Perspectives & Recommendations on Safe AI Development in Sensitive Healthcare Data

Lewis Hotchkiss

AI Risk Evaluation Community Group

Swansea University Prifysgol Abertawe
Dementias Platform UK | Data Portal
University of Cambridge
University of Oxford
SeRP
DARE UK

These recommendations were developed as part of a DARE UK funded initiative to set up an AI Risk Evaluation Group to bring together members of the public, researchers and data providers 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

Swansea University Prifysgol Abertawe | Dementias Platform UK | Data Portal | University of Cambridge | University of Oxford | SeRP | DARE UK

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

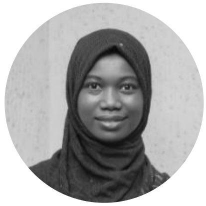
DPUK Operations Coordinator

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

Results from the Public Workshop

Do you have any concerns about AI?

Bias

Built in negative bias

Social Bias

Impact on vulnerable people

Leave people out

Identification

AI being used to discriminate

How will my identify be protected if data is leaked

Higher risk of identification If you have a rare condition

Manipulations of decisions at group level

Over Reliance

Become too reliable

Over reliance

Reduced specialised training

Result in less capable specialists

Trust

Reduced trust

Public Trust

Regulation

Lack of Regulations

Needs to be regulated

Speed of development vs regulation

Regulation

Other

Owned by big corporations

Data privacy

Lack of public involvement in development stage

AI Risk Evaluation Community Group

Results from the Public Workshop

Do you have any concerns with the use of your health data?

Identifiability

How easily can my data be used to identify?

Will my data be anonymous?

Need restrictions on what data is used

Cant de identify at population level

How could a data leak affect me?

No Control

Once initial consent is given, no updates are given after. What happens if things change.

No control of data

Who can access the data

What if my data is incorrect

Cant choose whats shared and whats not

Study consent might have been given before AI was a concern

The Unknown

How are the results going to be fed back to participants

Clarity on why the data is being shared and what its used for

Need to have trust in people sharing with

Not knowing how their data is being used

Selling Data

Corporations making huge profits from data

Concerns of selling data

Other

Using data to manipulate elections

Ethical concerns

AI Risk Evaluation Community Group

Concerns of AI Models

Attacks & Vulnerabilities

Inversion Attack
Where an attacker is able to reconstruct the original training data that was used in the model

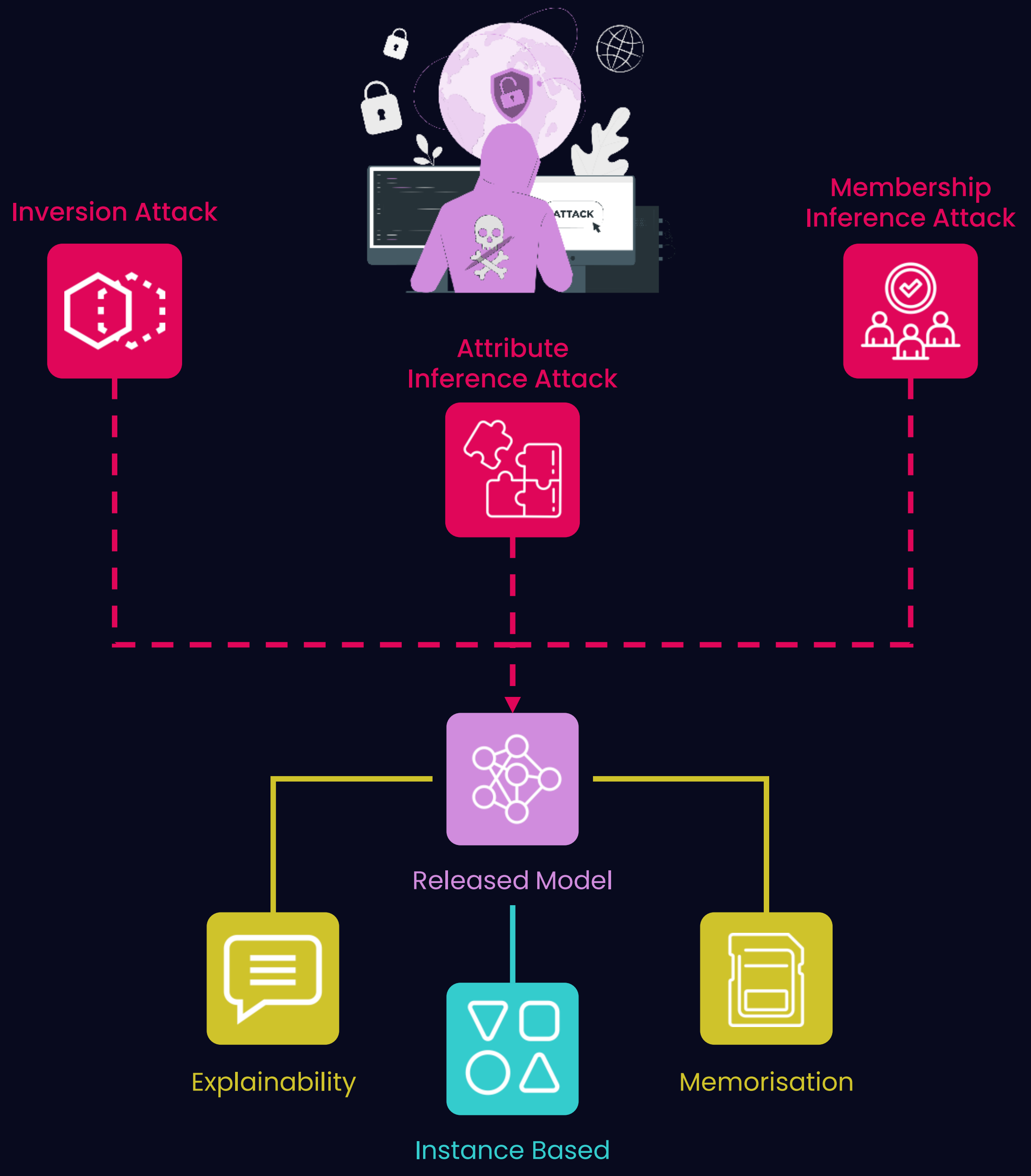
Membership Inference Attack
The attacker trains an attack model to predict whether a particular data point was part of the training dataset

Attribute Inference Attack
An attacker is able to infer unknown attributes from an individual that they might already know

Explainablity
Methods which allow users to understand and interpret predictions made by AI

Data Memorisation
Where an AI model is unable to generalise well on unseen data and instead overfits on the training data

Instance-Based Models
Uses the dataset as the model to compare unseen data to the data points in the dataset



AI Risk Evaluation Community Group

Privacy-Preserving Techniques

Protecting Patient Data

Homomorphic Encryption
Using encrypted data to train AI models.

