DARE UK

GRAIMatter

Guidelines and Resources for AI Model Access from TrusTEd Research environments (GRAIMatter)

Professor Emily Jefferson (On Behalf Of The Team)

The Team

DARE UK

GRAIMatter

Professor Emily Jefferson (PI): Director of HIC TRE

Professor Felix Ritchie: 5 Safes and Disclosure Control

Professor Jim Smith: AI Models

Dr James Liley: Assistant Professor in biostatistics

Maeve Malone: Lecturer in Intellectual Property law and Healthcare Law and Ethics

Professor Angela Daly: Regulation and governance of digital technologies, data protection, AI ethics

Dr Francesco Tava: Applied ethics, privacy and trust

Law and Ethics

Dr Esma Mansouri-Benssassi: Senior Research Fellow AI

Dr Christian Cole: Senior Health Informatician

Dr Simon Rogers: Principal Engineer - AI Models

Professor Josep Domingo-Ferrer

Dr Alberto Blanco Justicia:

International experts

Antony Chuter:

Jillian Beggs:

PPIE Co-leads

UWE Bristol University of the West of England

University of Dundee

Universitat Rovira i Virgili

Durham University

NHS National Services Scotland

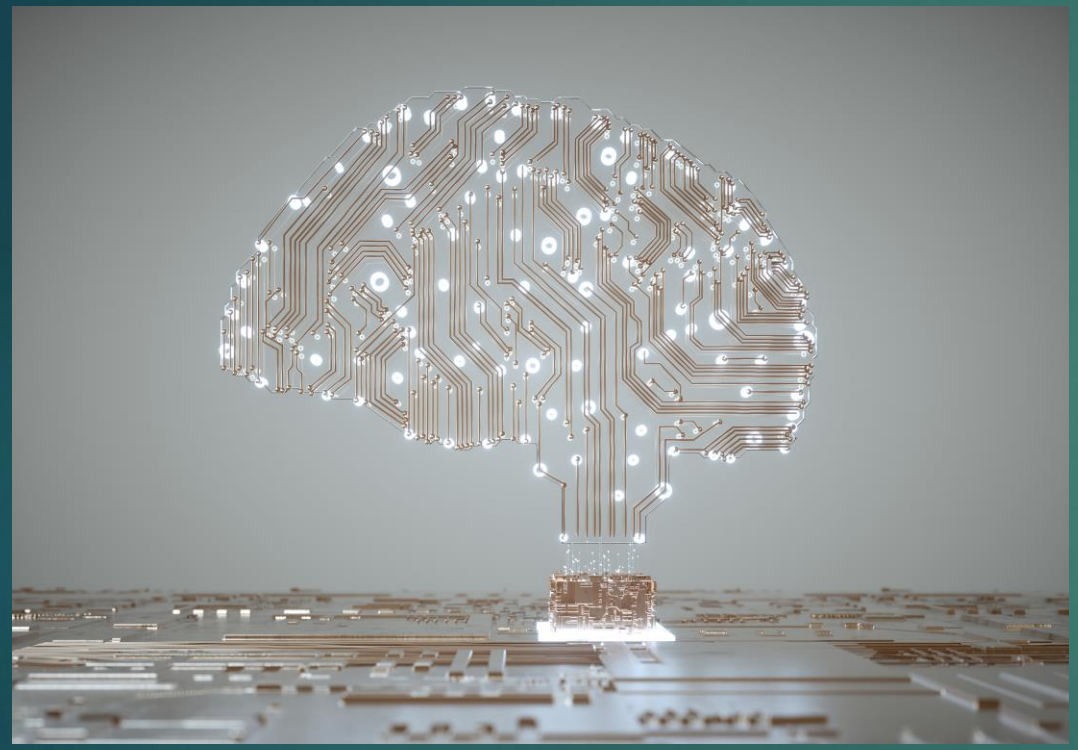
HDRUK Health Data Research UK

Health Informatics Centre

Problem statement

DARE UK

GRAIMatter

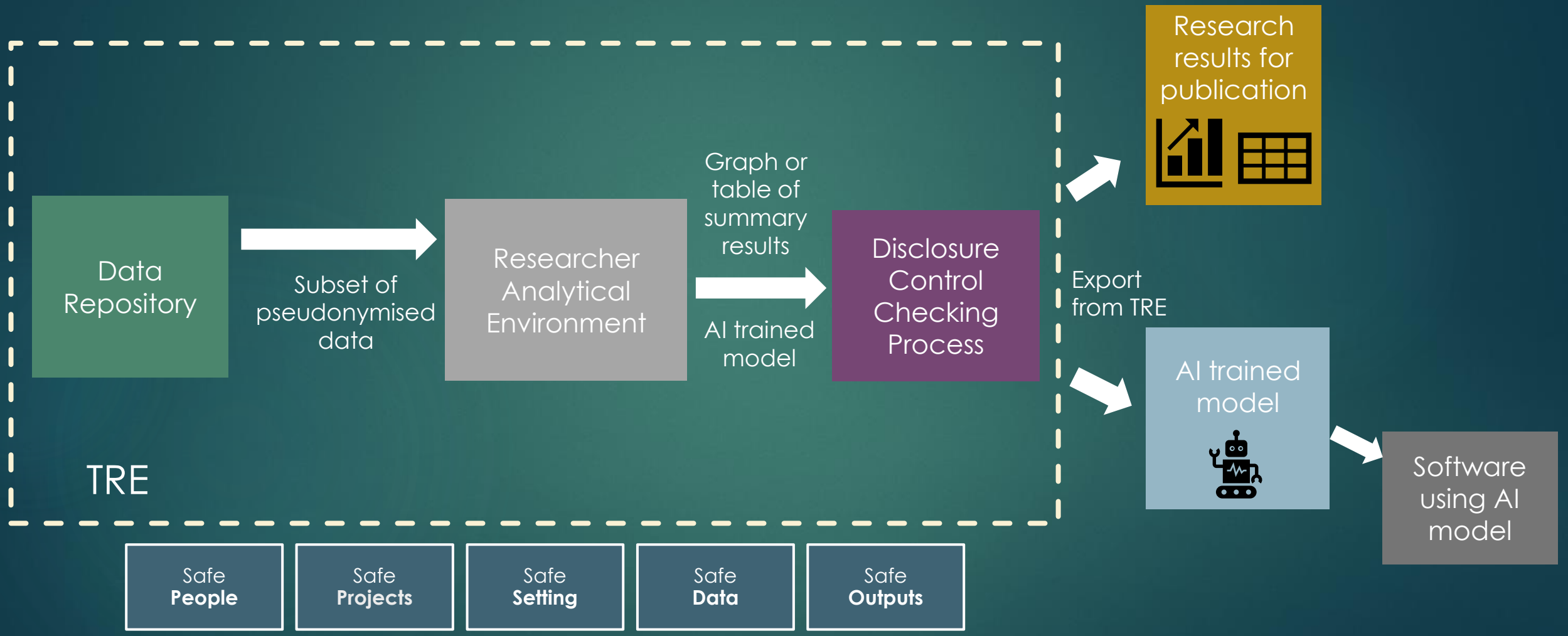


- Artificial Intelligence (AI) used in
  - spotting human errors
  - streamlining processes
  - helping with repetitive tasks
  - supporting clinical decision making
- AI training typically requires extensive real-world confidential data, best provided within a TRE
- **TREs do not have mature processes, tools or an understanding of disclosure control for AI algorithms**
- **TREs aren't engaging with AI because they don't know what to do**

Overview

DARE UK

GRAIMatter



DARE UK

GRAIMatter

Vision

To publish a Green Paper providing a set of guidelines and **implementable** recommendations supporting TREs to securely check trained AI exports for disclosure control

