# Recommendations for disclosure control of trained Machine Learning (ML) models from Trusted Research Environments (TREs)

UK Research and Innovation

HDR UK Health Data Research UK

ADR UK Data-driven change

An output from the Guidelines and Resources for AI Model Access from TrusTEd Research environments (GRAIMATTER) DARE Sprint Project

# 1 Contents

## 2 Current Status of Recommendations

This is the second draft of our recommendations. We are seeking feedback on this draft from the community. Please send any comments to erjefferson@dundee.ac.uk. Your time on this is greatly appreciated.

The text in *italics* throughout the document is a note to the reader, rather than part of the document text. We are using this format to explain the context to the user or (for example) to indicate that further work is ongoing on the particular topic.

## 3 Recommendations Team

Professor Emily Jefferson (Principal Investigator): Director of Health Informatics Centre Trusted Research Environment (HIC TRE) and Professor of Health Data Science, University of Dundee

**Emily Jefferson** has run a Safe Haven/TRE for a decade and a major area of her research portfolio is the development of methods to enhance Safe Havens to support next generation capabilities (such as handling of big data, scaling and performance as well as support for AI).

Dr Smarti Reel (Project Manager):

**Smarti Reel** is a postdoctoral researcher in the Health Informatics Centre (HIC), School of Medicine at the University of Dundee. Her research interests include machine learning and its use in the domain of medicine, imaging, and social media. She has worked on multi-omics biomarker discovery, multi-view image synthesis, broader image processing and other computing-based applications. She is experienced in the design, development, and evaluation of STEM projects.

### Work Package 1 - Risk assessment of AI models

Dr Christian Cole: Senior Lecturer in Health Data Science, University of Dundee

**Christian Cole** is a senior health informatician and co-leads the research team on the development of TRE Federation in Scotland and the development of next generation capabilities enabling genomics, genetics and large data projects within HIC's Cloud TRE. He has 15 years of research experience in bioinformatics and data science.

Professor  Josep Domingo-Ferrer

**Josep Domingo-Ferrer** is an international expert on information privacy and security and their interplay with AI: how to use AI to improve data protection/statistical disclosure/identifiability control, and how to ensure privacy and security in machine learning. He is a distinguished professor of computer science at Universitat Rovira i Virgili, Tarragona, Catalonia.

Dr Simon Rogers: Principal Engineer - AI Models

**Simon Rogers** is currently a principal engineer within the NHS National Services Scotland Artificial Intelligence Centre of Excellence, where he supports the development and deployment of AI solutions across NHS Scotland. Prior to this, he spent over 10 years as a lecturer / senior lecturer researching the development and application of ML methods.

Dr James Liley: Assistant Professor in biostatistics

**James Liley** is an assistant professor in biostatistics and has extensive experience in developing and ML models in TRE environments through two years' work with the Alan Turing Institute and Public Health Scotland.

Dr Alberto Blanco Justicia:

**Alberto Blanco-Justicia** is a senior postdoctoral researcher at Universitat Rovira i Virgili, Tarragona, Catalonia. His research interests include data privacy and security, privacy enhancing techniques, and ethically-aligned machine learning, specifically including robustness, privacy, security and interpretability of (distributed) ML models.

Dr Esma Mansouri-Benssassi:

**Esma Mansouri-Benssassi** is a senior research fellow at HIC, the University of Dundee working on several research topics such as dementia prediction, safe disclosure of machine learning models, and medical images feature extraction.

Alba Crespi Boixader:

**Alba Crespi Boixader** is a postdoctoral researcher at HIC, University of Dundee. She carried out her PhD research in bioinformatics at the School of Informatics, University of Edinburgh, applying machine learning methods to genomic data.

## Work Package 2 (WP2) - Assessment of tools

Professor Jim Smith: AI Models, the University of the West of England

**Jim Smith** has extensive theoretical and applied research experience in AI. He has worked closely with the Office for National Statistics for over 15 years, increasing their understanding of uncertainty/risk management, and developing AI-based tools they use for the SDC of published statistics.

Professor Felix Ritchie: 5 Safes and Disclosure Control, the University of the West of England

**Felix Ritchie** is the author of the 'Five Safes' model and advises on data governance across the world. He is an internationally recognised expert on output SDC, having developed the original theory and first generic guide for research environments in 2006, and subsequently leading both theoretical and operational developments. He devised and delivers the national training programme in output SDC.

Dr Richard Preen:

**Richard Preen** received the B.Sc. (Hons.) and M.Sc. degrees in computer science and the PhD degree in artificial intelligence from the University of the West of England, Bristol, U.K., in 2004, 2008, and 2011, respectively. He is currently a Research Fellow with the Department of Computer Science and Creative Technologies, the University of the West of England.

Andrew McCarthy:

**Andrew McCarthy** is a Research Fellow in Machine Learning with the Department of Computer Science and Creative Technologies at the University of the West of England, Bristol. He holds a B.Sc. (Hons) in Computing for Real-time Systems and an M.Sc. (Distinction) in Cyber Security. Currently, he is nearing completion of his PhD. His research interests include: Privacy and fairness of AI; Cyber Security; Secure machine learning; and methods for improving trust and reliability of AI models.

## Work Package 3 (WP3) - Legal and Ethical implications

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

**Angela Daly** is an international expert in the regulation and governance of digital technologies, in particular data protection, AI ethics, intellectual property, and medical device regulation. She is Professor of Law & Technology at the University of Dundee and the Chair of the Independent Expert Group to the Scottish Government on Unlocking the Value of Public Sector Data.

Maeve Malone: Lecturer in Intellectual Property law and Healthcare Law and Ethics, University of Dundee

**Maeve Malone** is the recipient of the Scottish Universities Law Institute (SULI) Early Career Academics grant for work on the legal and ethical framework on machine learning models.

Dr Francesco Tava: Applied ethics, privacy and trust, the University of the West of England

**Francesco Tava** works at the intersection of applied ethics, political philosophy, and phenomenology. He is interested in issues concerning data and business ethics such as privacy, trust, and solidarity.

Dr Charalampia (Xaroula) Kerasidou:

**Xaroula Kerasidou** works as a post-doctoral Research Assistant at the School of Medicine at the University of Dundee. In her research, she explores the ethical, legal, social and political aspects of new technologies such as AI.

## Work Package 4 (WP4) - Patient and Public Involvement and Engagement (PPIE)

Jillian Beggs:

**Jillian Beggs** is an experienced public/patient advocate, driving the patient engagement work package. She is a lay co-applicant on GRAIMATTER and has led similar work packages for several other large research projects.

Antony Chuter:

**Antony Chuter** is an experienced public/patient advocate, driving the patient engagement work package. He is a lay co-applicant on GRAIMATTER and has led similar work packages for several other large research projects.

# 4 Stakeholder Engagement

*To fill in once we have had input from other groups on this first draft.*

# 5 Funding

This work was funded by UK Research and Innovation Grant Number MC_PC_21033 as part of Phase 1 of the DARE UK (Data and Analytics Research Environments UK) programme, delivered in partnership with HDR UK and ADRUK. The specific project was Guidelines and Resources for AI Model Access from TrusTEd Research environments (GRAIMATTER).

# 6 Public Summary

Trusted Research Environments (TREs) provide a secure location for researchers to analyse data for projects in the public interest. TREs are widely and increasingly being used to support statistical analysis of personal data across a range of sectors (e.g., education, police, tax and health) as they enable collaborative research whilst protecting data confidentiality. TREs are often virtual environments where researchers can use their own laptop/desktop to access the remote secure TRE and the data relevant to their project. To ensure that individuals' personal data is protected, TREs apply a range of controls, such as:

- Projects are assessed to make sure they are in the public interest
- Only validated researchers who have undergone data governance training are allowed access to the data
- Only the data needed to answer the specific research question are provided to the researchers; and then only after being pseudonymised
- There is no access to the internet from within the TRE so researchers cannot inadvertently import or release any data without the appropriate checks
- Researchers must ensure that no one else can see their screens when they are working within the TRE and are not able to copy and paste from the environment
- Researchers must sign declaration forms covering their responsibilities when using the TRE and accessing data
- Normally, only aggregate, summary level data such as a trend or a graph can be removed from the TRE and used for purposes such as research publications. TRE staff perform manual and automated checks on outputs before release to prevent disclosure of individuals' personal data. This process is called disclosure control.

TREs have historically only supported the analysis of data using classical statistical methods. These result in outputs using descriptors such as trends, averages, and counts, for which disclosure control methods are well understood. There is an increasing demand to also facilitate the use of Machine Learning (ML) techniques for data analysis. Machine learning is broadly defined as training a machine to perform complex tasks in a way that is similar to how humans solve problems. The result of using Machine Learning to analyse data is a (typically complex) piece of software called a trained model. The role – and benefit – of a trained model is to make a prediction when provided with a new example. ML models have been trained for many valuable applications e.g., spotting human errors, streamlining processes, helping with repetitive tasks and supporting clinical decision-making. To realise those benefits in practice, the trained models need to be trained on data held within TREs. They then need to be released from TREs for use in the outside world. However, releasing trained ML models from TREs introduces an additional risk for the disclosure of personal data, including special category data under data protection laws, such racial or ethnic origin, or genetic, biometric, or health data. To meet legal requirements for data protection and ethical standards on fairness, accountability and transparency, particular care is needed for the safe release of such models. It is here that the size and complexity which give ML models their power present three significant challenges for the TRE's traditional disclosure-checking process:

First, even models that are 'simple' in ML terms are usually too big for a person to view easily.

Second, our research shows that a person can't say whether a model is disclosive simply by eyeballing it.

Third, and most significantly, ML models may be susceptible to external hacking using methods that reverse engineer the learning process to find out about the data used for training. These attacks can have greater potential to re-identify personal data than they would for conventional statistical outputs. This means that ML models trained on TRE data may be considered personal data(sets) and therefore fall under data protection laws.

The combination of, on the one hand, growing demand and potential benefits, and, on the other, significant challenges presented by using ML, creates a need to develop output-checking solutions specifically targeted at ML models.

We evaluated a range of tools and methods to support TREs in assessing output from ML methods for personal data. We also investigated legal and ethical implications and controls. We have developed recommendations for evaluating risk, clearing models, and ensuring good practice when developing ML models. These recommendations have been developed with input from public representatives through a series of five workshops and input from two lay co-leads within the core team.

When the recommendations are finalised, a lay summary will be added to this section. We will also add relevant contacts.

# 7 Scope

Releasing trained ML models from TREs introduces an additional risk for the disclosure of personal data, including special category data under data protection laws, such as racial or ethnic origin, genetic, biometric, or health data. This is in comparison to the mitigated risks from the disclosure of aggregate level results from classical statistical models, which are relatively well understood and managed in TREs, and usually do not include identifiable personal data, triggering the application of data protection laws. The key challenge is responsibility for a possible data breach resulting from the release of a trained ML model, where the output checker does not have access to human-readable outputs. The model inversion and membership inference attacks may lead to the identification of personal data, thereby potentially rendering the model a personal dataset, to which data protection law would apply [1].

These recommendations provide additional disclosure controls for the safe release of trained machine learning models from TREs to protect personal data. These are different from the controls required for aggregate level results from classical statistical models. However, **the scope assumes that TREs already apply the high-level "Five Safes" controls** [2], **and implement more granular level controls** such as those recommended in the Health Data Research (HDR) UK Principles [3] and Best Practices for Trusted Research Environments paper [4] and the Scottish Safe Haven Charter [5].

The 'Five Safes' is a popular way to structure thinking about data access solutions in the UK. Originally used mainly by UK statistical agencies and social science academics, in recent years it has been adopted more widely across the

UK government, health organisations, and private sector bodies. We briefly describe the 'Five Safes' framework, how it is used to organise and simplify decision-making, and how it helps to address the concerns of different constituencies. We show how the framework aligns with recent regulations, anticipating the shift towards multi-dimensional data management strategies. We provide several practical examples as case studies for further information. We also briefly consider what issues the 'Five Safes' does not address, and how the framework sits within a wider body of work which challenges traditional data, access models.

