing reports, and be more accessible to organizations with limited resources. They can also run locally on a device, instead of the cloud, minimizing latency and maximizing privacy.

Small-language models can be developed using a combination of publicly available web data and synthetic data created by a large-language model. Using this data together enables small-language models to reach similar performance levels to much larger models. The model is first trained on web data to teach it general knowledge and language understanding. It is then trained on more specific web data and synthetic data to teach the model logical reasoning and more specific skills. When the large-language model creates the synthetic data, to ensure it is of high quality, the output is repeatedly checked, filtered and fed back into the model until the required quality is reached.

CASE STUDY 3: Synthetic data to improve accuracy of neural networks[15]

Customers are nowadays able to make payments, access loyalty rewards and verify their age, using their palm. At select locations, including entertainment venues, convenience stores and grocery shops, customers have the option to use a device that will connect their palm to a payment method or account by recognizing the shape and contour of their palm, as well as the veins under their skin. However, when deploying such a system, accuracy and security are of utmost importance to ensure the correct payment method or account is identified and the privacy of the user is protected. In order to train the AI model behind it and improve accuracy, generative AI was used to produce millions of synthetic images of palms including the blood vessels under the surface, palms in different light conditions, in different poses, and palms with different conditions, such as with a medical plaster on. It was also used to help the model detect fake hands. This created a highly accurate model and at the same time reduced the amount of real-world data needed for training. This approach eliminated the need to use the real-world data in production.[16]

CASE STUDY 4: Synthetic data for developing AI reasoning in mathematics[17]

AI systems often struggle with complex problems in mathematics due to the need for large amounts of training data and high-level reasoning skills. Synthetic data can be used to overcome these issues. By generating 100 million unique examples for training data, synthetic data has enabled the development of an advanced AI system in mathematics.

Humans learn geometry by examining diagrams and using their knowledge to discover geometric properties and relationships. To emulate this process, 1 billion random diagrams of geometric objects were generated where the AI derived the relationships between the points and lines for each diagram. This data set was then filtered to exclude similar examples, resulting in a final training data set of 100 million examples of varying difficulty. As a result, synthetic data helped to address insufficient training data and facilitated the development of the AI system able to solve complex mathematical problems.

CASE STUDY 5: Synthetic data for training frontier open-source AI models[18]

Synthetic data can also improve the model development process by accelerating and scaling training efforts. In developing one of the largest open-source language models, synthetic data played a key role during post-training. Synthetic data allowed for rapid, iterative rounds of alignment using supervised fine-tuning. Multiple data processing techniques were also used to filter this synthetic data to the highest quality. This allowed for efficient model improvements while supporting scalability and diverse applications.

CASE STUDY 6: Synthetic data generation from proprietary data[19]

MOSTLY AI's recently launched open-source synthetic data toolkit is a key example of how organizations can generate high-quality, privacy-safe synthetic datasets from their sensitive, proprietary data, all within their own infrastructure. This innovative solution allows businesses to leverage their valuable internal data for AI training without the concerns of privacy risks or compliance challenges. By facilitating the secure use of proprietary data, this toolkit paves the way for the development of more accurate and contextually relevant AI models by empowering organizations to fuel their AI models with authentic and meaningful data.

Considerations for use of Synthetic Data

The use of synthetic data brings certain challenges to be considered before deployment:

- Models used to generate synthetic data must be designed carefully to mitigate biases or inaccuracies that can then impact the synthetic data sets generated from them. Biases present in real-world data will be replicated in synthetic data unless identified and addressed during the data-generation process. There are a variety of tools and techniques for generating synthetic data, each carrying its own risks for introducing bias. It is essential to understand and document these biases and classify and address them based on a risk-based approach. This will improve synthetic data generation across industries, ultimately enhancing the overall quality, fairness, and effectiveness of the data.
- Similarly, as synthetic data relies upon real-world data, where the real-world data is inaccurate or incomplete, it can negatively impact the synthetic data being generated. The real-world data must be carefully selected and checked for accuracy beforehand.
- Synthetic data can pose re-identification risks through "singling-out attacks", where synthetic instances

closely resemble real individuals; "linkability attacks", where synthetic data is linked to real identities by matching attributes across data sets; and "inference attacks", where sensitive information can be deduced from statistical patterns. Additionally, real-world data can be leaked if the synthetic model captures outliers from the original data, allowing malicious actors to infer information about the real data set. To mitigate these risks, methods such as removal of direct identifiers and identified outliers, along with the additional application of differential privacy, should be employed, and the synthetic data's proximity to the real data must be monitored to ensure sufficient privacy protection.[20]

- Repeated model training on synthetic data can cause "model collapse". This term describes a situation where performance of the model deteriorates after being repeatedly trained on synthetic data, causing the model to forget the statistical properties of the original real data. Researchers have demonstrated that this process can reduce the diversity of the model outputs.[21] Model collapse is particularly important in the context of large-language models: There is a risk that as the use of synthetic data replaces web-scraped data, an increasing number of models may be trained on synthetic data generated by other large-language models available on the web. Furthermore, this problem is not limited to text. Models trained on successive cycles of synthetic images has been shown to produce glitches and distorted images.[22] Although model collapse cannot be avoided when training on synthetic data alone,[23] when real data is used together with synthetic data it can be mitigated.[24] Therefore, to address this issue, it is recommended that models, including large-language models, also have access to real-world data,[25] and that synthetic data is continuously assessed for accuracy. Furthermore, implementing traceability and watermarking techniques could allow researchers to track and identify synthetic data, reducing the risk of over-reliance on this type of data resulting in model performance issues.
- Generating synthetic unstructured data, such as conversations or doctor's notes can present significant challenges. These include capturing the context of the data and ensuring it remains realistic. For example, conversations must reflect natural interactions, while doctor's notes need to accurately represent medical conditions and treatments. The lack of structure in this type of data can make it difficult to generate high-quality, reliable synthetic data sets that are useful for AI models.
- While synthetic data can be useful for developing AI models in some instances, it may not be suitable for all use cases, particularly when data from real individuals is required. For example, in healthcare, highly accurate data is needed to properly diagnose and treat patients. Even small inaccuracies can result in misdiagnoses or improper treatments, and in some circumstances, synthetic data may not sufficiently reflect the full range and complexity of real-world medical conditions.

b) Homomorphic Encryption

Homomorphic encryption enables encrypted computations to be performed on data without first having to decrypt them.[26] This avoids privacy risks associated with non-homomorphic encryption schemes where data must be decrypted before any computations can be performed on it. By keeping data hidden at all times, homomorphic encryption can be used in various ways when developing and deploying AI systems to protect privacy and security. Additionally, homomorphic encryption can be used to encrypt models themselves, ensuring that both the data and the models remain private during collaborative model training, model validation and model deployment.