Work Packages

DARE UK

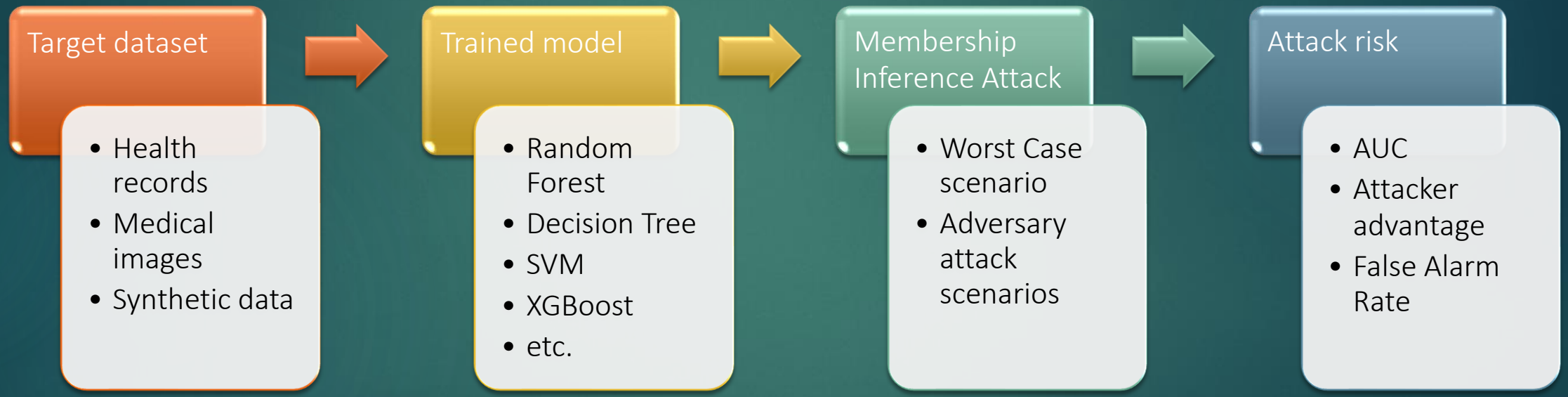
GRAIMatter

- WP1 - Risk assessment of AI models
- WP2 - Assessment of tools
- WP3 - Legal and Ethical implications
- WP4 - PPIE