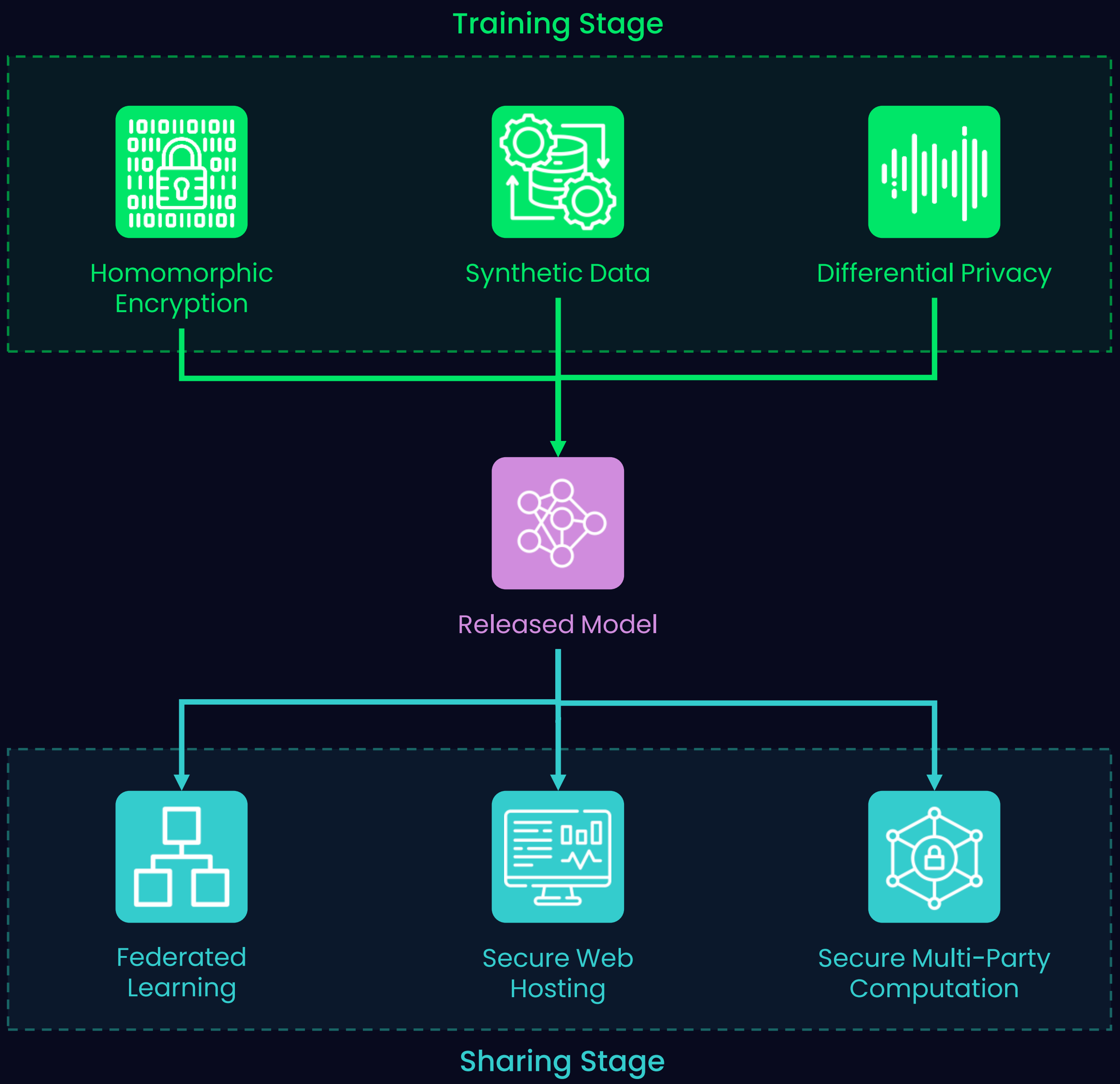
Synthetic Data
Data artificially created from real-world data which is statistically similar.

Differential Privacy
Statistical noise is added to the data which still describes the patterns of the group while protecting information about specific information.

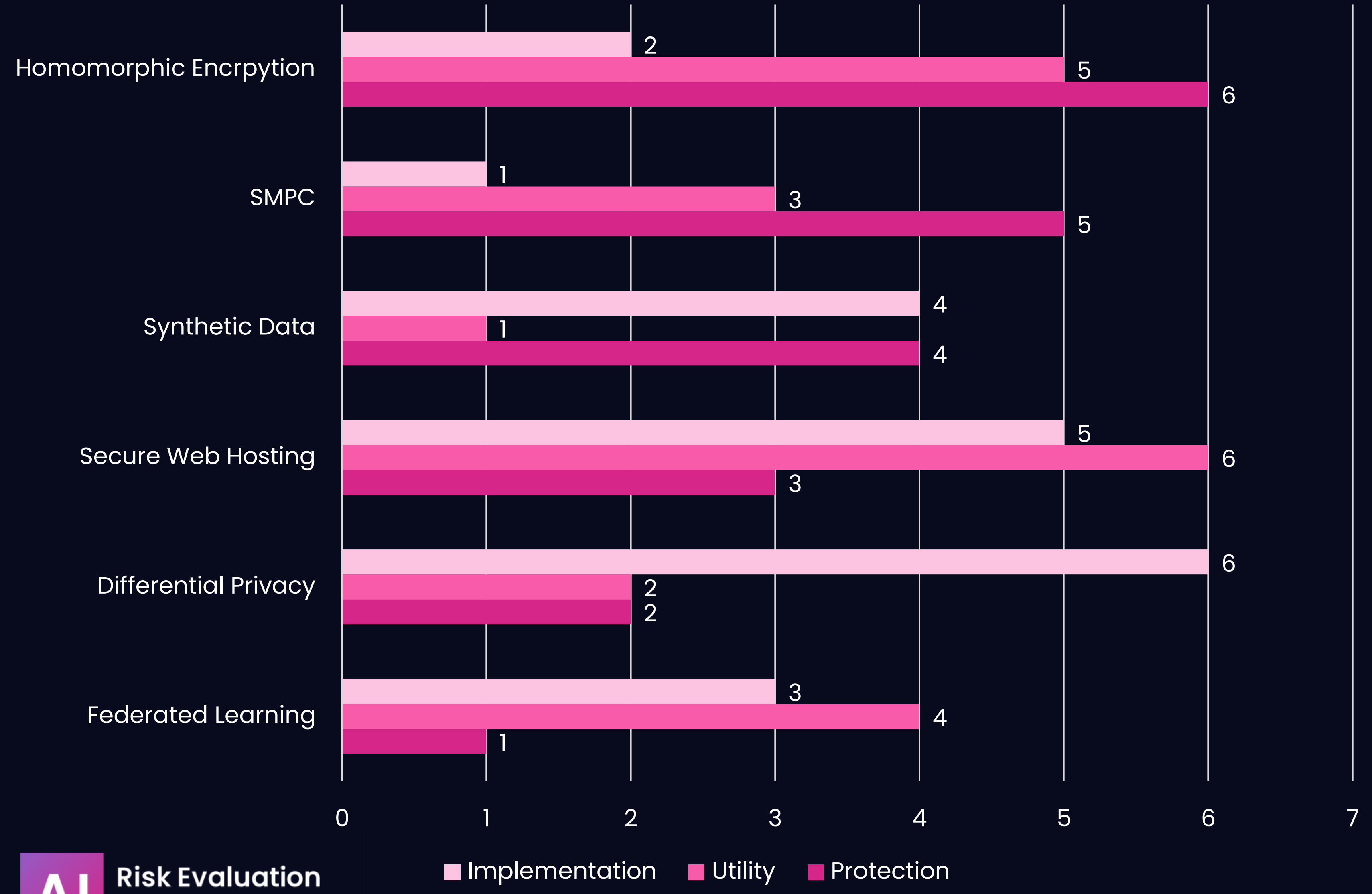
Secure Web Hosting
Hosting the AI model on a secure web service with authorisation and limits on number of queries.

Federated Learning
A distributed, decentralised approach to training AI models where local data doesn't need to be shared.

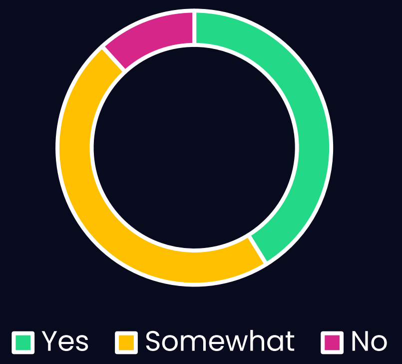
Secure Multi-Party Computation
Allows multiple parties to jointly compute a function on encrypted data.



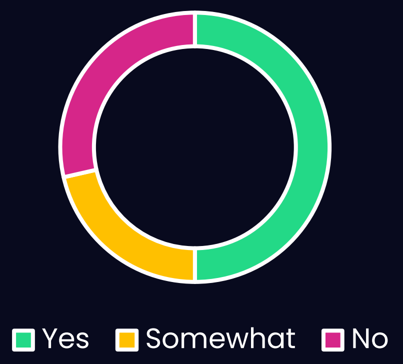
AI Risk Evaluation Community Group



Were you aware of these techniques?



Do neuroimaging & genomics pose unique challenges?



AI Risk Evaluation Community Group

Sequence data more identifiable

Genetic data more risky

Genomics have greater impact and potential repercussions

Neuroimaging surprisingly non predictive compared with genetics

If you are trying to find out more information about someone than a brain scan less likely than genomics

There is more AI in this space

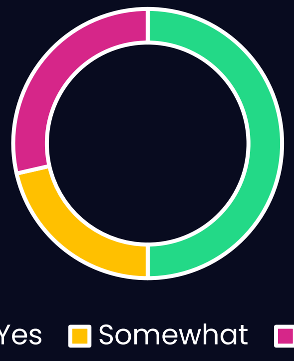
Dangerous to assume one type of data is safer than another

Genome sequencing harder to implement privacy techniques

Like fingerprinting – if have access to family data then can look at matching for genomics

Depends if using raw or derived

Do neuroimaging & genomics pose unique challenges?