Model Training

Homomorphic encryption consequently can be used to train models on encrypted data. This allows the data owner to encrypt the data before it is accessed by data scientists or developers and fed into the model. This means data scientists and developers never see the original or raw data, and there is no risk of unauthorized access or disclosure. Organizations shared with us that they use homomorphic encryption most commonly when model training needs to be outsourced to external infrastructure providers.

Furthermore, using homomorphic encryption during model training can also provide effective security against model inversion and membership inference attacks, and is therefore often used in combination with federated learning.[27] Such attacks aim to use information from the model outputs to learn either more information about individuals, or to determine whether information about an individual is present in the training data. Training the AI model on encrypted data protects against these attacks because attackers are unable to see the model outputs. Only the model owner has access to the decryption key and the model outputs. Consequently, attackers are not able to reverse-engineer or infer sensitive information from the model outputs.

Security

By leveraging homomorphic encryption, organizations can ensure data is secure at all times. Generally, the data is encrypted before computations are performed, and it remains encrypted during transmission, while stored on servers or in databases, and during the computation itself. As a result, sensitive information is protected throughout the entire process. This means that even in case of a data breach, the adversary would not be able to access personal data. Furthermore, fully homomorphic encryption is post-quantum, meaning it is resistant to attacks from quantum computing devices. The first post-quantum encryption algorithms were recently standardized by NIST and will begin to replace other encryption algorithms that are not as secure.[28]

Collaboration

Homomorphic encryption enables different parties to collaborate on training an AI model while their data remains hidden. Each party can encrypt their data before model training, ensuring each participant retains control over their data. Additionally, homomorphic encryption allows both the data and the workload to be encrypted, which is particularly powerful for collaboration purposes. Even the workload being executed, such as making inference requests, calculating feature importance, or performing data aggregations, can remain confidential, preventing any exposure of these operations to the participating parties.

Model Validation

Similarly to model training, homomorphic encryption can be used to keep data encrypted throughout AI model validation. The model will be able to perform computations directly on the encrypted test data, eliminating the need for direct access to this information. This enables data scientists and developers to securely evaluate model performance without accessing plaintext data.

Model Deployment

Where models have been deployed, homomorphic encryption can allow users to input their data into the model without the risk of exposing it to the model owner. The user's input data can be encrypted before it is processed by the model and remains encrypted until the user decrypts the final output. Consequently, the model owner only interacts with encrypted data and does not see the user's data.

CASE STUDY 7: Homomorphic encryption and federated learning for machine learning[29]

In traditional machine learning, participating entities are required to share their data with each other in order to train a combined machine-learning model. However, this involves privacy risks. Federated learning and homomorphic encryption can be used together to enable privacy-preserving collaborative machine learning.

In federated learning, each participant trains their own model and then sends model updates to inform a single global model. Although this means that no raw data is shared, attackers may be able to learn information about the data if they are able to intercept these model updates. As a result, homomorphic encryption can be used to encrypt the model updates before they are sent to the central server where the model updates from the different participants are aggregated and used to inform the global model. The global model is also returned to each participant encrypted for added protection in case it is intercepted. Therefore, by using homomorphic encryption with federated learning, participants can collaboratively share models with improved data privacy and security.

CASE STUDY 8: Homomorphic encryption and differential privacy for searching photo libraries[30]

Homomorphic encryption can allow users to search their photo libraries for landmarks and points of interest securely. To achieve this, the user's device privately queries a list of landmarks and points of interest to find approximate matches for places depicted in their photo library.

First, an on-device machine-learning model analyzes photos to identify regions of interest, potentially containing landmarks. If a region of interest is detected, the device creates a compact digital summary of that area, known as an embedding. Homomorphic encryption is then used to encrypt this embedding, send it to a server, and search how the embedding compares to its database of global landmarks without revealing the actual data. Furthermore, to ensure greater privacy, differential privacy is used to issue fake queries alongside the real query, and requests are routed through a secure relay that hides the user's identity.

When the server finds potential matches, it sends encrypted results back to the user's device. The device decrypts these results and uses another on-device model to select the best match. Once a match is identified, the photo is tagged with the landmark's name, allowing the user to easily search for it later. This combination of homomorphic encryption and differential privacy allows users to search their photo libraries accurately, without compromising privacy.

Considerations for use of Homomorphic Encryption

Homomorphic encryption also comes with certain challenges:

- The most powerful type of homomorphic encryption, fully homomorphic encryption, is computationally intensive, making some deployments expensive and time-consuming. However, with technological progress, computational power is increasing while decreasing in cost.
- Homomorphic encryption requires specialized knowledge and expertise to implement. To support wider use, leading organizations are making the technology more accessible. For example, Google's open-source general-purpose compiler for fully homomorphic encryption enables developers to write code and transform it into a form that can run on encrypted data.[31]

c) Differential Privacy

Differential privacy is a technical solution where random "noise" (often represented by the Greek character epsilon, Ɛ) is added to data to preserve privacy while potentially reducing the trade-off with data accuracy.[32] The purpose of differential privacy is to alter the data in a way that prevents the identification of any individual's data.

Model Training

By adding noise to the training data, differential privacy can reduce the risk that the training process reveals individual-level sensitive information. Differential privacy therefore enables data scientists and developers to work with training data without directly accessing or handling sensitive information.

Moreover, where differential privacy is used for model training, it also offers strong protection against both model inversion and membership inference attacks.[33] By adding sufficient noise to query responses or model outputs, the contribution of individuals' data will be hidden. This prevents adversaries from extracting valuable information from the model outputs and from being able to determine whether an individual's data is present in the data set used for training. Organizations shared that this was one of their most frequent applications of differential privacy in AI.

Security

By adding noise to individual data points, differential privacy also helps address security challenges. It ensures that these data points cannot be accurately determined, thus protecting the privacy of individuals. Differential privacy also reduces security risks during data collection, transmission, and storage as in case of a data breach the adversary will not be able to identify any personal data directly.

Collaboration

Differential privacy allows for different parties to collaborate on training a model without directly disclosing their data. Each party adds noise to its own data before sharing. This means no party's data is revealed and it also hides the contributions of individual parties to the final model, providing a secure approach to model training.

