Centre for Emerging Technology and Security

RESEARCH REPORT

RUSI

Privacy and Intelligence

Implications of Emerging Privacy Enhancing Technologies for UK Surveillance Policy

George Balston, Marion Oswald, Alexander Harris and Ardi Janjeva

July 2022

About CETaS and RUSI

The Centre for Emerging Technology and Security (CETaS) is a policy research centre based at The Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to inform UK security policy through evidence-based, interdisciplinary research on emerging technology issues. Connect with CETaS at cetas.turing.ac.uk.

The Royal United Services Institute (RUSI) is the world's oldest and the UK's leading defence and security think tank. Its mission is to inform, influence and enhance public debate on a safer and more stable world. RUSI is a research-led institute, producing independent, practical and innovative analysis to address today's complex challenges.

All views expressed in this report are those of the authors, and do not necessarily represent the views of The Alan Turing Institute, the Royal United Services Institute, or any other organisation.

Foreword

Privacy Enhancing Technologies (PETs) are amongst the most exciting applications of encryption and data science today. These technologies are maturing rapidly, driving a vibrant discussion around their potential use both in safeguarding personal privacy, and in supporting broader national security. This independent report by The Alan Turing Institute and the Royal United Services Institute forms a welcome and timely addition to that conversation, an example of the UK's leading role in the international scientific and ethical debates surrounding emerging technologies. I look forward to the review spurring further expert and public dialogue over the coming year.

Dr Paul Killworth, Deputy Chief Scientific Adviser for National Security

Acknowledgements

This report has been reviewed by expert readers before being published. The reviewers were not asked to endorse the conclusions of the report, but to provide independent scrutiny of its content. Reviewers acted in a personal and not a representative capacity. The authors gratefully acknowledge the contribution of the reviewers.

Reviewers

Ant Burke, The Alan Turing Institute

Dr Andrew Glazzard, Royal United Services Institute

Conrad Prince CB, Royal United Services Institute

James Sullivan, Royal United Services Institute

Expert readers who provided comments on the draft report

Dr Benjamin Curtis, The Alan Turing Institute

Dr Chris Hicks, The Alan Turing Institute

Dr Rachel Player, Royal Holloway University of London

The authors would also like to thank all those who contributed to the development of this project, particularly through participation in research interviews.

Executive Summary

This report aims to establish an independent evidence base to inform future government policy development and strategic thinking regarding national security uses of Privacy Enhancing Technologies (PETs). The findings are based on in-depth consultation with stakeholders across the UK Intelligence Community (UKIC), Investigatory Powers Commissioner's Office (IPCO) and academic experts.

The research has identified new opportunities for the UKIC and partner organisations to apply PETs in three main areas: data acquisition and information requests; secure machine learning; and non-operational data sharing. These opportunities possess high potential on the condition that the UKIC can establish clarity regarding the motivation for using a PET in each circumstance and the safeguards to be applied.

The research explored a range of techniques relating to PETs, analysing their potential value to the UKIC, and opportunities for increased privacy and security. Key findings were as follows.

Key Findings

PETs may provide a less intrusive means of carrying out intelligence work. While PETs could reduce the degree of privacy interference with individuals in terms of the acquisition and analysis of bulk datasets, it is unlikely that this would outweigh the need to acquire bulk data where necessary for accurate analysis, particularly involving multiple datasets. In each case **the motivation for using the PET must be clear** to inform the legal considerations governing its use.

PETs could encourage the sharing of knowledge and capabilities between the UKIC and external partners who maintain datasets of interest to the UKIC, particularly for training Machine Learning models. This would, nevertheless, require assurances regarding data labelling and model performance, to ensure the accuracy and reliability of any shared capabilities.

Specific use cases identified in the research include:

- **Private Information Retrieval** to enable the UKIC to request information from data holders, whilst better protecting its own sensitive information, and collect less data of limited or no intelligence interest. Particular value is identified in the use of **private set intersection** for checking the existence of UKIC selectors within external databases without the need to disclose any sensitive details.

- **Federated learning** to allow the UKIC to better develop data-reliant capabilities in collaboration with external partners, without the need to collect large amounts of data. **Homomorphic encryption** and **trusted execution environments** could enable greater sharing of operational capabilities, both with trusted allies and data holders, whilst protecting the sensitivities and robustness of the capabilities.

- **Differential privacy** and **synthetic data techniques** to enable better data sharing for non-operational purposes, such as undertaking research and development work with academia,

without the need to release secret or sensitive data. However, as these techniques are predominantly focussed on removing identifiable information from datasets, they present a limited degree of utility for any operational use cases.

The successful use of these PETs will be closely linked to the levels of cooperation from third-party data holders and the approach to dealing with questions of legal liability. Clarity will be necessary regarding PETs' compatibility with data protection legislation and the responsibilities incumbent on data holders given the lack of knowledge or control over the queries being run over their data and systems.

There is a clear risk to deploying PETs without having developed appropriate levels of trust. This includes trust within the UKIC in the reliability and accuracy of the PETs techniques being deployed, and amongst third party data holders where misunderstandings could arise regarding the motivations for using PETs and their functionality. **Future policy will need to account for these concerns,** to ensure that organisations retain appropriate knowledge and control over the analytical process deployed via the PET, thereby ensuring trust and accountability throughout the full analysis pipeline.

1. Introduction

1.1 Research Rationale

The Alan Turing Institute's Centre for Emerging Technology and Security (CETaS) partnered with the Royal United Services Institute (RUSI) to conduct an independent research study into the use of Privacy Enhancing Technologies (PETs) for national security purposes. The overall aim of the project is to establish an independent evidence base to inform future government policy development and strategic thinking regarding national security uses of PETs.

The field of PETs is evolving at pace, creating new opportunities to derive insights from datasets whilst helping to mitigate unacceptable privacy risks. Recent years have seen significant achievements in PETs research and development, but there remain few real-world applications, particularly in the UK public sector. In the coming years, uptake of PETs is likely to expand across the economy, as organisations are increasingly required to conduct large-scale analysis while protecting sensitive data.

1.2 Methodology

The project sought to address the following research questions:

i. What are the potential uses of PETs for national security and intelligence, and where could this technology provide the greatest benefit?

ii. What are the legal and policy implications of deploying these use cases in the national security context?

iii. What new policies, processes and guidance are needed to ensure the UK Intelligence Community can maximise the full potential of PETs?

iv. What are the legal, practical, and reputational implications for third party data controllers arising from the use of PETs in a national security context?

The project adopted a qualitative, practitioner-focused research design. Data was collected through a synthesis and consolidation of existing research in a targeted literature review, and semi-structured interviews with UKIC and IPCO personnel and a limited number of technical academic experts. 36 respondents participated in research interviews for the project: 26 from the UKIC, 3 from the IPCO, and 7 academic experts possessing a range of practitioner experience across government and the commercial sector. The broad affiliation across active practitioners in the national security community and experts with an outside perspective is something which – to the authors' knowledge – has not been replicated in other UK-based publications on this research topic. This participatory research approach enables insight to be gained from the experience of stakeholders in the assessment of PETs use-cases in real-world contexts, identification of potential issues and consideration of policy requirements.

Interviews were conducted between January 2021 and April 2021 on a strictly anonymous basis. Interviews were conducted in a semi-structured format, enabling the research team to adopt a broadly consistent line of questioning in each interview, but allowing for flexibility to probe specialised areas of knowledge and experience in respondents. The interviews focused upon i) practical application of PETs in respect of operational data analysis processes; ii) practical barriers and limitations; iii) policy and legal implications; iv) divergences between notions of privacy in the context of PETs and within the legal framework; and v) whether new polices, processes and guidance were required regarding the use of PETs for national security.

Throughout this report, an anonymised coding system is used to refer to interview data. The following prefixes are used to indicate the category of research participant to which the interview data refers:

A = Academic expert

IC = UKIC respondent

O = Oversight body

To preserve the anonymity of individuals and organisations, precise dates and locations of interviews are not referenced in this document. The views and responses expressed in this report reflect the opinions of these individuals and should not be interpreted to represent the official position of any government department, agency, or other organisation. The findings may not be generalisable to other agencies, either in the UK or overseas. Furthermore, the report does not refer to any classified tradecraft or capabilities that may have implications for the practicality and feasibility of PETs. It may

be the case that other classified information is relevant to the issues discussed in this report, but it is not possible to elaborate further within the parameters of the project. Finally, the report raises a number of legal issues and questions but should not be interpreted to constitute legal advice.

1.3 Definitions and Scope

There is ongoing debate about how a PET is defined and this was reflected in our research interviews. For this paper PETs are defined as a range of technologies designed to address and mitigate privacy risks through encryption, data minimisation, anonymisation, and pseudonymisation.

In 2016, the European Union Agency for Cybersecurity (ENISA) released a PET-specific technology readiness scale[1]:

1) *Idea* – lowest level of readiness.

2) *Research* – at least one academic paper has been published on the PET in question.

3) *Proof-of-concept* – the PET has been implemented and tested for certain properties, such as complexity.

4) *Pilot* – the PET has been used at a small scale with real users.

5) *Product* – the highest readiness level, where the PET is incorporated in 1+ products and used by a significant number of users.

