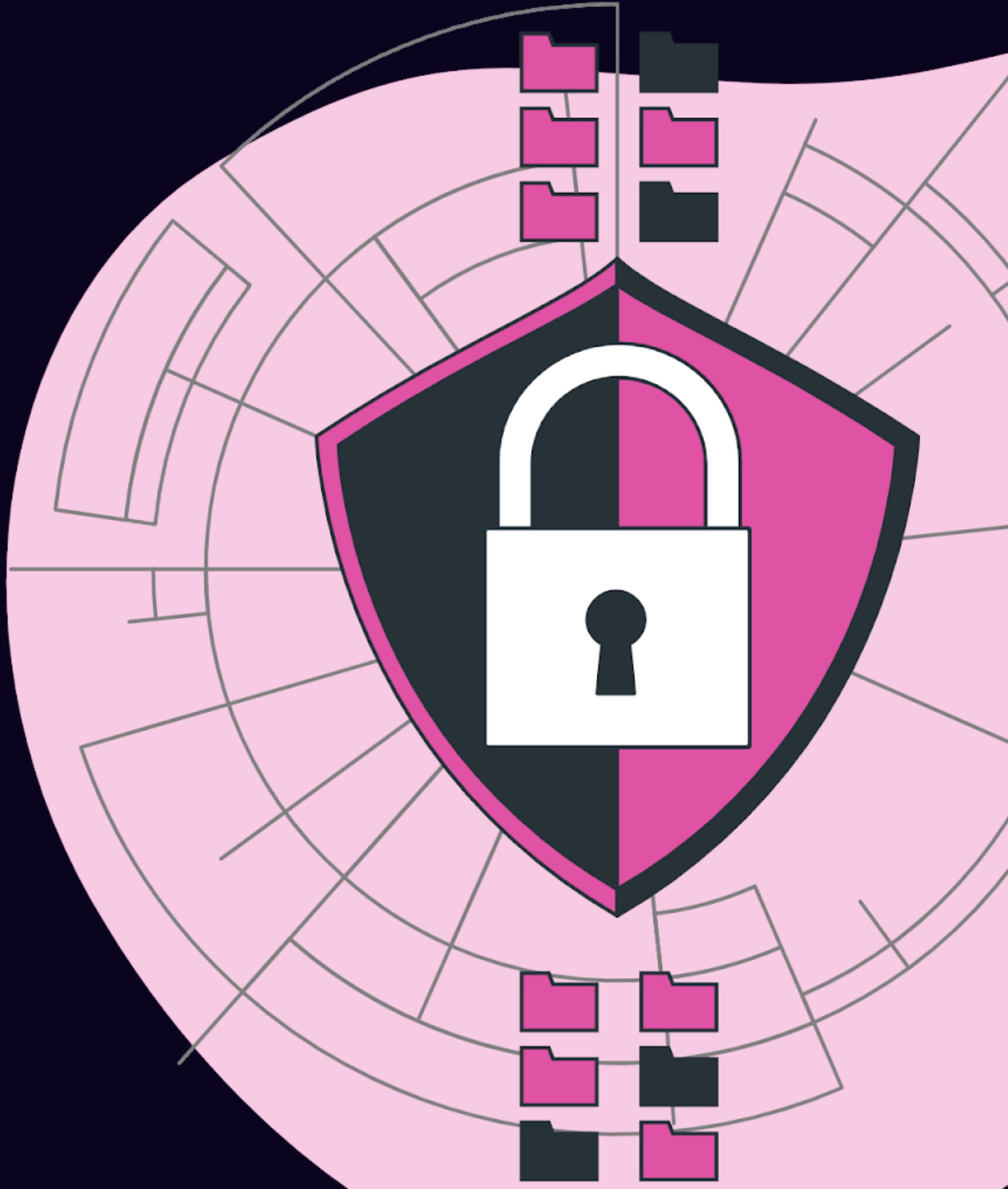
AI Risk Evaluation Community Group

Perspectives & Recommendations on the Development of Safe AI in Sensitive Healthcare Data

Lewis Hotchkiss, Emma Squires, Kafayat Adeoye, Alieyeh Sarabandi, Sharon Heys, Elen Golightly, Catrin Morris, Timothy Rittman, John Gallacher, Simon Thompson

DARE UK



AI RISK EVALUATION COMMUNITY GROUP

This report was prepared as part of a DARE UK funded initiative to set up an AI Risk Evaluation Group to bring together a range of stakeholders to understand perspectives of AI development / release from Trusted Research Environments (TREs), and the unique challenges posed by complex multi-modal data. The main goals of this group were to understand:

- What are the public most worried about with the use of their data for training AI models
- What are the unique challenges that neuroimaging and genomics present in AI disclosure control
- What is the actual risk of a person being identified if their data were released from an AI model
- How do researchers feel implementing privacy-preserving techniques in their research
- What is the risk appetite of data providers and do they agree with our recommendations
- How can we help researchers implement these privacy-preserving techniques in their research
- How can we build a framework to allow the safe development and release of AI models trained on complex data
- How can we help data providers quantify risk and assess these models for safe release

Foreword

Professor Simon Thompson

DPUK Deputy Associate Director
SeRP Co-Director

> With the rapidly evolving landscape in the development of AI models on sensitive healthcare data, it has never been more important to consider the risks that these models pose to patient privacy and what role Trusted Research Environments (TRE's) have in ensuring the responsible development of these models and the safe release of them. This report paves a way forward in allowing important AI research to take place within TREs while still ensuring that we do everything we can to protect the privacy of individuals and giving the right support to researchers to help facilitate that.

Contributors

Co-Chairs

Prof Simon Thompson

Simon is Chief Technology Officer at SeRP and Professor of Health Informatics at Swansea University.

Lewis Hotchkiss

Lewis is the Neuroimaging Research Officer at DPUK and leads work on responsible AI research in neuroscience.

Prof John Gallacher

John is the Director of DPUK and Professor of Cognitive Health at the University of Oxford.

Dr Timothy Rittman

Tim is a Senior Clinical Research Associate at the University of Cambridge and leads the QMIN-MC study.

Emma Squires

DPUK Programme Manager

Catrin Morris

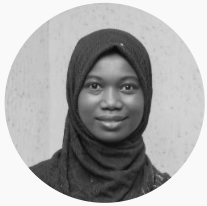
DPUK Operations Coordinator

Sibil Gruntar Vilfan

DPUK Administrative Assistant

Elen Golightly

DPUK Data Scientist



Kafayat Adeoeye

DPUK Data Scientist

Alieyeh Sarabandi

DPUK Data Scientist

Sharon Heys

Legal Advisor

organisations & universities represented

Swansea University Prifysgol Abertawe, Dementias Platform UK Data Portal, University of Cambridge, University of Oxford, SeRP, DARE UK, prevent dementia, Brains for Dementia Research

Bitfount, DataLoch, The Alan Turing Institute, DATALINK, biobank uk, UK Longitudinal Linkage Collaboration, The Francis Crick Institute, generation scotland

université de Bordeaux, The University of Edinburgh, King's College London, U Lisboa Universidade de Lisboa, University of Surrey, Imperial College London, University of Bristol

Executive Summary

These series of workshops highlighted the important role that TREs can play in the safe development and sharing of AI models in sensitive healthcare data. From our first workshop, we found that members of the public overwhelmingly preferred TREs to be used for developing AI models on their health data, but stressed the importance of public involvement at the decision stage to ensure models are being developed and shared in the public benefit.

From the researcher workshop, it was clear that most researchers aren't aware of the risks their AI models pose and how to appropriately mitigate these for safe release and deployment into the real world. They also acknowledged challenges with complex data types, such as neuroimaging and genomics, for disclosure control but also for implementing mitigations. Therefore, training and resources are crucial to enable AI models to be developed responsibly in health data. Researchers also felt the need for tools to help generate and evaluate safe data for training AI models.

Furthermore, in the data owner workshop, they felt that they lacked the expertise to comfortably assess AI model projects and relied on the TRE to guide them to help make decisions. They expressed the need to quantify risk of releasing AI models in various scenarios and to have researchers fill out AI risk impact assessments to evaluate AI projects appropriately. Additionally, it was clear that data owners felt that running attack simulations on AI models provided adequate assurance that they could be released from a TRE.

Ultimately, when it came down to who's responsible for potential risks in AI models, it was agreed that it should be a shared responsibility between the researcher, data owner, TRE and the funder to ensure that models are being developed responsibly.

From the results of these workshops, we put together a series of recommendations, tools and materials for assessing AI models being released from a TRE. However, it was decided that the need for releasing AI models from a TRE should be reduced as much as possible, without hindering scientific research.

There are three main reasons why a researcher may want to export an AI model – (1) to publish on a platform such as GitHub for open and reproducibility reasons, (2) to further train the model on external datasets, or (3) to deploy into clinical practice. In all three cases, the model doesn't necessarily need to leave the TRE.

In cases where (1) the researcher wants to publish an AI model, they should be published via the TRE where researchers can apply and access that AI model in the same way that derived data is. This allows the AI model to stay secure within the TRE while allowing access for reproducibility and validation. It is important to note that governance around model ownership would need to be clearly defined to ensure appropriate access protocols were in place to specify who makes the approval decision.

When external data is needed for training or validation (2), this can be achieved through secure federation of data environments to enable access and training on other datasets.

If an AI model is ready to be deployed into the real-world (3), then secure hosting offers the most suitable solution, where the model can stay within the portal, and the TRE offers ways to securely query it and receive predictions externally. This enables the safe translation of AI models into clinical practice while ensuring utility isn't affected.

However, if the researcher still requires an AI model to be released outside of the TRE, then privacy-preserving techniques should be implemented by the researcher and rigorously evaluated by the TRE to ensure that it is safe for release. As part of this work, we developed tools to help researchers assess the privacy/utility trade-off in creating safe data for AI models, as well as methods to help quantify risks in releasing an AI model to help data providers make informed decisions.

These perspectives, recommendations and materials help TREs move forward to enable important AI research to take place while still protecting the data they have been entrusted with. It is important to note that AI risks are constantly developing, and so too should the recommendations.

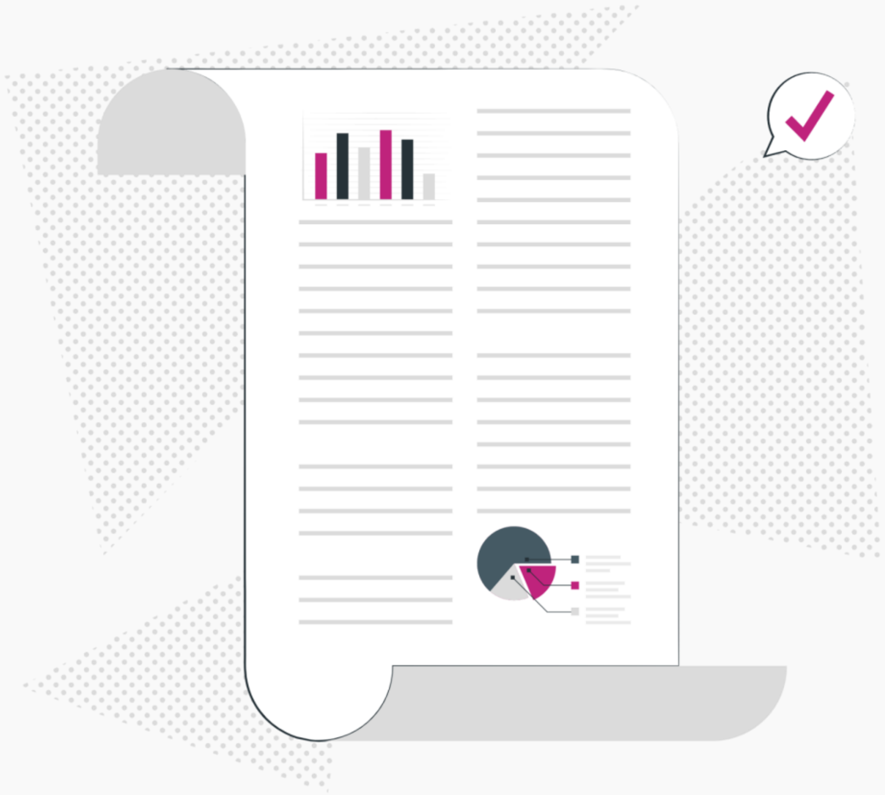
Lewis Hotchkiss
Neuroimaging Research Officer

Open and reproducible science has long been an issue for AI research, and the privacy risks posed to releasing these models play a big role in this. Therefore, **a key recommendation from this report, is the utilisation of the FAIR framework for AI models** to ensure that they can be accessed in a secure manner, while still enabling open and reproducible science.

By hosting AI models through TREs, rather than openly accessible platforms such as GitHub, **we can enable AI to be FAIR, while also fulfilling our duty to protect the data that was used to train those models**. This also significantly reduces the time, effort, and resources needed to evaluate AI models for release. However, we recognise the need to release AI models in certain circumstances, which is why we developed recommendations and materials to help in making this possible.

Overall, this work further showcases the role of TREs and the importance of them to protect data while enabling beneficial AI research to take place.

Table of Contents



CHAPTER 1

Background

AI Model Risks & Concerns with Complex Healthcare Data

AI models can be prone to several privacy attacks which allow for inference or reconstruction of the training data used to develop them. There are also various characteristics in these models which can make them more vulnerable to these privacy attacks. In this section we will explore these attacks and vulnerabilities as well as the unique risks that neuroimaging and genomics pose to re-identification.

AI Model Risks

Privacy attacks in AI models have been identified which allow inference of individuals or reconstruction of the training data. These very much rely on the types of data used and how the model was trained to be able to perform these attacks on released models.

Membership Inference Attack
Membership inference attacks enable an attacker to determine whether specific individuals' data was included in the training of AI models. This poses a significant threat, as discerning the inclusion of individual data can lead to unintended exposure of sensitive medical histories. By exploiting the models outputs, usually by observing a confidence score, attackers can infer whether specific individuals' data was included in the training set, therefore exposing whether an individual has a certain disease for example in something like a treatment response model.

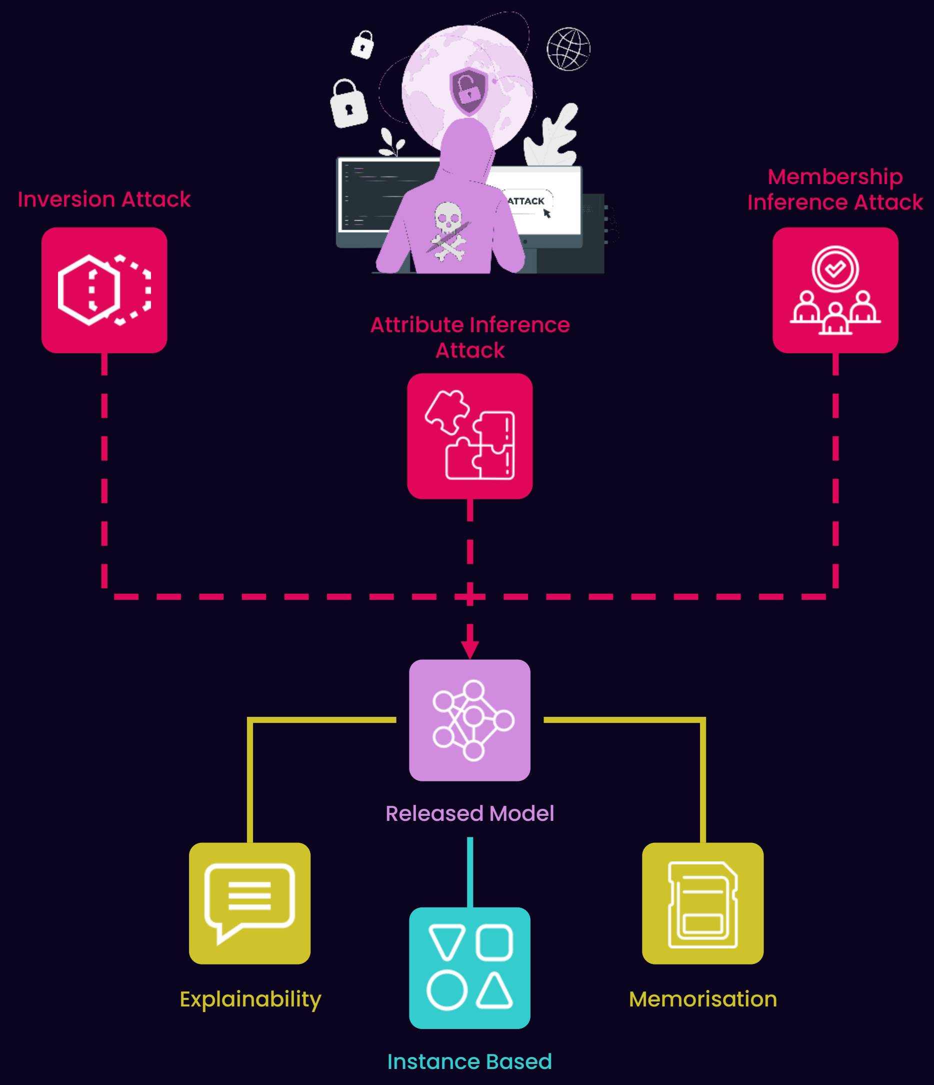
Attribute Inference Attack
In an attribute inference attack, an adversary may have partial knowledge of an individual and access to a model trained on records including that individual. From this, they can infer the unknown values of features in those records. The adversary uses the accessible model to make predictions for the instances they have partial knowledge of and by analysing the model's output, they attempt to deduce or infer the unknown values for those instances. This is usually an iterative process, refining their understanding of the model's behaviour and adjusting their queries to the model. Higher confidence in predictions will boost the attacker's confidence in the accuracy of their inferred values.

Inversion Attack
Model inversion attacks represent a sophisticated challenge, enabling the reconstruction of sensitive information from the AI model itself. This technique goes beyond merely identifying data participation or specific attributes; it seeks to reverse-engineer the model to reveal potentially sensitive details about individuals and reconstruct the original training data.

Data Memorisation
Data memorisation, also known as overfitting, is a common challenge in training AI models which occurs when the model becomes too focused on the specific details of the training data, rather than learning the underlying patterns. This can typically occur when there are too many features and/or too few participants in the training data. Inversion and inference attacks can exploit this vulnerability to potentially reveal specific participant data.