The 'Five Safes' comprises:

**Safe People:** The researchers accessing the data through TRE are trained and authorised to use the data safely, follow guidelines, and report data safety concerns if any.

**Safe Projects:** TREs ensure that the research projects are approved by data owners, and that data are used appropriately and for public benefit.

**Safe Outputs:** TREs screen all outputs thoroughly and approve the release only after ensuring that it is non-disclosive of personal data.

**Safe Data:** The data are de-identified/pseudonymised before access is granted to researchers. It is ensured that researchers only see the data that they need to.

**Safe Setting:** TREs provide a safe environment to access personal data and prevent any unauthorised use.

**Many of our recommendations utilise and extend the controls already applied using the 'Five Safes' model**. For example, all research projects must have appropriate ethical approvals, but we have identified specific ways in which that process could usefully be amended. Rather than creating a new ethical approval process for assessing the implications of the additional disclosure risk posed by releasing trained ML models, the recommendations suggest that details of the additional disclosure risks be included within a standard ethical application. This will allow the risks to be balanced against the benefits of the activity along with the controls to mitigate the risks.

This report makes many recommendations, covering both technical and operational measures. At this stage, we are **not requiring (or indeed, expecting) that a TRE would adopt all of the recommendations**. We have analysed the use of ML from a conceptual, worst-case perspective, and actual ML modelling in TREs will likely provide evidence for what works in practice. For example:

- if most modelling in TREs does not approach the worst-case scenarios analysed here, then unsophisticated checks, carried out by staff with lower training may be more appropriate with only expertise brought in as needed; if, on the other hand, it appears that most ML models are high risk and bespoke (*sui generis)*, then an equivalent level of expertise in protection may be necessary.
- we have proposed multiple solutions to give TREs the flexibility to tailor their output checking regime to their particular needs; it may be that operational tests will show that; e.g. responses A and B together or response C on its own provides adequate protection but that A, B and C is unnecessary.
- we have not considered in detail the implementation of restrictions by researchers but have used our experience as researchers and conceptions of what would be an acceptable output checking response; perhaps researchers will express a strong preference for response X over response Y because X is easier to incorporate in their code.
- ML models required to be 'safe' will have less predictive accuracy or usefulness than models which need not meet safety requirements. When researchers are aiming to release a 'safe' model they will typically

wish to compare the 'safe' model with an unrestricted, potentially unsafe model within the TRE environment to assess attenuation in performance. The question of whether 'safe' models can maintain adequate levels of accuracy or usefulness as compared to potentially unsafe models is highly application-specific, so we defer the question to researchers rather than trying to answer it in general.

*Although it is too soon to have experience with these recommendations in practice, we welcome feedback on the likely implementability of our recommendations, singly or in combination.*

This report focuses on ML models exported from TREs. With a mandated point of release, the TRE's output checking routine, and compulsory output checking can be enforced, as well as other checks (e.g., for compatibility with the original ethical approval). However, the protection measures identified here are also usable in non-TRE situations. As ML modelling generates the same confidentiality risks for a given dataset, whether inside or outside a TRE, good research practice would suggest these checks are applied by non-TRE modellers as well.

*We welcome view on how feasible it is to enforce checking outside TREs. Should there be, for example, a register of 'approved' models?*

## 7.1 Out of scope:

The following are beyond the scope of our recommendations:

i. The ethics of ML model training and use of the trained model i.e., the legal and ethical implications of using the trained model such as developing and using a medical device in a clinical setting. Such topics are assessed via the standard ethical approval process, whether or not a TRE is involved.
ii. Medical device regulation, which comes into effect after a trained ML model has been released from the TRE.
iii. Artificial Intelligence (AI) techniques other than Machine Learning (ML). ML is just one type of Artificial Intelligence (AI). Here we consider ML only. Updating existing ML models, as this requires additional statistical work to fitting models for the first time [6].

Assumptions:

a) It is assumed that TREs run a service and do not have intellectual property (IP) over or intellectual input into ML development, which is done by researchers.
b) The data used for model training within the TRE are pseudonymised following good practice pseudonymisation methodologies.
c) The existing best practices for doing research in TREs are followed, as per Health Data Research (HDR) UK Principles and Best Practices for Trusted Research Environments paper [4].
d) The ML models are being trained entirely on data within the TRE as opposed to being adapted from models initially trained on external data (transfer learning using pre-trained models).
e) That ML models can be considered separate and apart from personal data, but can also be considered personal data [1].

## 7.2 Audience

The audience for these recommendations includes all of the different groups responsible for implementing these recommendations, i.e. researchers, TREs, data controllers, data governance committees and ethical committees. These groups need to understand the recommendations, with specific groups responsible for applying the recommendation. For example, data governance committees need to understand the controls which are used by TREs to assess a project application which describes these controls, researchers need to understand the controls to describe them in their project application, and a TRE needs to specifically apply the controls during the project. For each detailed recommendation, we have indicated the groups which need to understand the recommendation or are responsible for applying the recommendation.

# 8 Glossary

### TREs AND ACTORS

**TRE:** a trusted research environment sometimes called a 'data enclave', 'research data centre' or 'safe haven' is an analytical environment where the researchers working on the data have substantial freedom to work with detailed row-level data (see below) but are prevented from importing or releasing data without permission, and typically are subject to a monitoring and a significant degree of access control

**Researchers:** someone who has permission to use a TRE and has access to data within that TRE. For this work, researchers are assumed to be interested in building machine learning (ML) models that they will then wish to remove from the TRE. We use the term "researchers" to mean researchers who could be academic or from a commercial or government setting – all of whom could be training ML models to develop a solution in the public interest.

**Data controller:** a person(s) or body corporate, who alone or jointly with others determine the purposes for which and the manner in which any personal data are, or are to be processed (or who are the controller(s) by virtue of the Data Protection Act 2018, section 32(2)(b) (by means by which it is required by an enactment to be processed)), and all of whom are officially registered on the Information Commissioners Office Data Protection Public Register [7]. Data controllers need to approve the use of their data for research projects. In this document, we are using the term Data Governance Committee to represent the group/person by which data controller(s) approve applications to use their data for the research project. TREs can be a service which is run by the same organisation who is the data controller or can be a service which is run by a separate organisation to the data controller. We have separated the roles of data controller and TRE within these recommendations.

**Joint Data Controller:** a person(s) and/or body corporate, who jointly determine the purposes for which and how any personal data are or are to be processed (or who are the controller(s) under the Data Protection Act 2018, section 32(2)(b) (by means by which it is required by an enactment to be processed)) and all of whom are officially registered on the Information Commissioners Office Data Protection Public Register [7].

**Attacker or adversary:** a person or group of persons or an organisation who attempts to extract, from the trained ML model, some or all of the personal data that was used to train it.

**Personal data:** 'any information relating to an identified or identifiable living individual' as per section 3 of the UK Data Protection Act 2018 (which implements the EU's General Data Protection Regulation or GDPR). Personal data is broadly interpreted, especially via the concept of 'identifiable' which is further defined as 'a living individual who can be identified, directly or indirectly in particular by reference to (a) an identifier such as a name, an identification number, location data or an online identifier, or (b) one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the individual'.

**Special categories of personal data:** the term used in current UK and EU data protection legislation for certain more sensitive categories of personal data. The special categories are: race; ethnic origin; political opinions; religious or philosophical beliefs; trade union membership; genetic data; biometric data (where this is used for identification purposes); health data; data about sex life; and sexual orientation. There are additional legal requirements which must be met to process special category personal data.

**Synthetic data**: data artificially generated to replicate statistical properties of a given real-world dataset, ideally not containing genuine identifiable information i.e. no personal data.

**Row-level data:** a data set whose columns are different types of measurements/images/features, and where each row contains the record of an individual person or organisation. It is synonymous with '**record-level data**' or 'microdata', and in contrast to aggregated data.

**Aggregate level data**: summary data which is acquired by combining individual-level data and maybe collected from multiple sources and/or on multiple measures, variables, or individuals. It is synonymous with '**aggregate(d) data**' or '**statistical results**'.

### MACHINE LEARNING

**Algorithm:** A set of instructions to execute a task. This is a very general definition; algorithms may be deterministic (always giving the same answer when presented with the same input) or stochastic (giving different answers with various probabilities). Algorithms are not necessarily run by a computer; humans also use algorithms implicitly when making decisions. We will usually use the term to mean a set of instructions which only refer to data generically, rather than a specific dataset. An example of an algorithm is the ordinary least squares (OLS) method for fitting linear models.

**ML model:** Some computer code which implements an algorithm that, when presented with some input data, processes it in some way, and produces some output. There are many possible ML models, differing by the particular type of process they implement, and how they implement that process. For example, a particular ML model might process images (the input data) to assign them into a category (the output), which might be useful when attempting to build an ML system for diagnosing disease from a medical image. To successfully perform a particular task, ML models must be trained. This process involves providing them with many examples of input data and the corresponding correct output (or just input data in some cases), from which they learn the patterns that enable them to generate meaningful output for inputs that they have not seen before.

Formally, a model is a set of candidate distributions over the domain of a given dataset (we leave differentiation between classical statistical models and ML models unspecified at present: there is no general distinction, and whether a model constitutes an ML model is best determined on a case-by-case basis)[8].

**Trained ML model:** ML models are not usable until they have been trained. Training involves presenting the model with data that is relevant to the task at hand and modifying any parameters within the model to optimise its performance in the task of interest. For example, an ML model that is to be used for diagnosing breast tumours from mammograms will be trained with mammograms (input) with known tumour status (output). The training process will modify the parameters within the ML model such that the number of mistakes it makes on this "training" data is minimised. Once trained, the ML model can be used to generate an output for inputs that were not part of the training data: for example, to predict the tumour status for a new mammogram.

Formally, a trained ML model is one of the candidate distributions of an ML model.

**Predictions:** The usable output of an ML model when given some data. Typically, this is the estimated chance of something happening given a set of inputs, where the estimation is made by the model. In the example above, the prediction would be the chance that the mammogram shows a real tumour.

Formally, a prediction is a conditional distribution derived from a trained ML model.

**Features:** independent variables, often organised in columns in a given dataset used to train the model; e.g., age, sex, medical history, heart attack incidence.

**Target model:** an ML model (untrained, trained or being trained) that is the target of an attack.

**Instance based models:** Models which, to be able to make predictions, must 'remember' one or more training data samples exactly, rather than just summary data. These provide an immediate security risk, since specifying the model entails specifying individual samples. Such models are sometimes able to be made private by transforming training data samples randomly and only remembering the transformed samples [9].

**Ensemble methods:** the use of multiple methods, usually with their outputs combined through some form of the voting process, done to improve overall performance. This may involve the use of the same method on different parts of the dataset (e.g., random forests [10] gradient boosting methods (e.g. XGBoost [11])) or different methods applied to the same dataset (e.g., super-learners [12]).

### Trained ML behaviours

**Machine learning model architecture:** The ML architecture specifies the various layers involved in the machine learning cycle: data acquisition, data processing, model engineering, execution and deployment. Broad categories of architecture are supervised learning, unsupervised learning, and reinforcement learning. Within each category, the architecture specifies the learning algorithm (e.g. neural networks, random forests, etc.) and its internal structure (number and type of layers). A trained ML model is saved to a computer-readable file. Such a file could be loaded and used to make predictions or loaded to inspect the properties of the model.

**Hyper-parameters:** High-level parameters that can control the aspects such as the model architecture (number of layers in a neural network, maximum depth of a decision tree, etc) and the learning process through which one particular trained model is chosen from all the possibilities. These are typically fixed by the researchers and are not learnt during training.

**Generalisation:** The ability of a machine learning model to make predictions on data that it did not see during training.

