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

Professor Emily Jefferson (On Behalf Of The Team)
The Team

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
Dr Francesco Tava: Applied ethics, privacy and trust
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
Professor Josep Domingo-Ferrer
Dr Alberto Blanco Justicia:

Dr Esma Mansouri-Bensassi: 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:
Antony Chuter:
Jillian Beggs:

Law and Ethics

International experts

PPIE Co-leads

University of the West of England
University of Dundee
Durham University
NHS National Services Scotland
Health Data Research UK
HDR UK
Health Informatics Centre
Problem statement

- 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

Data Repository → Subset of pseudonymised data → Researcher Analytical Environment → Graph or table of summary results → AI trained model → Disclosure Control Checking Process → Export from TRE

AI trained model → Software using AI model

Safe People → Safe Projects → Safe Setting → Safe Data → Safe Outputs
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

- WP1 - Risk assessment of AI models
- WP2 - Assessment of tools
- WP3 - Legal and Ethical implications
- WP4 - PPIE
WP1: Membership Inference Attack Simulation Framework

Target dataset
- Health records
- Medical images
- Synthetic data

Trained model
- Random Forest
- Decision Tree
- SVM
- XGBoost
- etc.

Membership Inference Attack
- Worst Case scenario
- Adversary attack scenarios

Attack risk
- AUC
- Attacker advantage
- False Alarm Rate
WP1 results

- 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.

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

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

- 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

- **ETHICAL IMPLICATIONS**
- **DRAFT RECOMMENDATIONS**
- **BRIEFING PAPER (for 21 April)**
- **RESPONSIBILITY SCENARIOS**
- **GLOSSARY**
REQUIREMENT OF A PRINCIPLES-BASED APPROACH
Compliance with highest legal and ethical principles needed in algorithm export from TReS in order to maintain trust and ensure legality

Decision-making process for TReS vis-a-vis algorithm export
- What is being exported?
- Is the export human-readable or not?
- What are the risks for data security and privacy?
- What other risks are there?
- How can these risks be mitigated by the TRE and the user before export?

RISK ASSESSMENT

WORKING ON...
- Contractual terms for TRE user agreements re this risk, mitigations and security
- Regular ethics audits
- How to involve the public and maintain public trust

Aligns with risk-based approach underpinning data protection, medical device and AI ethics best practice (see e.g. proposals for AI Act in EU)
PPIE Discussions and Outcomes

- 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
Recommendations: Technical

- **Safe wrappers** for each machine learning model should be used by researchers in TREs
  - GRAIMatter is developing proof of concept wrappers building on Python scikit-learn/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

- **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.
Recommendations: Legal and ethical

- 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 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 outweigh the risks included in the application assessed by reviewers
Recommendations: Training

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

- 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!
Thanks for listening!