DARE UK

GRAIMatter

WP1: Membership Inference Attack Simulation Framework

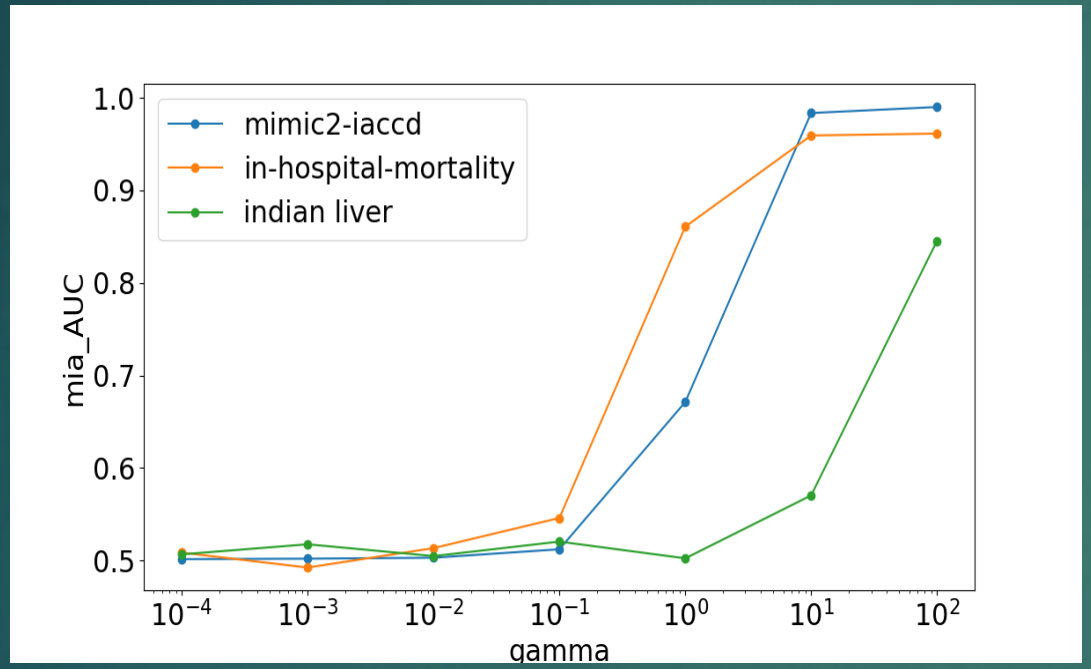


WP1 results

DARE UK

GRAIMatter

Y-axis indicator of success at a membership inference attack (higher = more successful attack)



X-axis - increasing the value of a hyper-parameter for a Support Vector Machine

- A level of attack simulation is automatable and could be run by TRE staff.
- Preliminary results suggest that there are model + hyper-param combinations that are dangerous and should be avoided (by e.g. SafeModel wrappers).
- Holding some data from researchers would make attack simulation for TRE staff more reliable.

DARE UK

GRAIMatter

WP2: SafeModel wrappers + TRE-defined “safe” values

Python wrappers around common algorithms

- Set parameters to “safe” values when model is created.
- Researcher uses them just like the version they are used to
- But then calls requestRelease()
  - Checks for common user errors
  - Produces report for TRE output checkers