**Overfitting:** Situation in which a model fits and remembers the training data too well and does not generalise well for unseen data. Overfitting can facilitate membership attacks. Typically, small or unrepresentative training datasets can lead to overfitted models, especially if the data points have many features. A bad choice of hyper-parameters can also lead to overfitting (for example, excessively big neural network for simple classification tasks). Detection of subtle overfitting is difficult and a fundamental area of ML theory. More egregious overfitting can be readily identified by non-experts.

Methods to reduce overfitting include: increasing the training dataset size, possibly using data augmentation techniques; using 'regularisation' techniques during training, which penalise candidate models for complexity; and optimising the choice of hyper-parameters, possibly with cross-validation. In neural networks, it is often beneficial to include dropout layers, which randomly deactivate neurons during training (effectively making the training procedure noisier). Differentially private optimizers (such as DP-SGD) add noise during the optimization steps/training process and often lead to better generalization.

**Data augmentation techniques** generate training samples from existing samples. In the case of images, a typical technique is to resize and rotate images in the original training set to generate new samples.

**Federated learning** is a technique that allows a machine learning algorithm to be trained on data that is stored in a variety of servers, devices, or TREs. The trained algorithm parameters (not data) are pooled into a central device which aggregates all individual contributions into a new composite algorithm.

### Attack types

**White box:** a type of model attack where the attacker knows some information about the training data, the target model classifier, architecture and learned parameters of the target model. For example, this might include knowing the weights of a neural network, or the decision thresholds in a rule or tree-based model.

**Black box:** a type of model attack where the attacker has only query access to the model. That is, they can present input data to the model and observe the predictive outputs that the model makes. For example, in a model that detects the presence/absence of a tumour in an x-ray image, the attacker can present an image to the model and will receive the probabilities that a tumour is present or not. Black box attacks do not have access to the interior of the model.

### Risk types

We have identified 3 major risk types (RT):
RT1: non-malicious researchers training models inside the TRE, naive to possible threats faced by those models and saving inappropriate information within the disclosed model file.
RT2: malicious researchers deliberately hiding data inside disclosed files.
RT3: an external attacker with access to disclosed model files after release.

### Disclosure Risks

**Membership Inference** is the risk that an attacker (of whatever coloured box) can create systems that identify whether a given data point was part of the data used to train the released model. This risk is far more likely to be disclosive of special category personal data in cases of medical data (X was part of a trial for a new cancer drug) than it is for other forms of data TREs might hold (Y was part of a survey on educational outcomes).

**Membership Inference Attacks (MIA)** is a type of attack where an adversary wants to predict whether row data, which belongs to a single patient, was included in the training data set of the target model.

**Attribute Inference** is the risk that an attacker, given partial information about a person, can retrieve values for missing attributes in a way that gives them more information than they could derive just from descriptions of the overall distribution of values in the dataset.

**Attribute Inference Attacks (AIA)** is a type of attack where the adversary is capable of discovering a few characteristics of the training data.

**Individual Disclosure** occurs when outputs from an analysis segment the participants in such a way that one sub-group has only a few members (for example, the mean income of left-handed vegetarian professors of underwater knitting was £Y). This is especially risky if external factors might make it possible for one person who knew they were part of a small group to identify the others. For traditional statistical analysis, where outputs might be tables designed by the researchers, a common threshold rule might be "don't release cells with data from fewer than 3 people". The creation of AI models automates an equivalent process to the researchers hand-designing a table, so the same risks apply. It remains to be seen exactly how the traditional disclosure rules relate to membership inference.

### Metrics and technical terms

**Mechanism** is a term describing a procedure which takes a dataset and outputs some information about it. Usually, this is random-valued and considered for fixed data. As an example, given data D=(X1,X2,X3), a mechanism M might return a 'noisy mean' M(D)=(X1+X2+X3)/3+λ, where λ is a random variable.

**Differential Privacy** is a measure assigned to an output mechanism roughly stating how similar outputs can be when their training data differs by only a single sample. Limiting differential privacy ensures that a malicious attacker with access to all-but-one of the training samples would have difficulty inferring the values of the last training sample. For a full definition and treatment see [13], particularly chapters 2 and 3. Differential privacy can be quantified as:

- **(Epsilon) ε-differential privacy**: the probability/density of seeing some output of the model never changes by more than a factor of **ε** when changing one sample.
- **(Epsilon,Delta) (ε,δ)-differential privacy**: the probability/density of seeing some output of the model usually does not change by more than a factor of ε when changing one sample, except for a set of values which have probability δ (delta) of being observed. The lower ε and δ, the more private the mechanism. Typically, lower ε and δ correspond to less useful mechanisms, in that they contain less useful information about the true dataset.

### Ethical and Regulatory terms

**Data Governance Application:** An application form sent to a Data Governance Committee which explains how the data will be used for the research project. Details of the risks and controls to personal data need to be articulated within the form along with the benefits of the project.

**Data Governance Committee:** The group/individual who assesses the data governance application and approve/disapprove. Such committees include the Data Controllers responsible for approving the use of their data for the project. In some instances, this could just be a single responsible individual data controller who approves the application e.g. Caldicott guardian in the case of health data. In other instances, the committee may include representatives such as lay members of the public.

**Data Governance Approval Process:** The process by which researchers fill in a data governance application and send this to a Data Governance Committee for approval/disapproval.

**Ethics Application:** An application form sent to an Ethical Board which explains how the data will be used for the research project. Details of the risks and controls to personal data need to be articulated within the form along with the benefits of the project (e.g. individual, commercial, societal etc.)

**Ethical Committee:** The group/individual who assesses the ethical application and approve/reject. Such committees include local research ethics committees (LRECs) based in specific universities and research centres as well as multicentre research ethics committees (MRECs), which were introduced in the UK following the publication of Department of Health guidance HSG (97)23 to deal with multicentre research.

**Ethical Approval Process:** The process by which researchers fill in an ethical application and send this to the ethical board for consideration.

**Contractual Agreement**: A binding legal document, signed, sealed and delivered by contracting parties.

**Linked Agreement**: This is an existing or contemporaneously agreed contractual agreement between contracting parties. A linked agreement can be incorporated into the terms of a contractual agreement. This might be necessary if the identity of the initial data controller or processor has changed, or if there was to be a change to the permitted purpose of the trained ML Model, as outlined in the original data governance and ethics approval. An example of a permitted purpose assigned to a trained ML Model might be: functions designed to protect members of the public. An example of a linked agreement would be the initial (or updated) data governance and ethics approval process for the development of ML model processing personal data. Another example of a linked agreement is a data sharing agreement.

### OTHER

**Safe Wrapper:** code that unobtrusively augments the functionality of existing software for machine learning. Typically, when a safe wrapper is applied, the model will retain the 'look and feel of its original version whilst adding functionality to:

- Automate the running of various attacks to assess the vulnerability of a trained model.
- Assist researchers in meeting their responsibilities, by warning when their choices for hyper-parameters or components are likely to result in models that are vulnerable to attack - and make suggestions for alternative choices.
- Detect when researchers have either maliciously or inadvertently changed important parts of a model (or hyper-parameters) between training and requesting release.
- Produce reports for TRE output checking staff summarising the above, to assist them in making good decisions about whether to release trained models.

**GitHub:** An open online platform that lets people work collaboratively on projects/software codes from anywhere while tracking and managing changes to software code.

**Encryption:** A process which protects personal information by scrambling the readable text into incomprehensible text which can only be unscrambled and read by someone who has access to a specific decryption key.

**Restrictive software:** Software that is subject to conditions or limitations imposed by a technology licensor or supplier and requires consent before its disclosure or assignment to third parties.

**Public release:** This means making the content of a work public through publication, presentation, broadcast or other means.

**Kernel based methods:** Group of model types that are used for pattern analysis. They use similarities between observations to build the model rather than the observations themselves. They are almost always instance-based methods, meaning that at least some of the training data must be saved within the trained model.

**Public representatives:** People who can represent the public interest and protect the integrity interests of the individual.

**Release** (when referring to a model): To export a trained machine learning model outside the TRE for deployment and for making predictions. Another term for this is "egress".

**Grant of a License or Transfer** (when referring to a model): To release a model to specific researchers, by way of the grant of a license. A license is legally and contractually binding.

**Deploy** (when referring to a model)**:** To set up the trained model within an environment where it can be efficiently used to make predictions.

**License:** a type of contractual agreement to authorise the granting of a license, to enable the use or release of a machine learning model. It can be perpetual or non-perpetual. Examples of open-source licenses that could be considered to accompany and apply to a trained ML model are the MIT License. *Another ethical license we are researching further is the RAIL License* [14].

**Reproducibility:** means achieving a high degree of reliability or similar results when the study/experiment/ statistical analysis of a dataset is replicated.

**Identifiability Controls:** Controls on privacy achieved exclusively through controlling aspects of the trained ML model, under the assumption that unlimited prediction queries may be made using the model by an attacker.

**Model Query Controls:** Controls on privacy achieved by restricting access to or use of the trained ML model after release.

**End User:** A person who uses the ML model outside the TRE by obtaining access to the model either through access to a resource storing the code (e.g. a version control system such as GitHub), software that implements the model, or a webservice that allows the model to be queried. This person is often a different person from the researchers who trained the model.

# 9 Executive Summary

## 9.1 Background

Trusted Research Environments (TREs), also sometimes referred to as safe havens, data enclaves or research data centres, have become widely used to support observational research on sensitive pseudonymised linked data within a secure virtual environment. Within TREs researchers can access row level data but cannot release this data; only aggregate level results (e.g., graphs, summary tables, regression models) can be released. There is a range of controls implemented by TREs, adhering to legal requirements (especially under data protection legislation) and the 'Five Safes' model [2]. These controls were expanded upon by the Scottish Safe Havens [5] the Turing Institute

[15] and Health Data Research UK [4]. TREs assure data controllers that their data can be securely shared for research purposes without risking individual or patient confidentiality, resulting in a more scalable, streamlined process for population-level statistical studies using health and administrative data.

With the emergence of AI (a subset of which is Machine Learning, or ML), researchers are now pushing the capabilities of TREs to support the development of trained ML models. Applications of such models include clinical decision support, streamlining public sector processes and spotting overlooked issues in routine healthcare provision or other sectors. To support ML development, TREs need to provide additional tools and computing resources. They also need to ensure that the trained ML models to be released from the environment do not contain any record level, personal data i.e. disclosure control needs to be applied to the trained ML model. The disclosure control guidelines developed for 'traditional' analyses are inappropriate for ML models [17]. ML model disclosure control is also needed to support Federated Learning across TREs.

Appendix G provides some real-world examples covering 5 different scenarios. The examples are written for a lay audience, with links to code evidencing these examples for technical readers:

- **Finding out unknown personal data about a famous person:** we show that if a model has been trained using data from a famous person - for example a Member of Parliament (MP) who attends a hospital in an area which provided the training data - and a hacker already knows some information about the famous person from data in the public domain, they may be able to query the trained ML model to find out the other data relating to the famous person. In our example, the hacker knew the MP suffered from cancer, is diabetic, asthmatic, smokes, and is 62 years old. The hacker could query the model many times in a systematic, automated way and find out that it is highly likely that the MP was also in an overweight BMI category and had slightly high blood pressure.
- **Finding personal data from publicly available data:** Some models, if trained with inappropriate parameters or overtrained, can contain data that can easily be extrapolated to the participants of the study just by knowing information that most people are happy to share publicly like birthday. The example illustrates a case to predict which drug users are at high risk of insolvency, and an agency that hires people found out about this model. If the agency can prove that any job interview candidate took part in the model that means that they are drug users, and therefore not given the job.
- **Identifying if someone famous has suffered from cancer:** Often, some details about the health of famous people are well known, either they admitted themselves or someone leaked it to the media: for example, smoker, asthmatic, etc. If an attacker realises that the training data from a machine learning model to predict the outcome of cancer treatment means only people with cancer were included in the study then the attacker could use the publicly available health status to prove a famous person had cancer.
- **Successful candidates in a job interview:** In this case, we demonstrate that sometimes the outcome of the model, can have easily identifiable patterns depending on whether a person was included or not in the training data, and it could have unexpected consequences if misused. We illustrate how researchers want to help drug users when they are at high risk of insolvency. If an employer finds out a candidate was part of this research, and therefore is a drug user (regardless of economic problems or not), the employer is likely to refuse to employ this person.
- **Hospital admission survival:** This example demonstrates how instance-based models contain data of some patients that were part of the training data and they can be retrieved from the model. In this case, the machine learning model was to predict the chances of surviving for patients admitted to the hospital.