6) *Outdated* – the PET is not used anymore because it has been superseded technically or is no longer required.

The majority of technologies discussed in this paper exist between proof-of-concept and product. Despite encouraging developments in the field, especially across the commercial sector, further technical research is required to develop practical applications appropriate for operational deployment.

These technologies vary widely in approach and capability. Possible applications of PETs include secure access to and analysis of sensitive datasets, the training of machine learning models across datasets owned by multiple data-owners, and the mitigation of the risk of privacy attacks. Selecting the 'right' PET (or combination of PETs) requires significant considerations around governance and policy, such as 'acceptable' privacy risk thresholds or safeguarding mechanisms. Broader challenges in implementation and deployment are also well-recognised. PETs can be computationally intensive, require significant technical expertise, and may only be relevant in certain contexts.

In the UK, surveillance policy has evolved to regulate the collection of datasets and restrict how particular collection, surveillance, and analysis capabilities (such as interception, equipment interference, and bulk powers) are authorised and deployed. Despite existing technical limitations,

[1] Hansen, Marit, Hoepman, Jaap-Henk, Jensen, Meiko. and Schiffner, Stefan. "Readiness analysis for the adoption and evolution of privacy enhancing technologies: methodology, pilot assessment, and continuity plan." *Technical report: ENISA* (2015).

the implementation of PETs could have significant operational, legal, and policy implications for the present landscape. In this context, two aspects of privacy were identified as particularly relevant: the minimisation of collateral intrusion into individuals of no intelligence interest; and the minimisation of disclosure of sensitive selectors or those conducting the queries.

This paper focuses on the following five PETs deemed to be of highest potential value to the UKIC. This section includes a brief overview of each technique, while the following sections include detailed illustrations of specific use cases.

1) Homomorphic encryption – a property of certain encryption schemes which allows for particular types of computation on encrypted data;

2) Secure multi-party computation – a set of protocols which enable computation and/or analysis on combined datasets, without individual parties revealing their inputs or data;

3) Trusted execution environment – a ringfenced processor within a computing system that enables the confidential execution of code, resulting in a secure and private code execution environment;

4) Differential privacy – a description of a system that permits the analysis of patterns and trends without releasing meaningful information about individuals within a dataset;

5) Synthetic data – the use of techniques to generate artificial data that mirrors the statistical properties of real data with the purpose of preserving privacy or augmenting datasets.

Homomorphic Encryption

Homomorphic encryption is a property of certain encryption schemes which allows for particular types of computation on encrypted data. It can be a useful mechanism for securing and analysing data in circumstances where all or part of the computational environment is not trusted.

Different implementations of homomorphic encryption allow for different levels of computation to be achieved. Fully homomorphic encryption (FHE) allows for the computation of unlimited additions and multiplications but is computationally complex and still practically inefficient. Partially Homomorphic Encryption (PHE) and Somewhat Homomorphic Encryption (SHE) are implementations which have a reduced computational complexity compared to FHE, but both have trade-offs – PHE allows for only a single type of operation, and SHE allows for only a limited combination of operations.

Practical deployments of homomorphic encryption are often still reliant on being heavily tailored to particular use cases and cannot easily accommodate a range of real-world tasks.

Deployments are limited due to the substantial inflation in data size following encryption, and the high computational costs associated with calculations on encrypted data. While progress is being made to boost efficiency across these areas, homomorphic encryption has remained limited to smaller and less ambitious applications. Additionally, significant technical expertise is required to tune, anticipate, and

manage the multiple parameters required for homomorphic encryption[2] – even small changes can result in significant reductions in performance, security, or privacy guarantees. However, the field has seen a rise in promising initiatives such as the Homomorphic Encryption Standardization[3] which aims to define a standard to assist in the setting of these parameters.

Secure Multi-Party Computation

Secure multi-party computation is a set of protocols which enable analysis to be carried out on multiple datasets without individual parties revealing their inputs.[4] Inputs are encrypted before being sent to other parties, who are able to process these inputs without needing to decrypt the values. Most secure multi-party systems will involve three main users: input parties; result parties and computing parties.[5]

Secure multi-party computation removes the need to share information with a central trusted authority.[6] The entity executing the computation can be replaced with several entities that perform computations in a distributed way, with each unaware of the nature of the input data and intermediate results.[7] Secure multi-party computation can also be used to securely distribute the training of machine learning models across many devices. By taking a potential single point of failure and distributing trust among numerous parties, any potential attacker is faced with a more complex multi-party attack surface.[8]

Secure multi-party computation involves computation on encrypted data, which gives rise to significant communication and computational requirements, rendering it impractical for most real-time computations. The lack of visibility across its inputs also raises challenges around the reliability and integrity of its outputs. Despite these challenges, secure multi-party computation is a useful mechanism in which sensitive data remains encrypted – therefore significantly reducing the chance of privacy attacks – and greater scope for collaboration between multiple data-owners and datasets is made possible.

Trusted Execution Environments

A trusted execution environment (also referred to as a 'secure enclave') is a ringfenced processor within a computing system that enables the confidential execution of code, resulting in a secure and

[2] Agrawal, Nitin, Reuben Binns, Max Van Kleek, Kim Laine, and Nigel Shadbolt. "Exploring design and governance challenges in the development of privacy-preserving computation." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-13. 2021.
[3] Further information: https://homomorphicencryption.org/introduction
[4] The Royal Society. "Protecting privacy in practice: the current use, developments, and limits of Privacy Enhancing Technologies in data analysis." p.13. 2019.
[5] Archer, David W., Dan Bogdanov, Yehuda Lindell, Liina Kamm, Kurt Nielsen, Jakob Illeborg Pagter, Nigel P. Smart, and Rebecca N. Wright. "From keys to databases – real-world applications of secure multi-party computation." *The Computer Journal* 61, no. 12 (2018), p. 1749-1771.
[6] The Royal Society. "Protecting privacy in practice: the current use, developments, and limits of Privacy Enhancing Technologies in data analysis." p.28. 2019.
[7] Pettai, Martin, and Peeter Laud. "Combining differential privacy and secure multiparty computation." In *Proceedings of the 31st Annual Computer Security Applications Conference*, pp. 421-430. 2015.
[8] Archer, David W., Dan Bogdanov, Yehuda Lindell, Liina Kamm, Kurt Nielsen, Jakob Illeborg Pagter, Nigel P. Smart, and Rebecca N. Wright. "From keys to databases – real-world applications of secure multi-party computation." *The Computer Journal* 61, no. 12 (2018): 1749-1771.

private code execution environment.[9] Data is operated on 'in-the-clear' before it is re-encrypted and sent back to the owner. This means that processes can run on that part of the processor without their memory or execution state being revealed to any other party, guaranteeing 'input privacy'.[10] Unlike fully homomorphic encryption or secure multi-party computation, trusted execution environments do not present a cryptographic solution, but rather rely on physical hardware segregation between processors.

One advantage of a trusted execution environment is that they are already commercially available. Microsoft Azure and IBM's cloud service both offer this capability, as do some Apple products. This means that if an Apple product were to be comprised by malware, for example, an attacker would still be unable to access contents of the trusted execution environment.[11] This makes trusted execution environments well suited for implementing biometric authentication methods, such as facial recognition, fingerprint sensors, and voice authorisation, in addition to securing financial and payment information.

A trusted execution environment is also relatively low cost – as computation is performed on data in unencrypted form, it requires no added noise. However, this can also form one of its main risks where adversaries may attempt to manipulate code running within an enclave so as to exfiltrate its secrets.[12] Another challenge is the lack of memory in a trusted execution environment, where only limited amounts of data can be processed at any one time. Additionally, the fact that the memory content of a trusted execution environment is inaccessible from the outside makes detecting if it has been compromised – for example, where an attacker has initiated a keylogger attack – more challenging.[13]

Trusted execution environments are often discussed in conjunction with other PETs rather than being a solution on their own – potential research combinations have included combining trusted execution environments with homomorphic encryption or secure multi-party machine learning.[14]

Differential Privacy

Differential privacy is a system that permits analysis of patterns and trends without releasing private or sensitive information about individuals within a dataset. Differential privacy was formally introduced in 2006 when Dwork and colleagues[15] proposed a mathematical definition – named epsilon (ε) – that mitigated the risk of revealing sensitive information via repeated queries. ε can be set as a limit – a 'privacy budget' – following which a user cannot perform any further queries on a database.

[9] The Royal Society. "Protecting privacy in practice: the current use, developments, and limits of Privacy Enhancing Technologies in data analysis." p.14. 2019.
[10] Big Data UN Working Group. "UN Handbook on Privacy-Preserving Computation Techniques." p.41. 2021.
[11] Hoffman, Chris. "Your smartphone has a special security chip. Here's how it works." *How-to Geek* (2018).
[12] Ning, Zhenyu, Fengwei Zhang, Weisong Shi, and Weidong Shi. "Position paper: Challenges towards securing hardware-assisted execution environments." In *Proceedings of the Hardware and Architectural Support for Security and Privacy*, pp. 1-8. 2017.
[13] Ibid.
[14] The Royal Society. "Protecting privacy in practice: the current use, developments, and limits of Privacy Enhancing Technologies in data analysis." p.38. 2019.
[15] Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis." In *Theory of Cryptography Conference*, pp. 265-284. Springer, Berlin, Heidelberg, 2006.