Explainability
Some methods to make AI models explainable can aid in adversaries conducting attacks on AI models. Not only can the explanations themselves help to reveal sensitive data, but it can also give adversaries information on how the model works and how to exploit it. These explanations can also give further details to exploit when it comes to running these attacks.

Instance-Based Models
These types of models use the training dataset as the model to compare unseen data to the data points in the dataset. This works by making predictions based on similar examples in the training data, compared to learning the patterns of that data. This means that these types of models actually store instances of the data.



Concerns with Complex Healthcare Data

Table 1

Gender	DoB Year	DoB Month	DoB Day	Marital Status	Postcode District	Risk
Male	1962	May	28	Never Married	SA1	
✓	✓	✓	✓	✓	✓	98%
✓	✓	✓	✗	✓	✓	53%
✗	✓	✓	✗	✓	✓	33%
✓	✓	✗	✗	✓	✓	6%
✗	✓	✗	✗	✓	✓	3%
✓	✓	✓	✓	✓	✗	2%
✗	✗	✗	✗	✓	✗	1%

Likelihood of being re-identified in an anonymous dataset given the combination of attributes (attributes vs risk)

Table 2

Gender	DoB Year	DoB Month	DoB Day	Marital Status	Postcode District	Risk
Male	1962	May	28	Same-Sex Partnership	SA1	
✓	✓	✓	✓	✓	✓	100%
✓	✓	✓	✗	✓	✓	99%
✗	✓	✓	✗	✓	✓	99%
✓	✓	✗	✗	✓	✓	93%
✗	✓	✗	✗	✓	✓	87%
✓	✓	✓	✓	✓	✗	82%
✗	✗	✗	✗	✓	✗	1%

Effect on likelihood given a less common attribute

The Problem of "Anonymised Data"

TREs typically anonymise data by stripping out personal identifiable information (PII) such as names, addresses, and dates of birth. Within the TRE, the security measure (such as virtual desktops and disclosure control methods) ensure the data is functionally anonymised, in that it cannot be combined with other data. If the same datasets were publically available outside of the TRE environment, then that data could potentially be identifiable due to privacy attacks and linkage to other data outside of the TRE. One study demonstrated that individuals can still be identified through the combination of attributes in anonymised data [1]. In this study, they were able to create a statistical model to evaluate the likelihood of an individual being identified, within an anonymous dataset, given 7 attributes about them. Table 1 shows the effect of different combinations of attributes on the risk of being identified. If all attributes are available, then there is a 98% likelihood of being identified within an anonymous dataset, but you can see that once attributes like postcode district and day of birth are removed, then this risk decreases for this particular example. However, if the same example is used, but with the marital status changed from never married to same-sex partner, then we can see a drastic difference in the likelihood of being identified (table 2). So although a fairly average person in a population may not have to worry about being identified, people from minority groups are at a much higher risk of being identifiable through a combination of their attributes. Of course this likelihood also increases given more attributes about a person, which is especially likely in cohort data collections. Another study, was able to successfully re-identify patients in an Australian de-identified open health dataset just by using publicly available information to name the individuals [2].

Unique Risks in Genomic Data

Genomics research is a promising field for the progression of dementia research; with recent advancement in the development and accessibility of technologies at lower costs, the ability to gather and store genomic data on a wider scale makes genomic research much more viable. With this growing availability however, comes greater concern regarding data privacy and disclosure control. The genomic information of an individual is inherently disclosive and unique to that individual. Moreover, this data is the same across all cells and remains relatively static throughout a person's lifetime. Due to the nature of genomics data, even a small percentage of an individual's genomics data in aggregate form has the potential to be identifiable and disclosive. This risk further increases when genomics data is linked with phenotypic or demographic data.

Genomics data is predominantly susceptible to two main types of attacks; identification, and phenotype inference attacks [3]. If the attacker is in possession of part of the genetic information of an unknown member of a cohort (DNA phenotype, linked data, genealogy data), they can exploit this information to gain the identity information of the target via identification attacks. An example would be if an attacker has the DNA phenotype of an individual, then they could construct an approximation of various observable phenotypes such as hair colour, eye colour and height. These observable features could lead to identification, particularly if these features are in the minority of the cohort and the attacker has some information regarding the context of the cohort.

If the attacker already knows the identity of their target however, they can leverage genomic information In order to gain knowledge of certain characteristics of their target. Genotype imputation can allow them to infer the kinship predisposition of the relatives of the target, and linkage disequilibrium can also be used to uncover masked genome markers of the target. With this data or other available genomics data, the attacker can infer phenotype information of the target.

Unique Risks in Neuroimaging Data

The sharing of neuroimaging data is usually facilitated by the defacing of scans to remove identifiable features of an individual. However, depending on how this defacing is done, the faces of these scans still have the potential for facial reconstruction if some identifiable features such as eyes are still kept [4]. Additionally, due to the effect that defacing can have on the results of analysis, some researchers request the original non-defaced scans to use in their processing/analysis. However, this poses significant privacy concerns as faces from these scans could be used to identify individuals. Identifiable features are not the only concerns of imaging data. It has been shown that the brain is unique like a fingerprint, meaning that if you have access to a scan of the individual in another dataset, then you could link the two together [5]. Its not just structural scans prone to this too, but functional connectivity matrices have been shown to also be unique to an individual [6]. So, as with genomics data, the concerns around the use of these types of data centre more around what data is linked to that or what external data could potentially be linked.

CHAPTER 2

Perspectives of the Public

Risks & Concerns of Complex Health Data in AI

Our public workshop brought together a diverse group of individuals, spanning a range of backgrounds and age groups, to gain insights into their concerns surrounding the utilisation of their data for training AI models in healthcare. The primary objectives of this workshop were:

Objective 1: Understanding Public Perception of AI Models
Assess the public's perception of AI models in healthcare, determining whether they view them positively or negatively. Explore their opinions on the potential benefits and risks associated with the use of AI technology in healthcare settings.

Objective 2: Identifying Concerns with Data Usage in AI Research
Identify and analyse the specific issues that the public has regarding the use of their data in AI research. This includes examining their concerns about data privacy, security, and potential biases or discrimination that may arise from AI algorithms.

Objective 3: Exploring Data Types with Privacy Considerations
Determine the types of data that the public would be most concerned about being released publicly. Understand their reasons for these preferences and explore any potential trade-offs between data sharing and privacy concerns.

Objective 4: Investigating Privacy Concerns with Healthcare Data
Investigate specific privacy concerns that the public holds concerning their healthcare data. Examine their views on data ownership, control, and the potential consequences of data breaches or unauthorised access to their health information.

Public Perception of AI Models in Healthcare

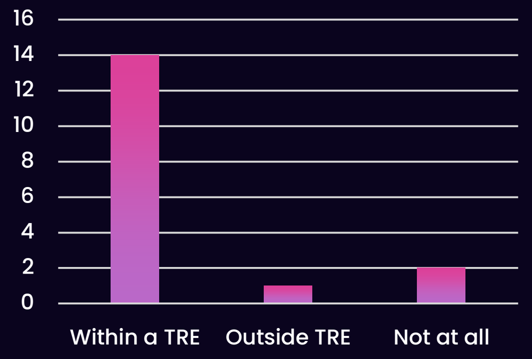
As part of this public workshop, we presented a series of informative and engaging talks to the participants to delve into how AI models work, the benefits that AI models offer in the healthcare domain, showcasing their potential to enhance diagnostics, treatment planning, and patient outcomes, but also what the potential risks of AI models are. After each talk, we facilitated interactive discussions to explore various topics related to the theme of the talk in greater depth. These discussions provided a platform for participants to express their views, concerns, and insights regarding the use of AI in healthcare. Additionally, to capture the participants' feedback and opinions in a structured manner, we provided them with a survey to complete throughout the workshop session. This survey covered a range of questions designed to gauge their understanding of AI models, their perceived benefits and risks, and their attitudes toward the use of their data. Through this combination of informative talks, engaging discussions, and surveys, we aimed to create a comprehensive and interactive experience for the participants to empower them to make informed decisions and contribute to the discussions in a meaningful way.

At the beginning of the workshop, before the talks & discussions commenced, we wanted to get an understanding of what the public's perception of AI models are. A majority of participants had a very basic understanding of what AI is and understood that it involved the use of data to understand patterns. And they were able to name some real-world examples such as medical diagnosis, breast cancer analysis, voice assistants like Alexa, ChatGPT, and flight control. We also wanted to understand whether people thought AI had a positive or negative impact in healthcare.

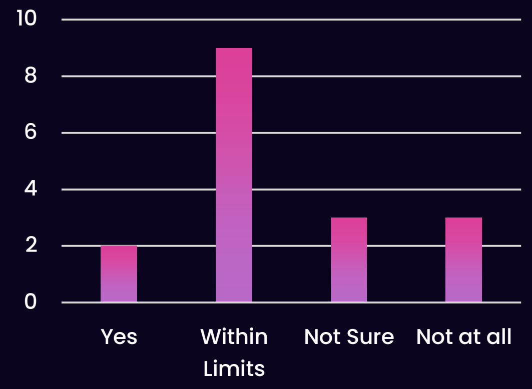
A majority of participants (64%) said that AI has its uses, with 18% saying that it was good, 18% saying unsure, and no one saying bad. By the end of the workshop, there was a shift towards a more positive attitude, with 56% of participants saying AI was good with the remaining 44% saying it has its uses. So after explaining more about what AI is, how it works, and what the benefits are, as well as detailing the risks involved, participants had a more positive attitude towards the use of AI in the healthcare sector.

After explaining what TREs are to the participants, we asked how they would feel about their data being used to train AI models inside of a TRE, compared to outside. A majority of participants (82%) said that they were happy for their data to be used to train an AI model within a TRE, with only 12% being happy their data being used outside a TRE and 0.06% not at all. This clearly shows that the participants trust TREs to protect their data while allowing it to be used for AI research. In the discussions, it was also noted that people have contributed their data for a reason and there is a moral duty to use it for the good of improving people's health. 65% of participants said that they would be willing to sacrifice their privacy, completely or within limits, for the benefit of society. So, there is a motivation to make sure that their data is used for public good and that we make sure we can allow this important AI research to take place as it could have wide ranging benefits to society. Participants also recognised the potential benefits that AI brings to the health sector, with increased speed of diagnosis and improved treatment being key themes which emerged from the discussions.

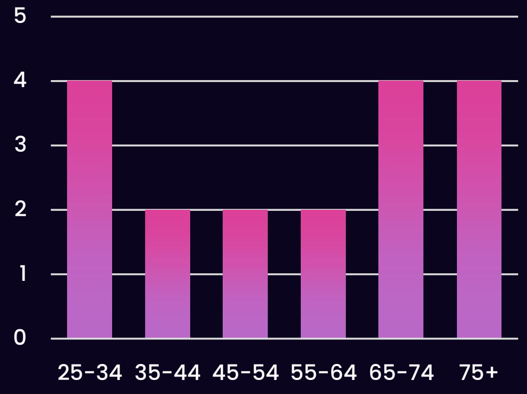
Would you be happy with your data being used for AI research?



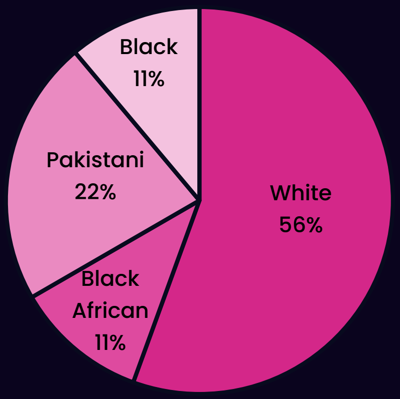
Would you sacrifice your privacy for the benefit of society?



Age of Participants



Ethnicity of Participants



Concerns of Health Data in AI

During the discussions, we aimed to gather people's concerns with the use of their data in AI models and as part of this, several key themes emerged. Firstly, before data is even shared, people felt that there were problems at the consent stage regarding no explicit mention of the use of their data for training AI models.

> *“Once initial consent is given, no updates are given after. What happens if things change, or consent was given before AI was a concern.”*

Of course, you can never predict what future concerns there might be, but people felt that consent shouldn't just be a one-off and that they should be kept up to date with how their data is being used, who its being shared with, and to keep them involved in the decision-making process to ensure these AI models are being developed in the public good.

This lack of control over their data fed into concerns around selling data to corporations or insurance companies, and not knowing how this data is being used and could potentially affect them. Interestingly, they acknowledged that their social media history could be a lot more disclosive, but one person said:

> *“Social media may contain more disclosive information but at least we have control of what we put out there.”*

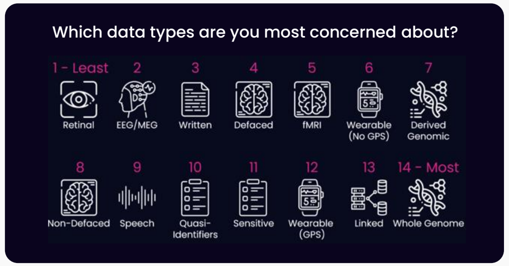
So, control seemed to be an important aspect for members of the public to have to make sure they feel confident in their data being shared and used in the public benefit. This is why people felt that there should be public involvement in the decision process.

AI models present unique concerns around identifiability of individuals and people felt that some groups were more at risk than others due to rare conditions or coming from an ethnic minority background.

> *“I have concerns around people like myself who have rare conditions. People could be at higher risk of identification as there's less people in the population like you.”*

This is a key consideration when assessing privacy concerns in AI models as minority groups are at a higher risk of identification (see chapter 1).

Concerns around regulation in AI research was also a key theme which emerged from the discussions. Most people felt that there was a *“lack of regulations”* in this space and that they need to *“keep up with the speed of development”* as AI is rapidly evolving and posing new challenges to safety and privacy. People felt that there is not enough regulation in the UK to protect them against the risks that AI models pose and felt that researchers should take the responsibility on themselves to ensure that they do everything they possibly can to ensure they protect the individuals that they are using to train their AI models.



Which data types are you most concerned about?

How is AI being Regulated in the UK?

At present, the regulation of Artificial Intelligence (AI) primarily relies on existing laws and regulations, such as the General Data Protection Regulation (GDPR) for data protection and the Equality Act 2010 to prevent discrimination by AI systems. These existing laws provide a foundation for addressing some of the challenges posed by AI, but they may not be sufficient to effectively govern the rapidly evolving field of AI. In recognition of this, the UK government published its AI White Paper, which outlined an approach for regulating AI in the UK [7]. The White Paper emphasises the need for a proportionate and risk-based approach to regulation that balances the promotion of innovation with the protection of individuals and society. Five principles were identified, one of which was:

safety, security and robustness: applications of AI should function in a secure, safe and robust way where risks are carefully managed

Twelve AI governance challenges were further identified by the Science, Innovation and Technology Committee, with one of these being the privacy challenge, where the committee was told that regardless of the sector, privacy should be “... an integral part of the balance of interests that you consider when you are deploying artificial intelligence” [8].

Beyond regulatory measures, researchers seek guidance from organisations such as the Information Commissioner's Office (ICO) for insights on data protection and privacy issues pertaining to AI models [9]. The ICO plays a crucial role in providing expert advice, issuing guidelines, and promoting best practices to ensure that AI systems are developed and deployed in an ethical and private manner. Furthermore, the ICO developed an AI and Data Protection Toolkit for AI developers to identify the potential risks to individuals' rights and freedoms and how to take steps to mitigate those risks and comply with current laws around data protection. The Ada Lovelace Institute also created an Algorithmic Impact Assessment, specifically for the NHS AI Lab, for the use of neuroimaging data to assess possible societal impacts of an algorithmic system before the system is deployed [10]. However, this is not focused on privacy and focuses more on ethical and discrimination considerations.

CHAPTER 3

Perspectives of Researchers

Evaluating the Suitability of Privacy-Preserving Techniques

This workshop brought together researchers who are actively involved in AI research to discuss the suitability of privacy-preserving techniques in their research. The main aims of this workshop were to:

Objective 1: Researcher Awareness and Concerns
Understand what privacy concerns and vulnerabilities in AI models researchers were aware of and their levels of concern regarding different modalities of data.

Objective 2: Assessing Privacy-Preserving Techniques
Evaluate their confidence in using privacy-preserving techniques and their willingness to incorporate them into their research. Identify the most suitable methods taking into consideration privacy, utility and ease of implementation.

Objective 3: Identifying Challenges and Barriers
Identify what the barriers are to incorporating privacy techniques in their research and what unique challenges exist with complex data such as neuroimaging and genomics.

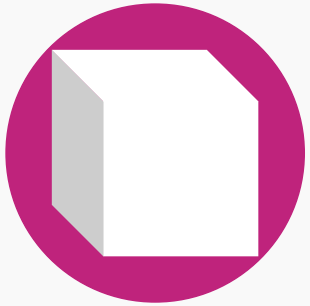
Objective 4: Evaluating Model Release Scenarios
Assess AI model release scenarios from TREs and how to effectively evaluate privacy in models to ensure that they are safe to release.

Privacy-Preserving Techniques

Currently, there are various measures available to safeguard privacy in AI models, either during the training phase or by imposing constraints during deployment. These can protect against a range of different attacks but should be carefully chosen depending on how that model is going to be used or shared. For instance, when sharing or releasing an AI model, there exists a vulnerability to white-box privacy attacks, where the attacker possesses full access to the model, enabling direct inspection and a wider range of attacks to be performed. In such cases, it's imperative to employ privacy-preserving techniques that safeguard the training data to counter these threats.

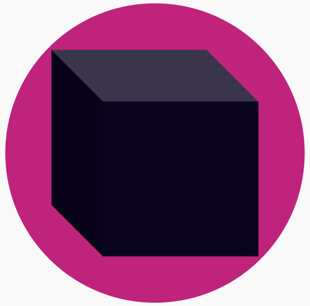
Conversely, in scenarios where the model is inaccessible but can be queried, it becomes susceptible to black-box attacks. Here, it could be more beneficial to enforce access/query limitations on the model or to employ privacy-preserving techniques during inference.

At the start of the researcher workshop, we explained a range of these privacy-preserving techniques to gauge researchers opinions on using them in their research and to find out what the barriers/challenges are to implementing them.