Model Validation

By adding noise to test data during model validation, differential privacy guarantees that individual data points remain protected during the evaluation of AI models. This approach allows data scientists and developers to assess model performance without the need to directly interact with sensitive data.

Model Deployment

Differential privacy prevents model owners from accessing any personal or sensitive information users may need to input where it is obfuscated by the addition of data noise. This gives users security that their data remains hidden. The noise will also ensure that individual data points do not overly influence the model's parameters. Therefore, neither the model owner nor adversaries will be able to learn the information about users from the model parameters.

CASE STUDY 9: Differential privacy in large-language models to create privacy-preserving synthetic text[34]

In order to create synthetic text with a formal privacy guarantee, researchers have demonstrated that by using differential privacy, they can generate synthetic data sets that at the same time ensure individuals in the source data cannot be identified. For example, differential privacy was used in the process of fine-tuning a large-language model where noise was injected into the model's updates during training, greatly reducing the risk of privacy leakage.

CASE STUDY 10: Differential privacy and secure multi-party computation for machine learning for choosing key photos to display to users[35]

To display a users' photos which may be significant to them on-device, machine learning with differential privacy is used to learn about important people, events and places based on the user's photo library, and to select the key photo based on popularity across all users.

Firstly, when a user takes a photo it is annotated using a model running locally on the user's device, which assigns common categories such as sky, person or recreation to the photo. If the user enables the improvement feature on their device and precise location for photos, a random location-category pair is selected, and random noise is added. This noise provides differential privacy assurance, protecting the user's privacy throughout the process. This output is then split into two shares, where each share on its own has no meaning. The shares are then encrypted independently and uploaded to a server.

At the server-level, both shares are separately decrypted and aggregated using corresponding shares from other devices. This technique, called secure aggregation, is a form of secure multi-party computation. Both aggregates are finally combined, allowing the model to learn the most popular location and category pairs and select the key photo to display to users.

CASE STUDY 11: Differential privacy and synthetic data for creating training data for on-device safe content classification[36]

Generating differentially private synthetic data can help provide data that resembles real-world data but at the same time is artificial and offers a mathematical guarantee that personal data is protected. Differentially private synthetic data is used to train a classifier which monitors the output of a large-language model used on devices to ensure it is appropriate for users. This is important as the decisions of the safety classifier must not reveal information about the users' data which were included in the classifier's training data set.

To create the differentially private synthetic data for the classifier, a large-language model is trained on internal data. This model is then fine-tuned using differential private data to protect the privacy of the users to whom the data belongs. Since differential privacy introduces noise, fewer parameters are trained to reduce the added noise, ultimately leading to greater accuracy and improving the quality of the synthetic data. The fine-tuned large-language model is finally used to create a synthetic data set that resembles the sensitive data, which is used to train the safety classifier.

Considerations for use of Differential Privacy

Implementing differential privacy involves several potential challenges that must be carefully considered:

- There is no one-size-fits-all approach on how much noise should be added.[37] This is a case-by-case decision,[38] with the sensitivity of the data and balancing accuracy and privacy in the specific context as determining factors.
- Differential privacy can be sensitive to outliers. Outliers can cause developers to add more noise than they otherwise would have to protect individuals' privacy. This can reduce the accuracy of the results for the majority of the data points that are not outliers.
- By design, differential privacy works by sacrificing accuracy, but excessive noise can degrade data quality, leading to less accurate AI models. Researchers are continually developing new methods to improve the accuracy-privacy trade-off. For example, for machine-learning models, the amount of noise applied to each attribute could be dependent on the feature's importance and data type.[39] IBM's differential privacy library allows users to explore the impact of differential privacy on machine-learning accuracy, using classification and clustering models.[40]

- Differential privacy requires expertise to implement correctly. For example, deciding the appropriate value for epsilon can be difficult, as explained above. Initiatives such as OpenDP are helping address this challenge. This open-source project develops tools and algorithms for differential privacy that are ready for use, as well as offering user guides and video tutorials for beginners. It also provides methods of analysis for the researchers who study the data.[41]

d) Federated Learning and Federated Analysis

Federated learning is a technique that enables different parties to train a shared machine-learning model without sharing their data.[42] The significance of federated learning is that, in contrast to traditional machine-learning model training, data is neither collected nor stored in one location. Federated learning enhances privacy because the raw data is never shared or moved.

Federated analysis is a method where AI models or analytics are applied to data distributed across multiple devices or platforms without the need to centralize the data. Instead of gathering and moving the data to a central server for processing, the analysis itself is brought to the locations where the data resides. The data remains on the distributed devices or platforms, and only aggregated insights or analysis results are shared centrally. This method enables the collection of insights from a wide array of distributed data sets without compromising data privacy.

Model Training and Collaboration

Federated learning is designed to facilitate model training on different parties' data while maintaining privacy. Each party trains the model in its own environment using its own data, then sends model updates to inform a single global model. Secure aggregation, a secure multi-party computation technique, is used to encrypt and send these updates to the global model and aggregate the encrypted updates to ensure that an individual party's update cannot be seen before aggregation. This is important as model parameters might otherwise leak information about the processed data. Federated learning therefore enables multi-party model training without sharing raw data. Organizations shared with CIPL that federated learning is frequently used in situations where data transfers are not allowed or convenient, for example due to data sovereignty laws or the size of the data.

Security

Federated learning and federated analysis enhance security by eliminating the need for centralized data storage, which is a common target for cyberattacks. In traditional machine-learning systems, sensitive data is often stored in large, centralized servers or data warehouses, making it a prime point of failure. Centralized systems are vulnerable to data breaches, hacking, and unauthorized access, with all data residing in one location. By contrast, during federated learning and federated analysis the data stays on local devices, reducing the exposure risk.

Model Deployment

Federated analysis, on the other hand, can play a significant role in the context of model deployment. It can help to reduce the need for large-scale data transfers, which can be resource-intensive and introduce privacy and security risks. By using federated analysis, AI models process data locally on user devices instead of sending data to a central server for analysis. Only the processed results or outputs, rather than the raw data itself, are transmitted to a central location. Unlike federated learning, which involves the training of a model across distributed devices, federated analysis focuses on local processing and sharing insights.

CASE STUDY 12: Federated learning and trusted execution environment for training a cancer detection model using data from multiple sources[43]

In order to develop an accurate cancer detection model, large amounts of data are required to be collected from multiple hospitals and medical centers. However, pathology images are highly sensitive and cannot necessarily be freely shared without introducing privacy and security challenges.

