International Journal of Population Data Science

Journal Website: www.ijpds.org

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

Emily Jefferson[1], Christian Cole[1], Alba Crespi Boixader[1], Simon Rogers[2], Maeve Malone[1], Felix Ritchie[3], Jim Smith[3], Francesco Tava[3], Angela Daly[1], Jillian Beggs[4], and Antony Chuter[4]

[1]University of Dundee

[2]NHS Scotland

[3]University of West of England

[4]PPIE Co-I

Objectives

To assess a range of tools and methods to support Trusted Research Environments (TREs) to assess output from AI methods for potentially identifiable information, investigate the legal and ethical implications and controls, and produce a set of guidelines and recommendations to support all TREs with export controls of AI algorithms.

Approach

TREs provide secure facilities to analyse confidential personal data, with staff checking outputs for disclosure risk before publication. Artificial intelligence (AI) has high potential to improve the linking and analysis of population data, and TREs are well suited to supporting AI modelling. However, TRE governance focuses on classical statistical data analysis. The size and complexity of AI models presents significant challenges for the disclosure-checking process. Models may be susceptible to external hacking: complicated methods to reverse engineer the learning process to find out about the data used for training, with more potential to lead to re-identification than conventional statistical methods.

Results

GRAIMatter is:

- Quantitatively assessing the risk of disclosure from different AI models exploring different models, hyperparameter settings and training algorithms over common data types
- Evaluating a range of tools to determine effectiveness for disclosure control
- Assessing the legal and ethical implications of TREs supporting AI development and identifying aspects of existing legal and regulatory frameworks requiring reform.
- Running 4 PPIE workshops to understand their priorities and beliefs around safeguarding and securing data
- Developing a set of recommendations including
  - suggested open-source toolsets for TREs to use to measure and reduce disclosure risk
  - descriptions of the technical and legal controls and policies TREs should implement across the 5 Safes to support AI algorithm disclosure control
  - training implications for both TRE staff and how they validate researchers

Conclusions

GRAIMatter is developing a set of usable recommendations for TREs to use to guard against the additional risks when disclosing trained AI models from TREs.

https://doi.org/10.23889/ijpds.v7i3.2005