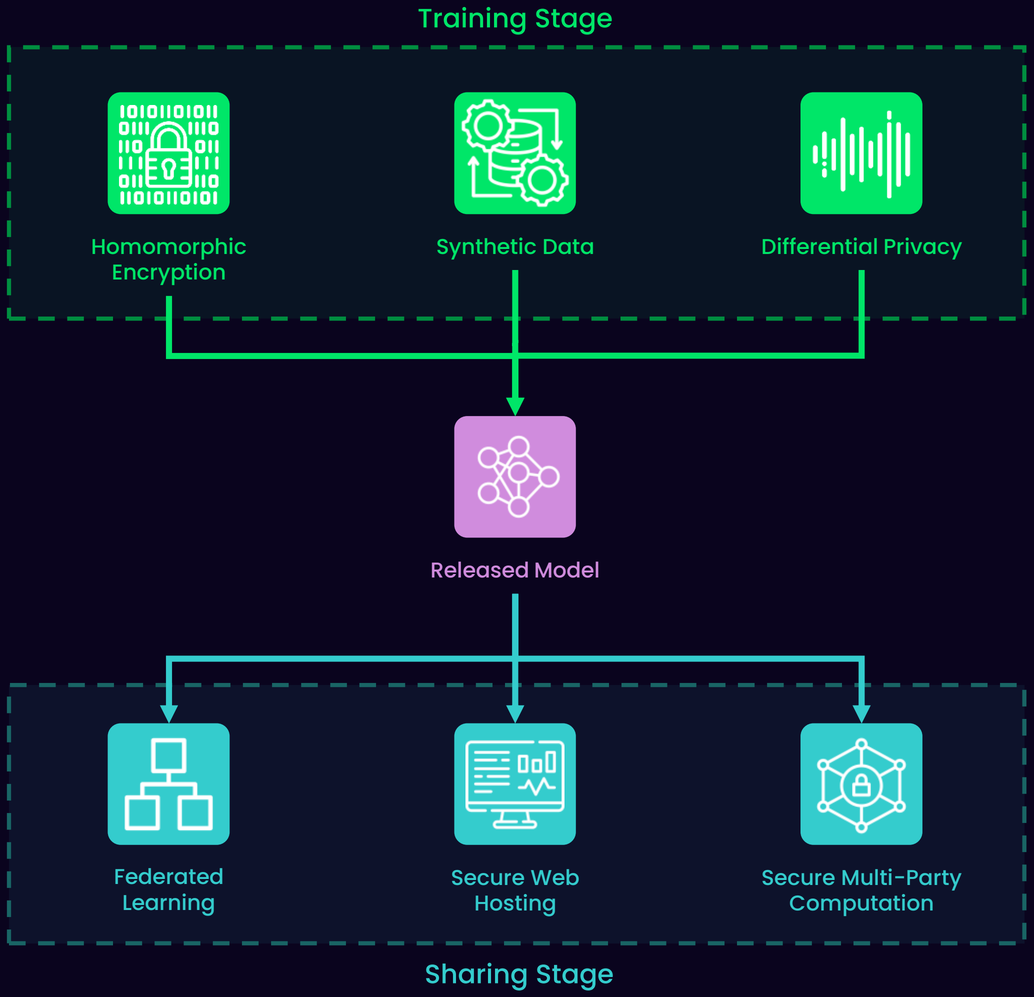
White Box Attack

Where the adversary has full access to the model, therefore enabling them to have additional information on the type of model and parameters used to help them perform their attacks.



Black Box Attack

Where the adversary has no additional knowledge of the AI model, and is only able to query the model and observe the relationships between inputs and outputs.



Privacy-Preserving Techniques

Differential Privacy (DP) in AI

Summary

Differential privacy works by adding noise either to the data, or the response of the model, to ensure that an adversary can't determine with confidence that information about an individual is present in the data. This level of noise is determined by epsilon, also known as the privacy budget, which controls the privacy guarantee of the data. However, differential privacy involves a trade-off between privacy and utility due to the effect of adding noise. Because of this addition of noise, this can reduce the accuracy of an AI model, so researchers have to carefully consider this trade-off and the level of noise suitable.

DP isn't a method itself, but instead a privacy guarantee that data or algorithms must meet [11]. That guarantee being that a given output shouldn't depend too much on any singular record. This is useful for analysis which looks at more general population scale trends rather than detecting detailed patterns within the data. But there are many ways that DP can be applied and two types of DP which exist - local and global.

In local DP the noise is added to the training data before any processing takes place whereas, in global DP noise is added to the gradient updates during the training process or added to the result at the prediction stage.

The Privacy Budget

The amount of noise added is determined by epsilon, otherwise known as the privacy budget. Setting this parameter too high will unlikely be sufficient enough to protect against privacy attacks, whereas setting it too low will significantly reduce the performance of the AI model. Because of this privacy-accuracy trade-off, it can be challenging trying to identify a good balance between the two, and is not consistent across datasets. This means that there are currently no standards for setting this value as it varies across different datasets. However, there is general consensus that:

- ε values (0 < ε < 1) will provide a strong privacy protection
- ε values (1 < ε < 10) can also provide robust protection in certain settings
- higher values (ε > 10) are unlikely to provide robust protection but may still provide some level of protection given certain settings

But this value highly depends on the data and method used so needs to be evaluated carefully.

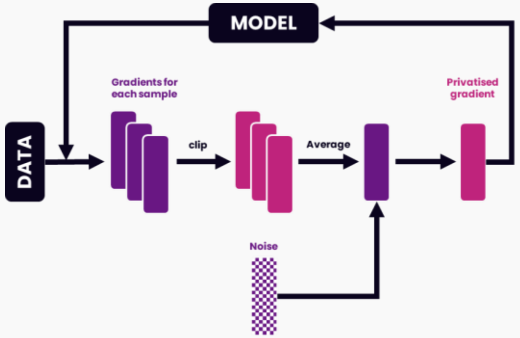
Smaller ε
More Privacy
Less Accuracy

Larger ε
Less Privacy
More Accuracy

DP at Training Stage

There are three different stages where DP can be applied in AI models – at the data stage, at the training stage, and at the prediction stage.

At the training stage of AI, there are typically two approaches of doing this. One method is called differentially private stochastic gradient descent (DP-SGD), which adds noise to the gradients before updating the model parameters [12].

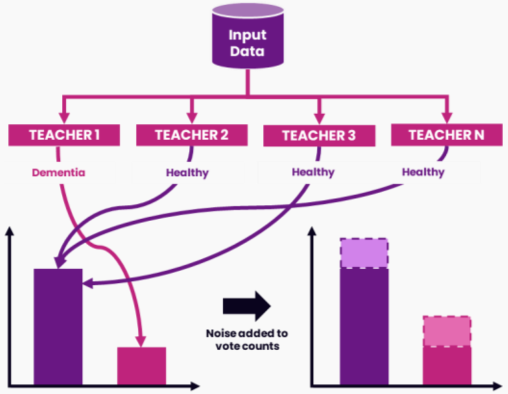


This process typically involves:

1. Randomly selecting a batch of samples
2. Clipping the gradient for each sample
3. Adding random noise to each gradient
4. Updating the model weights using the average noisy gradients in the batch

Another method is based on the Private Aggregation of Teacher Ensembles (PATE) framework [13]. Rather than adding noise to gradients, this method trains many non-private models (teachers) on subsets of the data, and then "votes" on the correct prediction using DP aggregation. Here, the noise is added to vote counts which avoids revealing the votes of any individual teacher.

The labels of the noisy aggregated predictions are then used to train a student model which is then the one which is shared. This method generally has less of an impact on accuracy compared to DP-SGD but relies on a student model being trained on public data and therefore mainly focuses on the privacy of teachers' training data and fails to protect the privacy of the students' data.



Privacy-Preserving Techniques

Differential Privacy (DP) in AI

There are of course more examples on the use of DP in the training stage for specific types of AI methods. In decision-tree based models for example it has been proposed that DP can be added by creating random decision nodes or adding noise to split thresholds.

Practical Implementation of DP

DP methods have developed into a somewhat mature space recently with companies such as Google and Apple using such techniques [14,15]. This means that there exists a range of open-source packages to be able to support the implementation of DP in AI models utilising frameworks such as TensorFlow or PyTorch. And, because of the utilisation of current frameworks, it means that these techniques can often be easily implemented by swapping out a non-private model for its private equivalent.

TensorFlow Privacy

Opacus

Google Differential Privacy Library

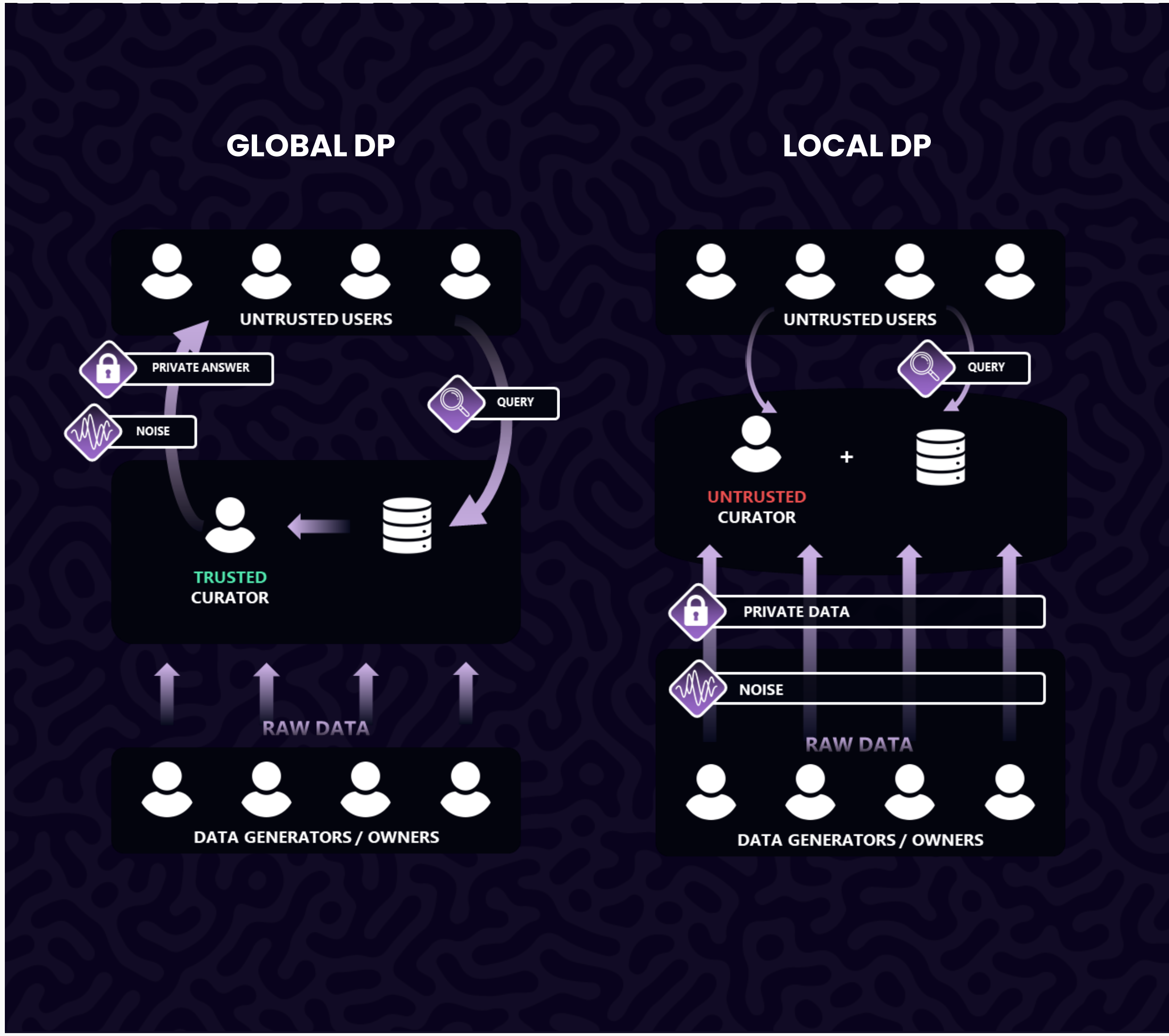
IBM AI Privacy Toolkit

Challenges in Neuroimaging & Genomics

Most demonstrations and applications of DP have been on tabular data where it is relatively easy to be able to add noise. However, in more complex data such as imaging or genomics, this becomes a greater challenge. This for one is predominately due to the size of these types of data which means that it adds extra complexity and computational requirements to be able to sufficiently use techniques such as synthetic data. Therefore, it is recommended that global DP is used for these data types as local DP is less feasible. However, in genomics data, although the original definition of local DP is not sufficient, a version named (ϵ,T)-dependent local DP was developed for genomics data [16].

Examples of DP in Practice

The use of DP-SGD in imaging data has been demonstrated several times, with one paper implementing a PyTorch DP framework for chest radiography classification and segmentation of computed tomography scans [17]. In the segmentation task, ROC-AUC performance between the private and non-private were on par with each other, whereas in the classification task, performance was reduced from 0.96 to 0.85. However, with more relaxed DP guarantees, this only reduced to 0.88. This shows how DP can be affected by the level of privacy guarantee and that it can vary between datasets and tasks so has to be carefully considered.



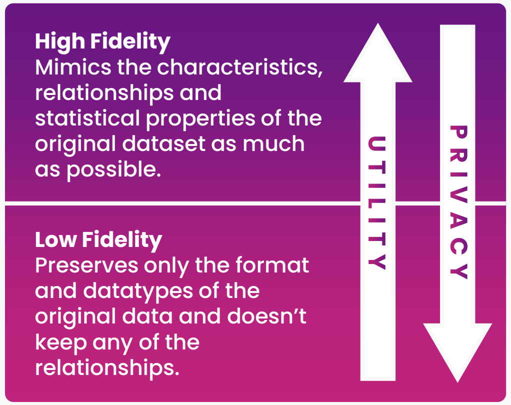
Privacy-Preserving Techniques

Synthetic Data

Summary

Synthetic data aims to generate artificially created data which replicates the statistical properties and patterns of the real data. This is usually done through training a generative model on some real data to learn the characteristics and structure of that data to be able to create new samples from it. Analysis of this type of data should produce similar results compared to using the original data but this depends on the level of synthetic data generated, and like differential privacy, there is a trade-off between privacy and utility depending on the fidelity of the synthetic data. The more the synthetic data mimics real data, then the more likely it is to reveal individuals' data.

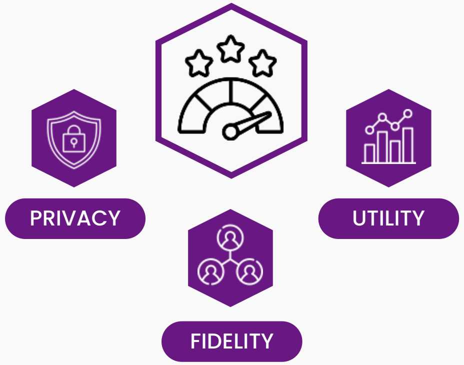
Synthetic data offers a way to create a private dataset which maintains the properties and relationships of the original data, but provides a privacy guarantee for individuals from the original data. This is usually achieved through generative AI models, such as Generative Adversarial Networks (GANs) or Variational Auto-Encoders (VAEs), which use AI techniques to learn from the original data to produce synthetic samples which mimic that data. In the case of GANs, they are composed of a generator and a discriminator. The generator aims to create synthetic data which cant be distinguished from the real data, whereas the discriminator aims to differentiate between the real and synthetic data. These compete with each other to refine their abilities to be able to create high quality samples which mimic the real data as much as possible. However, the more that synthetic data mimics the original data, the more likely it is to still reveal individuals' data from that dataset which is why additional techniques are often added to adjust the fidelity of the data generated. Fidelity in synthetic data refers to how closely it resembles the original data where low fidelity means that relationships aren't preserved between any of the columns in the data but still retains the structure, whereas high fidelity does capture those relationships and patterns .



High Fidelity
Mimics the characteristics, relationships and statistical properties of the original dataset as much as possible.

Low Fidelity
Preserves only the format and datatypes of the original data and doesn't keep any of the relationships.

Of course, when developing AI models, you still need high utility in your data to be able to have good performance which is why low utility synthetic data is of no use. So, a suitable trade-off between privacy and utility has to be found to be able to generate data which is both private but also statistically useful.



Evaluating Synthetic Data

There are three factors to consider when evaluating synthetic data – fidelity, utility and privacy and there are different ways of measuring these. Metrics to measure fidelity usually include some form of statistical similarity, boundary preservation and correlation similarity to evaluate the quality of the data generated and how similar it is to the original. To evaluate the utility of a synthetic dataset, we can evaluate its performance in something like an AI prediction task and compare it to using the original dataset. And to measure privacy, we can check if there are any rows which match, how novel new samples are, but also by running inference attacks to measure how well it protects against these. All of these factors have to be taken into consideration when generating synthetic data to ensure that it is usable, but also private.

Synthetic Data Tools

Thankfully, there are several tools available to generate and evaluate synthetic data. These typically utilise variations of GAN models which users can use to train on their data and generate synthetic samples. Some tools also allow for temporal synthetic data generation which is useful for time-series and longitudinal data.

YData-Synthetic

Gretel Synthetics

Synthetic Data Vault

Privacy-Preserving Techniques

Homomorphic Encryption (HE)

Summary

Homomorphic encryption provides high protection while retaining utility as it enables computations to be performed on encrypted data without the need of having to decrypt it. Although this is the most ideal solution, this method is currently very limited in its abilities in AI.

HE works by using a public key-generation algorithm where the public key is used to encrypt the data and the private key is used to decrypt the result. Typical encryption algorithms such as AES and RSA are not able to be used in HE as computations cannot be performed on this type of encryption, however there are some common HE schemes which can be used depending on the computations you want to perform. These can be performed at the data stage, training stage, or to actually encrypt the model itself.