Some light relief …

DARE UK

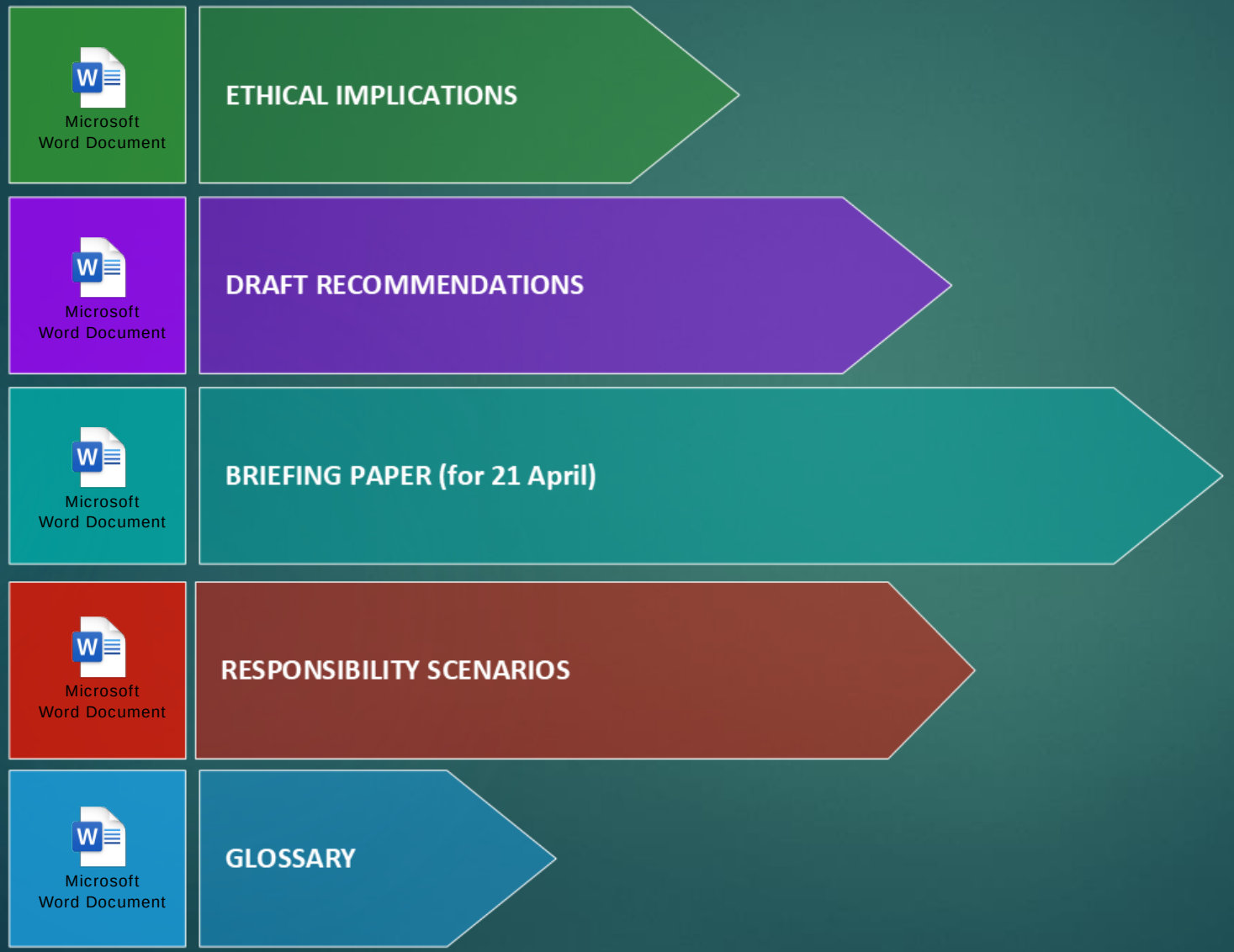
GRAIMatter

- Our intuition about how to make algorithms safe
  - Seems to be reflected in WP1 results
- Aiming to allow TREs to customise their risk appetite
  - As an institution
  - For particular data sets