AI Risk Evaluation Community Group

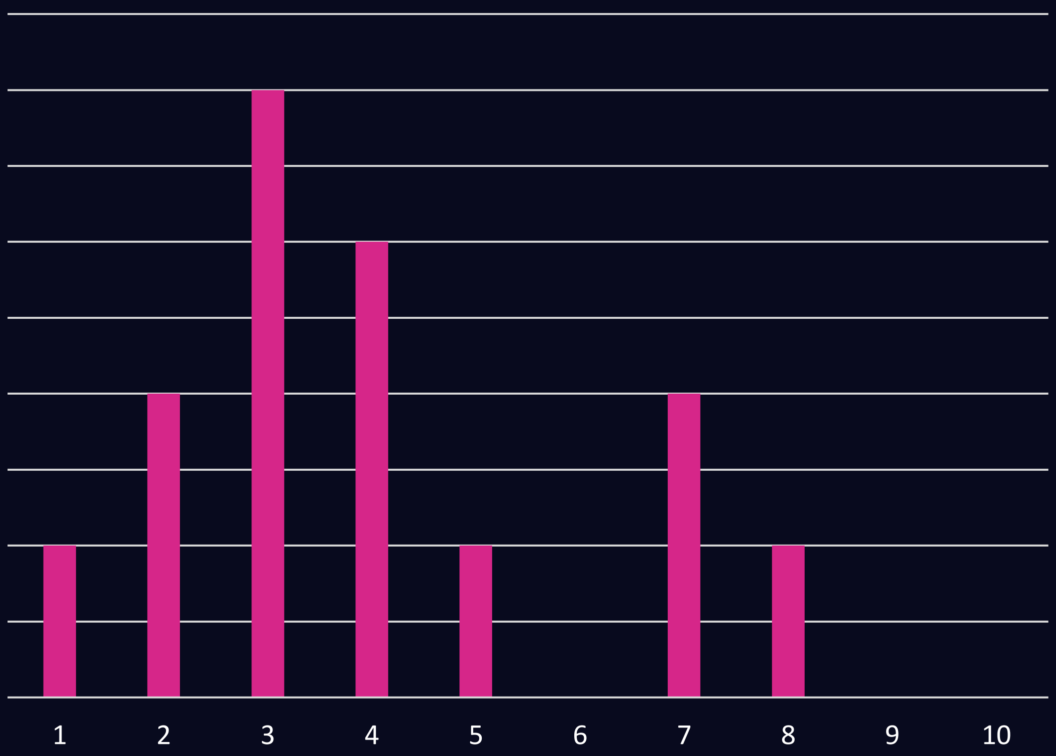
Concerns of AI Models

Data Types

1. Whole Genome
2. Linked Data
3. Derived Genomic
4. Non-Defaced Structural Scans
5. Questionnaire Data
6. Functional Imaging
7. Wearable Data
8. Retinal Scans
9. EEG/MEG
10. Defaced Structural Scans

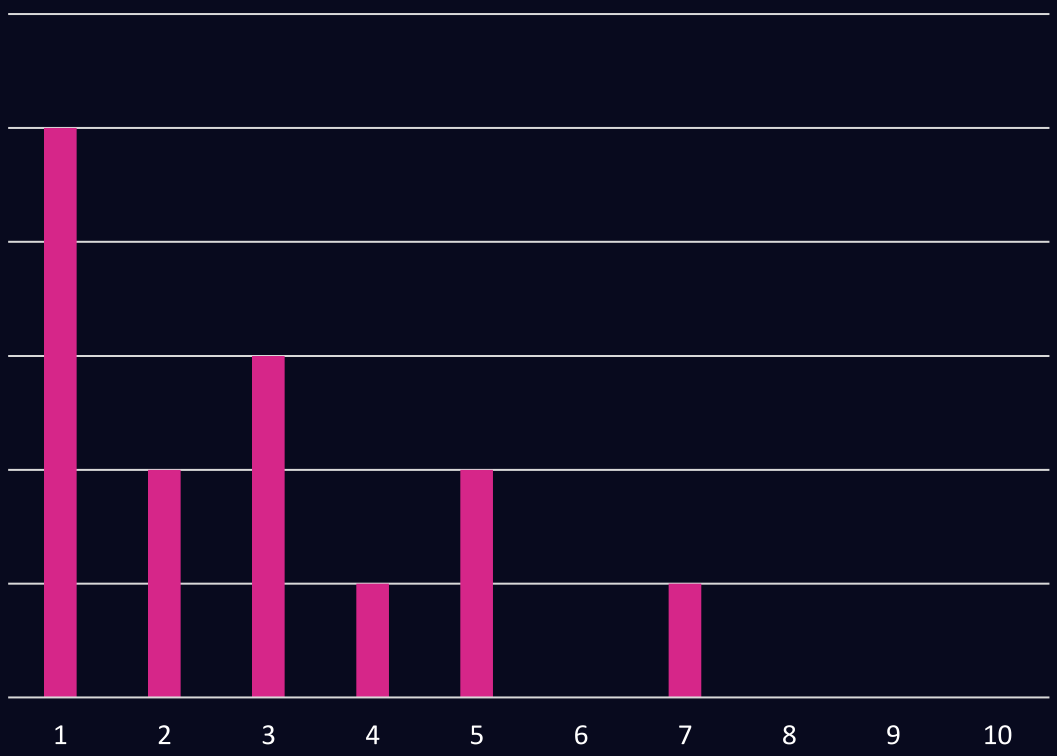
Implementing Privacy-Preserving Techniques

Barriers



■ Do researchers have the expertise/knowledge to implement

(1 = not at all, 10 = completely)



■ Are there enough resources and training available?

(1 = not at all, 10 = completely)

AI Risk Evaluation Community Group

Implementing Privacy-Preserving Techniques

Barriers

Need to be an expert in the field

Lack of experience

Out of date resources which don't keep up with speed of development

Should it be on the TRE to provide safe data

Need to outsource

Don't have expertise to judge level of noise acceptable

Would need guidance on best techniques to implement

Do researchers have the expertise to implement these techniques?

Not common practice

Need to have a shift in perspective

Not a mature space yet and not widely adopted

Worry about increasing time working on project

Need to have training

Need to educate researchers

If it is not a requirement then less likely to do so, need to have incentive from funders

Guidelines would be useful

AI Risk Evaluation Community Group

Privacy has to be preserved so the accuracy is what it is. Might push for more robust models

If accuracy is so low then it is pointless, need a balance

If data leaked is not disclosive then why take this hit, so depends on the data

Robust models should be able to handle some noise in the data anyway

How do you feel about having to trade off accuracy for privacy?

Do we just need to accept it? Need to find the right techniques which don't affect accuracy that much

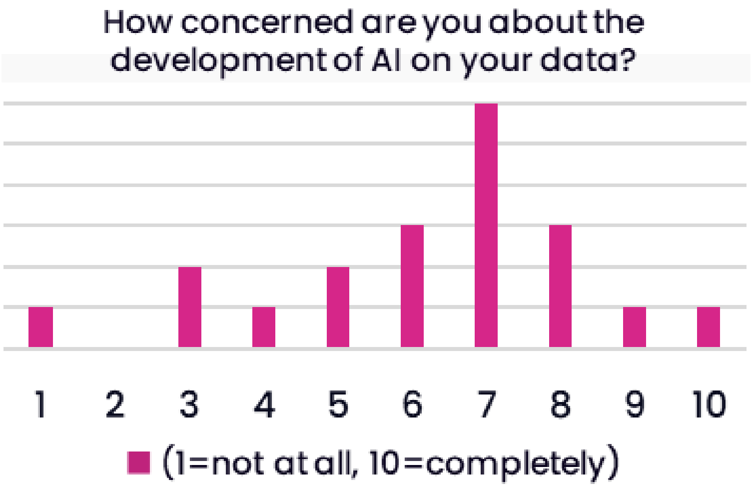
Need to find the right balance between privacy and utility.