Assuming ε is set correctly, it offers a provable privacy protection from inferences or reidentification by adversaries.

The privacy budget can be set by a central trusted authority which can add 'noise' to the data, either at the point of collection or information disclosure. This method works by altering some of the responses in the dataset following each query which preserves a level of privacy. However, increasing the amount of noise (and therefore privacy) reduces accuracy – referred to as the 'privacy-utility' trade-off. As such, the technique is better suited for large datasets where the addition of noise has less overall impact. While practical guidance on how ε should be most efficiently set is currently limited, encouraging developments are underway.[16]

The utility of differential privacy in the analysis of data is dependent on both the sensitivity of the data and the amount that is intended to be released. To optimally benefit from the use of differential privacy, a data owner should be aware of the type of queries that might be requested in order to set the privacy budget appropriately.

Despite these challenges, the strength of differential privacy is in its resilience to a range of potential data misuses. Unlike other techniques, differential privacy requires no modification as new forms of attack and vulnerabilities are discovered. Differential privacy is a promising technique in both mitigating the risk of privacy attacks and in assisting those entrusted with safeguarding sensitive data.

Synthetic Data

Synthetic data is data that is artificially generated to mirror the statistical properties of real data. If there is a good understanding or model of the real data, the generated data should maintain all the valuable properties of the real data whilst excluding any real information.

A primary benefit of using synthetic data includes enabling data-sharing across departments or organisations. As synthetic data contains little in the form of identifying features, testing and validation can be undertaken in less secure environments, which lowers a main obstacle to collaboration. Synthetic data can have particular utility in generating samples around rare events, or in improving imbalances within datasets, contributing to the development of predictive models.

Algorithms for generating synthetic data continue to mature and allow for the preservation of complex and multivariate relationships within data. Recent success with this approach was demonstrated with the Financial Conduct Authority's Digital Sandbox Pilot[17], where synthetic data was provided to participants for training and testing. The dataset included details of seven million fictional individuals, 400 million banking transactions made via five fictional banks, and five million devices used for electronic payments.[18]

Synthetic data has several limitations. Despite advances in the field, constructing accurate synthetic data remains extremely difficult. In some use cases, a rough degree of accuracy may be acceptable

[16] The Alan Turing Institute project, 'DiffPriv', hopes to address this challenge. Further information: https://www.turing.ac.uk/research/research-projects/diffpriv-enhancing-privacy-secure-computation
[17] Further information: https://www.digitalsandboxpilot.co.uk
[18] Further information: https://www.turing.ac.uk/research/impact-stories/supporting-innovation-fintech-sector

(i.e., general trends and insights can still be extracted[19]), but this is unlikely to suffice for particularly complex data analysis. In addition, before synthesis, a data generation model must be created to maintain the meaningful statistical components of a dataset. After synthesis, any analysis will be constrained to the statistical components which have initially been identified and accounted for within the data generation model. This may reduce the usefulness of synthetic data in scenarios which require exploratory analysis.[20]

[19] Further information: https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data

[20] Stadler, Theresa, Bristena Oprisanu, and Carmela Troncoso. "Synthetic data–anonymisation groundhog day." *arXiv preprint arXiv:2011.07018* (2021).

2. Exploration of Utility

Three main areas of utility were identified for the application of PETs in the UK national security context, within which specific scenarios for use were also explored. The three main areas of utility are introduced in this section, while the following section describes specific potential use cases.

2.1 Data Acquisition and Information Requests

National security agencies routinely acquire data and make requests to external organisations to help answer investigative questions in pursuance of their statutory functions. The extent to which PETs could improve these processes emerged as a key theme when considering opportunities for the use of PETs within the UKIC.

Simple requests for information may involve working with trusted intermediaries within external organisations to answer questions such as 'what is this individual's current address?', or 'does this individual hold an account with your organisation?'. More complex scenarios involve combining data from many different sources to answer more detailed questions such as 'does this individual possess an extremist mindset?', or 'where has this individual been over the past 24 hours?'.

PETs could provide utility in answering both simple and complex analytical questions whilst disclosing a reduced amount of sensitive data, and in certain circumstances diminish the requirement for the acquisition of Bulk Personal Datasets (BPD), discussed further in the following section.[21]

The scenarios discussed below involve different uses of Private Information Retrieval (PIR), a term for a suite of techniques that allow one party to query a second party's data, without the second party seeing either the query, or the results of the query. PIR may make use of homomorphic encryption, secure multiparty computation, or a mixture of both technologies.[22]

Simple Information Requests

One simple form of PIR is known as Private Set Intersection (PSI). PSI is a secure multiparty computation technique which allows two parties to check whether they hold any items in common, without sharing information about any non-matching items.

For example, using PSI, two hospitals in different regions could check whether any patients were registered in both regions, without revealing the identity of their patient records. Each hospital would only learn about anyone who was registered at both hospitals, and not reveal any other information. If there were no patients with records at both hospitals, each hospital would solely learn that no patients were registered at both hospitals.

Within a national security context, agencies regularly make simple requests for information about an individual or a selector against data held by other agencies, departments, or companies. Often, on the

[21] O1.
[22] Chor, Benny, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. "Private information retrieval." In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 41-50. IEEE, 1995.

balance of investigative priority, details on subjects of interest can be shared directly with trusted intermediaries within these other organisations.[23]

The use of PSI in these scenarios offers a more discreet alternative: rather than sharing the details of a specific individual with a number of external organisations and asking whether they hold any more information on the individual, an agency could first check the existence of a match before sharing any information. A specific PSI use case is explored in detail in section 3.1.

Complex Information Requests

More complex investigative questions require more sophisticated queries. Information requests to external organisations can be complex, for instance referencing many different columns within a database. More complex queries can result in an increase in the amount of sensitive information that is shared.[24]

Using advanced PIR techniques, whole databases can be encrypted in such a way that an external party can directly query the database, and receive answers to the query, without the database owner having the ability to interpret the query. This typically involves processing data whilst in its encrypted state using homomorphic encryption, although similar results can be achieved using trusted execution environments.

In a national security context, PIR could enable complex queries to be run on datasets held by data providers, and the results of such queries to be collected, without the requirement to collect whole databases. As such, an agency could increase the complexity of information requests made to external partners, while reducing the risk associated with the sensitivity of the request.

As an example, a transport company may hold a database of travel information associated with specific travel cards. Using PIR, and with the company's agreement, an agency could ask complex queries of this data, such as 'which individuals travelled between X station and Y station at 12:04 on 5 October 2021', without exposing this query to the company.

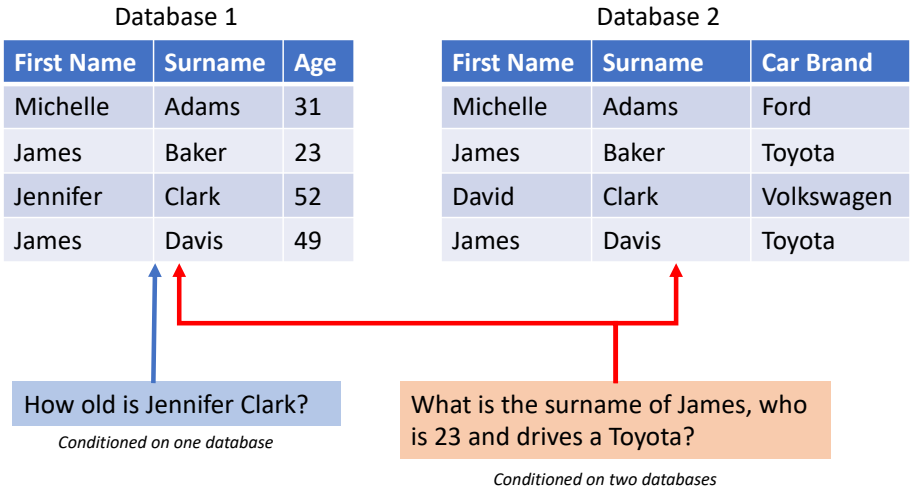
Bulk Personal Data

The UKIC has powers to collect and retain Bulk Personal Datasets (BPD) when it is necessary and proportionate to do so to achieve their statutory functions. The benefits resulting from the collection and analysis of BPD are twofold – firstly, sensitive queries can be protected from third parties, and secondly analysts can find links between separate datasets or undertake analysis requiring access to multiple datasets. PIR, as discussed in the previous section, can enable the UKIC to encrypt queries which are sent to data holders, whilst still obtaining answers to them. However, the utility of PIR on BPD will greatly vary depending on the extent to which datasets are combined in the data analysis process.

[23] IC2, IC20, IC22.
[24] IC20.

For investigative questions that are conditioned on queries from individual datasets, PIR could reduce or even remove the need to acquire these BPD. Instead, encrypted queries could be sent as filters to organisations which provide BPD to the UKIC, and filtered subsets of the BPD could be returned.