Challenges and Limitations

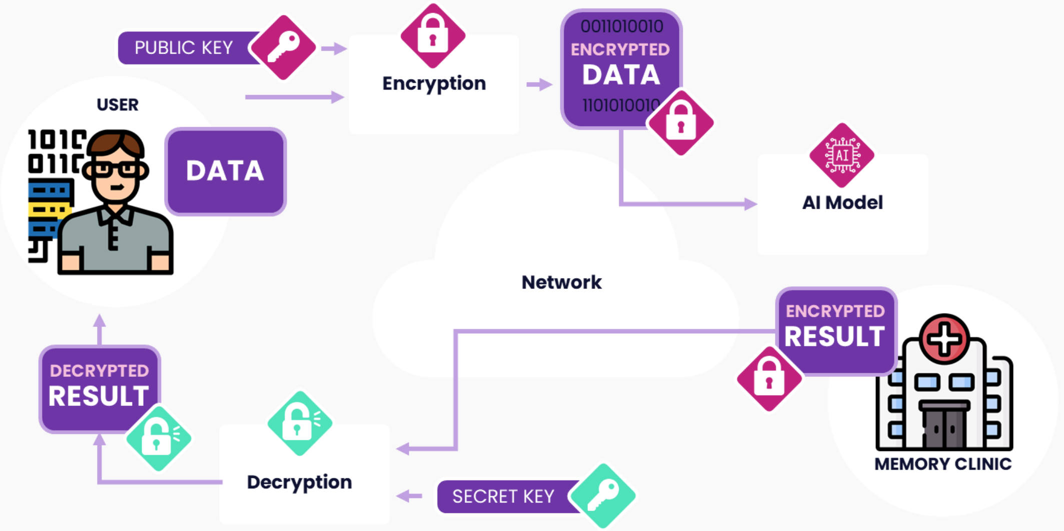
FHE techniques are relatively new and emerging so currently are often very slow and require significant computational resources. Additionally, they can also be difficult to implement as they require specialist knowledge and are often limited in the types and amount of operations that can be performed efficiently. This is why so far, HE tools to protect training data has only really been demonstrated in logistic regression and neural network models.

However, where HE is used more often is in encrypted inference to provide protected predictions on encrypted inputs. This allows an AI model to be queried without sharing data to get a prediction, but means that the model itself is still vulnerable to attacks as the training data is still unprotected.

Type of HE	Description
Fully HE (FHE)	Allows an unlimited number of computations but can increase resource and time required.
Somewhat HE (SHE)	Allows a limited number of additions and multiplications.
Partially HE (PHE)	Only allows either addition or multiplication and not both. Something like Paillier encryption can be used.

HE Tools

- TensorFlow Encrypted
- Microsoft SEAL
- Pyfhel
- ZAMA

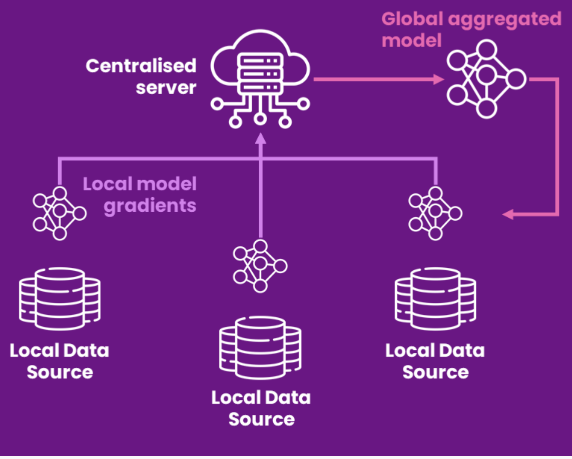


Privacy-Preserving Techniques

Mitigations at the Sharing / Release Stage

Federated Learning

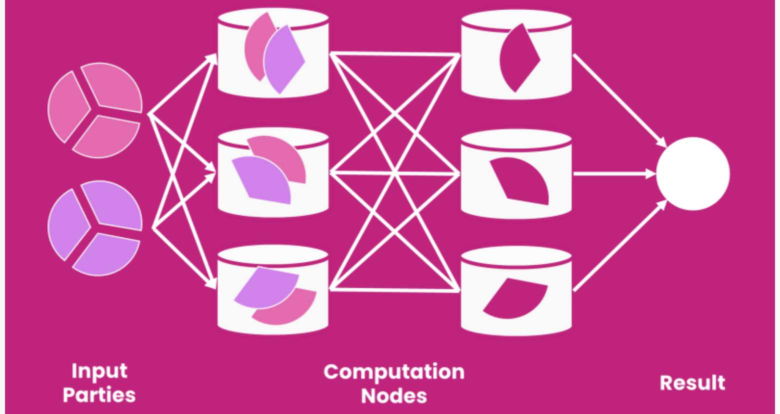
Federated learning (FL) allows multiple parties to train AI models on data from multiple sources, without having to share their local data. In centralised FL, gradients from each of the local AI models are sent to a server, where they are aggregated into a single global model, which is then sent back to the local sources to further develop. This process is repeated iteratively until the global model is refined and improved. In decentralised FL, instead of having one singular coordination server, the gradients are sent out to each local source where they all update the global model directly. Each version has its advantages and disadvantages. In centralised FL for example, you only need to trust the one server, whereas in decentralised FL you need to trust all parties involved. However, in decentralised FL there is no single point of failure. But, no matter what version is chosen, the AI model gradients and updates are still being shared between servers, and therefore is still vulnerable to privacy attacks unless additional privacy-preserving techniques are implemented as well.



Secure Multi-Party Computation

Secure Multi-Party Computation (SMPC) allows parties to jointly train an AI model on private inputs without revealing those inputs to the other parties. This is often achieved through "secret sharing" where the data is divided and distributed among all parties or used in combination with homomorphic encryption.

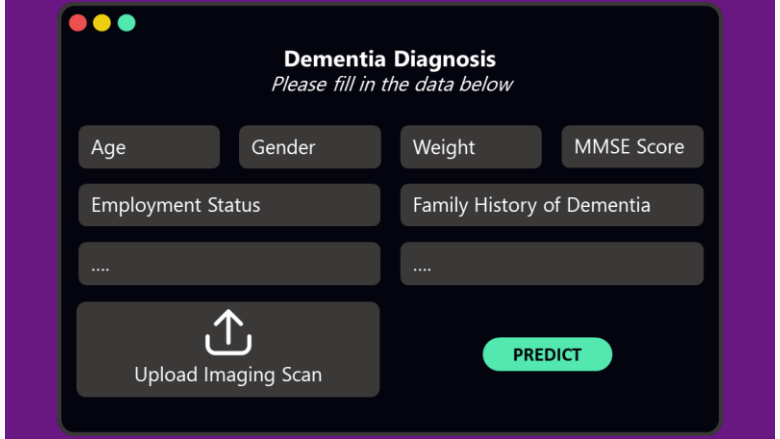
However, just like HE, this requires the expertise and resources to be able to sufficiently implement. Additionally, depending on how it is implemented, if the number of shares is above a certain amount, then the input data could potentially be reconstructed. But again, just in FL, it all depends on trust with the other parties.



Secure Web Hosting

If an AI model is ready to deploy, one option could be to host that model with restricted access and queries. This would mean that the AI model would stay within the TRE, and could only be queried through a web interface or an API. By imposing access and query controls, it means that the model can only be used by approved users, and attacks are prevented because of the query restrictions.

If an adversary did somehow manage to be able to query the model. Then they would only be able to run black-box attacks as they wouldn't have direct access to the model. This makes attacks more difficult to perform as the adversary only has the outputs from the model to attack, therefore limiting their capabilities.



Dementia Diagnosis

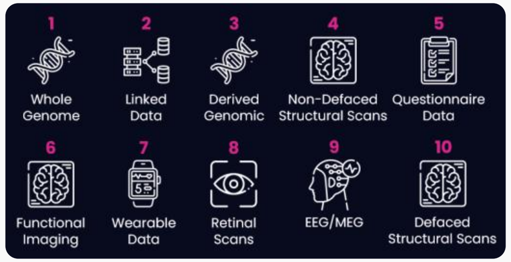
Please fill in the data below

Age	Gender	Weight	MMSE Score
Employment Status		Family History of Dementia	
....		....	

Researchers Opinions on Risks & Mitigations

The researcher workshop brought together researchers working in the field of AI with the aim to assess their awareness of privacy risks associated with AI model development, their levels of concern regarding data types, their confidence in using privacy preserving techniques and their willingness to incorporate privacy preserving techniques into their AI model developments.

The workshop started with talks around AI model risks and the consequences of AI model release. After the presentations researchers were asked a series of questions to determine their concerns around AI model development and awareness of risks. Regarding the data types used in AI model development, the researcher group felt that whole genome sequencing data was the riskiest type of data for AI model development and defaced structural imaging scans were of least concern with questionnaire/assessment data falling within the middle.

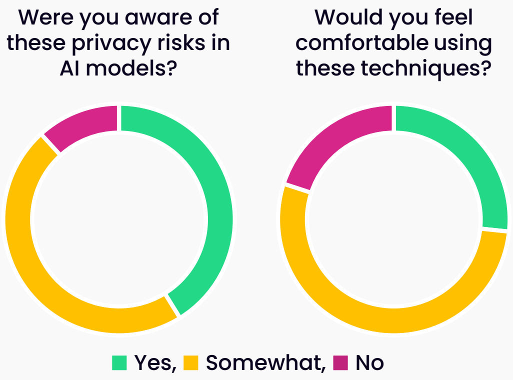


From the discussions, researchers said that *"If you are trying to find out more information about someone than a brain scan is less likely than genomics"* and that *"Genomics have greater impact and potential repercussions"*, however it *"depends if using raw or derived"*.

Some researchers felt that it was *"dangerous to assume one type of data is safer than another"* and that the risk of these data types is linked to other data such as if you *"have access to family data then can look at matching for genomics"*.

Regarding types of AI models, they felt that instance based models were of greatest concern and that linear and logistic regression models were of least concern. So, the group seemed to recognise the different risks involved in different AI models.

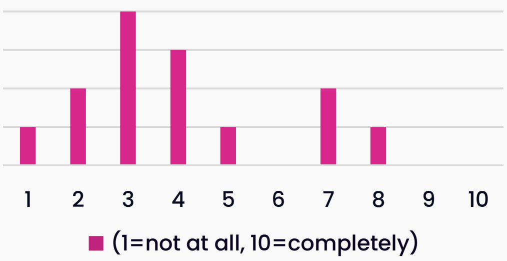
Due to the theme and agenda of this workshop, it was likely to attract researchers who have at least some understanding of privacy concerns prior to their attendance. Despite this, ~11.9% of participants had no awareness of privacy concerns in AI models and their potential vulnerabilities, and ~47% had a vague understanding. However, ~41.1% of participants confirmed that they were aware of some of these prior to their attendance.


Were you aware of these privacy risks in AI models?

Would you feel comfortable using these techniques?

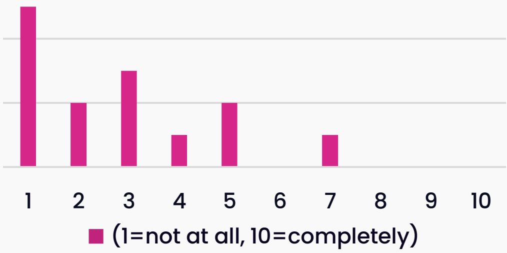
However, most researchers felt that they weren't comfortable enough to be able to implement privacy-preserving techniques to mitigate these risks. Additionally, we asked if they thought researchers in general had the expertise to implement these techniques and found that ~78.6% scored 5 or under. During the discussion, this was put down *to "a lack of experience"*, it being *"not common practice currently"*, and not being a *"mature space yet"*.

Do researchers have the expertise/knowledge to implement mitigations?



These results suggest that awareness could be a major barrier to overcome for researchers in order for them to begin implementing privacy-preserving techniques, along with developing expertise in implementation.

Are there enough training and resources available for researchers?



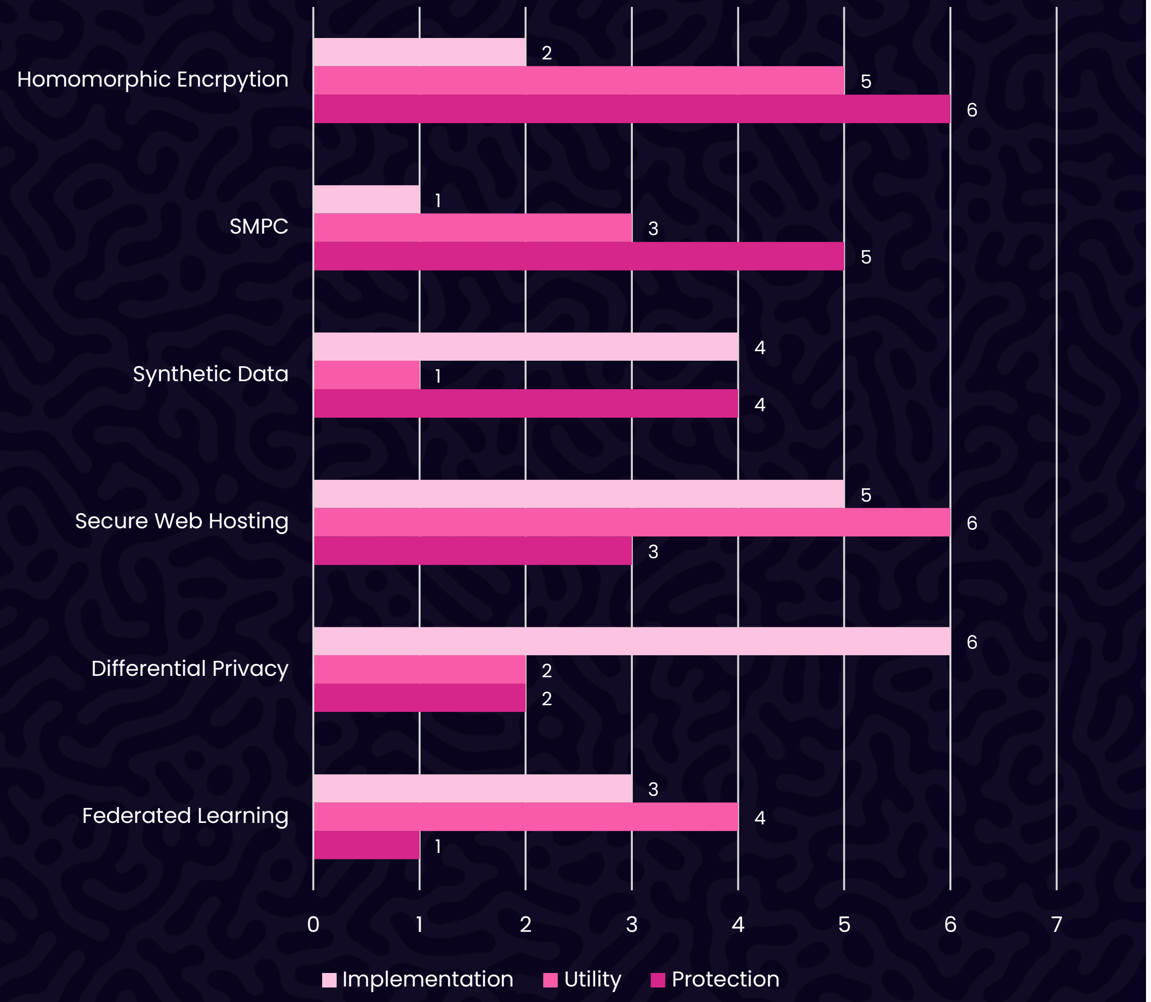
This is further backed up from the results showing that researchers don't think there is enough training and resources available to learn about implementing these techniques.

During the discussions, researchers discussed several barriers and challenges for them to implement mitigations in their AI models. One researcher said that they *"don't know which techniques are best to use given the type of model and type of data"* and another said they would find it *"hard to know how to balance privacy/utility"*. This was mainly put down to *"Guidance and resources not being available"*.

There were also concerns around the effect that these mitigations may have on their research. Several researchers were worried if *"extra computational power is needed"* then *"would researchers need to pay more for extra compute"* to be able to implement some of these techniques. They were also worried about the extra time and effort that it may add to a project.

The effect on the utility of their AI models was also a key concern with several researchers saying – *"There is no point having a private model with no utility"* and *"If accuracy is so low then it is pointless, need to have a balance"* however, others had the opinion that *"robust models should be able to handle some noise in the data anyway"* and that *"privacy has to be preserved so the accuracy is what it is. Might push for more robust models"*. So some researchers thought that the implementation of these techniques would also lead to more robust solutions.

Researchers Opinions on Risks & Mitigations



Average ranking of three different aspects in privacy-preserving techniques

We asked researchers to rank various mitigations based on three different aspects – implementation (1 = hardest, 6 = easiest), utility (1 = lowest, 6 = highest) and privacy protection (1 = lowest, 6 = highest) to gain an understanding of what techniques they thought were most effective in these different areas and the feasibility of them.

Regarding privacy protection, techniques such as homomorphic encryption (HE) and SMPC were ranked the highest, whereas federated learning and differential privacy were ranked the lowest. HE and SMPC also ranked highly regarding utility as they often don't have an affect on the model performance, however, these methods ranked the lowest regarding implementation as they were considered the hardest to implement.

Secure hosting was ranked the highest for utility as performance isn't affected in this scenario, whereas techniques such as synthetic data and differential privacy ranked the lowest as these typically reduce the performance of models. However, these techniques were ranked as some of the easiest to implement.

This shows the trade-off that needs to be considered regarding ease of implementation and the level of privacy protection that these models give. From these results, although HE and SMPC offer some of the highest privacy protection, they are a lot more difficult to implement, meaning that techniques such as synthetic data and differential privacy may be more feasible for researchers to use.

As part of this work, we developed three different AI models – an SVM, Random Forest and Neural Network, and implemented differentially private versions of them to test how utility was affected but also how well they protected against membership inference attacks.

	SVM		Random Forest		Neural Network	
	Normal	Safe	Normal	Safe	Normal	Safe
Classification F1-Score	0.92	0.6	0.89	0.79	0.8	0.735
Attack Metric	0.85	0.57	0.76	0.74	0.52	0.49

This shows the effect that implementing privacy-preserving techniques can have on the utility of a model. However, it is also important to consider how well it protects privacy for a given model as for the Random Forest, the performance of the attack didn't significantly reduce enough to protect privacy. As for the SVM model, the attack performance did significantly reduce, but so did the accuracy of that model to the point where utility is severely impacted. So the type of privacy-preserving techniques a researcher should implement really depends on the type of model and data used. Researchers wished to see more research in this area to evaluate different mitigations against different attacks for a range of models and data to have a comprehensive overview of the best techniques for a given situation, so this is something that we will be working on extending.