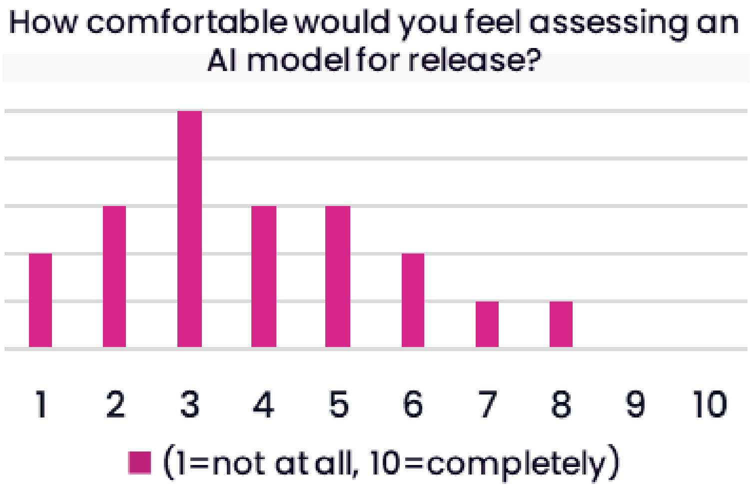
There is no point having a private model with no utility

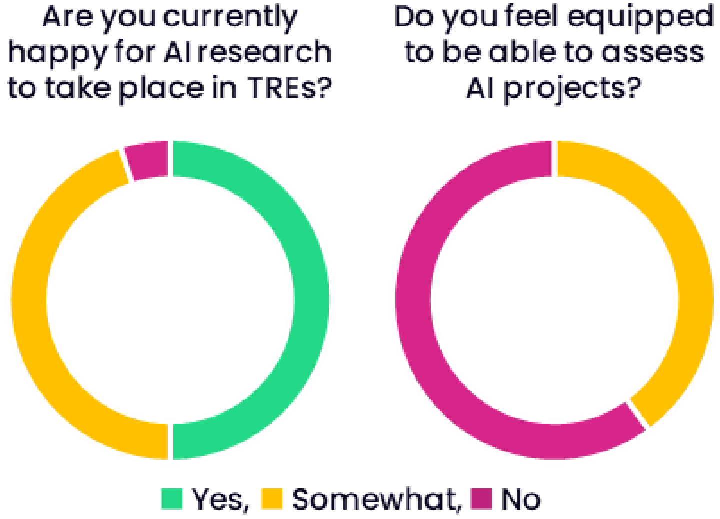
AI Risk Evaluation Community Group

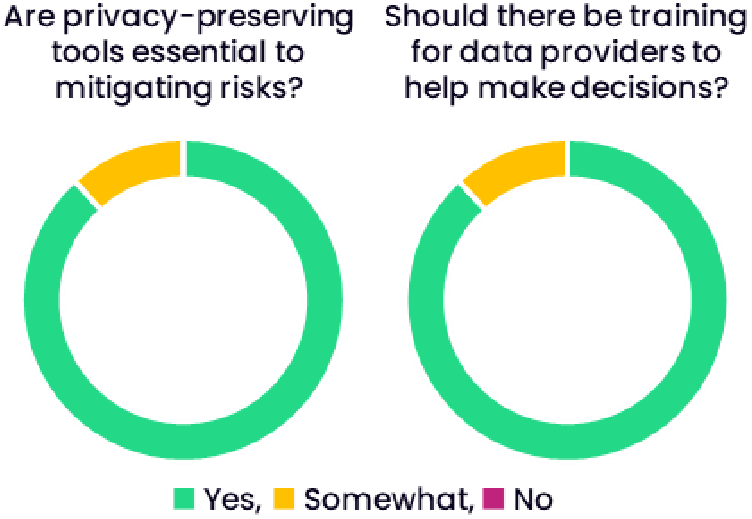
Assessing AI Models

Data Provider Perspectives





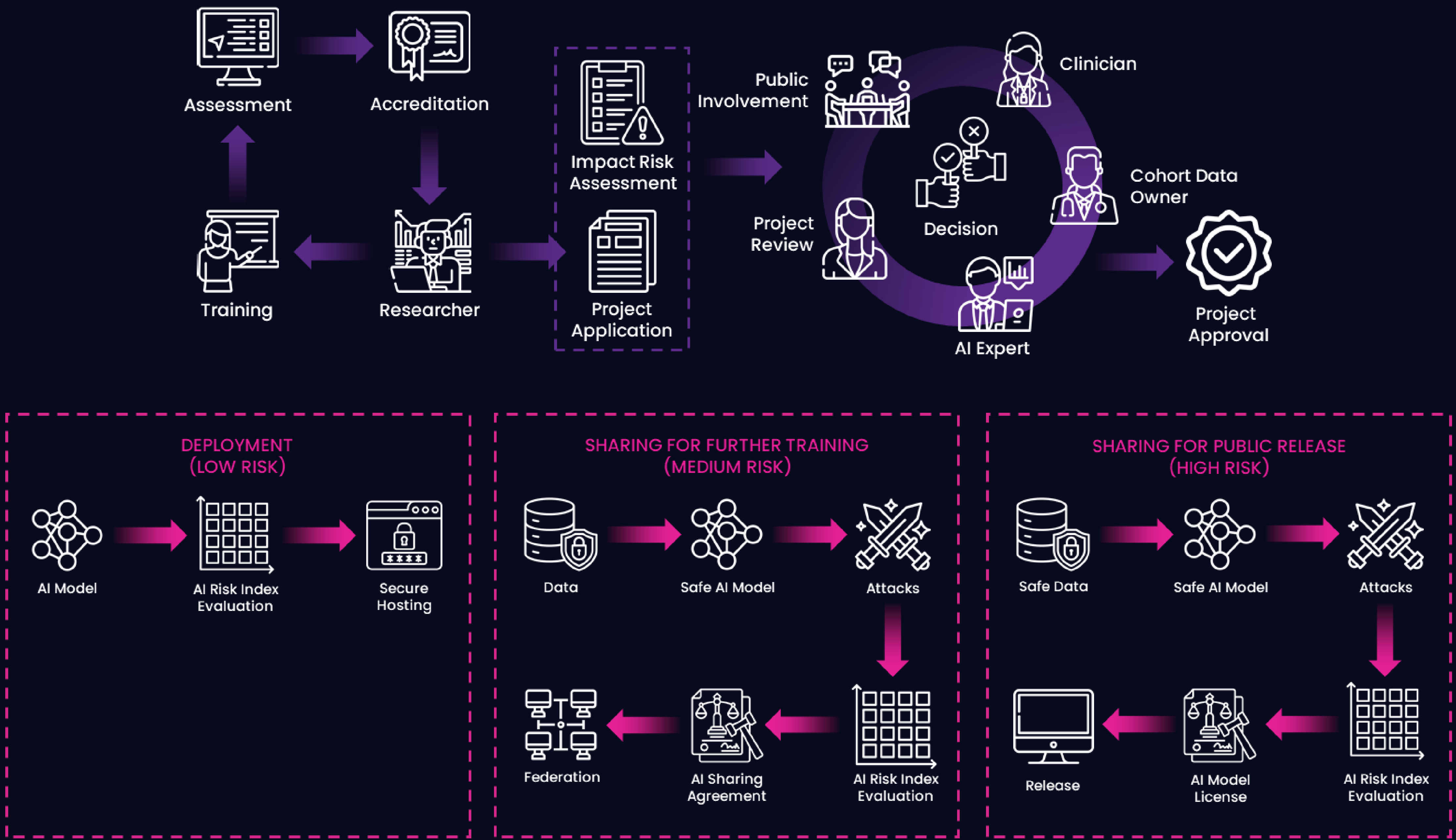




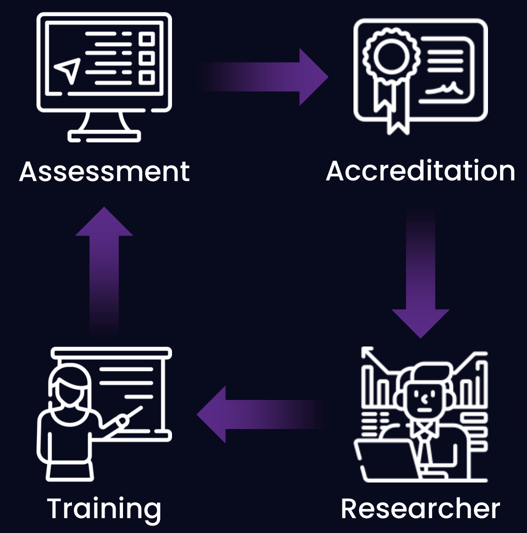
Recommendations

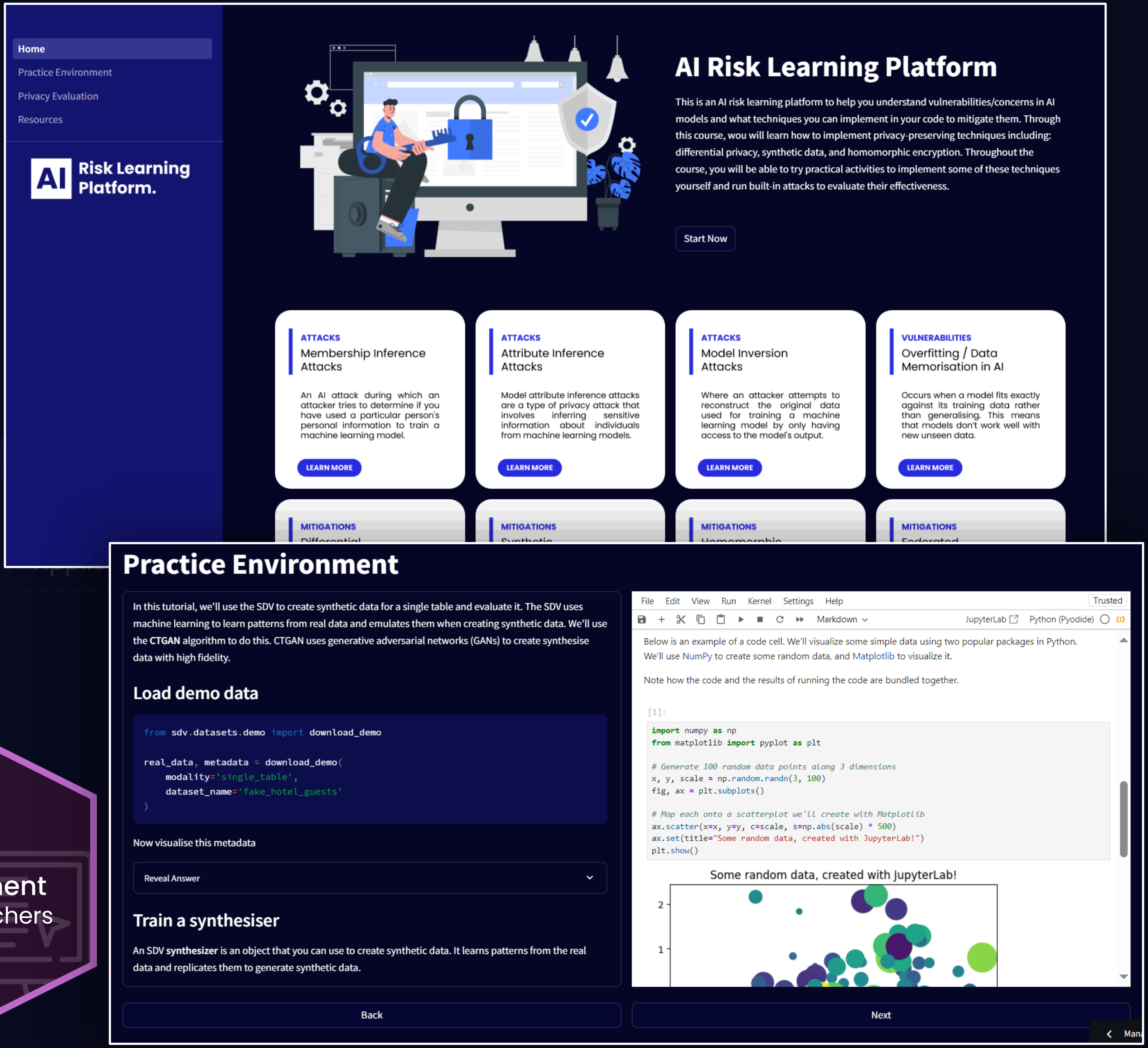
How to Allow the Safe Development & Release of AI

- Assessment
- Public Engagement
- Safe Data Tools
- Attack Simulations
- Risk Index Evaluation
- Training & Resources
- Accreditation
- Risk Impact Assessment
- Secure Hosting Service
- User Legal Agreement



Assessment
Accreditation
Training
Researcher
Impact Risk Assessment
Project Application
Public Involvement
Clinician
Decision
Cohort Data Owner
Project Review
AI Expert
Project Approval
DEPLOYMENT (LOW RISK)
AI Model
AI Risk Index Evaluation
Secure Hosting
SHARING FOR FURTHER TRAINING (MEDIUM RISK)
Data
Safe AI Model
Attacks
AI Risk Index Evaluation
AI Sharing Agreement
Federation
SHARING FOR PUBLIC RELEASE (HIGH RISK)
Safe Data
Safe AI Model
Attacks
AI Risk Index Evaluation
AI Model License
Release





AI Risk Evaluation Community Group

Datacise
Open Learning

Precision training for data science in **research**

In affiliation with

Swansea University
Prifysgol Abertawe
Population Data Science
Research & Innovation Institute
Gwyddor Data Poblogaeth
Sefydliad Ymchwil ac Arloesi