However, for investigative questions that are conditioned on queries from more than one dataset, the suitability of PIR diminishes. Interviewees noted that such queries are common.[25] For queries which rely on joining or combining more than one BPD, the complexity of a system designed to enable PIR greatly increases. In these circumstances, acquiring these datasets remains the best option to ensure reliable and timely results to queries.



Intelligence Community Data Sharing

Aside from external uses, PSI could also be used internally to improve confidentiality regarding sensitive subjects of interest (SOIs). The identities of sensitive SOIs within individual teams in the UKIC are very closely guarded, protected by the 'need-to-know' principle. Risks can arise when one sensitive SOI may be subject to some form of interaction with multiple teams within the UKIC.

If two departments are either investigating or cooperating with the same individuals, the use of PSI provides an opportunity to securely share this highly sensitive information, enabling teams to deconflict. One team could check whether their SOI appeared as an SOI in another investigation. This could minimise risks to SOI confidentiality, and risks of multiple engagements with the same SOI.

As well as deconflicting SOIs, PSI could be used to check whether international partners were investigating the same individuals, without disclosing information from either side as to which individuals were of concern.[26]

[25] IC3, IC6, IC24.
[26] IC3.

2.2 Secure Machine Learning

For the most complex analytical tasks, machine learning can be used as part of the data analysis process. Machine learning describes a set of techniques that allow computers to learn from experience, without being explicitly programmed.[27]

Machine learning is rapidly entering a diverse array of real-world applications in areas such as healthcare, finance, and national security. Machine learning models can be used to automate complex or effort intensive tasks. Data is required within machine learning systems for two purposes – model training, where algorithms run over large volumes of data to train models to undertake certain tasks; and inference, where machine learning models are applied to data to produce a set of output inferences. Using automatic speech recognition as an example, training data would comprise thousands of hours of speech with associated human-generated transcriptions to teach the model how speech corresponds to text; inference would then involve running new speech data through the model to generate an output transcript.[28]

While there have been great practical developments in both cryptography and machine learning, these have largely been independent of each other. However, recent research has focussed on applying privacy techniques to machine learning, both for the purposes of training models requiring large amounts of data, and for securely making inferences from data.[29]

Inference

The UKIC trains machine learning models on data they collect and uses these and other models as part of the analytical process. These models are typically tailored for specific tasks. Due to the high costs of developing machine learning models, sharing models can be beneficial, whether between government departments, or between the UKIC and external organisations.

However, sharing models gives rise to security risks in relation to both the performance of the model, and the data used to train the model. Machine learning models inherit information from the datasets that they are trained on. Whilst the training data itself does not form a direct part of models trained on such data, it has been shown that models can leak important information about the training data. This is commonly referred to as model inversion.[30] Further, if adversaries gain access to models, they may also benefit from a knowledge of the outputs of these models, which provide information on the inferences the model is making. This knowledge can lead to the development of adversarial inputs – the subtle manipulation of input data to trick the model into highly confident but incorrect inferences.

[27] Samuel, Arthur L. "Some studies in machine learning using the game of checkers." *IBM Journal of Research and Development* 11, no. 6 (1967): 601-617.
[28] IC26, IC28.
[29] Al-Rubaie, Mohammad, and J. Morris Chang. "Privacy-preserving machine learning: threats and solutions." *IEEE Security & Privacy* 17, no. 2 (2019): 49-58.
[30] Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322-1333. 2015.

This poses a challenge to sharing models, from the risks of both the exposure of classified training data, and the manipulation of inputs to create false outputs.[31]

PETs can help to address these challenges. One technique which can address model inversion is to apply differential privacy to the training data, adding sufficient noise to the data to maintain a specific privacy budget. This comes at the cost of performance – as more noise is added, performance will degrade, and the outputs of the model become less reliable. Further, differentially private models provide no extra protection against adversarial attacks. An alternative and more comprehensive approach is the use of homomorphic encryption. In enabling the processing of data in its encrypted form, homomorphically encrypted machine learning models provide protection against both training data leakage and manipulation of input data, whilst benefitting from a similar level of reliability as unencrypted models.

A separate hardware-based approach is the use of trusted execution environments (TEEs). Like homomorphic encryption, TEEs allow for secure computation on data. However, as discussed in section 1.3, TEEs achieve this through physical measures rather than in software. Models could be moved to TEEs, which would also provide protection against model inversion and the manipulation of data, without loss of accuracy or reliability.

Encrypted models could be shared with data holders for them to run over data they have in their possession. The outputs of these encrypted models, which could be described as encrypted inferences, could be passed back to the model owner for decryption. In this instance, the data on which the inferences were made, held by the external partner, would not need to be transferred outside of their systems. In essence, this provides a way for one organisation to learn something about the data held at another organisation, without ever seeing their data. An example of this is explored in more detail in use case 3.2.[32]

There remain significant technical challenges to implementing homomorphic encryption systems for model sharing. They are computationally complex and can require large amounts of processing on the systems which hold data, which would usually be owned by external partners. Implementations can also still be reliant on high trust models – the data holder must typically encrypt their data using the model trainer's public key to enable the encrypted model to function. This provides a serious flaw in any untrusted implementation – the model owner has the capacity to decrypt the raw data if they obtain access to it. Research is ongoing to address issues around both performance and trust models.[33]

Federated Learning

As well as providing the ability to secure models which have already been trained, PETs can assist in protecting the data that models are trained on.

Machine learning models are trained by algorithms which learn through a process of iteration. In supervised machine learning, a model is first exposed to some training data, which are examples of input data and a desired output. For example, in speech recognition, the input data might be an audio

[31] IC1, IC28, O1.
[32] IC1, IC14.
[33] Foltz, Kevin, and William R. Simpson. *ERP Homomorphic Encryption Performance Evaluation*. Institute for Defense Analyses., 2019.

clip of some speech, and the desired output might be a textual transcript of the speech. Through a lengthy process of iteration, the model starts to learn how to translate speech as an input to text as an output. Successful machine learning techniques typically require large amounts of data to achieve high levels of accuracy. This gives rise to a common challenge – how to acquire large amounts of high-quality data with which to train machine learning models, particularly when the data may contain sensitive characteristics such as identifying information.

Federated learning, a form of secure multiparty computation, provides an alternative solution to the training of machine learning models. Instead of acquiring data, then training models on the data, federated learning techniques enable models to be sent to the data, updated with new data, and then returned. However, this results in a limited oversight of training data used in models, giving rise to risks relating to data quality and auditability. These techniques are used by companies such as Google and Apple to update the machine learning used in their voice assistants and word recommendation systems, without needing to extract the underlying data from users' devices.[34]

Differential privacy techniques can be used to further protect any training information in federated learning scenarios. It is possible that private information can be gleaned by comparing models before and after federated learning. To help prevent the leakage of information, noise can be added during learning which can provide security regarding the data used and ensure privacy regarding the individuals within such data. However, as previously discussed, adding noise to any model will result in some degradation to performance.[35]

2.3 Non-Operational Data Sharing

The UKIC regularly collaborates with other parties to research and develop capabilities, including the academic community, the commercial sector, and partner agencies. There are numerous challenges to collaborating on national security research, and chief amongst these are the high sensitivities that UKIC capabilities and data hold. Classified data cannot be shared with individuals without appropriate levels of security clearance. With the growth in data science and machine learning techniques, the specific inability to share data with research partners can result in research being less targeted and less effective in relation to UKIC challenges.[36]

As discussed in section 1.3, differential privacy techniques can provide a level of assurance that, for specific analysis of a dataset, the output of that analysis will be indistinguishable with respect to the removal of any one individual. This essentially provides a guarantee of confidentiality for any individuals within a dataset for a specific analysis of that dataset, providing technical assurance that this data cannot be used for any other purpose.

Alongside differential privacy, synthetic data techniques can also provide a similar result. Synthetic data, as discussed in section 1.3, is data that is created artificially to match as closely as possible to an

[34] McMahan, Brendan, and Daniel Ramage. 2017. "Federated Learning: Collaborative Machine Learning Without Centralized Training Data". *Google AI Blog*; Hao, Karen. "How Apple personalizes Siri without hoovering up your data." *Technology Review* (2020).
[35] A4.
[36] A2, IC30.

exemplar dataset. Using data science techniques, real data can be analysed to derive a data generation model which allows for the generation of data with similar statistical properties to the real data. This synthetically generated data can then be used for analysis, but if designed correctly will contain none of the sensitivity of real data. Synthetic data and differential privacy can also be combined to provide better assurance for data which has been generated synthetically.

Using either or both techniques, the UKIC could engineer data which holds similar statistical properties to real data, but none of the data sensitivities. This would not be beneficial in all cases, as often sensitivities can stem from data collection mechanisms and capabilities as well as the data itself. Nevertheless, for scenarios involving data collected through well documented collection mechanisms, differential privacy and synthetic data techniques could allow for typically sensitive data to be shared outside of the UKIC for research purposes. However, the trade-off for added data privacy is a degradation in the utility of the dataset for analysis, and as such the accuracy of any research outputs.

3. Example Use Cases

This section describes three exemplar use cases that demonstrate potential scenarios in which PETs could be used, as well as possible implementations of these use cases. A discussion of the legal and policy issues relating to the use of PETs follows in section 4.