Through the use of federated learning and trusted execution environments in combination, healthcare institutions are able to securely share their data to jointly train a model. Federated learning allows model training to take place locally, with model updates aggregated and sent to a global model. A trusted execution environment is used to provide further security by encrypting model updates during transmission, thus preventing unauthorized parties from intercepting and accessing the data.

CASE STUDY 13: Federated learning, synthetic data and differential privacy for the collaborative training of machine-learning models[44]

Many organizations do not have the resources to train their own model, do not have sufficient data, or do not have sufficiently diverse data to develop a high-quality model. Collaborative approaches can help address these issues and facilitate data sharing for the training of machine-learning models. To ensure this is done in a privacy-preserving way, a combination of federated learning, synthetic data and differential privacy is used.

In this approach, participants use federated learning to train their local model on synthetic data generated from their original data. The model parameters are then sent to a central aggregation server which orchestrates the federated learning rounds. Differential privacy is additionally deployed by adding noise to the model parameters, before they are sent from participants to the central server for aggregation. This additional step protects against potential data leakage, as the parameters might otherwise reflect certain characteristics of the data. Noise is also added after the averaging of model parameters, further protecting against attacks and preventing adversaries from learning information from the model. The use of these three PETs together ultimately enables the development of a single machine-learning model, trained on meaningful synthetic data generated from raw data of different organisations.

CASE STUDY 14: Federated learning and synthetic data for training a fraud detection model[45]

To build an accurate fraud detection model, multiple life insurance organizations use federated learning and synthetic data to increase training data while complying with privacy regulations. Each organization generates synthetic data using their own real-world data. This synthetic data is then used to train a centralized model based on federated learning. By using these PETs together, an effective model can be trained on the data of multiple organisations while protecting privacy and confidentiality. This is particularly important here, due to the sensitivity of the data in this context, such as medical details, age and financial data.

Considerations for use of Federated Learning and Federated Analysis

There are a number of challenges associated with federated learning and federated analysis that must be taken into account before deployment:

- The frequent communication of model updates between devices and the central server means communication overhead can be significant.
- Removing the influence of a party on the central model when they leave the federation remains a nascent technique (i.e., machine unlearning).[46]
- Researchers have demonstrated that by comparing the differences between a model before and after updates, information can be revealed about changes in the training data.[47] Model parameters could also be intercepted during transmission and can be used to learn information about training data. Using other PETs, such as differential privacy or homomorphic encryption, in conjunction with federated learning, can bolster privacy and security.

e) Trusted Execution Environments

Trusted execution environments are a secure and isolated area within a computing system providing a platform for running code and accessing data in a protected way.[48] Applications running outside the trusted execution environment cannot access data within it, but applications running inside the trusted execution environment can access the data outside of it.

Model Training and Collaboration

Trusted execution environments enable secure model training by allowing multiple parties to collaborate on model training without exposing their data. In this setup, each participant encrypts their data and sends it to the trusted execution environment, where model training occurs in a protected environment. The trusted execution environment ensures that no data is leaked or shared between parties during the process. This allows organizations to retain full control over their data while benefiting from collaborative training.

Security

Having collected data for the development of an AI model, organizations become custodians of large volumes of potentially valuable information. As a result, it is imperative that these large data sets are stored securely. Trusted execution environments can help address data security concerns by providing isolated environments for storing and accessing data securely.

Data is encrypted within trusted execution environments and only specified users and code have access to this data, thus preventing unauthorized parties from accessing or tampering with it. Trusted execution environments also provide attestation mechanisms, which allow external parties to verify the integrity of the trusted execution environment and the software running within it. This adds another layer of security by ensuring that only trusted code is executed on the stored data. While trusted execution environments provide strong security, they do not offer the same level of protection as post-quantum secure homomorphic encryption, but they are still considered very secure for many practical applications.

Model Validation

Trusted execution environments also facilitate model testing on confidential data by ensuring that the testing occurs within an isolated, secure environment. This process means that the model is tested on the raw data but only the aggregated validation results (such as performance metrics or error rates) are shared, removing the need to reveal any raw data used during the model validation process.

Model Deployment

In the context of model deployment, trusted execution environments offer a secure environment where models can be run without exposing data. When deployed in this environment, models can process data while ensuring that it never leaves the protected enclave and protecting it from unauthorized access. This not only protects the data, but also ensures that only authorized entities are able to interact with the model, providing an extra layer of control.

CASE STUDY 15: Trusted execution environment for developing tailored risk score models[49]

To create tailored risk score models for clients, such as banks, data providers (organizations which supply data or pre-trained models) need to access and train their model using the bank's data. This requires the bank to share its financial data or the data provider to share its model. However, both of these are sensitive.

By leveraging a trusted execution environment, the data provider is able to train their model with the bank's data without compromising privacy or intellectual property. The trusted execution environment keeps data encrypted and hidden throughout the entire training process, therefore allowing model training without exposing the sensitive model or the bank's data.

CASE STUDY 16: Trusted execution environment for private cloud computing for generative AI[50]

For users to enjoy the benefits of generative AI on their devices, the devices need to be able to communicate in a secure and privacy-preserving way with the generative AI models stored in the cloud. To enable this, a type of trusted execution environment called a secure enclave is used to protect the relevant code and decryption keys.

When a user wishes to use generative AI, their device makes a request that is sent to the model stored in the cloud. This request includes the user prompts, the model to be used and the inference parameters. This request is encrypted on the user's device before being sent to the cloud, protecting the data in transit. This end-to-end encryption means that user data sent to the cloud is not made available to anyone except the user – not even staff with administrative access to the cloud infrastructure.

In the cloud, the code is loaded by the secure enclave to ensure it is not tampered with. The secure enclave is also used to securely store the decryption keys used to decrypt the request. As a result, the secure enclave plays an important role in protecting user privacy and ensuring only the correct code is executed and is granted access to the data.[51]

CASE STUDY 17: Trusted execution environment for protecting privacy in AI model training and user prompts[52]

AI models and their weights can be considered sensitive information. They must be protected in use to prevent against model tampering and unauthorized access of training data or other details on how the model was developed. Trusted execution environments can be used to help protect models as they prevent unauthorized entities from accessing the model and changing its behavior.

User prompts also often include sensitive data. By leveraging a trusted execution environment, these prompts will not be visible to anyone besides the user as only the trusted execution environment will have access to it. Additionally, this data remains encrypted to protect against data leakage.