ADR UK

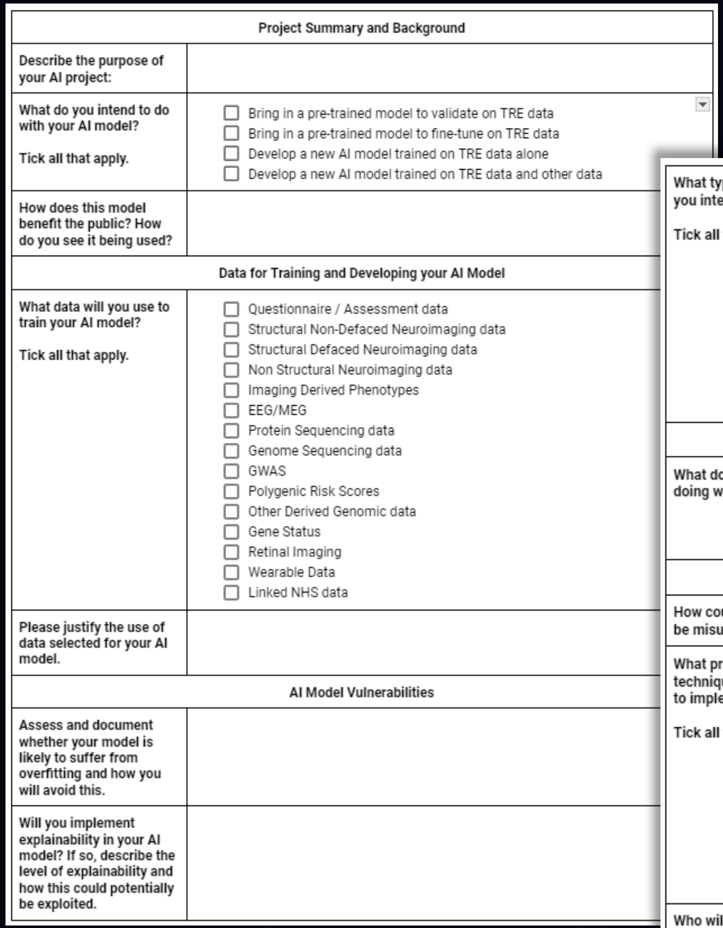
Protecting Patient Privacy in AI

Supported by

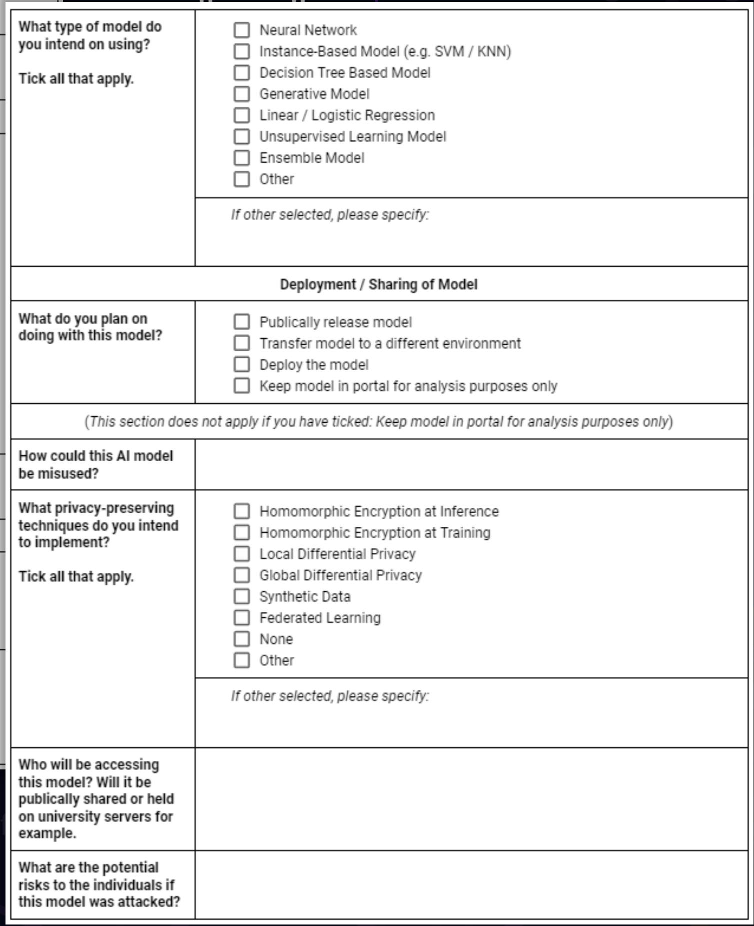
YDG CYMRU | ADR WALES

UKRI Economic and Social Research Council

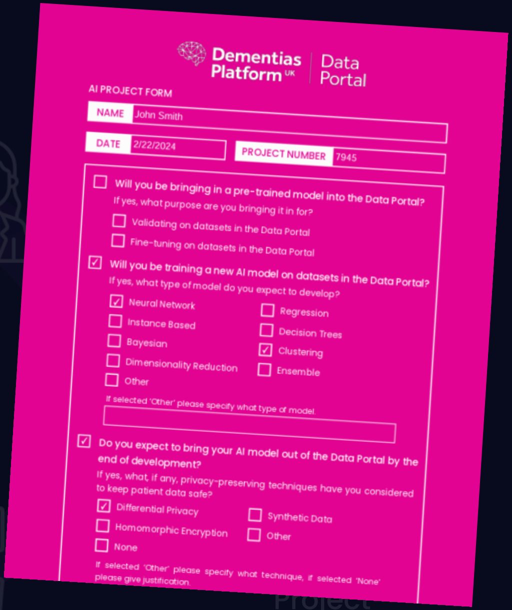
PRE-PROJECT STAGE



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?	



Dementias Platform UK | Data Portal

AI PROJECT FORM

NAME John Smith

DATE 2/22/2024 PROJECT NUMBER 7945

☐ Will you be bringing in a pre-trained model into the Data Portal?
If yes, what purpose are you bringing it in for?
☐ Validating on datasets in the Data Portal
☐ Fine-tuning on datasets in the Data Portal

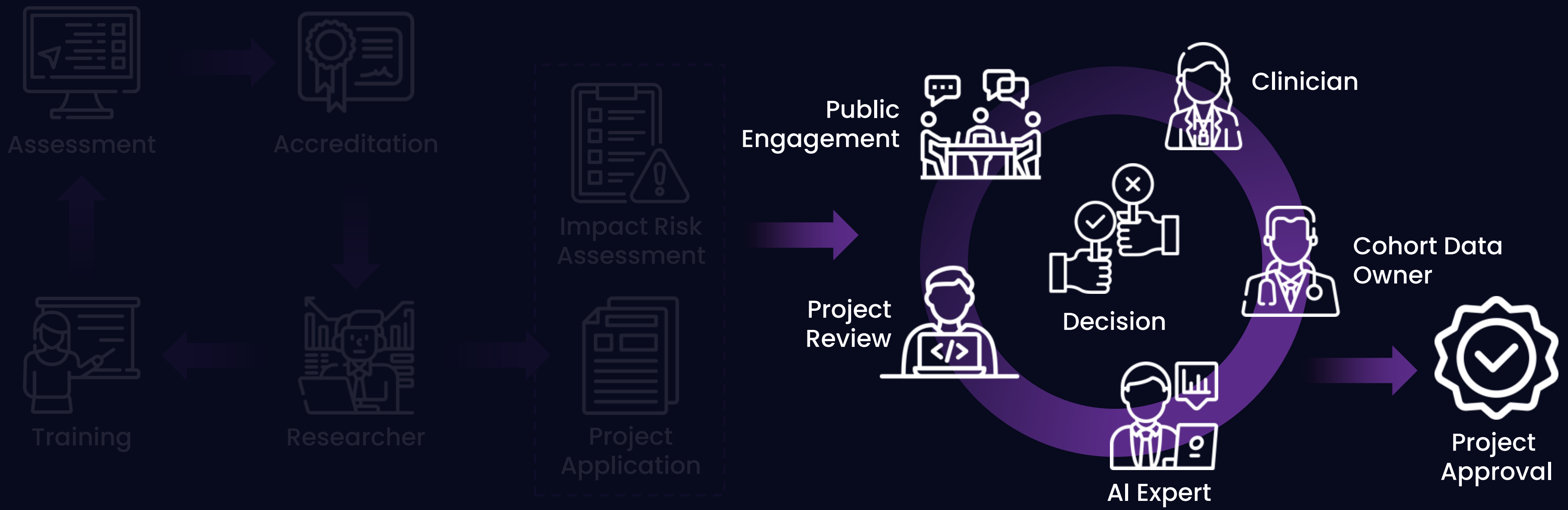
☑ Will you be training a new AI model on datasets in the Data Portal?
If yes, what type of model do you expect to develop?
☑ Neural Network ☐ Regression
☐ Instance Based ☐ Decision Trees
☐ Bayesian ☑ Clustering
☐ Dimensionality Reduction ☐ Ensemble
☐ Other
If selected 'Other' please specify what type of model.

☑ Do you expect to bring your AI model out of the Data Portal by the end of development?
If yes, what, if any, privacy-preserving techniques have you considered to keep patient data safe?
☑ Differential Privacy ☐ Synthetic Data
☐ Homomorphic Encryption ☐ Other
☐ None
If selected 'Other' please specify what technique, if selected 'None' please give justification.

Risk Impact Assessment

AI Risk Evaluation Community Group

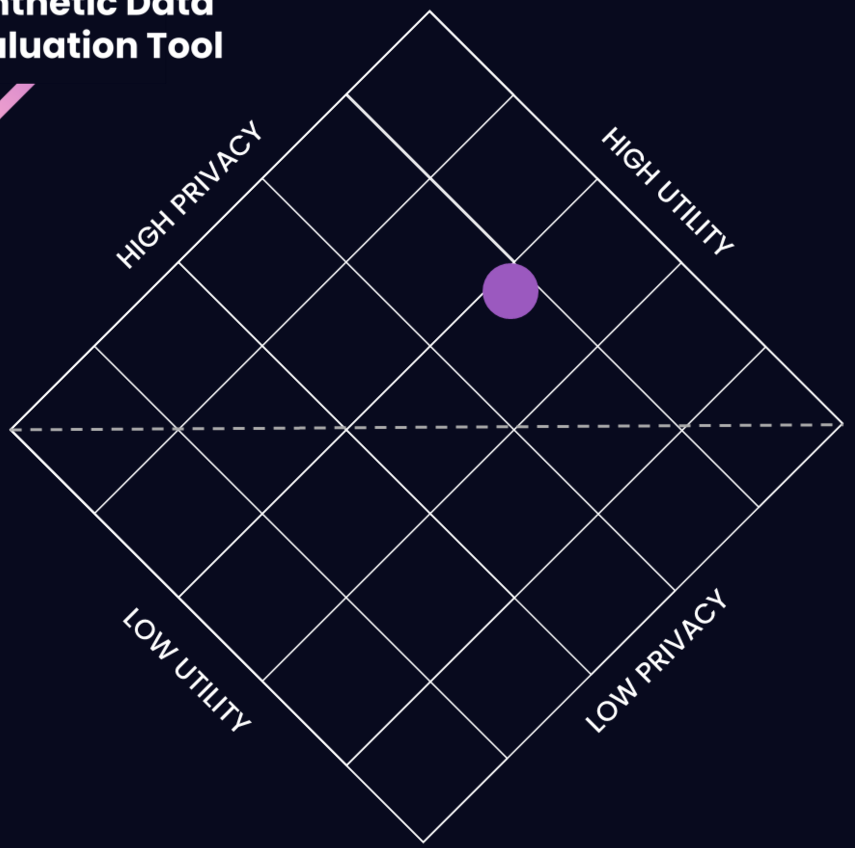
PRE-PROJECT STAGE



AI Risk Evaluation Community Group

PROJECT STAGE

Synthetic Data Evaluation Tool



Synthetic Data Generation

Upload a CSV Dataset

Drag and drop file here
Limit 200MB per file • CSV

Browse files

Dementia_Data.csv 50.0B

Select discrete columns: age × mmse × diagnosis ×

Select predictor column: diagnosis ×

Select identifier column: subject id ×

Select noise level: 6 (0 – 10)