3.1 Private Set Intersection for Information Requests

A common Private Set Intersection (PSI) protocol is based on the RSA cryptosystem, a public-key system that is widely used for secure data transmission.[37] The PSI protocol assumes two parties are collaborating to allow one party (the 'query provider') to query a separate party's data (the 'data holder') to find any common elements. The below method is based upon a Google tool allowing for the detection of compromised online credentials.[38]

This use case considers a situation where an intelligence agency wants to check whether three unique individuals under investigation hold an account with a specific high street bank (HSB). The agency does not want to disclose sensitive details forming part of an investigation, so uses a PSI technique to answer this question without divulging any information.

Both parties have their own data sets – one in the form of a query, and one in the form of a customer database.

Agency

Query

David Clark
Greg Evans
Beth Frank

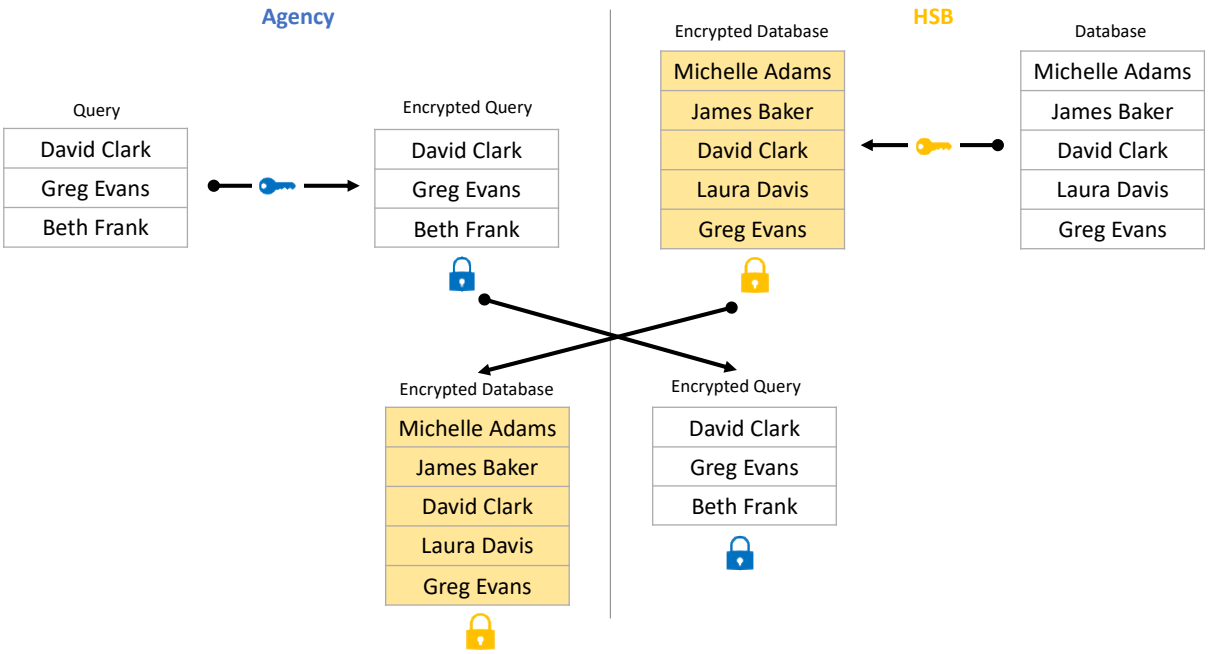
High Street Bank (HSB)

Database

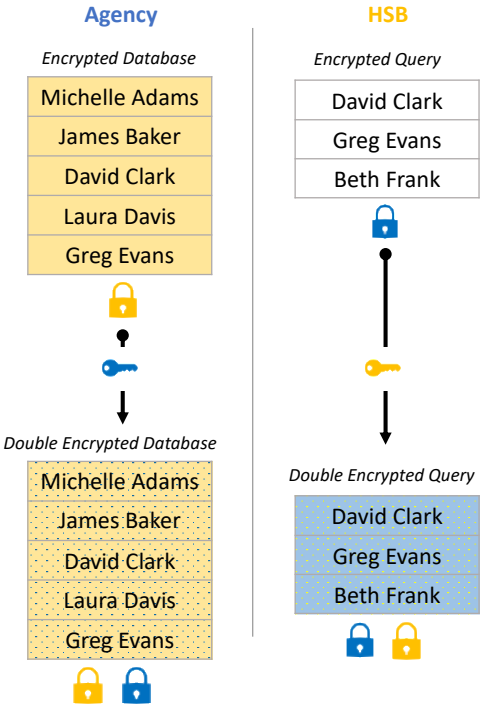
Michelle Adams
James Baker
David Clark
Laura Davis
Greg Evans

Both parties then encrypt their data with private keys. This means no other party without their private key can decrypt or read their data. The two parties then exchange encrypted data. Since neither party holds the other's private key, no interpretable data is shared.

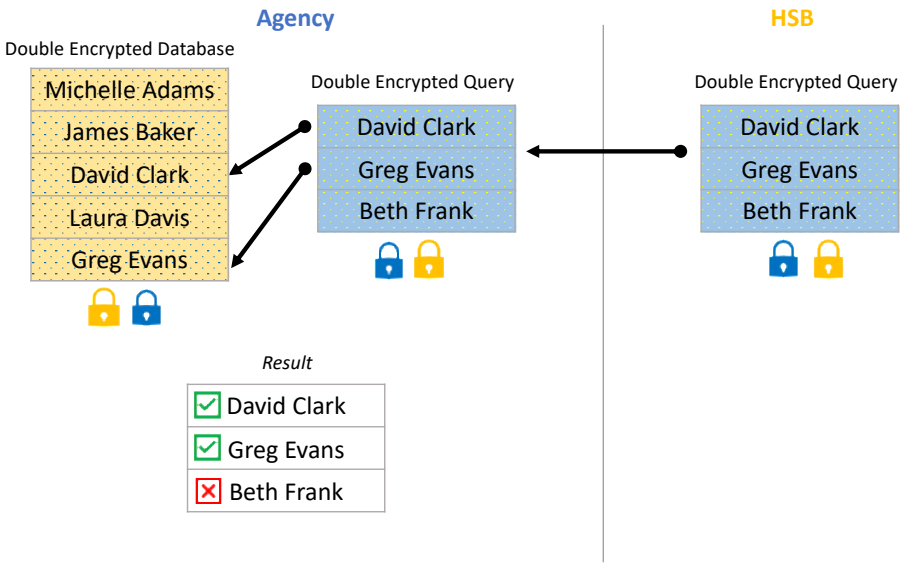
[37] Paar, Christof, and Jan Pelzl. "The RSA cryptosystem." In *Understanding Cryptography*, pp. 173-204. Springer, Berlin, Heidelberg, 2010.
[38] Walker, Amanda, Sarvar Patel, and Moti Yung. 2019. "Helping organizations do more without collecting more data". *Google Security Blog.*



Once in possession of each other's encrypted data, both parties encrypt the once-encrypted data again with their own private keys, resulting in double encrypted data.



The use of a specific protocol means that whilst in their double encrypted state, each party's data can be compared to find identical elements. The HSB passes its double encrypted data back to the agency, who can identify the matching elements, and obtain an answer to their query.



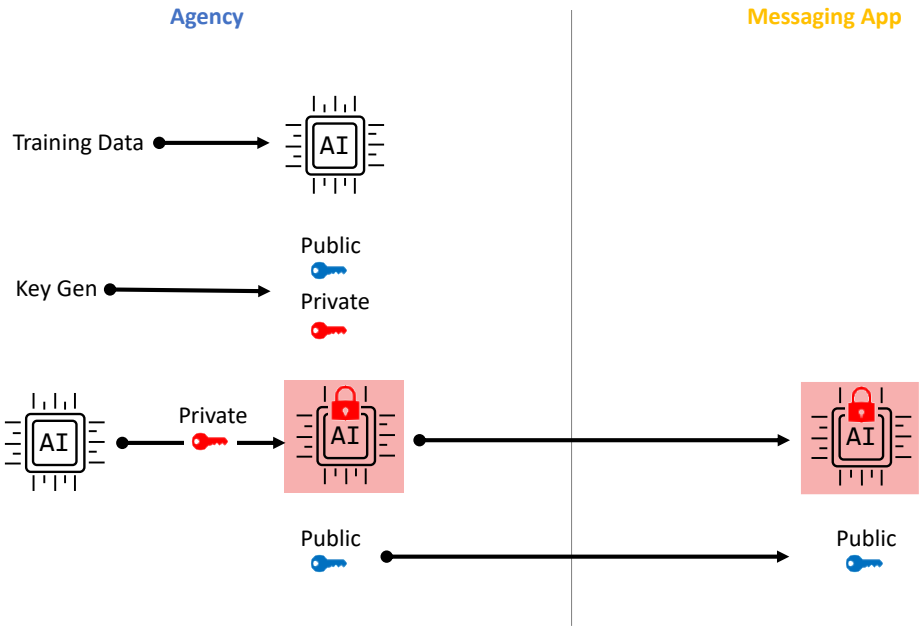
The agency has now learned that the HSB contains records for two out of the three individuals under investigation. Due to the encryption protocols used, the HSB has not learned any information about those under investigation by the agency.