Considerations for use of Trusted Execution Environments

Trusted execution environments also come with certain challenges that must be considered:

- Trust is placed in the manufacturer of the trusted execution environment and the cloud service provider or the specific computer system, rather than a mathematical formula that has guarantees. Attestation mechanisms can help prove the security of a trusted execution environment by confirming that the code is executing inside the secure environment of the PET.
- There are no formal standards that describe what a trusted execution environment is, how different trusted execution environments should interact with each other or the best attestation mechanisms. However, the Confidential Computing Consortium has started to bring together hardware vendors, cloud providers and software developers to develop and drive adoption of solutions and standards.[53]

f) Secure Multi-Party Computation

Secure multi-party computation provides a solution to allow multiple parties to compute on their combined data, without either party revealing any information about their input data.[54] It does this using encryption and secret sharing. Each party's data is encrypted or divided into different shares and distributed among the other parties. When split into shares, the data are no longer comprehensible unless combined with other, original elements. Where encryption is used, the computation is performed on the encrypted data of all parties before the final output is jointly decrypted. For secret sharing, each party computes on their shares and distributes the results to the other parties to help reach their target answer. By allowing different parties to collaborate, secure multi-party computation offers parties the ability to share data securely when developing AI systems.

Model Training and Collaboration

Secure multi-party computation enables multi-party collaboration for AI model training while protecting privacy. Each party encrypts or splits their data into shares and distributes it among the other parties. The parties then update the model parameters by performing computations on their encrypted or secret shares, without directly sharing the raw data with each other. After performing these computations, the parties share the results of their computations with each other. These results are aggregated to compute the combined update to the model parameters. The combined update is then applied to the global model parameters, completing model training without any party revealing its raw data.

Security

By using secure multi-party computation, organizations introduce another layer of cryptographic security to their data. A single share in a secure multi-party computation setup is indecipherable, protecting the underlying data from other participating parties or malicious actors. Secure multi-party computation technology is currently also capable of withstanding quantum attacks.[55]

Model Validation

Model validation can cause privacy risks where organizations wish to test models from different vendors. The organization may not wish to share their test data nor the vendor their model data. Secure multi-party computation offers a practical solution to this. In this setup, the different parties engage in an interactive protocol where they exchange encrypted values. The organization encrypts their input data and securely transmits this to the vendor. The vendor then inputs this encrypted data into the model to generate an encrypted output, which is sent back to the organization for decryption. Throughout this process, neither party learns sensitive information about the other's data beyond what is strictly necessary for computing the model's output, ensuring confidentiality for both the vendor's model weights and the organization's input data.[56]

CASE STUDY 18: Secure multi-party computation to measure brain activity[57]

Resting-state functional magnetic resonance imaging (rs-fMRI) measures brain activity by discovering changes in blood flow while the brain is at rest. This data can be used by machine-learning models to provide insights into brain function and diagnose neurological disorders. However, this data is highly sensitive as it can reveal information such as a patient's state of health, subconscious preferences (such as likes and dislikes), and even personality. As a result, secure multi-party computation has been used to protect the privacy of rs-fMRI data when performing machine-learning analysis.

In this application, the first party does not wish to share the weights of their machine-learning model while the second party wishes to protect their rs-fMRI data. The first party strips their model of its weights and shares the model architecture with the second party. This model architecture is then converted into two secure multi-party computation protocols, one for each of the two parties. These protocols will take the original machine-learning model and rs-fMRI data and through the exchange of encrypted pieces of data, output the result without the need for either party to share any of their confidential data.

Considerations for use of Secure Multi-Party Computation

Implementing secure multi-party computation may require addressing certain challenges:

- Secure multi-party computation can lead to high communication costs and, therefore, scalability issues. Data reduction techniques can reduce the size of the inputs or intermediate results that need to be communicated. For example, data compression algorithms can be used to compress the data before transmission to reduce communication costs.
- There is the risk of collusion. Secret sharing may make it possible for the input data to be reconstructed if some of the parties secretly communicate. To address this, auditing and accountability measures can help. Furthermore, homomorphic encryption can be used to ensure the data is kept hidden at all times.
- Implementing and deploying secure multi-party computation protocols correctly requires expertise that organizations may need to take additional steps to bring in-house or contract.

3. Privacy and Utility

PETs can play a critical role in safeguarding individual privacy during the development of AI systems. One continued point of tension when deploying any privacy protective means is often the balance between privacy protection and the continued utility of the data for the intended purpose. This is also true in the context of the AI life cycle.

While the importance of maintaining individual privacy rights cannot be overstated, we must also recognize that the quality and functionality of AI models hinges on their access to diverse, accurate and sometimes specific data. As organizations consider implementing PETs, it is vital to ensure that measures to protect privacy are also sufficiently balanced against data utility considerations. This requires a nuanced approach by all stakeholders, including regulators, where PETs facilitate, rather than impede, the development of accurate AI systems through data sharing.[58]

CASE STUDY 19: AI to anonymize data sets without sacrificing data utility[59]

One participating organization has developed an AI tool in-house to anonymize its data sets. AI is used to analyze data sets with large amounts of data input in order to first assess whether individuals are identifiable. During this assessment, the tool also examines other data sets held by the organization to ascertain whether individuals can be identified by combining sets. Where the AI determines that an individual is identifiable, it is set up to alter this data by replacing specific values rendering the individuals unidentifiable. This controlled addition of noise to the data ensures individuals cannot be re-identified, while the replacement values allow continued utility of the data (for statistical analysis for example).

4. Broader Advantages of PETs in AI

As demonstrated above, PETs offer a number of other advantages for use in AI, beyond protecting privacy or confidentiality:

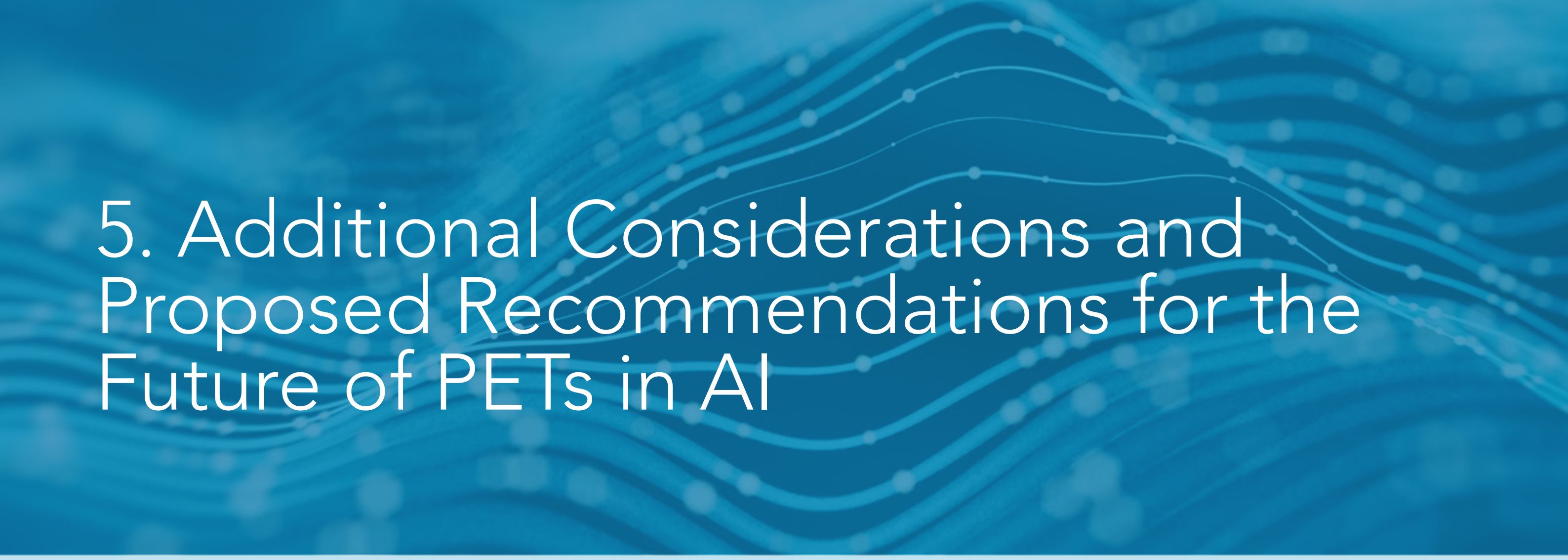
- **Improve data quality.**
  PETs can help generate high-quality, diverse and representative data sets that can be used to train models. For instance, synthetic data can be used to address a shortage of data, helping to address bias and leading to more fair and accurate outputs.

- **Enable collaboration.**
  PETs, such as secure-multi computation, enable multiple organizations to collaboratively develop AI models. This solution enables the training of models on data that they might otherwise not be able to access.

- **Establish trust.**
  By offering secure data sharing, PETs like homomorphic encryption allow data to be shared between parties while maintaining the integrity of the data used in AI systems.

- **Personalize services.**
  By allowing data to be collected and combined from multiple sources, PETs can allow organizations to personalize AI services for customers without compromising their privacy.

- **Enhance security.**
  Many PETs mitigate security risks, including trusted execution environments, by helping to prevent adversaries from accessing and tampering with data.



5. Additional Considerations and Proposed Recommendations for the Future of PETs in AI

PETs will undoubtedly play an increasingly central role in the development and deployment of accountable, privacy-friendly AI systems. High-quality data is essential to building accurate, fair, and effective AI systems, resulting in an ever-growing need to train models on the best available data. By enabling privacy-preserving data sharing, PETs remove barriers and facilitate the development of trusted AI systems. They are paving the way for innovative AI applications and unlocking new opportunities to realize the benefits of data and technology.

At the same time, there are potential challenges to the use of PETs in AI contexts that need to be considered carefully.[60] In addition to technical challenges we mention in respect of specific PETs above, in this section, we describe some of those broader challenges and offer recommendations to PET developers, deployers, regulators, and policymakers for addressing them.

- **Regulators should issue more clear and practical guidance to reduce regulatory uncertainty around the use of PETs in AI.**

  A number of regulators and government agencies have already recognized the particular utility of PETs in certain applications. For example, pseudonymization and encryption are acknowledged by the European Data Protection Board (EDPB) as effective methods for safeguarding personal data during transfers to other regions.[61] Similarly, the French Data Protection Agency has discussed the use of PETs such as synthetic data and differential privacy to help protect the security of AI systems, and describes the role of other PETs, including homomorphic encryption, secure multi-party computation, trusted execution environments and federated learning, in operationalizing the privacy by design principle.[62] The ICO also strongly supports PETs for a number of use cases and is clear that PETs can facilitate data protection by design and by default, and support data-protection principles, such as data minimization and data security.[63] Furthermore, NIST's guidance on generative AI recommends that organizations implement differential privacy to mitigate the risks of linking AI-generated content with individual human subjects.[64] It also suggests that organizations consider using synthetic data during the development of generative AI models to protect against the disclosure of personally identifiable information.

  While the guidance from these bodies is helpful, there remains a need for more clear and practical advice. For instance, the EDPB recommends that supervisory authorities evaluate the methods organizations use to reduce identifiability during AI model development, including the effective implementation of PETs.[65] However, it does not provide detailed guidance on what constitutes effective implementation or how PETs should be applied in practice. In contrast, the Personal Data Protection Commission (PDPC) in Singapore has issued a proposed guide on synthetic data, outlining key considerations and best practices in synthetic data generation for organizations to consider.[66] More comprehensive and practical support on the effective use of PETs would encourage even more organizations to implement these technologies. For example, regulators should avoid setting overly stringent standards for determining if a particular PET achieves compliance with a specific data-protection principle. Equally, they must be mindful of the trade-offs between data utility and data protection, understanding the technical limitations of PETs. Importantly, PETs should not be seen as a one-size-fits-all solution or a silver bullet for every data use case in AI.

- **Regulators should adopt a risk-based approach to assess how PETs can meet standards for data anonymization, providing clear guidance to eliminate uncertainty.**

  There is a broader question about whether various PETs could be used to meet legal standards for anonymization of data. In some circumstances and jurisdictions, there is a lack of clarity as to whether encrypted data could be considered sufficiently anonymized vis-à-vis the receiving party. For example, in Europe, Data Protection Authorities have traditionally taken a conservative interpretation on effective anonymization and have also assumed that encryption does not meet the standard of anonymization.[67] Indeed, according to the EDPB's recent draft guidelines on pseudonymization, data remains pseudonymous even if the additional information needed to attribute it to an individual is held securely elsewhere.[68] Meanwhile, the European General Court has ruled that in order to determine whether an individual is identifiable, account should be taken of all means reasonably likely to be used, and that this test must be performed from the perspective of the recipient/holder of the data.[69] This should mean that if the decryption key is inaccessible, then the data could be deemed anonymous vis-à-vis the recipient without access to the key. Regulatory clarity on this point and on the ability of PETs to anonymize data could be a powerful incentive for organisations to invest and use PETs more broadly. Regulators should take a risk-based approach to what the legal threshold for anonymization is, viewing it not as a reduction of the risk of re-identification to zero, but rather to a sufficiently low level, taking into account the context and purpose of the data processing.