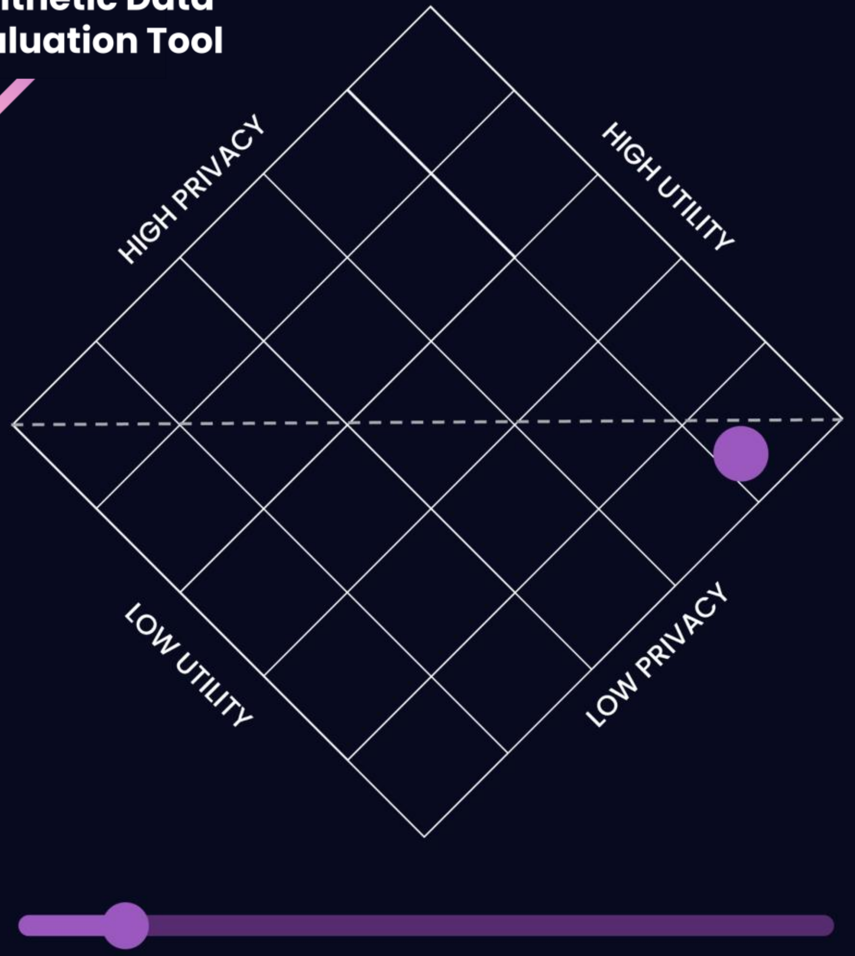
Submit

AI Risk Evaluation Community Group

Safe Data Tools

Synthetic Data Evaluation Tool

PROJECT STAGE



NOISE LEVEL

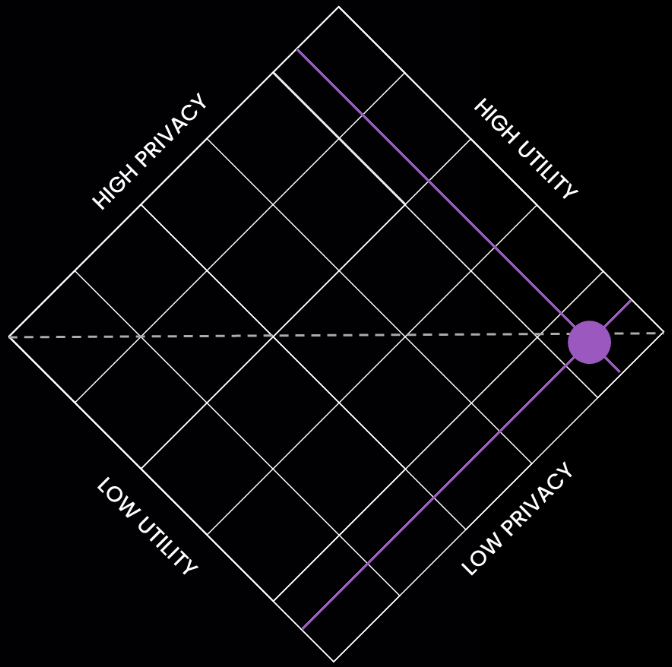
- Machine Learning Efficacy
- Privacy Against Inference
- Data Likelihood
- Data Matches
- Adherence
- Similarity

AI Risk Evaluation Community Group

Safe Data Tools

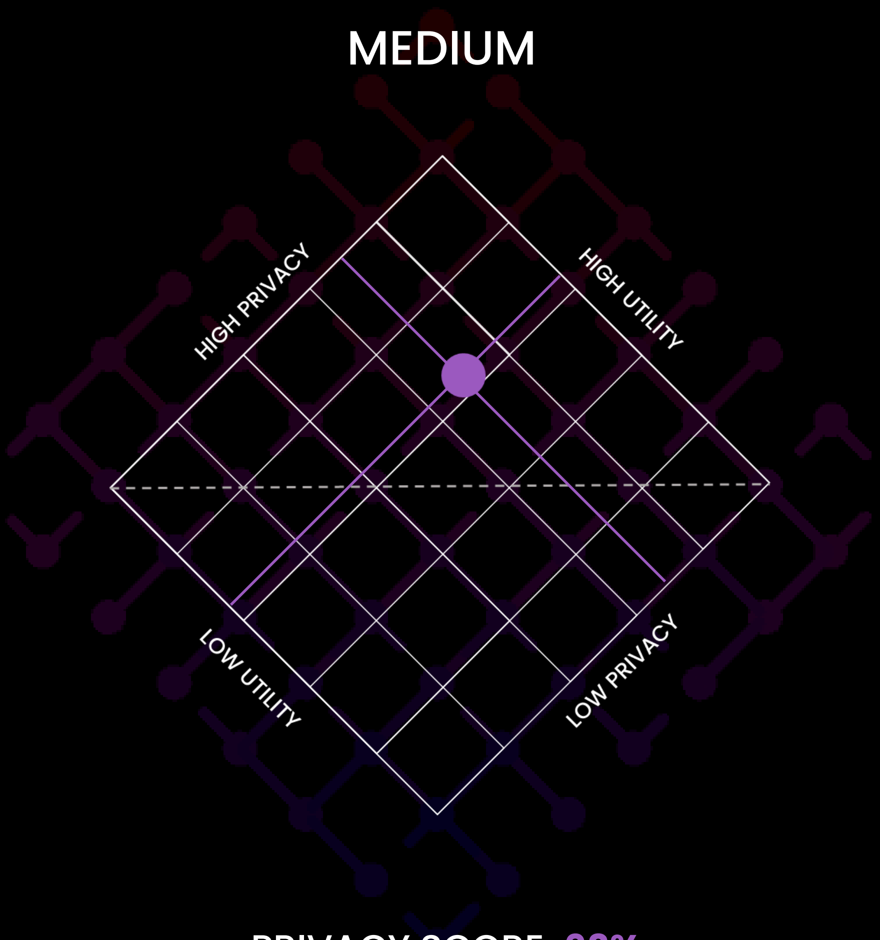
SYNTHETIC DATA

LOW



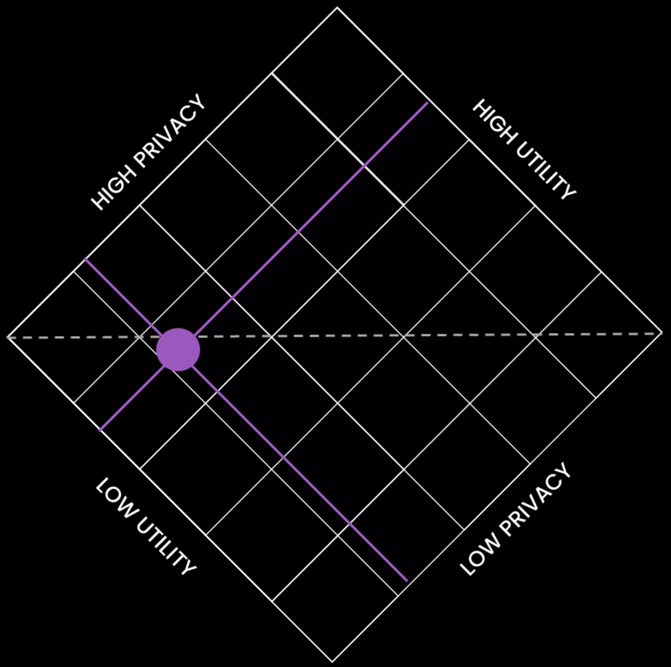
PRIVACY SCORE: **11%**
UTILITY SCORE: **87%**