CHAPTER 4

Perspectives of Data Providers

Assessing Risk Appetite of Data Providers

We brought together data providers and TREs to discuss the findings from the previous workshops and to refine recommendations for developing safe AI on their data. The main aims of this workshop were to:

Objective 1: Discussion on Workshop Findings
Discuss the results from the public and researcher workshops to find out what data providers think of AI model privacy risks and the implementation of mitigations.

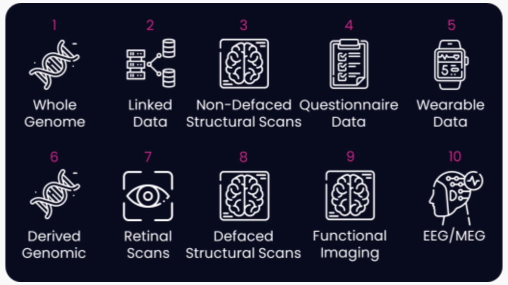
Objective 2: Assessing Risk Appetite
Assess the risk appetite of data providers and how they feel about AI models being trained on their data and released from a TRE.

Objective 3: Evaluating Assessment Strategies
Evaluate the most effective ways to assess AI models for release and how we can quantify risk to help data providers make informed decisions.

Objective 4: Refining Recommendations
Refine recommendations from the public and researcher workshops to create a framework for developing and releasing AI models which ensure data providers data is protected.

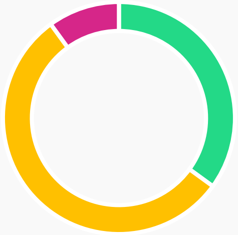
Risk Appetite & Concerns

During this workshop, we wanted to understand the risk appetite of data providers, and how they felt about AI development on their data. As we did in the previous two workshops, we asked data providers to rank data types based on their disclosure risk. The results from this show similar rankings to the public and researcher results where whole genome and linked data is ranked highly, whereas neuroimaging data was ranked fairly low.

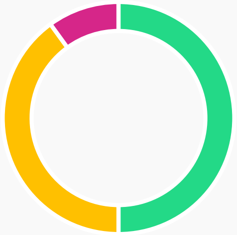


Additionally, when they were asked whether neuroimaging and genomics pose unique risks to privacy, a majority felt that they did.

Does neuroimaging pose unique privacy risks?



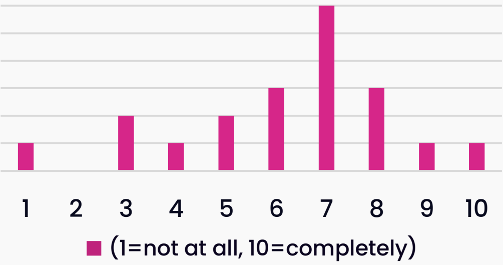
Does genomics data pose unique privacy risks?



🟩 Yes, 🟨 Somewhat, 🟥 No

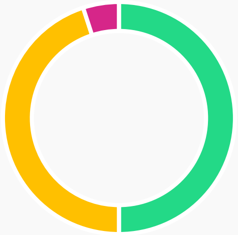
Despite a majority of data providers being concerned about the development of AI on their data, most of them were happy for AI research to currently take place within TREs.

How concerned are you about the development of AI on your data?

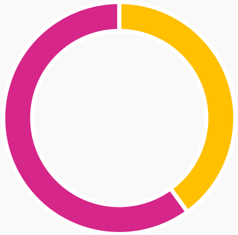


However, a clear majority felt that they weren't properly equipped to be able to assess AI projects as they didn't have the necessary expertise to judge the risks of these projects. During the discussions, it was apparent that data providers felt they need help in making these decisions from experts in the field to confidently assess AI projects to use their data.

Are you currently happy for AI research to take place in TREs?



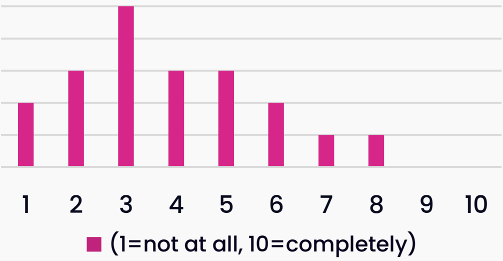
Do you feel equipped to be able to assess AI projects?



🟩 Yes, 🟨 Somewhat, 🟥 No

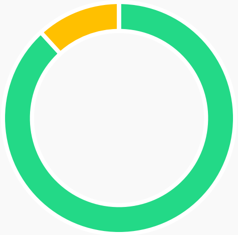
Regarding the assessment of AI models for release, it was also clear that most data providers weren't comfortable in confidently making decisions to allow this to happen.

How comfortable would you feel assessing an AI model for release?

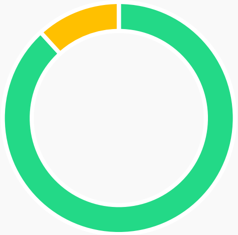


This was mainly due to not understanding the risks involved in releasing AI models and how privacy-preserving techniques mitigate these, which is why most data providers felt that there should be training and resources for them, as well as researchers, to help make decisions on AI model development and release.

Are privacy-preserving tools essential to mitigating risks?



Should there be training for data providers to help make decisions?



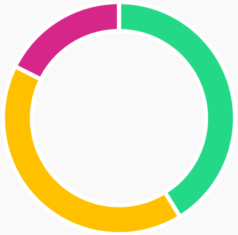
🟩 Yes, 🟨 Somewhat, 🟥 No

Furthermore, they acknowledged the role of privacy-preserving techniques in mitigating risks in AI models, with a majority saying that they are essential to ensuring models are safe for release. Data providers felt that privacy-preserving techniques would enable them to feel more confident in allowing AI models to be released, and therefore, should be necessary in AI projects using their data.

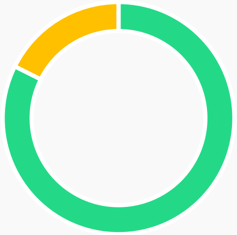
However, they also realised the potential challenges of employing these techniques for more complex data such as neuroimaging and genomics which could make the implementation of them less feasible.

Additionally, although most data providers favoured the use of these techniques, some had concerns around the effect they could have on the restriction of research. They felt that privacy concerns shouldn't hinder the ability to do research and that we should do everything we can to make it as easy as possible for the researcher to implement these techniques.

Does neuroimaging pose unique challenges in mitigations?

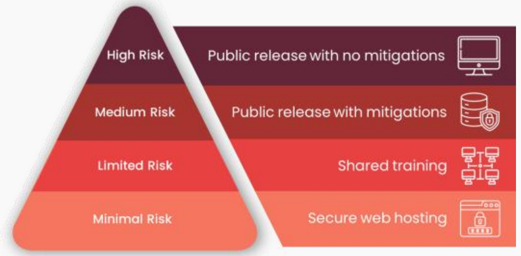


Does genomics pose unique challenges in mitigations?



🟩 Yes, 🟨 Somewhat, 🟥 No

Assessing AI Models & Privacy Risks



From the workshop, we identified different levels of risk depending on the release scenario of an AI model. High risk was determined to be a public release with no privacy-preserving techniques implemented, and even public release with mitigations was identified as medium risk. Scenarios such as federated learning were identified as having limited risk if parties are trusted, and minimal risk were scenarios such as clinical deployment of an AI model. These risk scenarios play a crucial role in determining the safety of releasing AI models from a TRE and was ranked as one of the more important aspects to consider when making a decision for release.

Considerations for Release	Rank
Attacking Model	1
Assessing Model Used	2
Assessing Release Scenario	3
Assessing Data Used	4
Risk Impact Assessment Form	5
Agreements & Licenses	6

The most important consideration when deciding to release an AI model was determined to be running privacy attacks to effectively quantify the potential risk. Data providers felt that this offered the best way to judge whether an AI model would be safe enough to release outside of a TRE as it provides assurances that it can protect against attacks.

Apart from attacks, they felt that it was important to consider the type of model and data used as these can have an impact on the potential risks involved in releasing a model. For example, if a decision tree model was developed on derived structural neuroimaging data then this wouldn't pose as much of a privacy risk compared to an instance-based model being trained on questionnaire and genomic data. Therefore, the types of data and models used have to be carefully considered when assessing the risk of an AI model. Ranked as the lowest consideration was whether there would be any agreements or licenses in place for the release of an AI model.

Confidence in Mitigations	Rank
Homomorphic Encryption	1
Synthetic Data	2
Secure Hosting	3
SMPC	4
Differential Privacy	5
Federated Learning	6

When asked to rank privacy-preserving techniques, its important to note that some data providers declined to answer as they felt that they still didn't have adequate knowledge to be able to rank the effectiveness of these. This further demonstrates how they require resources to help them make decisions on assessing effective whether AI models are safe enough for release. However, data providers felt that if researchers implemented techniques such as HE, synthetic data and secure hosting, then that would give them the confidence to allow that AI model to be released. Whereas scenarios such as federated learning wouldn't give them the necessary guarantees and additional privacy-preserving techniques would need to be implemented.

From the discussions, responsibility was also discussed regarding the potential disclosure risk of AI models. Ultimately, data providers are the ones who own the data and have responsibility of making the final decision, but additionally, it is the researcher who has created that AI model and will be using it, and the TRE which has been entrusted to look after that data. Therefore, it was determined that there should be a shared responsibility between all three to ensure that AI models are released safely.

CHAPTER 5

Recommendations

How to Allow the Safe Development & Release of AI

Assessment

Public Engagement

Safe Data Tools

Attack Simulations

Risk Index Evaluation

Training & Resources

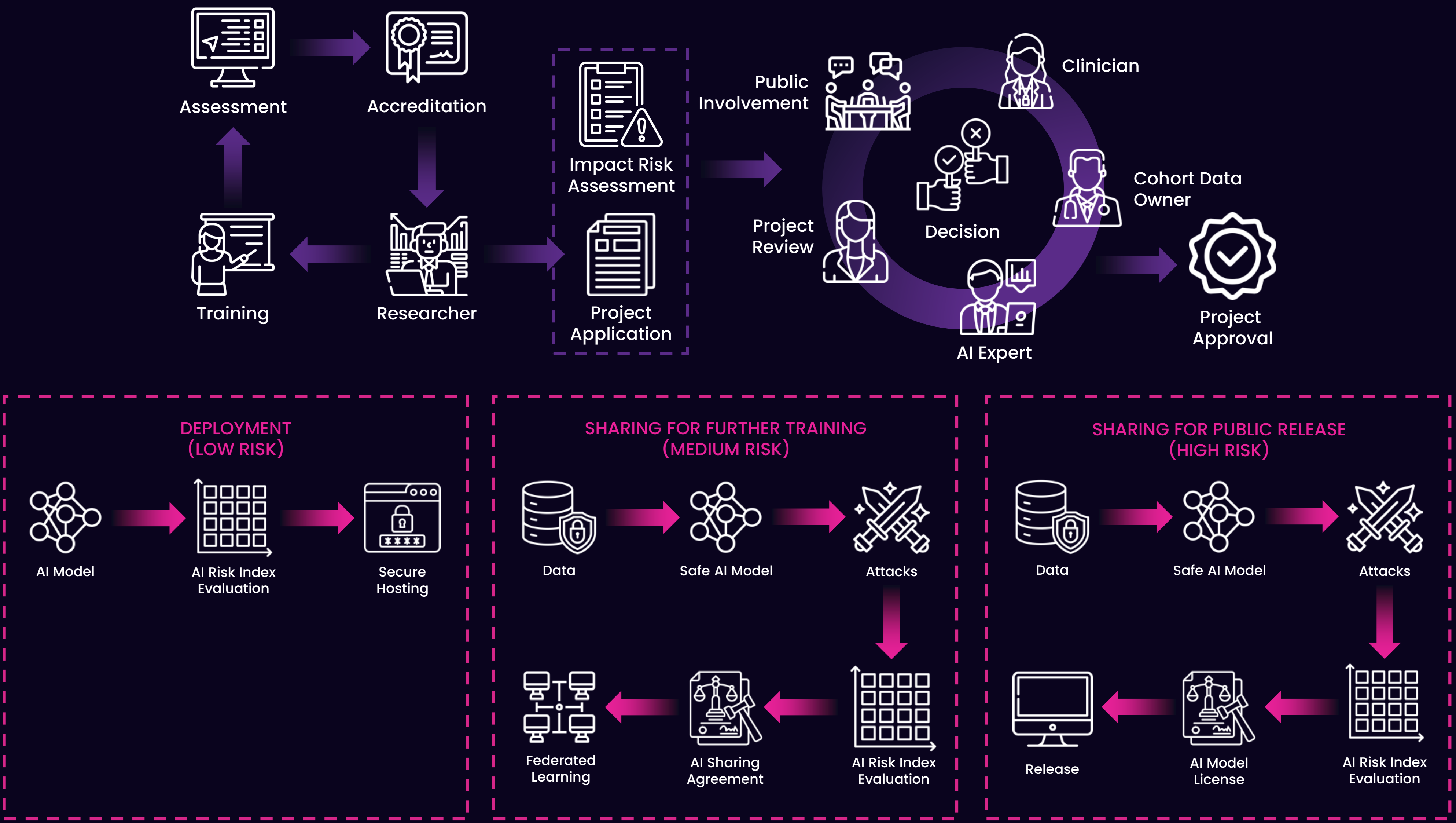
Accreditation

Risk Impact Assessment

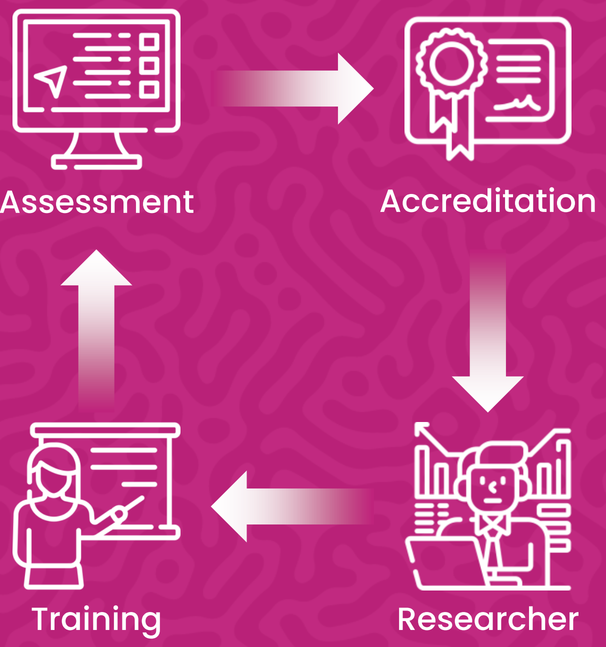
Secure Hosting Service

User Legal Agreement

Recommendation Framework



Recommendation 1
Training & Accreditation



From the researcher workshop, it was clear that researchers lack the knowledge and expertise necessary to be able to implement privacy-preserving techniques in their AI research. Training is therefore a key recommendation to enable researchers to learn about these risks and mitigations and to be able to implement them appropriately.

We found that most researchers were unaware of the privacy risks involved in AI models, and therefore awareness should be at the forefront of training so that researchers can gain an understanding of the potential risks in their research. Data providers also reported a lack of knowledge around AI model risks and felt that they would also benefit from learning resources to support them making their decisions on project approval. However, its not just researchers and data providers which lack knowledge, but also TRE staff on how to run attacks and evaluate AI models for safe release, so training should take into consideration all three of these distinct groups.

Additionally, researchers felt that currently there are not enough resources to learn about privacy-preserving techniques and how to implement them, demonstrating that this is a significant barrier. They especially felt that there were no resources for more complex data such as neuroimaging and genomics and how to select the best techniques given the data and model used in their research. Therefore, training and resources need to focus on giving researchers the skills on how to best select the most appropriate privacy-preserving techniques for their research. They also wished to see examples of these techniques in practice so that they could see how they can be implemented and used.

From this, we recommend that there should be some form of an AI Risk Learning Platform, where researchers can gain this knowledge and have access to an environment with synthetic data where they can test out different techniques and learn which ones may be best for them. We see this platform as an environment where researchers can run attacks on models to learn about the risks involved in AI and where they will gain skills to mitigate these risks through testing out different strategies and their effect on attacks. From this, researchers will be able to identify the most suitable techniques for their research and how to effectively implement them.

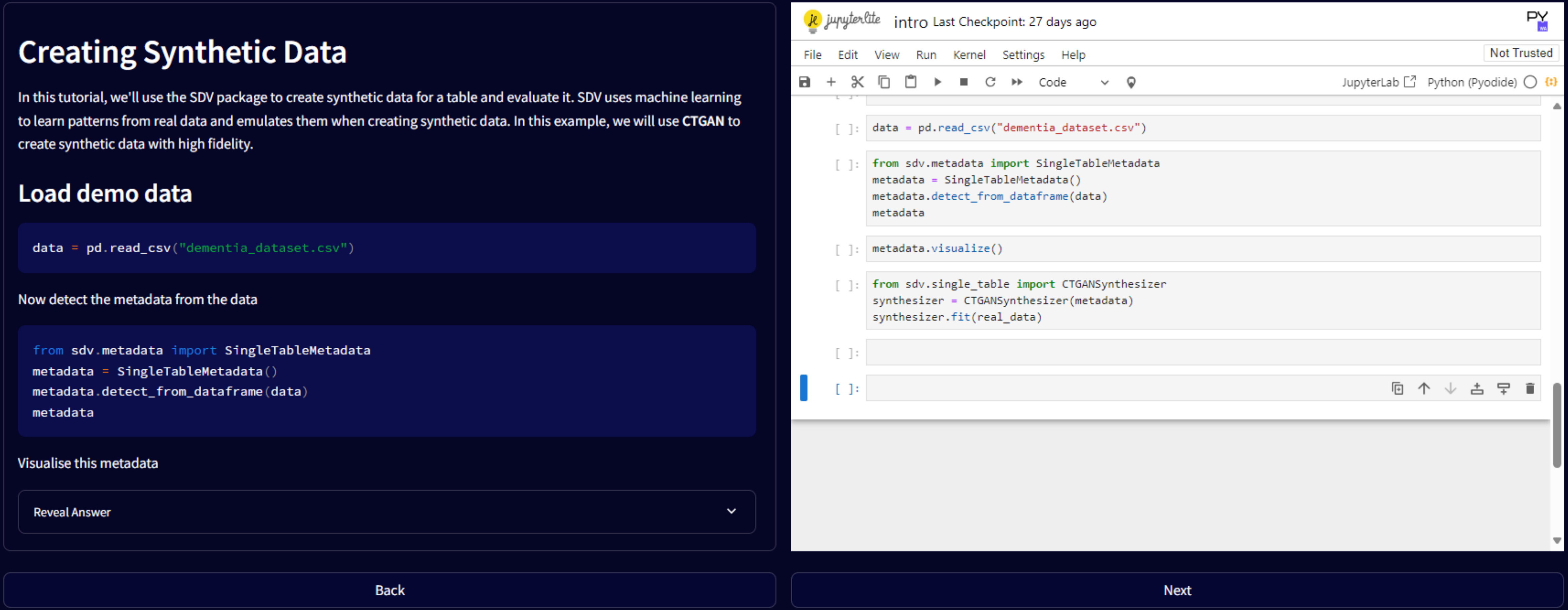
The establishment of an AI-specific accreditation, similar to the ONS Safe Researcher accreditation, was also proposed as a way to assess researchers' competencies in protecting patient privacy in their AI models. This accreditation would ensure that researchers possess the necessary knowledge and skills to identify and mitigate disclosure concerns in AI models before they are released.