Luckily, the TRE output checkers spotted the problem and did not allow this model to be released publicly, so the researchers had to decide on an alternative approach.

In the earlier work, we interviewed 14 UK and 6 international TREs [16] to discover how TREs check ML models for personal data before approving release. It was found that universally TREs did not have mature processes, tools or an understanding of disclosure control for ML models. The existing processes were manual and did not consider various risks. A recent work discussed a wide analysis of the risks associated with ML vulnerability at prediction time [17]. We also reviewed different types of ML models for the potential to encode individual-level data or personal data and assessed security threats and vulnerabilities [18], [19]. Three major Risk Types (RT) were found:

- RT1: non-malicious researchers training models inside the TRE, naive to possible threats faced by those models and saving inappropriate information within the disclosed model file
- RT2: malicious researchers deliberately hiding data inside disclosed files
- RT3: an external attacker with access to disclosed model files after release.

Attacks carried out by such actors can include: discovering whether someone's data was in the training set (Membership Inference), recovery of personal data given a partially complete record (Attribute Inference) or uncovering `whole records' likely to be in the training set (Model Inversion). This previous work proposed some very high-level mitigation strategies including restricting types of models used, avoiding overfitting, model/code inspections pre-disclosure, testing the model to be released on a subset of data withheld from the researchers, assessment of attack risks, and use of synthetic data and models.

The UK data protection regulator, the ICO, provides a great overview of the different types of risks and methods to mitigate these risks [20].

DARE UK (Data and Analytics Research Environments UK) has funded the GRAIMATTER (Guidelines and Resources for Artificial Intelligence Model Access from TrusTEd Research environments) sprint project to develop a set of guidelines and recommendations for how TREs should carry out disclosure control on ML models. DARE UK is a programme funded by UK Research and Innovation (UKRI) to design and deliver a more coordinated national data research infrastructure for the UK. **This recommendation green paper is one of the main outputs of GRAIMATTER.**

GRAIMATTER has assessed technical, legal and ethical, training, costing and PPIE (patient and public involvement and engagement) risks and controls. Section 9.2 provides a summary of the recommendations. A summary of the investigative work carried out within each category (Technical / Ethical and Legal Aspects (ELA) / Costing / PPIE / Training), is provided along with detailed recommendations in Sections 10 to 14. Recommendations for areas for future development and research are provided in Section 15.

High-level recommendations are numbered "R (n)". Detailed technical recommendations are numbered "TECH (n)", Legal and Ethical recommendations are number "ELA (n)", costing recommendations are numbered "C (n)", PPIE recommendations are numbered "P (n)" and training recommendations are numbered "TR (n)".

## 9.2 Summary of investigative work

The GRAIMATTER project investigated several areas:

- **Quantitative assessment of the risk of disclosure from different ML models:** We trained models across a range of parameter/hyper-parameter values and data types and assessed the performance of disclosive MIA in each case. We have formalised and implemented different types of AIA and repeated the large-scale testing for this risk. This enabled us to identify factors for assessing risks in disclosure of these model types, and parameter and hyper-parameter regimes to avoid. Finally, we considered the practical relevance of these risks.

- **Controls and Evaluation of Tools**: We evaluated a range of tools to determine their usefulness in the semi-automating assessment of disclosure risk. Our evaluation considered the current 'effectiveness', requisite level of support/maintenance, and the risk of these tools themselves becoming part of an 'arms-race'. We considered approaches, where a model fitted to synthetic data [21], [22] is released instead of the true model, which partly shifts [23] privacy considerations to the synthesis process. We developed Python 'wrappers', around commonly used modelling functions (scikit-learn/Tensorflow), automatically assessing disclosure risk and producing reports to assist the output-checking team.
- **Legal and ethical implications:** We investigated the legal and ethical issues accompanying ML model release from TREs. We identified and assessed how current UK legislation applies to TREs supporting trained ML model release and the extent to which the legislation addresses ethical issues pertinent to the release of ML models out of TREs, and what happens after that release, developed based on personal data, such as transparency, privacy, data protection and non-discrimination. We looked at the main legal obligations incumbent on researchers and TREs, including protecting the confidentiality and security of the processing of the personal data held by the TRE and used for training ML models. We took into account the duty of data controllers to protect confidentiality but also share data in the public interest. We drew from the field of AI ethics and governance from an international level (UNESCO Draft Recommendation on the Ethics of Artificial Intelligence [28]) and EU level (proposed Artificial Intelligence Act) to inform our analysis, especially in considering the reform of applicable UK frameworks.
- **Public Engagement:** We ran 5 PPIE workshops – chaired by our lay co-applicants and including 8 members of the public selected with consideration for diversity including, but not limited to, gender, ethnicity, age and geography. The workshop structure included explaining the challenge with machine learning and AI, outlining the legal challenges, presenting project outputs and agreeing on the next steps. All workshops were followed up with a one-to-one call to check understanding and offer additional support if needed. The feedback we received feeds into these recommendations.

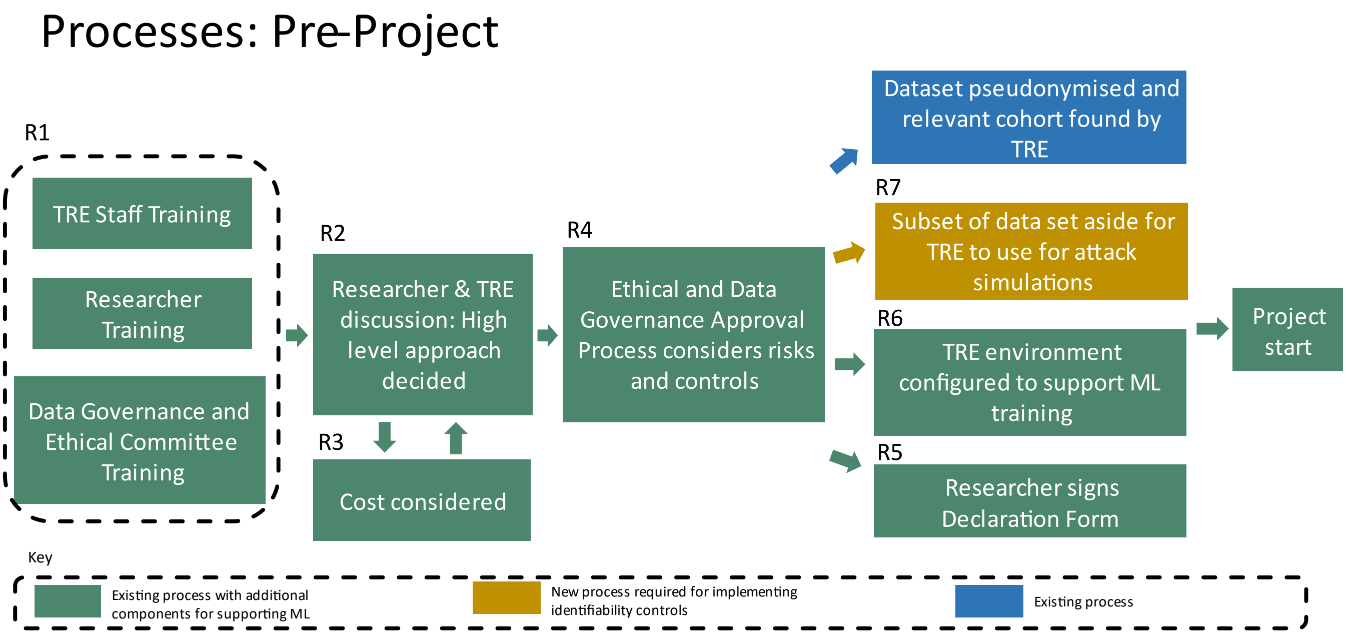
Although risks of disclosure from trained ML models are widely recognised and there is significant research in this area both by academic and commercial groups, during our literature search we did not find any examples of a data breach from trained ML models. There were many examples of data breaches of sensitive data including special category personal data from other sources. This is because ML training and use are in relative infancy – particularly with health data. It is appropriate that recommendations such as these are implemented before such breaches occur.

## 9.3 Summary recommendations

To summarise the recommendations, we have grouped them into the stage of the project where they occur and we have listed the relevant detailed recommendations which apply to each high-level recommendation. As explained within the Scope (Section 7) these recommendations utilise and extend the controls already applied using the 'Five Safes' model such as those recommended in the Health Data Research (HDR) UK Principles [3], Best Practices for Trusted Research Environments paper [4], and the Scottish Safe Haven Charter [5]. For example, the data provided to researchers is pseudonymised, there are checks on the individuals accessing them to ensure they are *bone fide* researchers, and the TRE is configured to limit open access to the internet.

'Researchers' could be either academic researcher(s) or researcher(s) from industry or government.

## 9.3.1 Pre-Project



Processes: Pre-Project

R1

TRE Staff Training

Researcher Training

Data Governance and Ethical Committee Training

R2

Researcher & TRE discussion: High level approach decided

R3

Cost considered

R4

Ethical and Data Governance Approval Process considers risks and controls

Dataset pseudonymised and relevant cohort found by TRE

R7

Subset of data set aside for TRE to use for attack simulations

R6

TRE environment configured to support ML training

R5

Researcher signs Declaration Form

Project start

Key

🟩 Existing process with additional components for supporting ML

🟨 New process required for implementing identifiability controls

🟦 Existing process

*Figure 1 - Pre-project process where the recommendations apply. Green boxes show existing process which take place for “traditional” TRE projects and where new recommendations apply to support ML training. The blue box shows a process which already takes place and where modifications to support ML training are not required. The mustard coloured box shows a new process which is required to support identifiability controls for ML models. The “R” labels correspond to the high level recommendations within the main body of text.*

### *9.3.1.1 R 1: All groups involved in the process should be appropriately trained.*

Prior to ML training projects commencing, all groups involved in the process (TREs, data governance committees, ethical committees, and researchers) should undergo specific training relevant to their group to ensure they are aware of the risks and controls which could be applied re disclosure control of personal data, the legal and ethical implications, and their responsibilities within the process (TR1, TR2, TR3, TR4, TECH 5).

There should be a group of TRE trained experts on ML disclosure control who can work across TREs, reducing the burden on each TRE (TECH 5, C5).

**Relevant Detailed Recommendations:** TR1, TR2, TR3, TR4, TECH 5, C5

### *9.3.1.2 R 2: Researchers should discuss their plans with TREs and decide upon a high-level approach which applies appropriate disclosure controls and mitigates risks*

Once the TREs and researchers are trained in the risks and controls which could be applied, they should have an informed discussion regarding the relative merits and challenges of different approaches for the specific research project, and they should agree on an approach and high-level project plan (TECH 1). The TRE can support the researchers to decide the approach to be taken. We are not recommending any one of these approaches over another; each has advantages and disadvantages. Both model query controls and identifiability controls, if applied correctly, will protect personal data. The decision will be a combination of researcher preference and whether the data controller is supportive of the approach.