3.2 Model Sharing with Third-Party Data Holder

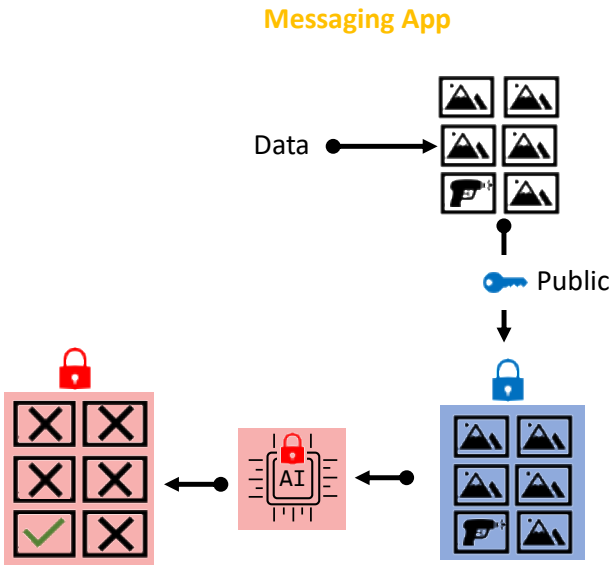
Within a national security context, agencies may use predictive models to make inferences from data, such as recognising certain objects within images. PETs such as trusted execution environments and homomorphic encryption can enable these models to be encrypted and distributed to third parties who can run the models on data that they hold. The third parties could pass back the inferences the models made, without having to share the data on which the inferences were made. This could enable agencies to ask questions of certain data sets, and receive answers to these questions, without having to collect the data, and whilst protecting the nature of the questions from third parties. However, it also gives rise to issues of accountability and auditability, in that the agency would not have oversight of exactly what data had been analysed.

This use case considers a situation where an intelligence agency has a predictive model for detecting images of weapons. It wants to send this to a company which administers a messaging app to run over the pictures shared across their platform. As the model itself contains sensitive elements at risk of being reverse engineered, the agency does not want to share the model directly. The agency and the messaging app agree to collaborate and use homomorphic encryption to enable the model to securely run over the messaging app's data.

The agency trains an AI model which can accurately detect images containing weaponry. They also generate a public/private key pair. Using their private key, the model is encrypted, and both the public key and the encrypted model are shared with the messaging app.

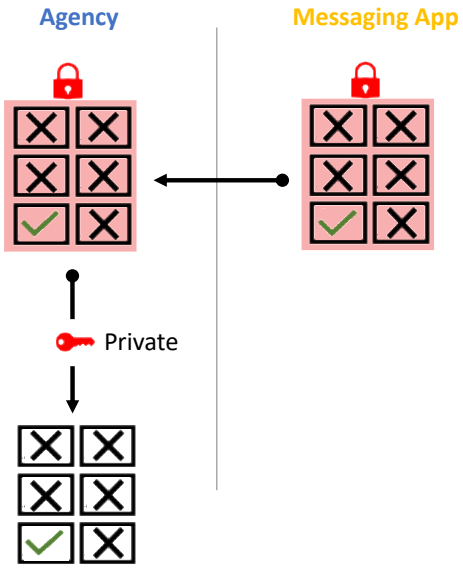


The messaging app collects images that are shared on its platform. These images are then encrypted with respect to the agency's public key and run through the encrypted model. The model's outputs are an encrypted version of the inferences of the model – for each image, this would correspond to either a 'contains weapons' flag or a 'does not contain weapons' flag. As the outputs are encrypted, the messaging app cannot interpret the outputs of the model. In this case, there are six images, one of which contains a weapon.



The messaging app then sends the encrypted inferences – the outputs of the model – back to the agency. Using their private key, the agency can decrypt the inferences and learn that one image does

contain a weapon. Importantly, the agency does not collect any of the original data – only the inferences made about that data by their secure model.



The agency has now learnt that one of the images shared on the messaging app contains a weapon. Due to the encryption protocol used, the messaging app has not learnt anything about the agency's model or what is does. Importantly, the messaging app has not had to share any data directly with the agency – only inferences made about that data. However, this use case raises major questions with respect to accountability, auditability, and the potential outsourcing of the agencies' statutory functions. These questions are discussed in section 4.

3.3 Differentially Private Synthetic Data for Research

Differentially Private Synthetic Data can be used to understand trends within data and generate new data which fits these trends for research purposes, without compromising the privacy of individuals within datasets.

This use case considers a situation where an intelligence agency wants to release some data to assist in the analysis of the geographic spread of their subjects of interest, to inform resource allocation and investigative strategies. In doing so, they want to ensure in their analysis that the privacy of their subjects of interest is maintained. It is based on an implementation described by NIST.[39]

Through electronic systems holding bulk personal data (for example, electoral roll data), the agency has access to a dataset containing addresses for the individuals that are of intelligence interest. The

[39] Further information: https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data

agency decides to undertake a differentially private analysis in order to generate synthetic data to share with researchers to help understand the geographic spread of the individuals, whilst retaining privacy for the individuals.

From the original data, a histogram is constructed by counting the number of individuals who live in each district. A level of granularity on a geographical level is chosen such that it is specific enough to be meaningful, but broad enough that each individual district has enough data to maintain the privacy guarantee. In this example, for illustrative purposes, only three locations are considered.

Name	Location
Michelle Adams	London
James Baker	Manchester
David Clark	London
Laura Davis	Birmingham
Greg Evans	Birmingham
…	

Original Data

Location	Count
London	263
Manchester	117
Birmingham	94
Total	*474*

Histogram

Next, the agency adds noise to the histogram to satisfy differential privacy. The amount of noise is calculated with respect to the level of geographic granularity to ensure differential privacy is guaranteed. Finally, each noisy count is divided by the total to determine what percentage of all individuals lived in each specific area. The result of this is a "one-way marginal distribution" – only one attribute of original data is considered, and any correlations between attributes are ignored.

Location	Count
London	263
Manchester	117
Birmingham	94
Total	*474*

Histogram

Location	Count
London	270.5
Manchester	109.2
Birmingham	99.8
Total	*479.5*

Noisy Histogram

Location	Noisy %
London	56.4%
Manchester	22.8%
Birmingham	20.8%

1-Way Marginal

This creates a distribution that can be used for the generation of differentially private synthetic data via sampling that could be shared. Unlike a simple histogram and distribution, this data could be shared safely with third parties, with a statistical guarantee that no individual details could be interpreted.

Using a simple implementation like this has some shortcomings – it will not preserve correlations between data attributes. For instance, it may be of interest to factor in age when considering location, to see where particular age ranges are clustered. Age may also be highly correlated with location – one might expect to see, for example, younger people clustering towards big cities such as London. To preserve correlations, a two-way marginal distribution could be calculated over both age and location. However, this would have more 'buckets' (i.e., all of the combinations between age and location).

Location	Age	Count
London	18-34	211
London	35+	52
Birmingham	18-34	68
Birmingham	35+	48
Manchester	18-34	81
Manchester	35+	13
Total		*474*

Histogram

Location	Age	Count
London	18-34	208.8
London	35+	56.1
Birmingham	18-34	63.7
Birmingham	35+	51.8
Manchester	18-34	80.1
Manchester	35+	14.3
Total		*474.8*

Noisy Histogram

Location	Age	Noisy %
London	18-34	44.0%
London	35+	11.8%
Birmingham	18-34	13.4%
Birmingham	35+	10.9%
Manchester	18-34	16.9%
Manchester	35+	3.0%

2-Way Marginal

The result of this is that differential privacy may be harder to guarantee with smaller sample sizes for certain histogram buckets. However, as an extra feature is maintained, a more useful and accurate analysis could take place over the resulting data. The preservation of correlations within data therefore requires a trade-off between the level of detail in observations and the utility of the data for analysis. This demonstrates a challenge with the use of synthetic data – knowledge of the desired analysis must be known ahead of time to ensure important statistical properties are maintained.

4. Discussion and Analysis

This section discusses the benefits of the PETs described above, along with potential challenges to implementation including considerations from a legal and policy perspective.

4.1 Less Intrusive Means

Using PIR in suitable circumstances may reduce or remove the requirement to acquire bulk datasets. Analysts can currently provide information to data holders to help them return more relevant and targeted data. However, this process is not widely scalable, as information that would aid data holders in filtering their data before sharing with the UKIC is often classified.

Sharing unencrypted queries often involves the sharing of personal details and identifying records such as phone numbers or addresses, giving rise to collateral intrusion when parties outside of the UKIC gain visibility to these details. This is a suboptimal process for the UKIC, as the classified data is no longer as tightly controlled as it is within UKIC systems. Whilst the risk in any one instance is minor, the aggregated risk will increase incrementally as more information is shared, and more parties are entrusted with confidential information.[40]

PIR can enable analysts to request information in such a way that the query itself is encrypted, meaning that the data holder can no longer see any information regarding the queries. This method may also reduce the requirement for the agencies to acquire bulk datasets containing the data of those of no intelligence interest. Further, the aggregated risks associated with many small releases of classified data is greatly reduced.