WP3: Legal and Ethical

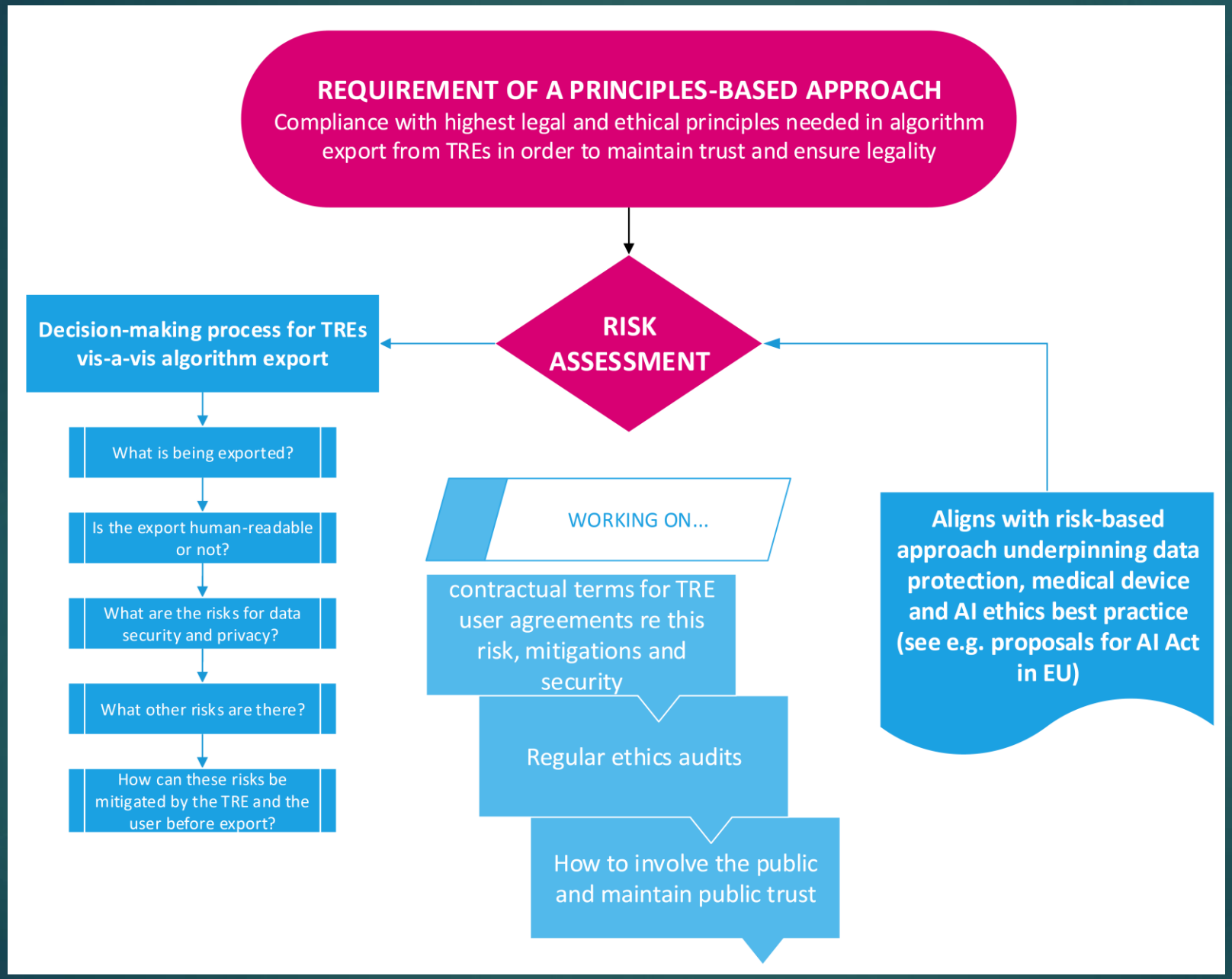
DARE UK

GRAIMatter



DARE UK

GRAIMatter



PPIE Discussions and Outcomes

DARE UK

GRAIMatter

- Two PPIE meetings held to date on 22.03.2022 & 19.04.2022
- 8 PPI participants from across the UK, many new to research
  - Introduced the concept of AI and machine learning in medical data sets
  - Explained how AI could work
  - Positioned the challenge of re-identification
- Upcoming meetings on: 24.5.22 &21.6.22
  - To position the legal challenges
  - To gain insights into the risks identified by the PPIE team

Antony Chuter:

PPIE Co-leads

Recommendations: Technical

DARE UK

GRAIMatter

- **Safe wrappers** for each machine learning model should be used by researchers in TREs
  - GRAIMatter is developing proof of concept wrappers building on python scikitlearn/tensorflow employing the safe parameters determined by WP1 experiments
  - Set of principles for safe wrappers so that the community can develop more instances e.g. R library versions
- TRE staff should run **attack simulations** – using a data set aside from training data
  - GRAIMatter has developed a proof of concept attack simulation suite which could be used by TREs
  - Set of principles for attack simulations so that the community can develop more instances

Recommendations: Legal

DARE UK

GRAIMatter

- A **risk assessment framework** is used by TRE staff
- Legal wording should be added to **user declaration forms:**
  - There is responsibility on the researcher/company to carry out due diligence of disclosive data within the trained model
  - Researcher/company agree that the TRE staff can run attack simulations on their model
- Clauses should be added to the **terms of use** of any resulting algorithm e.g. if an algorithm is embedded within a medical device the users are legally not allowed to hack the algorithm

GRAIMatter is developing suggested text

Recommendations: Legal and ethical

DARE UK

GRAIMatter

- In each research project ethical application researchers should describe the discloser risks
- Changes to the data governance approval process e.g. PBPP
  - Application forms should be modified to include information on the release of a trained model
  - Researchers should to make clear at the governance application stage that they want to disclose a model
  - Recognition of the longer term risk of discloser with justification of how the benefits out way the risks included in the application assessed by reviewers

Recommendations: Training

DARE UK

GRAIMatter

- Training courses and documentation is developed for:
  - **TRE staff** on AI, how to run attack simulations and the risks of disclosive data within trained models
  - **researchers** on the risk of disclosure control when training models – such training should be a requirement for access to the data for AI model training projects
  - **governance approval and ethical committees** to assess the applications considering disclosure risk from trained models

Next steps

DARE UK

GRAIMatter

- Team workshop next week to develop the recommendations in more detail
- Share the draft with the community
- Workshop in May to seek input into the recommendations from other sprint projects and other TREs who may be interested – please join us!
- Plan for Dare phase 2 – still loads more exciting work to be done in this area!

DARE UK

GRAIMatter

Thanks for listening!