The approach for managing disclosure control risk can be categorised into 4 high-level options: model query controls and 3 types of identifiability controls. These options are explained next:

- **Model Query Controls:** Rather than the model being released openly, researchers can choose to limit the queries on the trained model (TECH 16), e.g. the model could be hosted within a secure webservice which technically limits who can query the model and the number of times. In turn, such a webservice could be hosted by the TRE, a trusted third party or the research group. Another method of limiting queries is by embedding the trained ML model within software which technically applies controls. For example, if a trained ML model was incorporated into a medical device installed within hospitals, the software could constrain the number of requests and who had permission to query the software. However, it is recognised that such software-based technical controls are relatively easily hacked in comparison to a secure webservice.
  Model query controls limit the model to only black-box attacks, i.e. the model is not exposed to white-box attacks.
  The researchers could decide that they will also utilise legal constraints, e.g. within a software user licence or webservice user licence which would explicitly prohibit hacking of the trained model to determine potentially identifiable information.
  It may be that the model query controls sufficiently mitigate the risks such that the identifiability controls (see below) are not required (more likely for a webservice solution). E.g. TREs may not be required to run attack simulations or for researchers to use methods to protect personal data such as differentially private methods. However, it may be that for highly sensitive personal data some identifiability controls will also be required.
- **Identifiability Controls:** These are controls which reduce the risk that the ML model could be hacked to discover personal data (TECH 5-15). For each of the 3 high-level options of identifiability controls, TREs should run attack simulations on the trained model (mimicking possible hacking behaviour) to assess for vulnerabilities. The model should not be allowed to be exported if it is considered to be too vulnerable (based upon the thresholds agreed with the data controller during the data governance approval process - ELA 8). To run such simulations the TRE should be provided with basic information from the researchers regarding how the model was trained (TECH 9, TECH 7.2). TREs should agree to confidentiality agreements, should the researchers have concerns regarding their IP (ELA 10). TREs should also carry out other basic checks such as ensuring that the model is significantly smaller than the data used to train the model (to ensure that the full training dataset is not included within the model) (TECH 7.4) and eyeball the code used to generate the trained model (to ensure that the researchers are not maliciously trying to hide data) (TECH 7.6).
  - **Identifiability Controls – use of methods such as differential privacy or training on aggregate level data**: Researchers can choose to use methods which are designed to protect personal data such as differentially private methods or training on aggregate level data (TECH 12). These significantly reduce the probability that personal data will be encoded within the model and increase the chance that the model will pass the attack simulations run by the TRE. However, such methods have been shown to reduce the accuracy of trained models; hence, researchers may choose to adopt other controls instead. Methods such as differential privacy or training on aggregate level data are indeed highly recommended for instance-based models (such as KNN, SVM and deep learning). This is because the non-differentially private versions of these methods need all or some of the original data rows to be able to make predictions and thus personal data is highly likely to be encoded within the trained model. When DP methods or similarly randomised methods are used, randomisation should be dependent on a random seed, and the value of this seed corresponding to the released model should be chosen (randomly) by the TRE staff, rather than researchers.

- **Identifiability Controls – use of safe wrappers:** To increase efficiency over native training (below), researchers may wish to employ community developed Safe Wrappers (TECH 10). These Safe Wrappers provide the same functionality as the standard ML libraries from which they inherit but apply "safe" limits on the hyper-parameters of the model to reduce the risk the model encodes personal data. Using Safe Wrappers requires minimal changes to researchers' workflows while increasing the chances that the trained model will pass the attack simulations, reducing the effort of both the researchers and the TRE staff to run multiple iterations of training and running attack simulations prior to allowing export. Researchers will have increased confidence that they are not being disclosive and have avoided over-fitting. The disclosure control process for TREs should be streamlined and less costly.
- **Identifiability Controls – native training:** Researchers can utilise a method of their choice to train the model. However, this may not pass the attack simulations run by TRE staff, and the model may need to be retrained multiple times before it is considered "safe" for release. Such iterations are likely to be time-consuming and expensive.

**Relevant Detailed Recommendations:** TECH 1, 4-16. ELA 8, ELA 10

#### *9.3.1.3 R 3: Researchers and TREs should consider the costs of implementing the controls*

Most TREs work on a cost recovery basis. Researchers should ask TREs to estimate and charge for any additional work to undertake the identifiability controls of trained ML models such as running attack simulations (C 1) and the additional costs of maintaining the data and development pipeline to support legal requirements (C 4), e.g. for a certified medical device. TREs should consider outsourcing some of the highly technical work and obtaining estimates from these other sources if they do not have the skills in-house (C 5) to pass on the costs to the researchers.

The costs for limiting queries on the model should be considered (C 3), e.g. the costs of running a secure webservice or a software wrapper.

#### *9.3.1.4 R 4: Approvals processes should consider the risks and controls for disclosure control of trained ML models*

Once the approach has been decided upon, researchers should include details of the approach within their ethical and data governance application processes, providing sufficient detail to support the review process. Researchers should use standard text available from TREs which describe the controls which will be applied (ELA 9). This reduces the burden on researchers to draft such information and for TREs to review unfamiliar text as well as reducing the number of re-submissions of applications due to missing information.

Both the data governance and ethical approval processes should consider:

- the risks and controls associated with the training and release of ML models (ELA 2.1)
- the trustworthiness of the organisations involved (ELA 2.1)
- if legal contracts are proposed to cover the responsibilities of each party (ELA 1, 3)
- the length of time the model will be used prior to requesting new approvals (ELA 2.2)
- the requirement to keep the data and the pipeline available to meet legal requirements (ELA 2.3)
- the correct language to describe the process (ELA 2.4)
- if controls on the model are required after release from the TRE (ELA 3)
- the requirement that details of a released model are recorded on a data use register (ELA 4, 7)

- whether legal clauses will be added to the terms of use of any resulting trained ML model (ELA 6)
- whether researchers will sign additional clauses within researchers' declaration forms (ELA 7)
- whether confidentiality agreements are required to protect researchers' IP (ELA 10)
- whether agreements are required between the Data Controller and TREs and whether the TRE processes for applying for identifiability controls have been detailed within the application and approved by the Data Controller (ELA 11).

Data Controllers, TREs and researchers should consider that Data Protection legislation may apply to the trained model (ELA 1). Data protection legislation is generally considered not to apply to the anonymous aggregate level releases from classical statistical data analysis from a TRE, as appropriate controls ensure that these releases do not contain personal data. A trained ML model, in contrast, may be considered to potentially contain pseudonymised personal data, therefore requiring specific technical and organisational measures to ensure the processing is compliant with data protection law. In particular, it needs to be considered whether appropriate data security measures have been adopted to reflect and mitigate the risk of a data breach.

Researchers need to articulate the wider benefits of their work when they wish to release an ML model and explain how they will secure the models (ELA 2). This is especially important when ML models may be released as full consideration of benefits may add weight to the value of the research versus additional risks from disclosure. Approval processes may wish to consider a risk-based approach vis-a-vis personal data (ELA 8).

Public representatives should be involved in data governance and the ethical approvals process (P 1). It may be that the data governance committee or ethical committee requests that additional controls are placed on the process to mitigate the risks, e.g. the model cannot be released openly but has to be hosted within a secure webservice.

Figure 2 shows a decision flow chart, summarising the key questions that data governance and ethical committees should consider when reviewing an ML project application.

**Relevant Detailed Recommendations:** ELA 1-4, ELA 6-10. An understanding of all the TECH controls is also required for data governance and ethics committees to assess the applications and to support researchers in drafting applications.

#### *9.3.1.5 R 5: Researchers and the Principal Investigator should sign modified Researchers Declaration Forms*

Researchers must read and sign user/researchers declaration forms to analyse data within most TREs, in addition to the Principal Investigator and the TRE Manager. Such forms explain the researchers' responsibilities and behaviours to which they must adhere. We recommend additional clauses are added to researchers' declaration forms to cover the training and release of an ML model (ELA 7), including:

- Researchers must abide by the controls which have been approved by data governance and ethical committees.
- Researchers agree that details of the project will be recorded on the data use register and that researchers are required to update the TRE with any changes to the use of the trained ML model once it leaves the environment i.e. if it is incorporated into a product with a CE mark and is being sold commercially.

Researchers should sign these modified researchers' declaration forms if they are running an ML project.

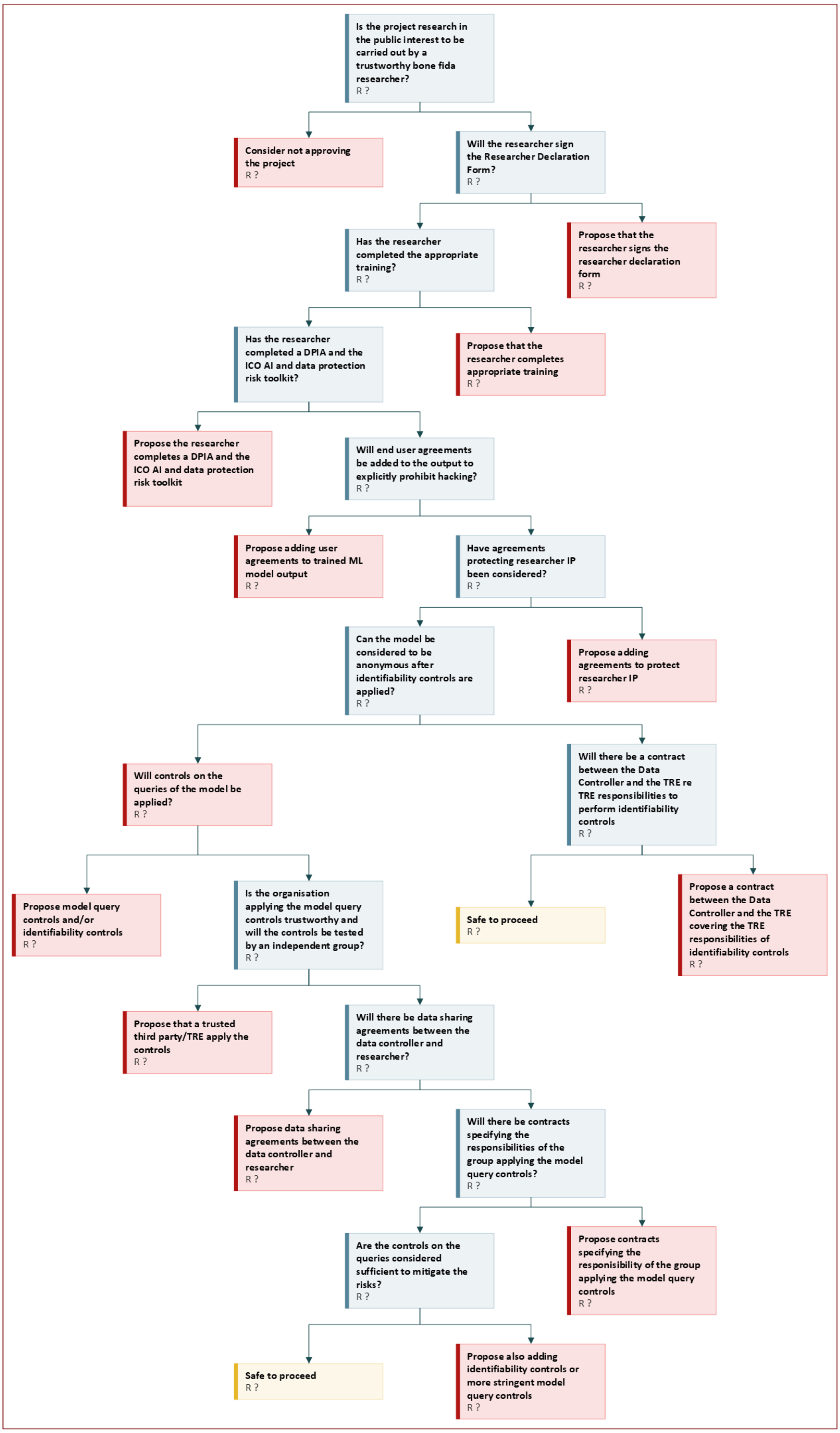

Is the project research in the public interest to be carried out by a trustworthy bone fida researcher? R ?
Consider not approving the project R ?
Will the researcher sign the Researcher Declaration Form? R ?
Has the researcher completed the appropriate training? R ?
Propose that the researcher signs the researcher declaration form R ?
Has the researcher completed a DPIA and the ICO AI and data protection risk toolkit?
Propose that the researcher completes appropriate training R ?
Propose the researcher completes a DPIA and the ICO AI and data protection risk toolkit
Will end user agreements be added to the output to explicitly prohibit hacking? R ?
Propose adding user agreements to trained ML model output R ?
Have agreements protecting researcher IP been considered? R ?
Can the model be considered to be anonymous after identifiability controls are applied? R ?
Propose adding agreements to protect researcher IP R ?
Will controls on the queries of the model be applied? R ?
Will there be a contract between the Data Controller and the TRE re TRE responsibilities to perform identifiability controls R ?
Propose model query controls and/or identifiability controls R ?
Is the organisation applying the model query controls trustworthy and will the controls be tested by an independent group? R ?
Safe to proceed R ?
Propose a contract between the Data Controller and the TRE covering the TRE responsibilities of identifiability controls R ?
Propose that a trusted third party/TRE apply the controls R ?
Will there be data sharing agreements between the data controller and researcher? R ?
Propose data sharing agreements between the data controller and researcher R ?
Will there be contracts specifying the responsibilities of the group applying the model query controls? R ?
Are the controls on the queries considered sufficient to mitigate the risks? R ?
Propose contracts specifying the responisibility of the group applying the model query controls R ?
Safe to proceed R ?
Propose also adding identifiability controls or more stringent model query controls R ?