- **Deployers should take steps to provide contextually appropriate transparency to customers and data subjects.**

  PETs add another layer to AI models and can be complex in their nature. This can make it difficult to interpret the behavior of AI models and explain their outputs to customers. For this reason, deployers should place emphasis on putting in place meaningful and contextually appropriate measures to help customers and data subjects understand how PETs work within their models.

- **Deployers should take care to ensure that clear mechanisms exist for data subjects to exercise their rights, where applicable.**

  PETs may modify, combine or obscure data to enhance privacy. This can create complications for fulfilment of data subject rights, such as the ability to access, remove, or correct data. However, when PETs such as differential privacy are used to successfully anonymize data, they can render certain data subject rights less relevant, as the data is no longer personally identifiable. Data subject requests must therefore be handled on a case-by-case basis, with the organization closest to the data subject responsible for informing the data subject about the use of their data and how the subjects can exercise their rights.

- **Policymakers and industry must work together to address the demand for large computing resources.**

  AI, especially generative AI, can demand collection, storage, and processing of large amounts of data. The application of PETs to these large data sets can significantly increase processing time. This problem may become less acute as technologies continue to progress and evolve. In the meantime, policymakers and industry will need to work together to ensure that computing resources are available to make the deployment of PETs at scale possible.

- **Regulators should incentivize proactive dialogue, further research, and experimentation with PETs within regulatory sandboxes.**

  PETs represent a rapidly evolving and dynamic field of research, with significant potential to advance data privacy and security. Encouraging collaboration within AI regulatory sandboxes – already established in jurisdictions such as Colombia,[70] Norway,[71] and Malaysia,[72] with others under development in Brazil[73] and Denmark[74] – would foster ongoing dialogue and mutual learning among regulators, researchers, developers, and deployers. In Singapore, the Infocomm Media Development Authority (IMDA) has operated a PETs-specific sandbox since 2022, and recently expanded its scope to include generative AI. Facilitating trusted exchanges of knowledge and best practices in this way would enable more effective and adaptable regulatory frameworks that keep pace with the development of PETs.

- **Stakeholders should adopt a holistic view of the benefits of PETs in AI.**

  While privacy and security are often the primary focus, it is crucial to recognize that PETs offer a range of additional advantages. These technologies can foster greater trust, enhance regulatory compliance, and open up new business opportunities by facilitating data sharing while protecting intellectual property rights and commercially sensitive data. It is essential that stakeholders do not overlook the diverse benefits of PETs in each case, ensuring that all advantages are duly considered when making decisions.

As discussed above, these challenges are addressable, and developers of PETs continue to work on ways to mitigate or overcome them. PETs are key to addressing privacy concerns across the various stages of AI development and deployment and may help satisfy regulatory requirements as they continue to enjoy broader adoption. However, the future success of PETs relies also on support and guidance from regulatory bodies, including privacy and AI authorities. These entities can create incentives and foster trust in PETs in ways that encourage integration of these technologies into organizations' AI and data governance frameworks.

6. Endnotes

[1] See, for example, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, European Data Protection Board. See also Information Commissioner's Office response to the consultation series on generative AI. See also CIPL paper, Applying Data Protection Principles to Generative AI: Practical Approaches for Organizations and Regulators. For more on the intersection of data protection, AI, and organizational accountability, see CIPL report, Building Accountable AI Programs: Mapping Emerging Best Practices to the CIPL Accountability Framework.

[2] We use these terms interchangeably in this paper.

[3] Emerging privacy-enhancing technologies: Current regulatory and policy approaches, Organisation for Economic Co-operation and Development (OECD), March 2023.

[4] For an overview of PETs, see CIPL's PETs paper, Privacy-Enhancing and Privacy-Preserving Technologies: Understanding the Role of PETs and PPTs in the Digital Age.

[5] In addition to the PETs discussed in this report, there are a range of tools and techniques that may be used to protect privacy in AI systems. Examples of these tools include the use of output filters in the context of generative AI or input controls for chatbots.

[6] For the purpose of this paper, we align with the OECD's definition of an AI system ("An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.")

[7] CIPL's PETs paper, *supra* note 4.

[8] Further papers will review the application of PETs in AdTech, FinTech and health-related research.

[9] The case studies in this paper were provided by participating organizations or drawn from publicly available sources. Organizations that participated in our research included BeiGene, Duality Technologies, Google, IBM, Mastercard, Meta, Nike, Telefónica and World Foundation.

[10] For a broader discussion of synthetic data, see page 41 in CIPL's PETs paper, *supra* note 4.

[11] Shuang Hao et al., *Synthetic Data in AI: Challenges, Applications, and Ethical Implications* (2024).

[12] The definition of sensitive data varies across different laws, but often includes personal characteristics such as race, religion, and sexual orientation.

[13] The technology behind InstructLab, a low-cost way to customize LLMs, IBM, June 7, 2024.

[14] Tiny but mighty: The Phi-3 small-language models with big potential.

[15] How generative AI helped train Amazon One to recognize your palm, Amazon, September 1, 2023.

[16] See more about how risk-based regulations can enable beneficial and safe uses of biometric technology in CIPL's Biometric Report.

[17] AlphaGeometry: An Olympiad-level AI system for geometry, Google, January 17, 2024.

[18] Introducing Llama 3.1: Our most capable models to date, Meta, July 23, 2024.

[19] Unlocking AI Training Data for All: MOSTLY AI Releases World's First Industry-Grade Open-Source Toolkit for Synthetic Data, MOSTLY AI, January 23, 2025.

[20] For how differential privacy can be used with synthetic data to mitigate data leakage, see Case Study 17 and Table 14 in CIPL's PETs paper, *supra* note 4.

[21] Yanzhu Guo et al., *The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text* (2024).

[22] Sina Alemohammad et al., *Self-Consuming Generative Models Go MAD* (2023).

[23] Mohamed El Amine Seddik et al., *How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse* (2024).

[24] *ibid.*

[25] Ilia Shumailov et al., *AI models collapse when trained on recursively generated data*, 631 Nature 755 (2024).