This would require some form of assessment to test researchers abilities and knowledge. To be able to give TRE's and data owners proper assurance, this should cover:

- The risks of sharing AI models
- The vulnerabilities of different types of data and models
- What privacy-preserving techniques exist
- Which techniques should be used in what scenario

If researchers have this level of understanding, then we can be sure that they will keep privacy in mind when developing their AI models.

Practice Environment



Creating Synthetic Data

In this tutorial, we'll use the SDV package to create synthetic data for a table and evaluate it. SDV uses machine learning to learn patterns from real data and emulates them when creating synthetic data. In this example, we will use **CTGAN** to create synthetic data with high fidelity.

Load demo data

```
data = pd.read_csv("dementia_dataset.csv")
```

Now detect the metadata from the data

```
from sdv.metadata import SingleTableMetadata
metadata = SingleTableMetadata()
metadata.detect_from_dataframe(data)
metadata
```

Visualise this metadata

Reveal Answer

```
[ ]: data = pd.read_csv("dementia_dataset.csv")

[ ]: from sdv.metadata import SingleTableMetadata
     metadata = SingleTableMetadata()
     metadata.detect_from_dataframe(data)
     metadata

[ ]: metadata.visualize()

[ ]: from sdv.single_table import CTGANSynthesizer
     synthesizer = CTGANSynthesizer(metadata)
     synthesizer.fit(real_data)

[ ]:

[ ]:
```

Recommendation 2

AI Risk Impact Assessment

An AI risk assessment is a crucial step that should be undertaken when an AI model has the potential to be released from the TRE at the end of its development. This assessment plays a crucial role in ensuring the safety and responsible use of AI models, as it prompts researchers to critically evaluate potential vulnerabilities and challenges. By conducting a comprehensive AI risk assessment, researchers are required to engage in a thorough analysis of their AI models' characteristics, intended applications, and potential impacts. This analysis helps them identify potential risks and vulnerabilities that could arise during the deployment and use of the AI models in real-world scenarios.

To effectively mitigate potential risks, researchers should be asked how they will implement robust mitigation strategies to prevent these risks. These strategies should include privacy-preserving techniques and establishing clear guidelines for the use and deployment of the AI model. Furthermore, researchers should be required to demonstrate how they will ensure the ongoing safety and responsible use of their AI models throughout their lifecycle.

This should be carried out at the pre-project stage, alongside a project application to go to the project application decision process. However, there should also be an additional impact assessment performed at the end of the project, once the model is ready to be released. During the project, their intended plans may have adapted or changed, so this will identify what steps researchers have actually taken to ensure that there AI models are safe and what vulnerabilities they may have identified from their final model.

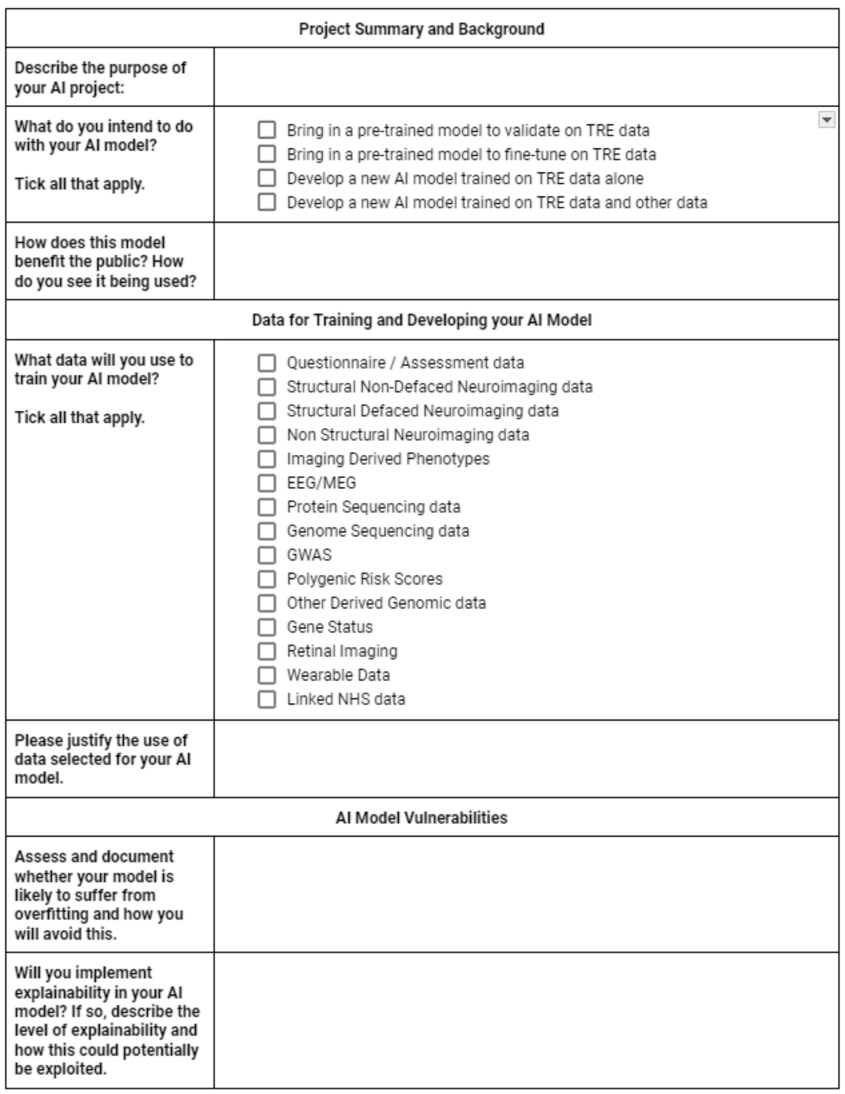
As part of this post-project assessment, TRE staff will also have a section to fill in with results from their evaluations and attacks of the AI model requested for release. Therefore, this form will provide a detailed overview on the safety of an AI model which can be used in the decision making process of whether to allow it to be released from the TRE.

Pre-Project Assessment Form	**Page 29**
Post-Project Assessment Form	**Page 30**

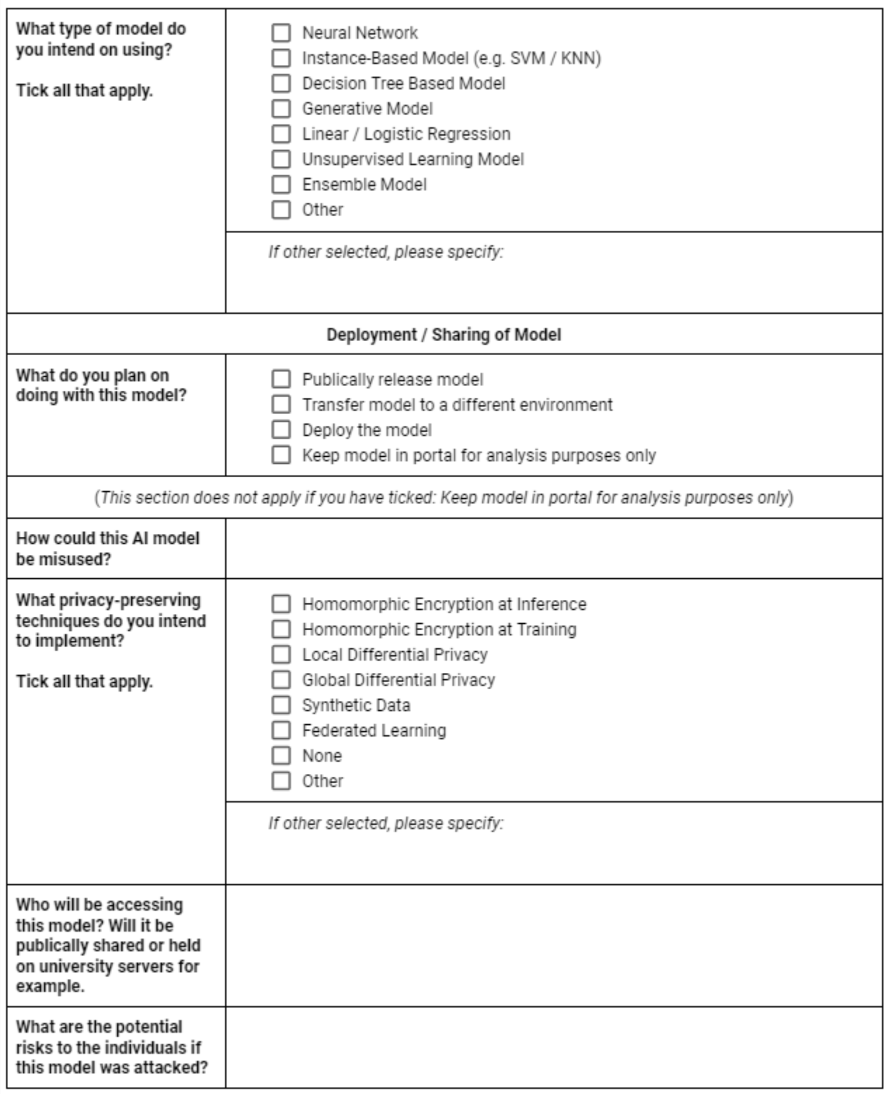
Recommendation 2

AI Risk Impact Assessment

Pre-Project Form



Project Summary and Background	
Describe the purpose of your AI project:	
What do you intend to do with your AI model? **Tick all that apply.**	☐ Bring in a pre-trained model to validate on TRE data ☐ Bring in a pre-trained model to fine-tune on TRE data ☐ Develop a new AI model trained on TRE data alone ☐ Develop a new AI model trained on TRE data and other data
How does this model benefit the public? How do you see it being used?	
Data for Training and Developing your AI Model	
What data will you use to train your AI model? **Tick all that apply.**	☐ Questionnaire / Assessment data ☐ Structural Non-Defaced Neuroimaging data ☐ Structural Defaced Neuroimaging data ☐ Non Structural Neuroimaging data ☐ Imaging Derived Phenotypes ☐ EEG/MEG ☐ Protein Sequencing data ☐ Genome Sequencing data ☐ GWAS ☐ Polygenic Risk Scores ☐ Other Derived Genomic data ☐ Gene Status ☐ Retinal Imaging ☐ Wearable Data ☐ Linked NHS data
Please justify the use of data selected for your AI model.	
AI Model Vulnerabilities	
Assess and document whether your model is likely to suffer from overfitting and how you will avoid this.	
Will you implement explainability in your AI model? If so, describe the level of explainability and how this could potentially be exploited.	

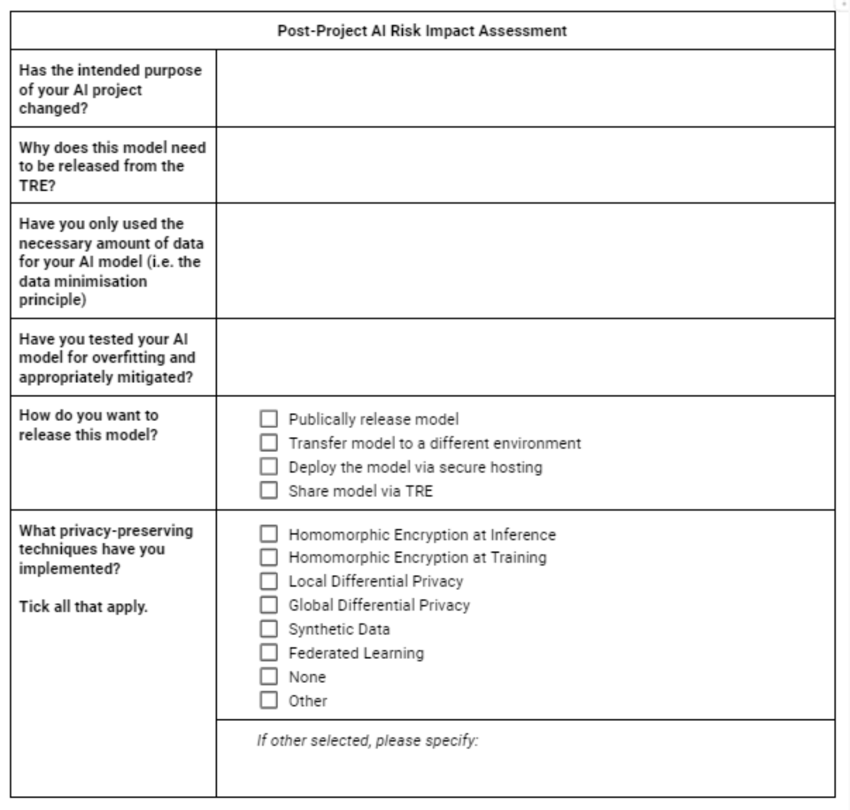


What type of model do you intend on using? **Tick all that apply.**	☐ Neural Network ☐ Instance-Based Model (e.g. SVM / KNN) ☐ Decision Tree Based Model ☐ Generative Model ☐ Linear / Logistic Regression ☐ Unsupervised Learning Model ☐ Ensemble Model ☐ Other
	If other selected, please specify:
Deployment / Sharing of Model	
What do you plan on doing with this model?	☐ Publically release model ☐ Transfer model to a different environment ☐ Deploy the model ☐ Keep model in portal for analysis purposes only
(This section does not apply if you have ticked: Keep model in portal for analysis purposes only)	
How could this AI model be misused?	
What privacy-preserving techniques do you intend to implement? **Tick all that apply.**	☐ Homomorphic Encryption at Inference ☐ Homomorphic Encryption at Training ☐ Local Differential Privacy ☐ Global Differential Privacy ☐ Synthetic Data ☐ Federated Learning ☐ None ☐ Other
	If other selected, please specify:
Who will be accessing this model? Will it be publically shared or held on university servers for example.	
What are the potential risks to the individuals if this model was attacked?	

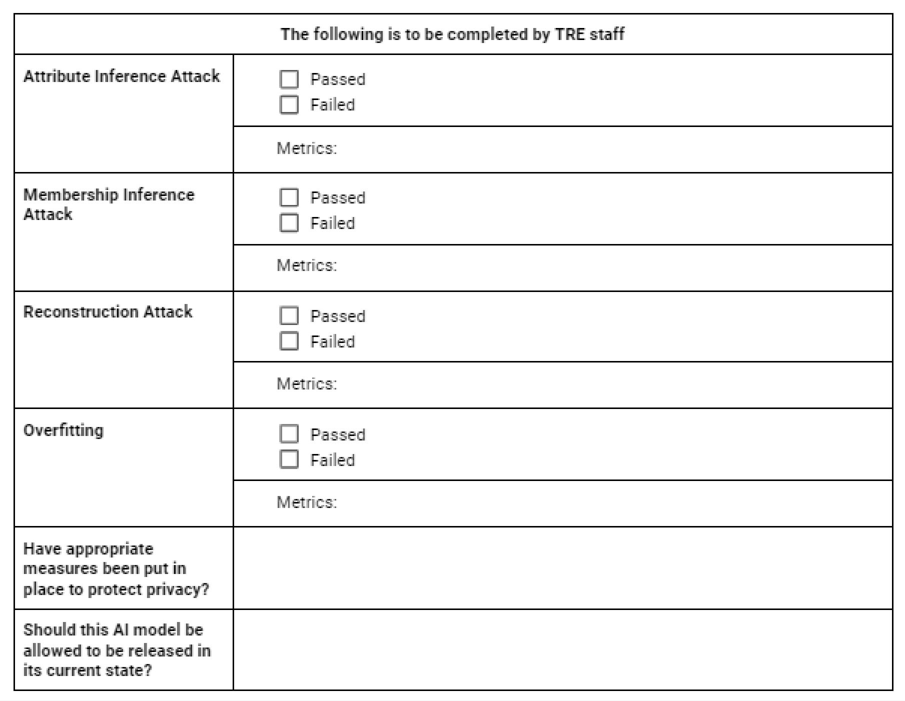
Recommendation 2

AI Risk Impact Assessment

Post-Project Form

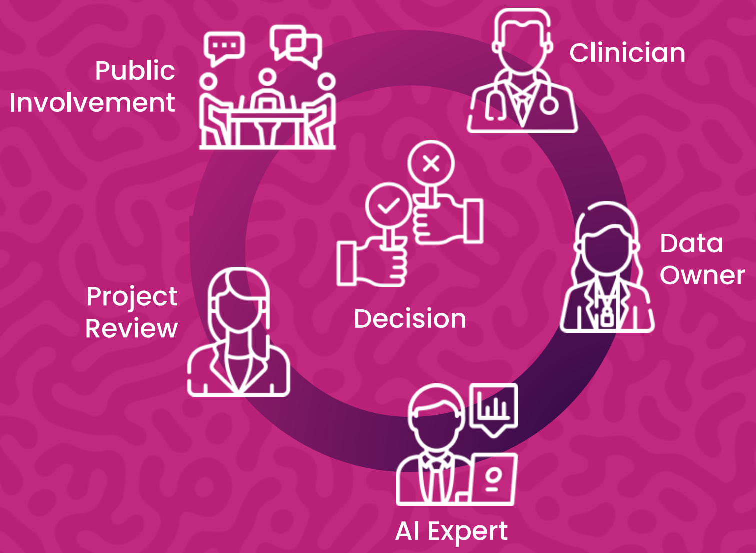


Post-Project AI Risk Impact Assessment	
Has the intended purpose of your AI project changed?	
Why does this model need to be released from the TRE?	
Have you only used the necessary amount of data for your AI model (i.e. the data minimisation principle)	
Have you tested your AI model for overfitting and appropriately mitigated?	
How do you want to release this model?	□ Publically release model □ Transfer model to a different environment □ Deploy the model via secure hosting □ Share model via TRE
What privacy-preserving techniques have you implemented? **Tick all that apply.**	□ Homomorphic Encryption at Inference □ Homomorphic Encryption at Training □ Local Differential Privacy □ Global Differential Privacy □ Synthetic Data □ Federated Learning □ None □ Other
	If other selected, please specify:



The following is to be completed by TRE staff	
Attribute Inference Attack	□ Passed □ Failed
	Metrics:
Membership Inference Attack	□ Passed □ Failed
	Metrics:
Reconstruction Attack	□ Passed □ Failed
	Metrics:
Overfitting	□ Passed □ Failed
	Metrics:
Have appropriate measures been put in place to protect privacy?	
Should this AI model be allowed to be released in its current state?	