*Figure 2 - Flow chart of decisions for data governance committees. The boxes coloured blue or yellow indicate a positive answer to the preceding question. The red boxes indicate a negative answer to the proceeding question*

**Relevant Detailed Recommendations:** ELA 7.

#### *9.3.1.6 R 6: TREs should be enhanced to support ML training*

To support ML training the TRE environment should be configured with the software required to train ML models (TECH 2).

**Relevant Detailed Recommendations:** TECH 2

#### *9.3.1.7 R 7: Data should be set aside for applying identifiability controls*

If the TRE is to run attack simulations as an identifiability control, some data should be kept to one side and not provided to the researchers for use in attack simulations (TECH 6).

**Relevant Detailed Recommendations:** TECH 6.

### 9.3.2 Project running within the TRE

There are 2 areas within the TRE:

**TRE researchers ML training zone:** This is the zone where researchers will carry out the training of their ML model once the project has been granted the appropriate permissions.

- **TRE export zone:** This is the zone where TRE staff check files for personal data prior to enabling them to be exported from the TRE if they are considered "safe" based upon any limits within ethical and data governance approvals.

Figure 3 shows these 2 zones and the processes within them.

**Processes : Project within TRE**

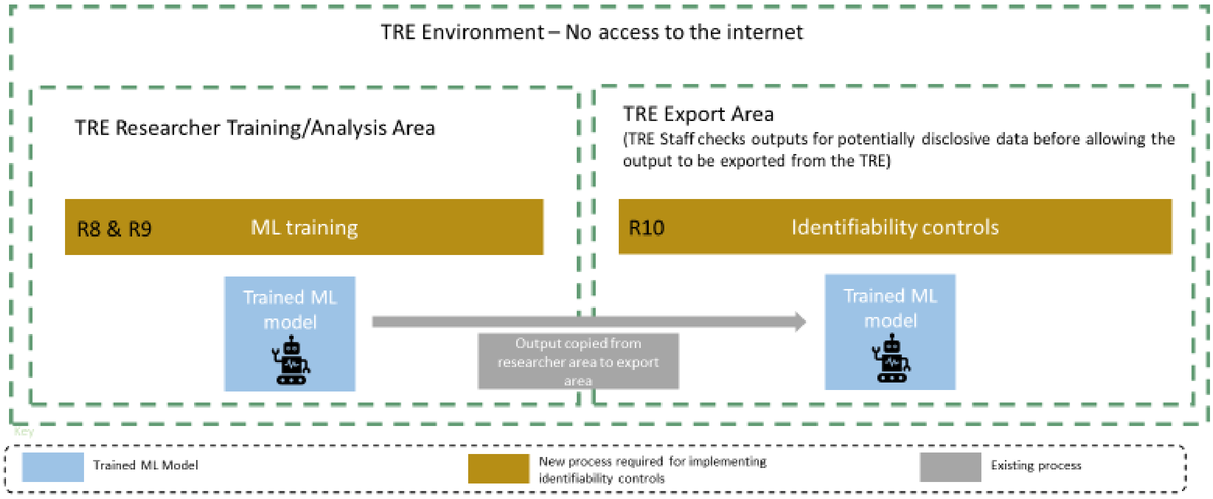

TRE Environment – No access to the internet
TRE Researcher Training/Analysis Area
TRE Export Area
(TRE Staff checks outputs for potentially disclosive data before allowing the output to be exported from the TRE)
R8 & R9
ML training
R10
Identifiability controls
Trained ML model
Output copied from researcher area to export area
Trained ML model
Key
Trained ML Model
New process required for implementing identifiability controls
Existing process

*Figure 3. TRE project process where the recommendations apply. The mustard-coloured boxes show a new process which is required to support identifiability controls for ML models. The grey box shows an existing process. The light blue box shows the trained ML model. The "R" labels correspond to the high-level recommendations within the main body of text.*

#### *9.3.2.1 R 8: Within the TRE researchers ML training zone, researchers should use differentially private methods or training on aggregated data for instance-based models if identifiability controls are required*

If identifiability controls are required, instance-based models (such as KNN, SVM and deep learning) should have additional controls such as training on aggregated data, using differentially private methods, or training on synthetic data generated by a differentially private mechanism (TECH 12).

**Relevant Detailed Recommendations:** TECH 10-13

#### *9.3.2.2 R 9: Within the TRE researchers ML training zone researchers may wish to use safe wrappers if identifiability controls are required*

If identifiability controls are required, to increase efficiency, researchers may wish to employ Safe Wrappers (TECH 10) which have been developed by the community based upon a set of principles (TECH 11).

**Relevant Detailed Recommendations:** TECH 10-11

#### *9.3.2.3 R 10: Within the TRE export zone TREs should apply a range of identifiability controls*

For projects where identifiability controls are a requirement based on the data governance and ethical approvals for the project, the TRE should apply a range of identifiability controls prior to the model being "approved" for release.

Researchers should provide a 'data dictionary' describing the inputs to their model in a standardised format (TECH 7.2, TECH 9). To assess the risk of a model containing identifiable data, TREs should run a range of checks on the model to be released including:

- utilising a risk assessment check-list (TECH 7.1)
- running the model to be released against set-aside data to ensure that it provides the expected result (TECH 7.3)
- checking the size of the trained model (TECH 7.4)
- utilising a tool that reads values from the data they provided and scans the released file for variables within the data (TECH 7.5)
- eyeballing the code used to train the model (TECH 7.6)
- checking the file type of the model to be released to check that it is contained within a list of accepted file types for release (TECH 7.7)
- running attack simulations using the set-aside data (TECH 7.8):
  - passing the training data through the trained model to obtain predictive probabilities.
  - passing the data held out by the TREs through the trained model to obtain predictive probabilities.
  - attempting to build a new model that can predict whether a particular example is in the training data or the held-out data.

This will result in a series of metrics describing the membership inference risk, that would be interpreted by the TRE staff.

It is recommended that TREs do not assume that a safe ensemble implies safe base models or vice versa (TECH 13). TREs should take care to ensure that both the overall model is safe, and the base models that constitute it are safe.

When using Federated Learning the final models should be tested for vulnerability (TECH 14). Synthetic data should still be considered to be potentially disclosive (TECH 15). If the generated synthetic data encodes relationships between the variables accurately enough to use for ML model training, then the relationships encoded could also provide enough information to be vulnerable to hacking to find personal data. The software/pipeline that was used to generate the final trained model which has been approved for release should be stored as a snapshot so that it can be reproduced (TECH 3).

**Relevant Detailed Recommendations:** TECH 3, TECH 7, TECH 9, TECH 13-15

### 9.3.3 Model release

**Processes: Model Release**

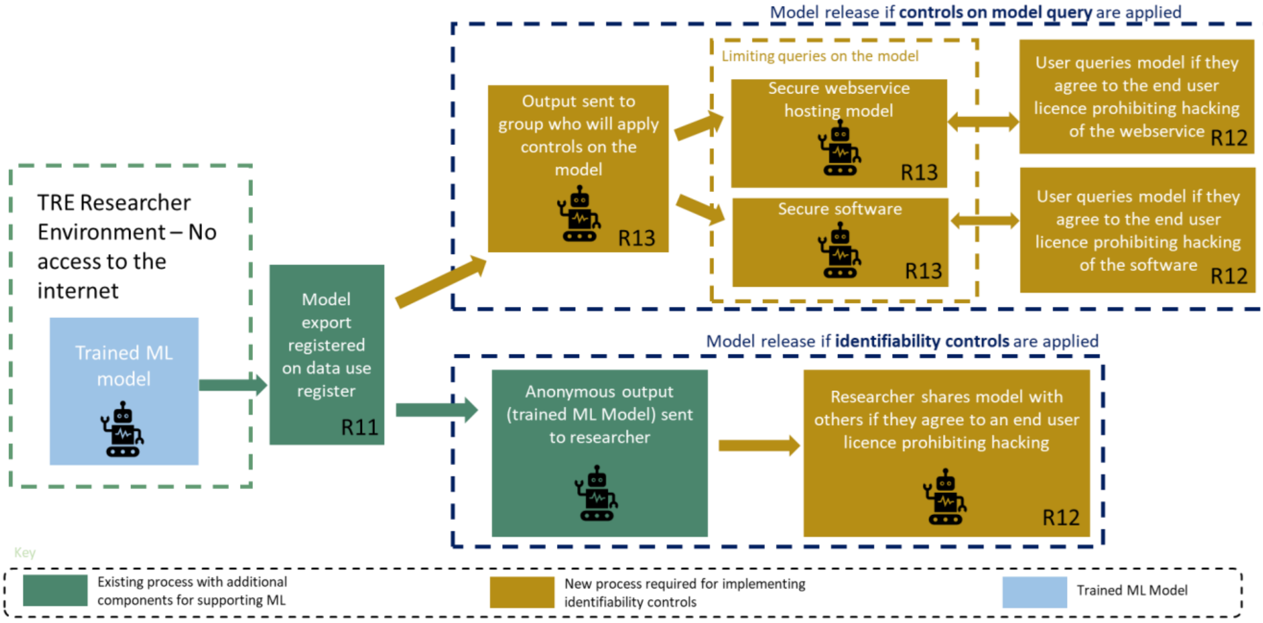

Model release if **controls on model query** are applied
Limiting queries on the model
Output sent to group who will apply controls on the model
R13
Secure webservice hosting model
R13
Secure software
R13
User queries model if they agree to the end user licence prohibiting hacking of the webservice
R12
User queries model if they agree to the end user licence prohibiting hacking of the software
R12
TRE Researcher Environment – No access to the internet
Trained ML model
Model export registered on data use register
R11
Model release if **identifiability controls** are applied
Anonymous output (trained ML Model) sent to researcher
Researcher shares model with others if they agree to an end user licence prohibiting hacking
R12
Key
Existing process with additional components for supporting ML
New process required for implementing identifiability controls
Trained ML Model

*Figure 4. Model release process where the recommendations apply. Green boxes show the existing process which takes place for "traditional" TRE projects and where new recommendations apply to support ML training. The mustard-coloured box shows a new process which which is required to support identifiability controls for ML models. The light blue box shows the trained ML model. The "R" labels correspond to the high-level recommendations within the main body of the text.*

#### 9.3.3.1 R11: When a model is exported from a TRE an entry should be made within a data use register

Data Use Registers should be extended to include information on trained ML models and details around their export and controls (ELA 4). The use of data for training ML models and the controls on the models should be visible to the public through searching data use registers (P 2).