[26] For a broader discussion of homomorphic encryption, see page 25 in CIPL's PETs paper, *supra* note 4.

[27] Dimitris Stripelis et al., *A federated learning architecture for secure and private neuroimaging analysis*, 5(8) Patterns 1 (2024).

[28] Announcing Approval of Three Federal Information Processing Standards (FIPS) for Post-Quantum Cryptography, National Institute of Standards and Technology, August 13, 2024.

[29] Guangze Su et al., *The Utilization of Homomorphic Encryption Technology Grounded on Artificial Intelligence for Privacy Preservation*, 2 WEP 52 (2024).

[30] Combining Machine Learning and Homomorphic Encryption in the Apple Ecosystem, Apple, October 24, 2024.

[31] Our latest updates on Fully Homomorphic Encryption, Google, June 14, 2021.

[32] For a broader discussion of differential privacy, see page 38 in CIPL's PETs paper, *supra* note 4.

[33] Yanling Wang et al., *Differential privacy in deep learning: Privacy and beyond*, 148 FGCS 408 (2023).

[34] Xiang Yue et al., *Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe* (2023).

[35] Learning Iconic Scenes with Differential Privacy, Apple, July 21, 2023.

[36] Protecting users with differentially private synthetic training data, Google, May 16, 2024.

[37] For more on the trade-offs associated with noise levels, see CIPL's PETs paper, *supra* note 4.

[38] Privacy-enhancing technologies (PETs), Information Commissioner's Office (ICO), June 2023.

[39] Assem Utaliyeva, Jinmyeong Shin and Yoon-Ho Choi, *Task-Specific Adaptive Differential Privacy Method for Structured Data*, 23(4) Sensors 1980 (2023.

[40] Diffprivlib: The IBM Differential Privacy Library

[41] OpenDP.

[42] For a broader discussion of federated learning, see page 34 in CIPL's PETs paper, *supra* note 4.

[43] Secured Collaborative AI for Oncology Research, Duality Technologies.

[44] FedSyn: Synthetic Data Generation using Federated Learning, JP Morgan, April 6, 2022.

[45] Taiwan National Institute of Cyber Security, 1 October, 2024.

[46] Hyejun Jeong, Shiqing Ma and Amir Houmansadr, *SoK: Challenges and Opportunities in Federated Unlearning* (2024).

[47] Analyzing Information Leakage of Updates to Natural Language Models, Microsoft, November 2020.

48 For a broader discussion of trusted execution environments, see page 30 in CIPL's PETs paper, *supra* note 4.

49 Participating organization, 27 June, 2024.

50 Private Cloud Compute: A new frontier for AI privacy in the cloud, Apple, June 10, 2024.

51 To see how secure enclaves can also be used for privacy-preserving AI evaluations, see the pilot experiment conducted by OpenMined in partnership with the UK AI Safety Institute and Anthropic.

52 How Confidential Accelerators can boost AI workload security, Google, June 18, 2024.

53 Confidential Computing Consortium.

54 For a broader discussion of secure multi-party computation, see page 28 in CIPL's PETs paper, *supra* note 4.

55 Tapaswini Mohanty et al., *Quantum Secure Protocols for Multiparty Computations* (2024).

56 For more detail on how secure multi-party computation can be used for model validation, see Case Study 5 in CIPL's PETs paper, *supra* note 4.

57 Leverage Secure Multi Party Computation (SMPC) for machine learning inference in rs-fMRI datasets, Microsoft, February 14, 2024.

58 While this paper is focused on the application of PETs in AI, Case Study 18 demonstrates how AI can also be used as a PET to help balance the protection of privacy and utility.

59 Participating organization, 27 September, 2024.

60 For a more detailed analysis of the obstacles to the development and adoption of different PETs, and how to overcome these, see CIPL's PETs paper, *supra* note 4.

61 Recommendations 01/2020 on measures that supplement transfer tools to ensure compliance with the EU level of protection of personal data, European Data Protection Board, June 18, 2021. However, the current version of the EDPB's draft guidelines on pseudonymisation does not discuss PETs directly, see Guidelines 01/2025 on Pseudonymisation, European Data Protection Board, January 16, 2025.

62 AI how-to sheets, Commission Nationale de l'Informatique et des Libertés (CNIL), June 7, 2024.

63 ICO's PETs Guidance, *supra* footnote 38.

64 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, National Institute of Standards and Technology, July 26, 2024.

65 Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, European Data Protection Board, *supra* note 1.

66 Privacy Enhancing Technology (PET): Proposed Guide On Synthetic Data Generation, Personal Data Protection Commission Singapore and Agency for Science, Technology and Research, July 15, 2024.

67 Opinion 05/2014 on Anonymisation Techniques, Article 29 Data Protection Working Party.

68 Guidelines 01/2025 on Pseudonymisation, European Data Protection Board, *supra* note 61.

69 *Single Resolution Board v. European Data Protection Supervisor* (Case T-557/20).

70 Sandbox on privacy by design and by default in artificial intelligence projects, Superintendence of Industry and Commerce.

71 Regulatory privacy sandbox, Datatilsynet.

72 The Launching of Artificial Intelligence Sandbox Programme Together with Nvidia, National Technology & Innovation Sandbox, April 18, 2024.

73 ANPD publishes notice for partnership in Artificial Intelligence and Data Protection Sandbox project, Autoridade Nacional de Proteção de Dados, November 22, 2024.

74 New regulatory sandbox for AI, Datatilsynet, March 5, 2024.

About the Centre for Information Policy Leadership

The Centre for Information Policy Leadership (CIPL) is a global privacy and data policy think tank within the Hunton law firm that is financially supported by the firm, 85+ member companies that are leaders in key sectors of the global economy, and other private and public sector stakeholders through consulting and advisory projects. CIPL's mission is to engage in thought leadership and develop best practices for the responsible and beneficial use of data in the modern information age. CIPL's work facilitates constructive engagement between business leaders, data governance and security professionals, regulators, and policymakers around the world. Nothing in this document should be construed as representing the views of any individual CIPL member company or Hunton. This document is not designed to be and should not be taken as legal advice. For more information, please see CIPL's website at:

http://www.informationpolicycentre.com/

CIPL
Centre for Information Policy Leadership
HUNTON

DC Office
2200 Pennsylvania Avenue
Washington, DC 20037
+1 202 955 1563

London Office
30 St Mary Axe
London EC3A 8EP
+44 20 7220 5700

Brussels Office
Avenue des Arts 47-49
1000 Brussels
+32 2 643 58 00