MEDIUM



PRIVACY SCORE: **63%**
UTILITY SCORE: **68%**

HIGH



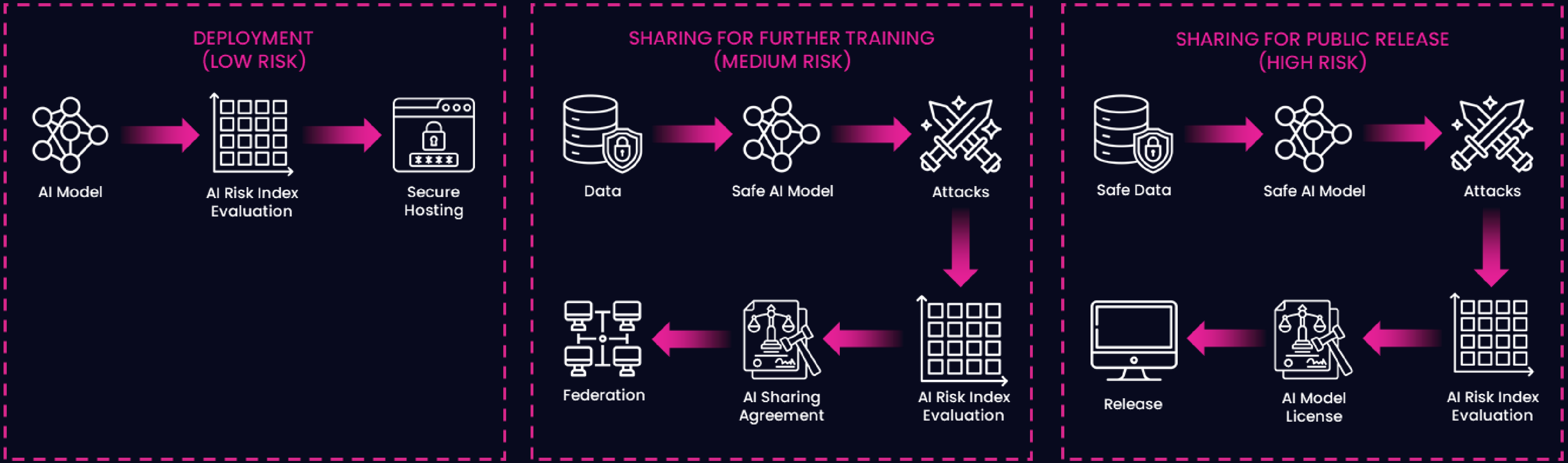
PRIVACY SCORE: **73%**
UTILITY SCORE: **23%**

POST - PROJECT STAGE

High Risk	Public release with no mitigations
Medium Risk	Public release with mitigations
Limited Risk	Shared training
Minimal Risk	Secure web hosting

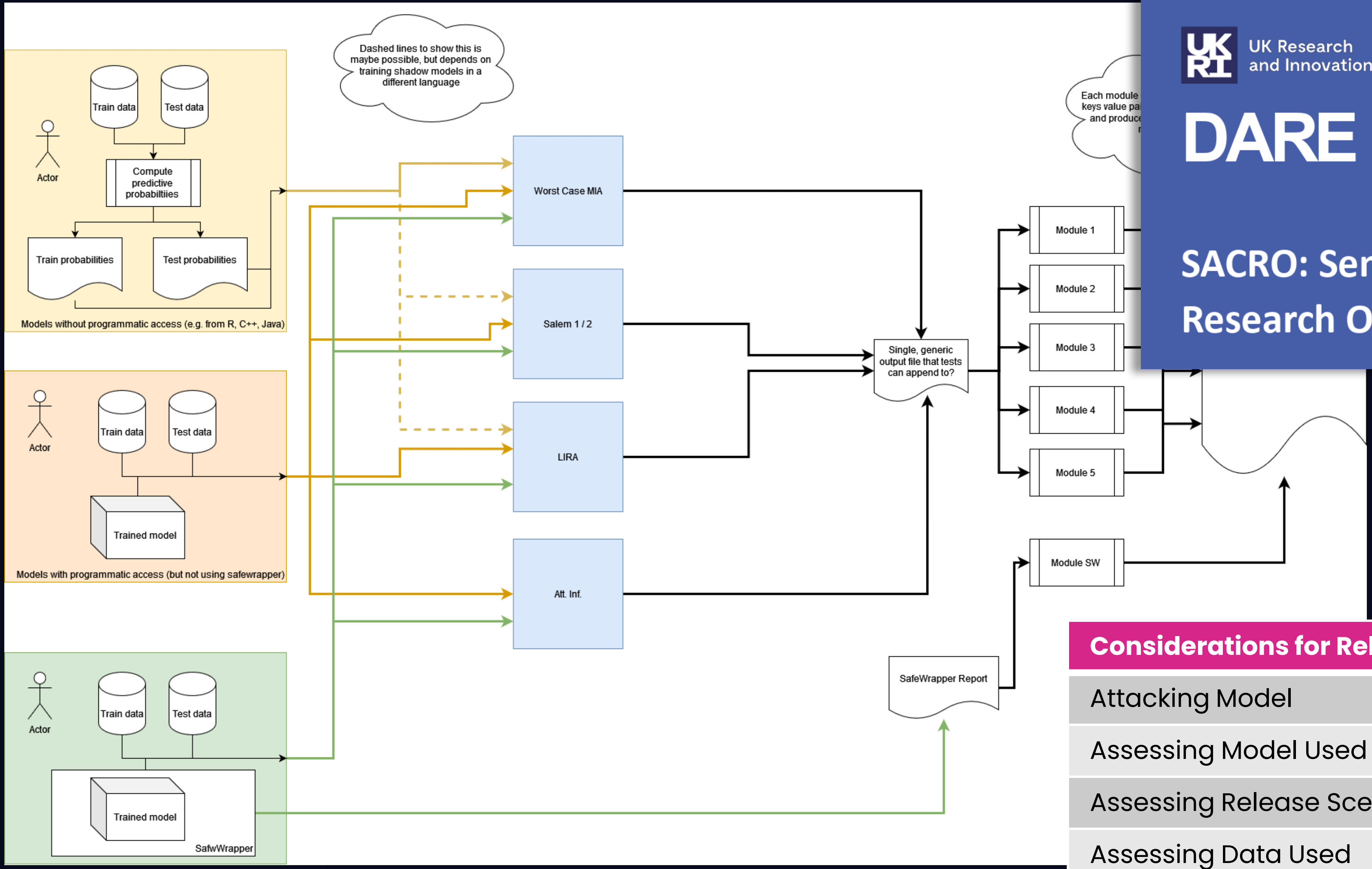
- Attack Simulations
- Risk Index Evaluation
- Secure Web Hosting Service
- User Legal Agreement

POST - PROJECT STAGE



AI Risk Evaluation Community Group

POST - PROJECT STAGE



UK Research and Innovation
HDRUK Health Data Research UK
ADR UK Data-driven change
SACRO

DARE UK

SACRO: Semi-Automated Checking of Research Outputs

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

Attack Simulations

POST - PROJECT STAGE

AI Risk Index

		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

AI Risk Evaluation Community Group

POST - PROJECT STAGE

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.

F Findable

A Accessible

I Interoperable

R Reusable

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.

Perspectives & Recommendations on Safe AI Development in Sensitive Healthcare Data

Lewis Hotchkiss

AI Risk Evaluation Community Group

Swansea University Prifysgol Abertawe
Dementias Platform UK | Data Portal
University of Cambridge
University of Oxford
SeRP
DARE UK