Recommendation 3
The Decision Process



From the public workshop, it was identified that there should be greater public involvement in the decision making process of AI project applications. This would ensure that the model being developed is in the public benefit and assure the public that researchers will be taking appropriate measures to protect patient privacy in their models. It would also give them that sense of control which they felt they previously lacked.

Public engagement therefore, should be implemented at the early stages of the review process to provide data providers with public opinion to inform their decision. So as well as typical technical and scientific reviews, applications, along with a lay summary, should be sent for public review by at least one public member to give their input. Public and Patient Involvement and Engagement (PPIE) is increasingly an important part of most TRE's and is an area of growth across the landscape. Integrating the need for PPIE involvement in AI risk management should be introduced into these frameworks so it become a mainstream activity and evolves at the same rate of other risk management efforts.

However, from the data provider workshop, they felt like this wasn't enough and that two additional inputs were required to help enable them make decisions on AI projects - an AI expert, and a clinician.

Data providers felt that they lacked the expertise to be able to properly judge and assess AI model applications and the privacy risks of them, but traditional TRE staff also lack this expertise. Therefore, applications should require AI expert input to help assess the risk of that project and the privacy implications. This input would give data providers the confidence that they need to be able to appropriately make decisions on AI projects using their data. They also felt that input from clinicians would also help them make a decision on whether the AI model had any real-world benefit and offer valuable insights into clinical relevance, potential impact on patients, and alignment to current best practices and healthcare standards.

By combining public, AI expert, and clinician opinions, we can enable a comprehensive evaluation of the impacts and risks associated with AI model projects, and provide data providers with valuable insights to make informed decisions.

Recommendation 4

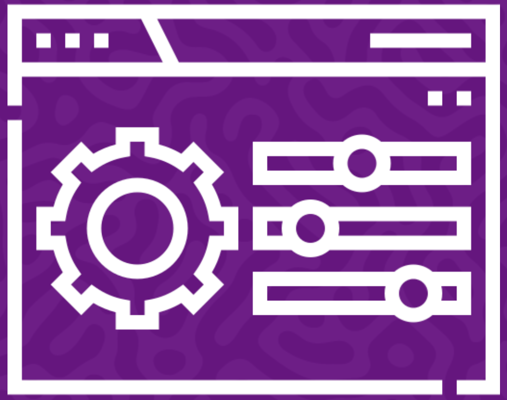
Generation & Evaluation Tools

The implementation of privacy-preserving techniques should be made as easy as possible for researchers to be able to incorporate into their models. From the workshop, it was identified that most researchers won't have the skills necessary to be able to implement these techniques, but also not be able to effectively evaluate their effectiveness either. This was particularly a problem when differential privacy and synthetic data was talked about as researchers felt they lacked the knowledge necessary to be able to determine effective trade-offs between privacy and utility. TRE's need to remove as many barriers as possible for researchers so that they can concentrate on the science and not have to spend a significant amount of time and resources trying to implement these techniques.

Therefore, we propose a set of tools which make the process of generating synthetic data, and data with differentially private guarantees, as simple as possible for the researcher, with ways to easily evaluate that data. As part of this work, we created a synthetic data generation and evaluation tool to be able to help researchers evaluate the privacy/utility trade-off in generating synthetic data. This tool allows them to select generative models and adjust certain parameters to generate synthetic data from a given dataset. The synthetic data can then be evaluated against the real dataset for a range of privacy and utility evaluation metrics, and combined into a score which can be visualised on a privacy/utility plot This visualisation easily allows the comparison between different synthetic datasets, so allows researchers to find a suitable trade-off.

With more complex data like genomics and neuroimaging, this becomes more difficult. However, Dementias Platform UK is currently working on methods to generate synthetic MRI scans which could also be useful for this purpose.

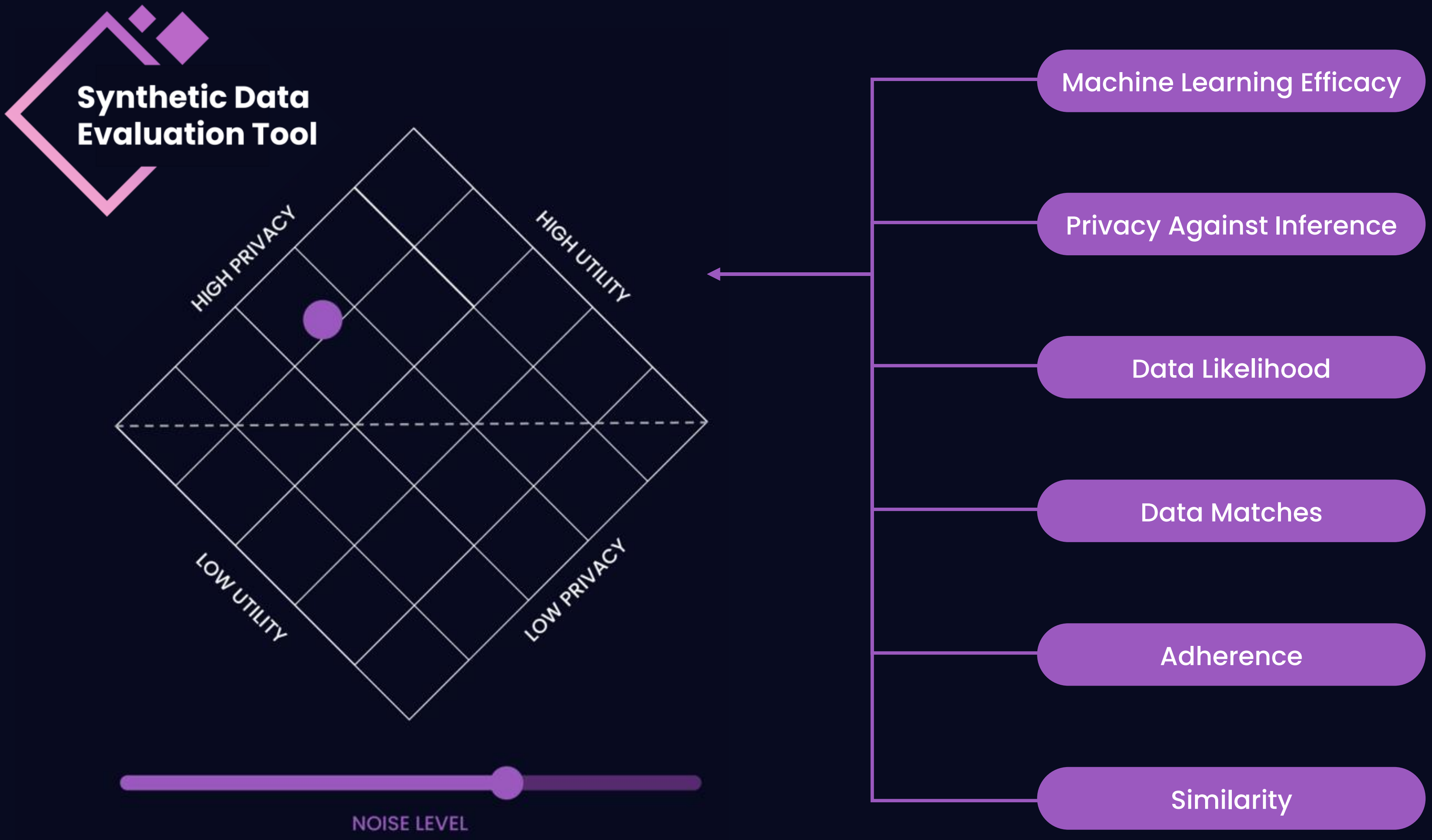
It was also suggested by researchers that TREs should offer "ready-made" synthetic data for each of their datasets which have already been evaluated and tested to ensure that they are private enough, but also offer the best utility. These "research ready" synthetic datasets make it easier for the researcher as they don't have to generate and evaluate the data themselves, and also gives the TRE confidence that the data being used is safe as it has already been subjected to rigorous testing and governance assurance.



Recommendation 4

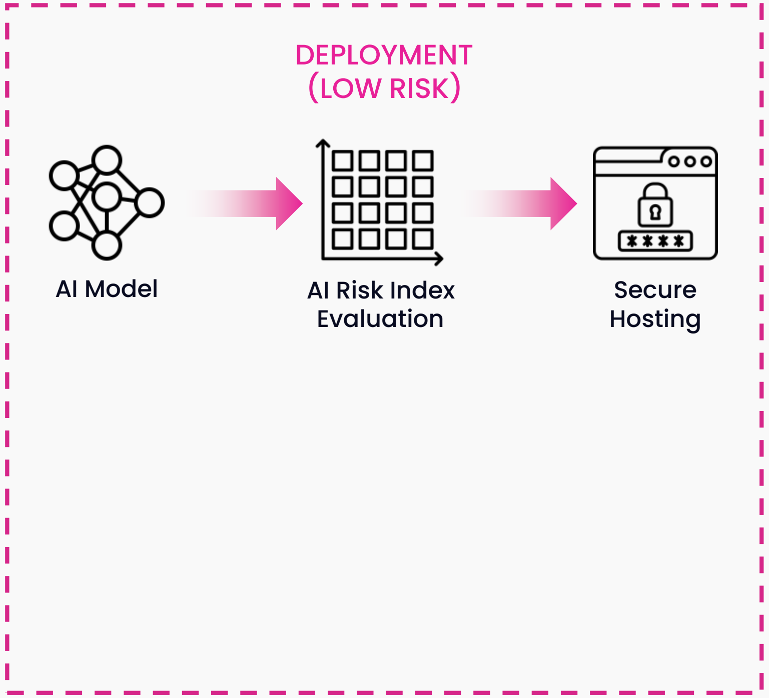
Generation & Evaluation Tools

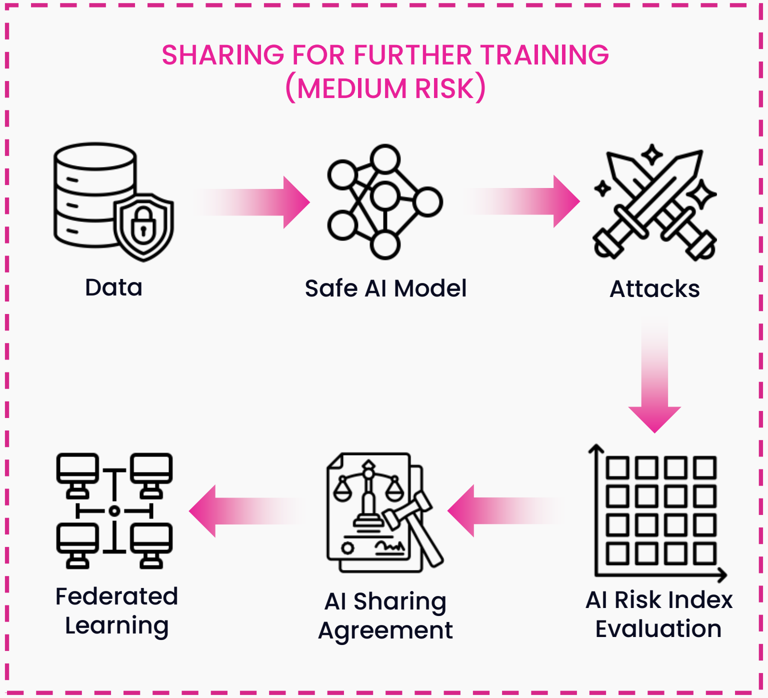
Synthetic Data Evaluation Tool

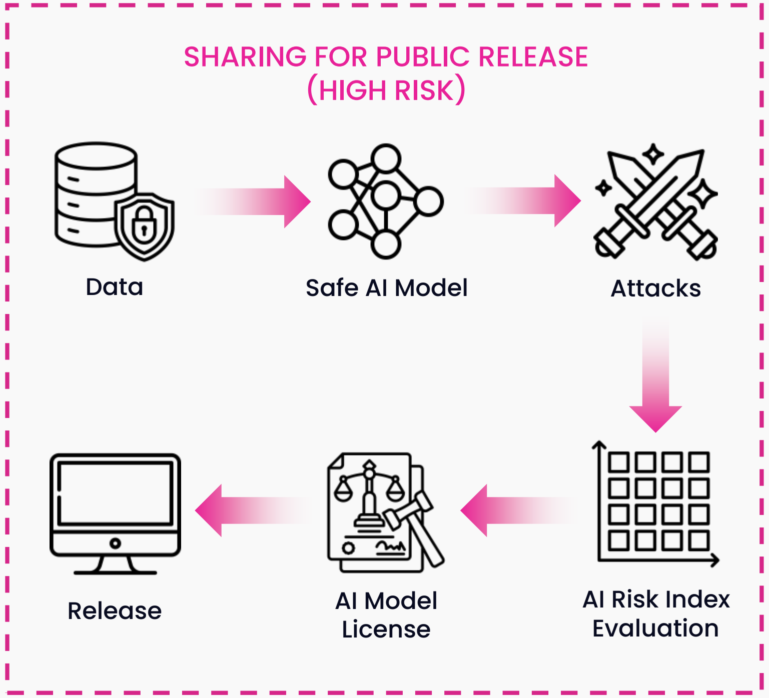


- Machine Learning Efficacy
- Privacy Against Inference
- Data Likelihood
- Data Matches
- Adherence
- Similarity

AI Model Release Scenarios & Evaluation







Three risk scenarios were identified from the workshops; deployment, shared training, and public release. AI models ready to deploy were seen as low risk if hosting via secure web services or an API where the AI model stays within the TRE, and is only able to be queried. In this case, there are no risks of white box attacks, and with the addition of access/query restrictions it means that black-box attacks are reduced. This is one of the safest options for allowing an AI model to be used, and doesn't affect the utility as privacy-preserving techniques don't need to be implemented.

If a researcher requires a model to be further trained or validated on external datasets (either through something like federated learning or transfer to another server), then this was deemed as medium risk as the model is still being shared in some form, but poses less of a risk compared to public release, as it would only be to other university or TRE servers. In this case, the model should still have some form of privacy-preserving techniques implemented to ensure it is safe, but most importantly, an AI sharing agreement should be signed to ensure that the model is not shared further outside of the original agreement and that there is no attempt to attack the model or use it inappropriately.

The highest risk scenario posed was publically releasing the AI model. This is because the AI model would be open to anyone to use and potentially attack. In this case, it would have to be proven that the researcher has sufficiently implemented privacy-preserving techniques in the AI model to ensure it protects against attacks. Additionally, for the model to be released, it should have an AI model license to govern use and redistribution.

Recommendation 5

Privacy Attack Simulations

From both the researcher and data provider workshops, attack simulations were seen as one of the most suitable methods for assessing an AI model for release. Several packages have been developed which allow attacks to be run on standard AI models which can be used to assess the privacy of them, but these are typically restricted to certain frameworks, and provide different privacy evaluation metrics which can often be difficult to interpret. So, to be able to use these attacks in TREs, they need to be simple to run, and have clear evaluation metrics which can be easily interpreted.

One package which offers this, is the SACRO AI-SDC package which has been developed to specifically meet the needs of TREs and offers a good solution for running and evaluating attacks [18].

However, as a community we need to ensure that this package is continually developed and updated to ensure that it meets a range of needs and keeps up with the development of AI methods and new attacks. We also need to be able to provide the necessary training to TRE staff to be able to effectively use these tools and interpret the results.

By being able to run privacy attacks on AI models, we can determine to some confidence that they are safe enough to release if they sufficiently protect against them. Therefore, these attacks should be run for any AI model if they are to be released from the TRE.

Package	Attribute Inference	Membership Inference	Reconstruction / Inversion	Supported Frameworks
Adversarial Robustness Toolbox	✓	✓	✓	TensorFlow, Keras, PyTorch, Scikit-learn
AIJACK		✓	✓	PyTorch, Scikit-learn
TensorFlow		✓		Tensorflow, Keras
SACRO	✓	✓		Tensorflow, Keras, Scikit-learn

Recommendation 6

AI Risk Index Evaluation

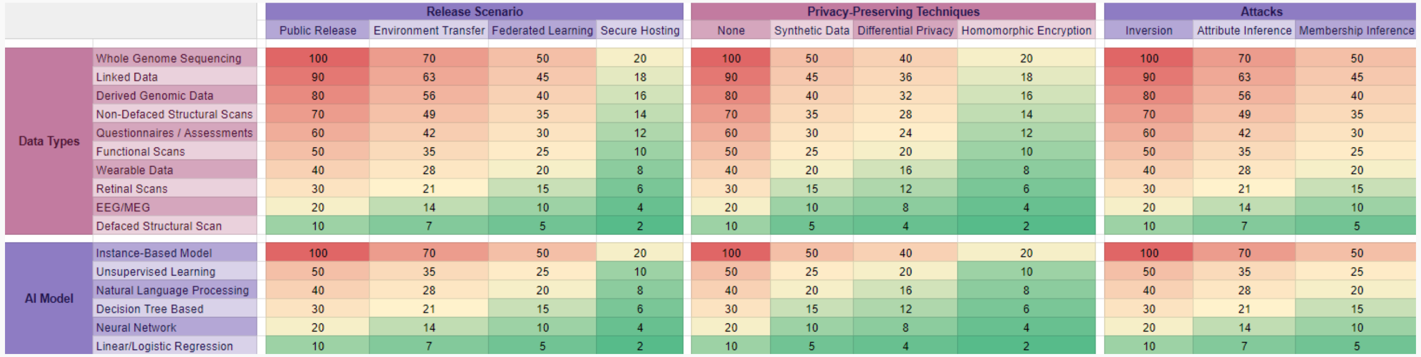
From the results of the workshops, we developed an AI Risk Index based on the rankings gathered on a range of different aspects of AI development and release. This takes into account the type of model developed along with the types of data used to train that model, as well as the release scenario, privacy-preserving technique used, and whether it fails any attack. This combines the rankings from all three workshops to provide a comprehensive risk score matrix which can be used to help evaluate the risk level of an AI model.