For a specific BPD which is predominantly or solely queried in isolation – that is, not conditioned on other acquired or internal databases – the ability to protect sensitive information by encrypting queries can challenge the justifications for acquiring that BPD. However, for BPD which generate value through combination with other datasets, the usefulness of PIR as a technique diminishes, and the justification for acquisition remains.

To understand the implications of this in legal and policy terms, it is first important to understand the centrality of the Investigatory Powers Act 2016 (IPA) in authorising intelligence activities by warrant. This includes the acquisition and analysis of bulk datasets where the majority of individuals within such datasets are not of intelligence interest.

In laying out the agencies' general duty in relation to privacy, s2(2) IPA states that the agency must have regard to 'whether what is sought to be achieved by the warrant, authorisation or notice could reasonably be achieved by other less intrusive means' (reflecting Article 8 of the European Convention on Human Rights, ECHR). As some interviewees pointed out, this requirement may suggest an obligation to consider whether the use of PETs could provide a less intrusive means of data acquisition and/or analysis:

[40] IC25.

'Anything which further mitigates and reduces the extent of the intrusion into those people who are not of interest… must all be to the good in terms of justifying and making clear that it's necessary and proportionate.'[41]

Is Use of a PET a Less Intrusive Means?

Focusing on bulk powers specifically, the recent ECtHR decision in *Big Brother Watch and Others v U.K.*[42] stated that the degree of interference with individuals' Article 8 rights will increase as the bulk interception process progresses (para 330) from acquisition to eventual analysis by an individual analyst. This emphasised the need for end-to-end safeguards:

'In order to minimise the risk of the bulk interception power being abused, the Court considers that the process must be subject to 'end-to-end safeguards', meaning that, at the domestic level, an assessment should be made at each stage of the process of the necessity and proportionality of the measures being taken; that bulk interception should be subject to independent authorisation at the outset, when the object and scope of the operation are being defined; and that the operation should be subject to supervision and independent ex post facto review' (para 350).

Interviewees recognised the importance of this big-picture view, with one commenting:

'There is a real risk that we look at this too much as a stage-by-stage process, we need to look at processes as a whole and ask what is the least privacy-intrusive [means] in terms of the [overall] objective.'[43]

The use of PETs could reduce the degree of privacy interference with individuals in terms of the acquisition and analysis of bulk datasets, while still enabling analysis to be carried out effectively. Bearing in mind the stipulations in s2(2) of the IPA, interviewees questioned whether they would be under a legal obligation or legal compulsion to deploy the technology on this basis. As public awareness of PETs increases, one might anticipate a shift towards expecting national security organisations to articulate why carrying out a particular intelligence activity using PETs would not be more appropriate than the status quo, if the proposed benefits to intrusion are deemed to be significant enough.

Machine Intrusion

Furthermore, it is worth noting that the extent of intrusion resulting from machine analysis remains a matter of debate, and it should not be assumed that analysis by way of PETs is automatically less intrusive than analysis by a human operator.

As has been noted in the context of artificial intelligence, 'Use of AI could result in additional material being processed which may not have previously been possible for technical or capacity-related reasons. This would need to be considered when assessing proportionality of any potential intrusion,

[41] IC1.
[42] Application nos. 58170/13, 62322/14 and 24969/15.
[43] IC20.

balanced against the increase in effectiveness of analysis that may result', an issue that is equally applicable to the use of PETs.[44] Interviewees also recognised that *'PETs might allow agencies to do more analysis than before, increase the level of data that can be obtained from BPDs'*, and questioned what legal or policy issues this would raise.[45]

To this end, one emphasised that:

'This needs to be an additional modus operandi rather than a replacement, there is a risk that retaining it (bulk data) is seen as bad, because there are alternatives that are less intrusive but less effective… one method should not be seen as better in all circumstances.'[46]

Acquisition and analysis of data via PETs will continue to require justification as being necessary and proportionate and falling within the agencies' statutory functions and general powers to use and disclose information. As one interviewee pointed out:

'Part of the justification of us bringing BPDs onto our system is precisely because one cannot interrogate them where they are in a safe way, because it gives away the question which one is asking. To the extent that there are techniques that enable one to do that interrogation without bringing it onto those systems, one can avoid the intrusion associated with doing that. So that is a technique that will be of assistance in terms of limiting the extent to which one might acquire things.'[47]

Motivation for Use of PETs

The legal assessment of each PET will therefore be highly dependent on the context in which it will be used, and the harm that it is intended to minimize. Interviewees highlighted both the potential for minimising collateral intrusion, and concern to minimise disclosure of sensitive selectors/search parameters that would reveal sensitive national security information such as the details of individuals who were of interest to the agencies (thus contributing to compliance with the agencies' information disclosure and security obligations). Although inter-related, these are two different aims, and it will be important to clarify in each case the motivation for using the PET, to inform the legal considerations governing its use.

4.2 Sharing and Development of Capabilities

The use of PETs could help the UKIC and its allies develop new capabilities and strengthen existing ones, through enabling the ability to better share and develop new data focussed techniques. As such, one interviewee for this research framed PETs as 'partnership enhancing technologies':

[44] Babuta, Alexander, Marion Oswald, and Ardi Janjeva. "Artificial Intelligence and UK National Security: Policy Considerations." (2020).
[45] IC2.
[46] IC6.
[47] IC1.

'The value of PETs really comes from forming partnerships, for example, if companies want to share attack vectors that they have seen in the wild. Their utility lies in bringing someone or something that has data to share that they would not usually want to share or combine.'[48]

Much of the discussion of PETs has hitherto focused on what the UKIC can share with third parties, but this perspective outlines the potential benefits to the UKIC of a more reciprocal approach. While the above comment refers to 'companies', the logic may similarly apply to intelligence agencies and other government or third sector bodies, domestically and internationally.

Sharing models with international partners could aid in the collaborative development of capabilities, particularly in areas where training data is sparse and hard to collect. The UKIC could benefit from the use of federated learning, collaborating with external partners to enable models to be trained on data without the requirement to acquire such data. This might include international allies, social media companies, or other organisations which maintain curated datasets of interest to the UKIC. Particularly for areas where data is hard to store or collect such as child sexual exploitation material or proscribed terrorist material, the use of federated learning can remove the requirement to share data in order to train models.

Whilst there are a number of potential benefits, relying on external data for the training of models poses some challenges.

When training machine learning models, the accuracy of training data is paramount, often represented in the phrase 'garbage in, garbage out'. In the federated learning scenario, the UKIC would have to seek assurance that the external data was labelled with sufficient accuracy to improve rather than degrade models, highlighting the issues of accountability, audit and provenance that were raised during the research interviews. As a minimum, the performance of models would need to be checked periodically to mitigate against these risks. Additionally, the UKIC would be reliant on the data holders using their own computational infrastructure to update the models, creating an extra cost for the data holder.

4.3 Accuracy of Analytical Functions and Data Providers

The nature of interaction with third party data holders presents an important challenge to the feasibility of using PETs for national security purposes.

The ability to encrypt machine learning models whilst retaining the ability to make inferences opens new possibilities for sharing models with third parties. Encrypted models mitigate against risks to both security and confidentiality. The inability of external parties to understand how models were trained, and the data they were trained on, decreases the likelihood of bad actors using this information to craft adversarial examples which trick models into making incorrect inferences with high levels of confidence. Further, the encryption of models prevents any sensitive data from being extracted from models, providing a confidentiality benefit to individuals whose data was used in model training. Both aspects lower the risk of sharing encrypted machine learning models with third party data holders and

[48] A2.

could allow them to run such models on their systems and pass back any concerning inferences without having to directly share data, as explored previously.

Using encrypted inference could provide a new mechanism for data holders to find content of concern within their systems, and flag this to the UKIC. However, the UKIC would face a major challenge in ensuring the accuracy and robustness of shared AI models. No machine learning model is perfectly accurate, and both false positives and false negatives are impossible to completely remove from such systems. Further, errors are rarely distributed equally within AI models and tend to cluster around those areas where training data is patchy or incomplete. These areas tend to cover those who are already underrepresented within society, providing a risk that such models may disproportionately affect those individuals.

Consequently, any content flagging system would require a significant degree of oversight to prevent inequitable outcomes, including audit mechanisms to track the source of data, its type, how it was selected, how it was obtained and who received it – a provenance chain to ensure a high standard of compliance and confidence to explain how an individual 'floated up to the top of the haystack in the first place'. [49] From an oversight perspective, understanding how the technology works on third party data will be crucial in respect of false positives and other errors which could themselves result in privacy intrusion.

The above discussion raises a fundamental question of whether PETs have the potential to alter the nature of intelligence agency relations with third parties, which could attract criticism regarding inappropriate delegation or outsourcing of the agencies' statutory functions[50]. An interviewee commented that:

'It would completely depend on what that organisation was doing. If one was giving it any element of discretion or judgement…one would be getting into potentially the territory of delegating our functions.'[51]

Future policy will need to account for these concerns, to ensure that organisations retain appropriate knowledge and control over the analytical process deployed via the PET. This would ensure that potential errors and uncertainties can be identified and assessed, and a provenance chain maintained.