#### 9.3.3.2 R12: When a model is exported from a TRE clauses should be added to the terms of use

Clauses should be added to the terms of use of any resulting ML model (ELA 6): for example, in the terms of use for webservice users or the terms of use for users of the software with ML models embedded.  Terms of use should also be included if the model is to be released as open source. A legal term outlining the "permitted purpose" of the trained ML model should be added.

#### *9.3.3.3 R13: Model query controls should be utilised if identifiability controls are not used/sufficient*

If identifiability controls are not being applied or are not considered to be sufficient to mitigate the risks, controls should be considered to limit the queries on a model once it is released from a TRE (TECH 16). E.g. the model could be hosted within a secure webservice or embedded within the software which technically limits who can query the model and the number of times. Such controls to limit the number of queries should be tested by an external party (TECH 17). The costs for limiting queries on the model should be considered (C 3).

Legal contracts should be considered to cover the responsibilities of each party if controls on the model are required after release from the TRE (ELA 3).

**Relevant Detailed Recommendations:** TECH 9, TECH 16-17, C 3, ELA 3

# 10 Technical

## 10.1 Technical Background

This section provides a high-level description of the field and the investigations undertaken by GRAIMATTER to inform the recommendations. More details regarding the investigations and experiments are provided in Appendices A, B and C.

Any model, whether hand-crafted by human researchers or created automatically by a machine learning algorithm, attempts to describe the underlying data in ways that facilitate the drawing of useful inferences on unseen data. In doing so, they risk disclosing information about individuals in the training set. These risks are well understood for tables of data – for example, no cell should just reflect the data from fewer than k people as that might allow an individual or his or her attributes to be disclosed (so-called k-anonymity) - and TRE staff are trained in spotting and rejecting disclosive outputs. However, Machine Learning (ML) models are typically far more complex, and rarely human-readable. For example, it is possible (albeit laborious) to manually check that no 'leaf node' of a single decision tree only refers to one person's data. However, this may be impractical or impossible when the model takes the form of a forest of such trees, each reflecting a partial view of the data, so it is the intersection of leaves from different trees that leads to disclosure. In some cases, it may be even worse, as trained models such as K-Nearest Neighbours and Support Vector Machines will typically directly encode row-level training data in the model output.

Thus, one aspect of the work undertaken has been to examine how the hyper-parameters that control the learning behaviour of different ML algorithms can result in models that pose different types of risk to an individual's privacy.

Since many trained ML models are represented as large arrays of numbers, another risk factor is that malicious researchers might attempt to manually edit models to 'hide' some of the training data within them – analogous to how text can be 'hidden' within an image (steganography).

Thus, the second strand of research has been concerned with exploring the role that different automated tools could play in helping:

- researchers (who may not be privacy experts) abide by best practices,
- TRE output staff make a rapid informed decision about the likely level of risks posed by different trained models for which release has been requested.

### 10.1.1 Quantitative assessment of the risk of disclosure from different AI models

Decisions on ML model disclosure from a TRE concern trade-offs between usability and privacy risk, particularly re-identification (RT3), necessitating prior assessment of risks [21]–[24]. Factors such as ML model types, learning algorithms, data types and researchers' expertise can (sometimes provably) influence the risk of disclosure. We performed experiments to assess the risks of different models, for different data types, within different training regimes (see Appendix A). We identified the following experimental factors as the best representatives of the risks and threats to ML models within TREs:

- Data types: different types of datasets including medical images, Electronic Health Records, and administrative data.
- Algorithm/training types: we have identified a range of representative model types spanning classical and AI settings: SVM, Decision Trees, Regression, K-means, MLP, CNN, and KNN. We have incorporated different training algorithms, including those based on Differential Privacy. We have simultaneously implemented methods for the generation of synthetic data [21]–[24]
- Attack types: we have used the most common attack types conferring privacy risk: membership inference attacks (MIA), attribute inference attacks (AIA) and model inversion [25].

We trained models across a range of parameter/hyper-parameter values and data types and assessed the performance of disclosive MIA in each case (detailed descriptions of the experiments are provided in Appendix A). We have formalised and implemented different types of AIA and are in process of repeating that large-scale testing for this risk. This enabled us to identify factors for assessing risks in disclosure of these model types, and parameter and hyper-parameter regimes to avoid. Finally, we considered the practical relevance of these risks (*to be covered in a later draft of this document)*. Context matters: we know from 'traditional' SDC that risks which are meaningful in some environments are irrelevant in others. Evidence also matters: we need to understand whether the information burden placed upon an attacker is likely to be feasible (that is, we are not interested in attacks which are theoretically possible but meaningless in practice), or has historical precedents.

### 10.1.2 Controls and Evaluation of Tools

We evaluated a range of tools to determine their usefulness in the semi-automating assessment of disclosure risk (details provided in Appendix B). Our evaluation considered the current 'effectiveness', requisite level of support/maintenance, and the risk of these tools themselves becoming part of an 'arms-race'. We considered approaches, where a model fitted to synthetic data [21], [22] is released instead of the true model, which partly shifts [23] privacy considerations to the synthesis process.

Open-source privacy attack and defence tools we evaluated include:

- TensorFlow Privacy https://github.com/tensorflow/privacy
- Adversarial Robustness Toolbox https://github.com/Trusted-AI/adversarial-robustness-toolbox
- Fawkes https://github.com/Shawn-Shan/fawkes
- Diffprivlib https://github.com/IBM/differential-privacy-library
- IBM AI Privacy Toolkit https://github.com/IBM/ai-privacy-toolkit
- ML-PePR https://github.com/hallojs/ml-pepr
- ML Privacy Meter https://github.com/privacytrustlab/ml_privacy_meter
- ML-Doctor https://github.com/liuyugeng/ML-Doctor

- PrivacyRaven https://github.com/trailofbits/PrivacyRaven
- CypherCat https://github.com/Lab41/cyphercat
- AttriGuard https://github.com/jjy1994/AttriGuard
- MemGuard https://github.com/jjy1994/MemGuard

Factors considered in our assessment of the tools:

1. Project license
2. Project documentation and tutorials
3. Project is based on peer-reviewed publication(s)
4. Project version is numbered and has a version tagged as a release
5. Project recency and update frequency
6. Project popularity (e.g., GitHub stars and contributors)
7. Project quality assurance (e.g., use of unit tests, static analysis, and continuous integration tools)
8. Project distribution: location and ease of installation with current platforms/software
9. Project ease of use: including output analysis metrics and/or visualisations
10. Number of and quality of other projects that use/depend on the project

We developed python 'wrappers', around commonly used modelling functions (scikit-learn/Tensorflow), automatically assessing disclosure risk and producing reports to assist the output-checking team.

From a researcher's perspective, these behave just like the familiar library code, only adding functions to:

- Suggest changes if the researchers initialise the model with hyper-parameters or algorithm variants likely to lead to disclosure risk.
- *Request_release()* of a trained model: which saves a copy to file and produces an automated report for the TRE output checkers.

These 'wrapper' functions are controlled by a human-readable 'config' file that embeds the TRE and dataset-specific risk appetite. For example, they can be configured to restrict parameter choices to 'safe' regimes, preventing unintentional researcher-introduced risk (RT1).

In investigating the possibilities for functionality invoked when a researcher calls *requestRelease()*, we have implemented and evaluated mechanisms for detecting whether changes had been made to trained models *after* the ML algorithms had run. The latter risk might occur through misunderstanding or unintentionally (changing a parameter but forgetting to 'retrain')(RT1). However, it might also result from malicious attempts to hide data or subvert 'safe' choices for training parameters (RT2).

We investigated but discarded as impractical the possibility of hiding potentially 'vulnerable' parts of models (e.g., 'support vectors' of a support vector machine, or layers of a neural network) from either malicious researchers (RT2) or external 'white-box' attack post-release (RT3).

### 10.1.3 Differential privacy

We investigated the use of differentially private (DP) methods as a way to reduce the risk that a trained ML model contains identifiable data. Within our context, differential privacy is a mathematical way of measuring [13] the susceptibility of a trained ML model to an adversarial attack. Strictly, differential privacy considers a 'mechanism' which is an algorithm to convert private data to a public release – e.g. ML model training. Such a mechanism may be differentially private to a given degree. In this sense, DP is a property of the *algorithm* used to generate a trained ML model, rather than the private data used or the specific TRE release; we may use (for example) a 'Differentially private linear model' mechanism to fit a linear model to some private data, to give us coefficients which we then release, but it is the *mechanism* which is DP rather than the data or coefficients (Figure 5 below).

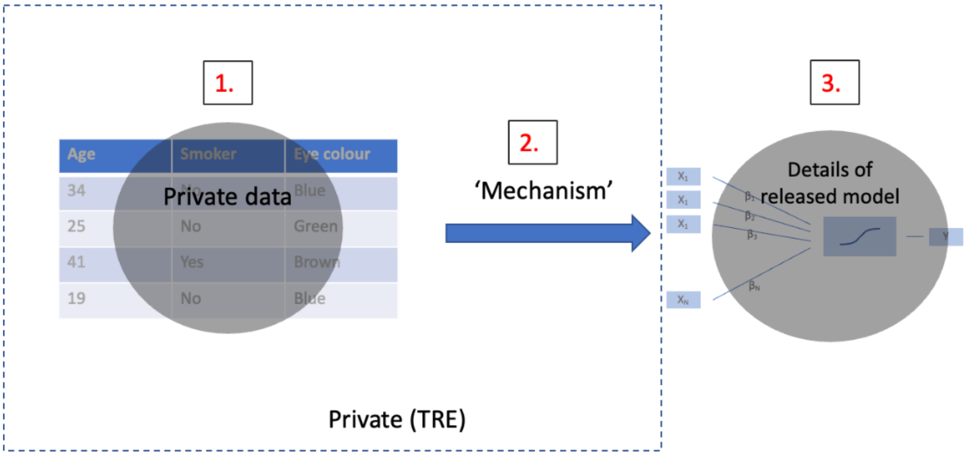


*Figure 5. Abstract description of the process of preparing a model for release. Differential privacy is a property of the 'mechanism' (object 2), whereas most output checks concern only the release itself (object 3).*

DP measures the worst-case susceptibility of a trained ML model to an attacker with the aim that non-worst-case settings (for instance, where an attacker has imperfect or no knowledge of any private data) are automatically covered. Suppose that we have some data for a range of samples within a TRE and we use some of this data to generate a public release of the trained ML model, using some mechanism. A hypothetical attacker has access to all of the data we have inside the TRE, except for one small thing: there are two samples, A and B, and only one of them was used in generating our release. The attacker knows everything about samples A and B except for which one we eventually used. They also know everything about the mechanism we are using to go from the TRE data to the release, and they can see the data we release. The attacker is interested in working out which one of A and B we used to generate our release. DP asks the question: to what extent can they work this out from looking at what we released publicly?

To see why this helps: let us suppose the attacker currently knows nothing about how we trained a model. Then the task of finding out whether A or B is in the dataset is essentially impossible, as there are so many other unknowns affecting the public release. If, on the other hand, the adversary knows 99% of what we did to train the model, the last 1% (whether A or B is in the dataset) is much easier. DP limits the ability of the attacker to find that last 1%. This also makes the already-difficult task of finding the first 99% much more difficult, thereby making the attacker's task much more difficult in general.

One important thing to notice is that if our mechanism is deterministic (that is, given the same input data, always gives the same output) then the attacker can work out which of A or B was included immediately – they would just run the mechanism with A included, then with B included, and see which of those matched the actual release (this also implies that releases like the mean or standard deviation of a dataset are **not** differentially private). Thus, differential privacy mechanisms introduce some randomness into the release.

A mechanism is differentially private if we can guarantee that the likelihood of seeing any particular release if we used A differs from the likelihood of seeing the same release if we used B by at most a particular multiplicative factor ε (epsilon, usually given on a logarithmic scale) and an additive factor δ (delta, usually given as a difference from 1). These factors specify the 'level' of differential privacy The lower epsilon is and the closer delta is to 0, the more private the mechanism is.