Take an example where a researcher has used derived genomic, questionnaire, and defaced structural scan data to develop a decision tree based model. They decide to implement synthetic data and want to request that the model is transferred onto their university server. When it came to running attacks, it failed one – the membership inference attack. The matrix can be used to add up the scores of these aspects to give a total risk score of 306 making it medium risk.

This provides a quantitative evaluation of the risk level for the release of that model which can be used to help aid decision making on whether it should be allowed to be released in this scenario.

As these scores take into account the opinions of the public, researchers and data providers, we can ensure that each groups concerns are taken into consideration in the release review process and allow data providers to make an informed decision.

However, although this already takes into account a range of perspectives, it should be an ongoing activity to feed data into this index to ensure it becomes more statistically robust overtime.



		Release Scenario				Privacy-Preserving Techniques				Attacks		
		Public Release	Environment Transfer	Federated Learning	Secure Hosting	None	Synthetic Data	Differential Privacy	Homomorphic Encryption	Inversion	Attribute Inference	Membership Inference
Data Types	Whole Genome Sequencing	100	70	50	20	100	50	40	20	100	70	50
	Linked Data	90	63	45	18	90	45	36	18	90	63	45
	Derived Genomic Data	80	56	40	16	80	40	32	16	80	56	40
	Non-Defaced Structural Scans	70	49	35	14	70	35	28	14	70	49	35
	Questionnaires / Assessments	60	42	30	12	60	30	24	12	60	42	30
	Functional Scans	50	35	25	10	50	25	20	10	50	35	25
	Wearable Data	40	28	20	8	40	20	16	8	40	28	20
	Retinal Scans	30	21	15	6	30	15	12	6	30	21	15
	EEG/MEG	20	14	10	4	20	10	8	4	20	14	10
	Defaced Structural Scan	10	7	5	2	10	5	4	2	10	7	5
AI Model	Instance-Based Model	100	70	50	20	100	50	40	20	100	70	50
	Unsupervised Learning	50	35	25	10	50	25	20	10	50	35	25
	Natural Language Processing	40	28	20	8	40	20	16	8	40	28	20
	Decision Tree Based	30	21	15	6	30	15	12	6	30	21	15
	Neural Network	20	14	10	4	20	10	8	4	20	14	10
	Linear/Logistic Regression	10	7	5	2	10	5	4	2	10	7	5

Recommendation 7

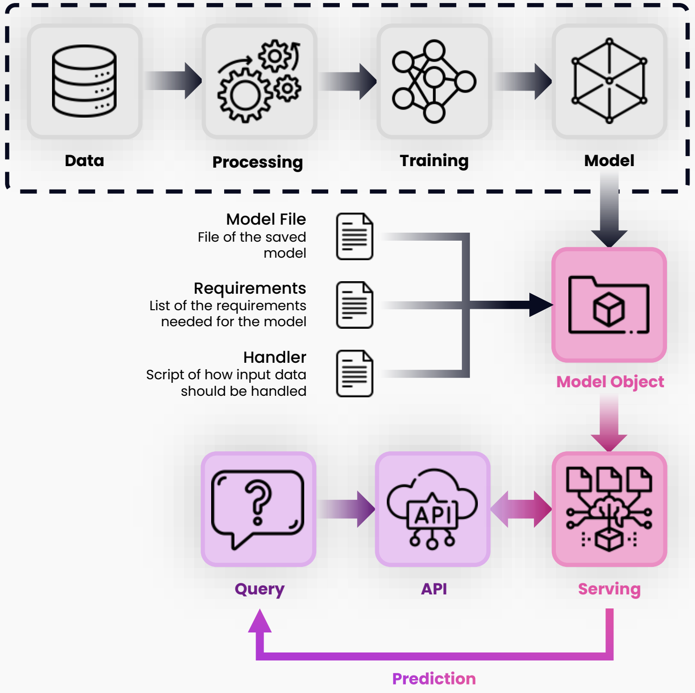
Secure Hosting Services

In regards to deploying an AI model, where it doesn't need to be further trained or validated, secure hosting was deemed as the safest way to allow for that model to be used in the real-world. This is because the model can stay securely within the TRE, but can be queried externally by authorised users with query/inference controls in place.

Typically, an AI model can be hosted online through a serving framework such as Tensorflow Serving, TorchServe or MLflow which allow an AI model to be deployed for inference via an API. This would allow clients to send a query to an API, which would send the data to a secure server where the serving framework is deployed within the TRE. This would include the saved model, along with scripts to handle the input data in case pre-processing is needed, as well as the requirements needed to run these and the model. Depending on the framework used, there are different ways of doing this, and in some cases Docker can be used. The model will then return a prediction via the API to the client. Restrictions can be imposed on top of this to ensure that only authorised users can query the model, and controls can be put in place to limit the amount of queries for example to reduce the risk of attacks. This is a service which should be offered by TREs to keep the model secure in their infrastructure, while still allowing researchers to access and use those models in practice.

TensorFlow Serving **TorchServe** **MLflow**

In cases where there may be concerns sending external data to query the model, the researcher may also decide to implement encrypted inference so that the model could be queried with encrypted data through homomorphic encryption. This would allow sensitive data from other sources to safely be sent to the server to receive a prediction.



Recommendation 8

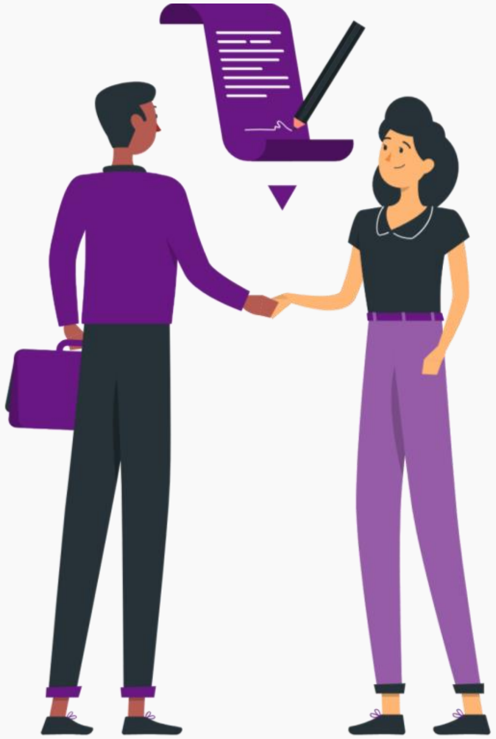
Researcher Agreements & AI Model Licensing

Agreements and licenses should play a crucial role in releasing and governing the use of AI models, and can be used for a variety of purposes and release scenarios. Just like how software licenses control the distribution and usage of software, AI model licenses impose controls surrounding the utilisation and redistribution of AI models. These licenses can specify the intended purposes for which the AI model can be used, as well as imposing restrictions to ensure that the user doesn't attempt to attack or gain access to the data.

As an AI model transitions from a controlled TRE environment to broader accessibility, ensuring adherence to the permitted purposes and restrictions becomes vital. In release scenarios, the establishment of clear license agreements serves as a way to ensure that users accessing and using the model comply with specified terms and conditions to safeguard the privacy of the training data.

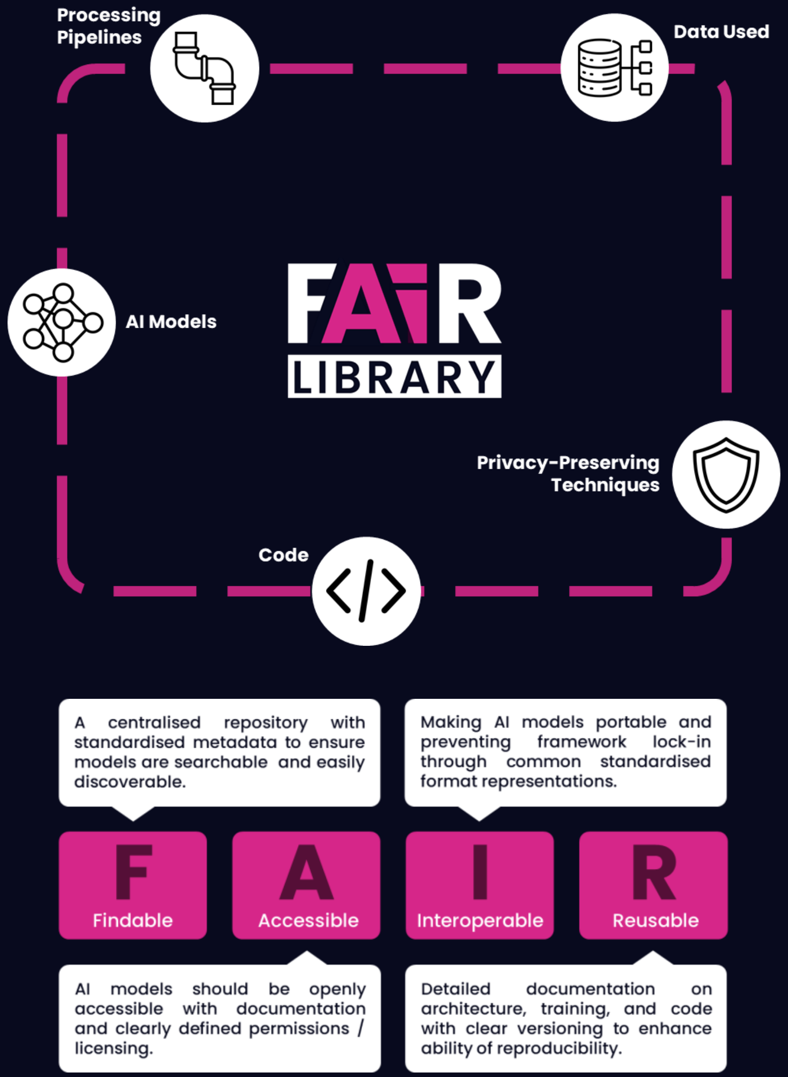
Additionally, agreements could also be implemented in federated learning, or environment transfer, scenarios, where parties can define clear guidelines and protocols for the collaborative training process. These agreements can be used to ensure that the other parties involved don't attempt to attack the AI model involved or to retrieve data.

However, despite the use of license agreements, there are significant challenges to the enforcement of them, particularly in the context of international deployment. The global nature of AI research and deployment means that enforcing compliance of a license across diverse jurisdictions becomes more complex. Furthermore, the evolving nature of AI technologies poses additional hurdles to enforcement, as adversaries continually seek novel methods to attack and exploit vulnerabilities in AI models.



Recommendation 9

AI Model Library



A centralised repository with standardised metadata to ensure models are searchable and easily discoverable.

Making AI models portable and preventing framework lock-in through common standardised format representations.

AI models should be openly accessible with documentation and clearly defined permissions / licensing.

Detailed documentation on architecture, training, and code with clear versioning to enhance ability of reproducibility.

Data providers asked the question - *"why would researchers need to bring these models out of a TRE if they are not being clinically implemented?"*. Data providers felt that unless these models were being deployed, then there was no need for allowing access outside of the TRE. However, certain funders or journals will require that researchers publish their AI model, and is encouraged for reproducibility reasons.

A solution for external validation and reproducibility was proposed, through the utilisation of the FAIR framework for AI models, which would treat AI models as derivatives of the original data that can be applied for and accessed via the TRE.

Each TRE should securely store AI models which have been developed using their data, along with details on pre-processing, datasets used, and accompanying code with proper documentation. Then, just like data, these models can be applied for by researchers to be able to validate, and fine tune.

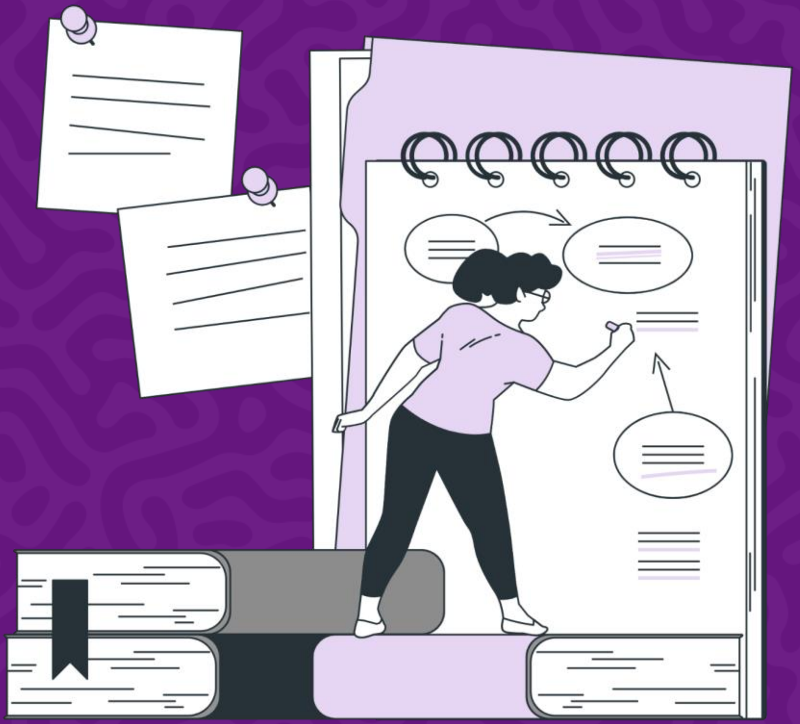
In situations where these models need to be validated on **external** datasets for example, then this should be done through some form of secure federation.

By implementing this extended FAIR framework for AI models, researchers would benefit from enhanced accessibility, reproducibility, and transparency, while also maintaining the security of models.

CHAPTER 6

Future Work

AI Risk Evaluation



These perspectives and recommendations explore the feasibility and challenges of implementing various mitigations in AI model research which highlights the need for training, and further research into effective methods for various scenarios.

To enable the safe development of AI within TREs, training and resources are paramount to ensure that researchers have the skills necessary. This training needs to focus on increasing awareness of privacy risks and effective ways to mitigate these depending on different scenarios.

We also need to ensure that AI privacy evaluation tools are further developed and improved to meet the needs of TREs and the rapidly evolving nature of AI. The risks of pre-trained models and Large Language Models also need to be assessed.

Additionally, from all three workshops, bias and discrimination was identified as one of the biggest risks in AI model development / deployment. Therefore, further work in AI risks should evaluate how TRE's can help researchers tackle these issues in their AI model research and how bias can be assessed.

[1] Rocher, L., Hendrickx, J.M. and de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications, [online] 10(1). doi:https://doi.org/10.1038/s41467-019-10933-3.

[2] Culnane, C., Benjamin and Teague, V. (2017). Health Data in an Open World. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.1712.05627.

[3] Bonomi, L., Huang, Y. and Ohno-Machado, L. (2020). Privacy Challenges and Research Opportunities for Genomic Data Sharing. Nature genetics, [online] 52(7), pp.646–654. doi:https://doi.org/10.1038/s41588-020-0651-0.

[4] Schwarz, C.G., Kremers, W.K., Wiste, H.J., Gunter, J.L., Vemuri, P., Spychalla, A.J., Kantarci, K., Schultz, A.P., Sperling, R.A., Knopman, D.S., Petersen, R.C. and Jack, C.R. (2021). Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. NeuroImage, 231, p.117845. doi:https://doi.org/10.1016/j.neuroimage.2021.117845.

[5] Puglisi, L., Barkhof, F., Alexander, D.C., Parker, G.J., Eshaghi, A. and Ravì, D. (2023). DeepBrainPrint: A Novel Contrastive Framework for Brain MRI Re-Identification. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2302.13057.

[6] Venkatesh, M., Jaja, J. and Pessoa, L. (2020). Comparing functional connectivity matrices: A geometry-aware approach applied to participant identification. NeuroImage, 207, p.116398. doi:https://doi.org/10.1016/j.neuroimage.2019.116398.

[7] GOV.UK (2023). A pro-innovation approach to AI regulation. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

[8] Parliament. House of Commons. (2023) The governance of artificial intelligence: interim report: Ninth Report of Session 2022–23 Available at: https://publications.parliament.uk/pa/cm5803/cmselect/cmhaff/635/report.html.

[9] ico.org.uk. (2023). AI and data protection risk toolkit. [online] Available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/.

[10] www.adalovelaceinstitute.org. (2022). Algorithmic impact assessment: a case study in healthcare. [online] Available at: https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/.

[11] Dwork, C. and Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. Foundations and Trends® in Theoretical Computer Science, 9(3-4), pp.211–407. doi:https://doi.org/10.1561/0400000042.

[12] Ziller, A., Usynin, D., Braren, R., Makowski, M., Rueckert, D. and Kaissis, G. (2021). Medical imaging deep learning with differential privacy. Scientific Reports, 11(1). doi:https://doi.org/10.1038/s41598-021-93030-0.

[13] Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K. and Erlingsson, Ú. (2018). Scalable Private Learning with PATE. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.1802.08908.

[14] Google, (2021). How we’re helping developers with differential privacy. [online] Available at: https://developers.googleblog.com/2021/01/how-were-helping-developers-with-differential-privacy.html.

[15] Apple Machine Learning Research. (2017). Learning with Privacy at Scale. [online] Available at: https://machinelearning.apple.com/research/learning-with-privacy-at-scale.

[16] Yilmaz, E., Ji, T., Erman Ayday and Li, P. (2022). Genomic Data Sharing under Dependent Local Differential Privacy. PubMed Central. doi:https://doi.org/10.1145/3508398.3511519.

[17] www.ons.gov.uk. (n.d.). ONS methodology working paper series number 16 - Synthetic data pilot - Office for National Statistics. [online] Available at: https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot.

[18] Smith, J., Preen, R.J., McCarthy, A., Crespi-Boixader, A., Liley, J. and Rogers, S. (2022). Safe machine learning model release from Trusted Research Environments: The AI-SDC package. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2212.01233.