4.4 Data Holders, Cooperation and Liability

PETs could enable the UKIC to work with a greater number of partners. Interviewees, however, raised questions as to the extent to which data holders are likely to be willing to cooperate in granting access to their data systems, with the international dimension of many data holders raising additional complexities. Several interviewees passed comment on this particular point, with one explaining:

'In a more traditional model, they know what we are asking for. There is a big reputational risk for data owners, and it [use of PETs] may not play well with stakeholders and staff.'[52]

[49] IC2, IC18..
[50] O2.
[51] IC1.
[52] IC17.

Other interviewees were more sceptical in their appraisal, for instance:

'What social media company is going to let an intelligence agency run an algorithm on its network? The legislation would need to be international. It's not going to work [otherwise].'[53]

Several interviewees anticipated that data holders would be concerned about data protection legislation, and whether they would retain their data controller responsibilities despite a lack of knowledge and control over the queries being run over their data and systems.

Although PETs could serve as a means for the agencies to minimise intrusion and/or protect sensitive selectors, controller responsibilities under data protection legislation and other confidentiality obligations would remain for third-party data holders. It will be important to engage with representatives of such third parties to investigate their likely willingness to cooperate with the deployment of PETs and their views on reputational and legal risk, including the potential need for additional legal reassurance.

4.5 Authorisation and Warrantry

Linked with addressing issues of third-party liability and cooperation mentioned above, several interviewees suggested that it would be necessary to ensure appropriate authorisation and warrantry for the use of PETs, bearing in mind the innovative nature of the technology and methods. However, it was noted that the warrantry regime under the IPA relates only to the specified activities governed by the Act. The IPA regime was not focused on data held by third parties outside the agencies, nor upon doing acquisition and analysis simultaneously. Interviewees queried whether an equipment interference and/or bulk personal dataset warrant under the IPA, and/or a property warrant under the Intelligence Services Act 1994 (ISA) would be necessary to deploy a PET on a third-party system. The alternative would be *not* to require a specific warrant for deploying a PET on a third-party system, but rather to employ the PET as a measure to increase the proportionality of already warranted activity. The research was inconclusive in this regard.

Although installation of a PET on a data holder's system would not appear to fall within the definition of equipment interference, this might depend upon the extent of the data holder's knowledge of how the PET would operate, and also the interaction with the accounts of the individuals using the data holder's system. In addition, section 5 of the ISA references property warrants, which are designed to authorise otherwise illegal interference with property, such as entry onto premises, search of goods, and manipulation of computer systems. Theoretically, these could be used to authorise 'interference' with a third party's system in circumstances where the third party is only partially aware of the PET method being used. However, there has been debate over the scope of s5 warrants in the context of digital data, with the High Court recently ruling that 'we do not accept the suggestion in [the interested parties'] argument that… powers conferred on the Secretary of State in statute, such as the power in section 5(2) of the 1994 Act, must be given the widest possible construction'.[54] Further consideration

[53] IC14.
[54] [2021] EWHC 27 (Admin).

would be needed as to whether the use of a s5 warrant to authorise interference or intrusions that resulted from use of a PET was sufficiently foreseeable.

In certain circumstances, the use of PETs may avoid the need for the agencies to retain BPDs as set out in the IPA definition. The requirement for authorisation by warrant applies where an agency wishes to retain, or retain and examine, a BPD (s200, IPA). The detailed operation of each PET would need to be examined, however, and consideration given to how the PET changes the nature of the data acquisition process, for instance if new data is still being acquired by targeted examination/analysis of a bulk personal dataset but without acquisition and retention of the complete dataset.

In the PETs context, where third parties may be uncertain about whether they would be breaching confidentiality obligations by collaborating with intelligence agencies, existing legislation may help to provide some clarity and reassurance. For example, s19 of the Counter-Terrorism Act 2008 ensures that a person may disclose information to any of the intelligence services for the purposes of the exercise by that service of any of its functions. Nonetheless, such provisions alone may not allay the wider range of concerns of third parties' customer bases, so any re-emphasis of these provisions would need to be situated within a broader strategy aimed at maintaining public trust in commercial sector-intelligence agency ties, including clarifying any additional warrantry and oversight considerations arising from the deployment of PETs on a third-party system.

4.6 Oversight

Research interviews highlighted that the changing role of third parties in intelligence activities involving PETs may alter the demands on oversight bodies. For example, one interviewee suggested that:

'You would need to have oversight that has visibility over both parties… it could be a change in scale of the challenge faced by oversight bodies.'[55]

It remains to be seen whether this proposed expansion of oversight responsibilities is either practical or desirable. Nevertheless, research participants highlighted the importance of a system of attestation of techniques and appropriate verification and audit by the oversight body, suggesting that clarity would be needed over storage of and access to audit logs.[56] This links to the points made above regarding the importance of a robust audit function and the maintenance of a data provenance chain.

If PETs are deployed on intelligence systems, reassurance will also be required that compliance monitoring is not undermined, for example ensuring that PETs do not allow adverse behaviour internally to be more easily hidden.[57] It was suggested, for instance, that encrypted queries sent to third parties would still need to remain auditable in their unencrypted state within UKIC systems.[58]

[55] IC3.
[56] IC13.
[57] IC11.
[58] O2.

To have confidence in the reliability and applicability of PETs to intelligence activities, one interviewee suggested that:

'There should be a way of judging if this technology has passed a test and some way of knowing that someone will be held accountable if these tests fail… (at the same time) you need good stories on how the technology has been used in a trusted way.'[59]

Developing these mechanisms will help to advance understandability of the technology, crucial in respect of uncertainties and errors which could undermine the effectiveness of investigatory actions and result in privacy intrusion. Interviewees generally proposed an expansion of oversight responsibilities as the most appropriate mechanism to ensure confidence in the reliability of new and emerging PETs, although it remains unclear whether this is a feasible course of action. Nevertheless, ensuring adequate technical expertise within IPCO will be vital to critically evaluate the operation of the technology – for example in the process of spot-checking – and to create a regime which would provide reassurance to third party data holders.

4.7 Public Trust

Finally, the research identified a perceived lack of general understanding about the functionalities and limitations of PETs:

'The reason we are not seeing increased levels of adoption is because people do not know they exist or do not trust them. Most people still think PETs are about security. PETs are actually about being able to facilitate who should know what. This is not possible with conventional technology which is still as the stage of "Can I share this dataset with you or not?" PETs allow for precision.'[60]

Concerns were raised that public perceptions about the purpose of a PET may bear little relationship to its actual functionality. Misunderstandings regarding both the motivations for using PETs and their functionality may lead to confusion about how PETs relate to terms like 'privacy', 'ethics' and 'fairness' in the broadest sense, emphasising the importance of developing clear definitions of what can and cannot be protected by a given technology. Such definitions could facilitate a more informed public debate about the use of PETs in national security.[61] Interviewees also recognised possible scenarios in which the public viewed PETs as a 'cloak of surveillance' for security agencies to obtain access to more data than they otherwise would have.[62]

In respect of this challenge, one interviewee argued:

'The narrative should focus on intentionally giving defence and security stakeholders less access to data for more precise queries. People will not be interested if you just focus on the algorithm, so you have to amplify this narrative and pick simple success stories that appeal to a broad range of

[59] A6.
[60] A5.
[61] A1, O1.
[62] A1, IC7.

individuals (…) pivot the story from being about restriction to being about going the extra mile to not learn things that you do not actually care about."[63]

Therefore, oversight arrangements should be cognisant of this risk and put in place methods of monitoring the longer-term cumulative effects of the use of PETs on the data acquisition practices of the agencies.[64]

Furthermore, public engagement relating to PETs should proceed hand-in-hand with a wider agenda of transparency and engagement. As with any emerging technology, discussions around how PETs might be deployed by intelligence agencies run ahead of the general public's understanding of what the technology is and the value it adds to society. This emphasises the importance of the intelligence community working closely with the scientific community, the commercial sector, and civil society to cultivate a more inclusive societal conversation around PETs.

[63] A5.
[64] O2.

About the Authors

George Balston is Programme Co-Director, Defence and Security at The Alan Turing Institute. He leads a portfolio of work applying data science and artificial intelligence to national security, cyber security, and defence challenges, working closely with national and international partners. Prior to joining the Turing, George worked within the UK Government as a Researcher, Data Scientist, and Technical Lead, focused on researching, developing, and deploying new analytical capabilities.

Dr Marion Oswald is a lawyer with over 30 years' experience spanning several contexts: law firms, international technology businesses, central government including national security, academia, and oversight functions. She researches the interaction between law and digital technology, with a particular interest in the human rights, ethics and use of data analytics within policing and intelligence agencies.

Alexander Harris is a Research Associate in the Defence and Security programme at The Alan Turing Institute, working across a research portfolio exploring the application of data science and AI to a range of national security and defence challenges.

Ardi Janjeva is a Research Associate in the Centre for Emerging Technology and Security at The Alan Turing Institute. He previously worked as a Research Fellow at The Royal United Services Institute (RUSI), leading the Institute's research on Technology and National Security.

Centre for Emerging Technology and Security

RESEARCH REPORT