Generally, DP algorithms are adjustable, so we can set the level of (ε,δ). There is usually a fundamental trade-off in that the more private a model is, the less useful it is; that is, there is a necessary trade-off between privacy and accuracy,[13], [21], [24].

It should be noted that as interest in Differentially Private Machine Learning increases, a range of technical problems is arising, and variants of DP are being proposed. DP is a useful metric of risk but should be interpreted exactly as per its definition. Furthermore, the extent of noise that needs to be added to guarantee acceptable privacy controls via differential privacy typically leads to an unacceptable compromise in accuracy; an acceptable compromise generally needs to be decided between TRE staff and researchers on a case-by-case basis, which is potentially unrealistic to do at the research planning stage [26]. In fact, for federated learning, it is shown in [27] that anti-overfitting techniques can offer a better accuracy-privacy trade-off than DP.

Differential privacy is not directly comparable to statistical disclosure control practices typically used in non-ML TRE releases; for instance, the common statistical disclosure control practice of avoiding summary statistics calculated on fewer than 5 samples does not confer ε-differential privacy for any finite ε.

### 10.1.4 Synthetic data

We considered the evaluation of synthetic data generators (SDGs) in two settings:

- SDGs may be used inside a TRE in a setting for which the aim is to release a predictive model. In this case, the SDG would be used to generate synthetic data inside a TRE and a model for release trained to this synthetic data instead of original data. Note that in this scenario, the performance of the final released model is heavily dependent on the quality of the generated synthetic data.
- The SDG itself could be released, either in addition to or instead of a predictive model, as an output.

In case 1, the SDG is used either instead of or in addition to a privatised predictive model and can be assessed for disclosure risk by simply assessing the predictive model that is released.

Case 2 is more difficult as an SDG is a model of the overall data distribution, which a predictive model is not. In other words, the maximal amount of disclosive information potentially contained in an SDG is higher than that in a predictive model. Released SDGs can be provably private (e.g. differentially private)[24], [28], [29].

Since synthetic data generators typically estimate a sampling distribution which resembles the distribution of the original data, they are not necessarily invulnerable to simply returning original data samples: indeed, if the estimated sampling distribution is sufficiently overfitted, this is exactly what they will do. For this reason, synthetic data generators should be treated with equivalent scrutiny to supervised machine learning models.

*We are currently analysing several SDGs for privacy for case 1 use in parallel to the assessment of predictive models. We are looking into options to assess SDGs for privacy in case 2. Our findings will be included in the next draft of this document.*

### 10.1.5 Instance-based models

We investigated row or instance-based models and found that particular care is required for such models. For example, the K-nearest neighbours (KNN) algorithm, when given a new data sample, begins by finding the K-closest training samples and thereby requires all training samples to work. Similarly, the commonly used Support Vector Classifier (SVC) requires a subset of the training samples (known as the support vectors) to operate. SVC is an example of a 'kernel' based methods (which requires comparison of a new data sample with training samples) and all are dangerous in this way (other examples include Support Vector Regression and Gaussian Processes).

Such models are generally not appropriate for release from TREs, as the row-based data is automatically disclosive and can be essentially read off. Some model-specific adaptations are possible; for instance, KNN problems can be converted to 'radius-neighbours' classification [30], in which points within a fixed 'closeness' are counted, and SVC can use random transformations to hide the original data points [9]. Both methods compromise the accuracy of the predictive model. These examples are provably differentially private.

### 10.1.6 Imaging data

Imaging data creates particular issues when it comes to disclosure. In many cases, it will not be the image *per se* that holds any personal data (although some MRIs and chest x-rays might identify individuals and therefore constitute personal data), but some artefacts within the image. For example, we will rarely care about pixel values in the background of an image. This leads to a range of interesting questions about how aspects such as membership inference should be defined and whether or not anonymisation is useful and possible with such unstructured data [31]. Promising reconstruction attacks of training datasets of images based on model updates have been proposed in [32], [33].

*We are currently exploring imaging data and aspects such as membership inference and will add these findings to the next version of these recommendations.*

### 10.1.7 Limiting Queries on the Model

We identified two broad categories of the use of the model which necessitate different recommendations across each of the different control areas: unlimited queries on the model (for example, when the model is distributed to other end-users) and limited queries on the model (for example, when end-users can upload data to the model and run it, but the models runs on the servers owned by the TRE).

Typical assessments of disclosure risk [34], [35] operate through assessing the *behaviour* of a machine learning model, rather than directly *observing* it. More formally, when we assess disclosure risk, we characterise a machine learning model as a function from data to outputs: two distinct models which implement the same function can have equal susceptibility to attack when considered in this way. To be able to fully 'see' the function; all we need to do is to be able to repeatedly evaluate it on a range of inputs.

In this sense, an important dichotomisation of TRE-released models concerns whether their associated functions can be queried arbitrarily by parties external to the TRE, or whether such queries are restricted. This is distinct from the question of whether a released model is white-box (internal operations visible) or black-box (internal operations hidden), although the question of whether queries are restricted need only be considered for black-box models, as the associated function in white-box models is directly observable. Restricted-query models may be implemented in several ways: for instance, hosted on secure servers with curation of queries to the model and restrictions around who may submit queries, or bundled with software which restricts queries and encrypted so that the model can only be accessed through the restrictive software. Restriction of queries may take the form of limiting the range of inputs for which the function can be evaluated (for instance, limiting a medical model to only use for data linked to real patients, rather than simulated data), or limiting the number of queries able to be made in a given time period.

An important appeal of restricting queries is the potential to reduce restrictions on the use of safe models, as restriction of model types can only weaken performance (see the section on differential privacy above). Careful curation of queries which can be made to a released model restricts the capacity of potential attackers to accurately characterise the function implemented by the machine-learning model, and hence to perform membership- or attribute- inference attacks. This capacity to resist model attack through restriction of queries rather than restriction of model type can potentially allow more accurate models to reach the stage of public use, ultimately improving prediction performance (which may, e.g., result in better outcomes for patients when ML is used in health applications).

The category of model release type (unlimited and limited queries on the model) is important in determining the appropriate recommendations for model security.

## 10.2 Technical Recommendations:

### 10.2.1 TECH 1: Risks of disclosure need to be discussed at an early stage of the project application between the researchers and the TRE

Empirical analysis of a model can answer questions such as “are we able to successfully attack this model?” and “can we infer the value of this feature, given values for the other features?”. It cannot determine whether this in itself represents a disclosure risk. TREs should therefore ensure that the risk pertaining to the particular dataset in question is discussed at an early stage and that researchers are aware of the kind of risks that will be checked for. An example of the kind of discussion that is required would be around attribute inference: by their nature, ML models learn correlations between attributes that can allow researchers to make predictions of one attribute based on the other. A released model may therefore allow people to make predictions of feature values for any individual, regardless of whether or not they were in the training set. A disclosure risk can occur when there is a discernible difference in the accuracy of those predictions for people in the training set, as opposed to those not in the training set. The TRE-researcher discussions should also include agreement on which attributes/features within a dataset are considered personal since not all will be.

Discussions between researchers and the TRE can help researchers to understand the risks and controls available to support their data governance and ethical approval processes (see ELA recommendations).

**Responsibility:** TRE and researchers. **Understanding:** Data governance teams.

### 10.2.2 TECH 2: TREs need to provide tools for training ML models within their environments

Many TREs have traditionally provided a relatively small number of software tools within the environment, e.g. statistical programs such as SPSS, STATA and R TREs [36]. Installation of such tools needs to go through security controls with regular patching  (as they already do as part of best practice for managing any software within the TRE). To support ML training, TREs will have to install a range of additional tools and assess them for security risks, e.g. python ML specific packages such as Tensorflow, scikit-learn etc.

**Responsibility:** TRE for implementation. **Usage & Understanding:** Researchers.

### 10.2.3 TECH 3: The software/pipeline that was used to generate the final trained model which has been approved for release should be stored as a snapshot so that it can be reproduced for audit purposes

The data and pipeline that was used to generate the model should be saved so that it can be accessed if there are found to be any issues in the future. This may be required for regulatory processes.

This is a matter of accountability and tracing. Additionally, if 'inside' attacks are believed to be feasible, then clearly stating that production mechanisms must be auditable can be a disincentive to malicious attackers. In some scenarios – such as were governed by the Financial Conduct Authority or Medical Device certification, being able to account for and justify decision support models may be mandatory.

TREs should consider the technical requirements of maintaining the data and development pipeline to support these legal and regulatory requirements

**Responsibility:** TRE for implementation & researchers to provide.

### 10.2.4 TECH 4: Two broad categories of controls should be considered: controls to reduce the risk of personal data stored within the trained ML model (Identifiability Control) and controls to limit the queries on the model (Model Query Control)

There are many advantages to releasing a trained model openly, including peer review, scientific reproducibility and re-use. However, to reduce the risk of releasing personal data within openly shared models we are recommending many different controls. Some controls, such as the use of differential privacy for instance-based models, may be viewed as too restrictive by some as they can also reduce the accuracy of models. Other controls require the TRE to understand the training process to run the required attack simulations which some researchers, especially those from industry, may be uncomfortable with regard to IP and trade secrets.

Rather than controlling the risk of releasing personal data within the model (identifiability controls), an alternative is to control the access to the model once it has been released by limiting the queries on the model (model query controls). For example, the model could be hosted within a webservice which constrains who can query the model and how often. It is likely that for many projects either Identifiability Controls or Model Query Controls are chosen. There may also be instances where a combination of Identifiability Controls and Model Query Controls are employed. A combined approach might be appropriate, for example, where the model will be embedded within a software program which limits the queries on the model. There is a risk that such software can be illegally hacked, in which case some of Identifiability Controls may also be utilised as additional controls.

We suggest that the balance between identifiability and model query controls be guided by practicality, with the understanding that disclosure control is enforced with either option. Exclusive use of identifiability controls could facilitate easy use of released models by end-users, but possibly poor predictive model. Conversely, exclusive use model query controls could allow potentially stronger predictive models, but mean they are harder for future researchers to use.

**Responsibility:** TRE/researchers for implementation. **Understanding:** Researchers & data governance

### 10.2.5 TECH 5: Identifiablity Control: Disclosure control of ML models needs staff within TREs who are ML experts/trained

Our experiments (see Appendix A) suggest that empirical testing needs to be performed on a model to determine how safe it is to be released (i.e. model disclosure risk cannot be reliably determined prospectively). This will require TRE staff to run experiments on the model. Although to a certain extent this can be automated, significant expertise is needed within TREs to perform these tests and critically evaluate these results. It is likely that the level of training required will be significant, i.e. not just a few days of training, but e.g. MSc level.

Each TRE does not necessarily have to employ someone directly with this advanced skill set. TREs could source this skill set from a pool of trained experts working across the network, or second an individual from another TRE to support a specific project.

It may be also that much of the checking could be carried out by less qualified staff. [37] categories traditional outputs for checking into 'runners' (outputs needing very simple approvals; can be done automatically), 'repeaters' (outputs requiring some human review at a low technical level) and 'strangers’ (outputs requiring review by expert statisticians). As the expected share of outputs of each type is something like 90%, 9%, and 1%, TREs use this to manage output checking resources.

*An unknown question at present is: what are the runners/repeaters in ML models? Should we treat everything as a 'stranger' needing expert review until we get more experience?*

**Responsibility:** TRE for implementation. **Usage & Understanding:** TRE staff & researchers.

### 10.2.6 TECH 6: Identifiability Control: TREs should keep a proportion of relevant data to one side and never give this data to the research team to be used to assess the model

Running attack simulations requires two data sets – one data set containing examples that were used to train the model, and one containing examples that were not. To ensure that examples exist that have not been used for model training, the TRE should hold back some data from the researchers. In addition, it is important that the TRE staff must know which rows in the data given to the researchers were used for training. Note that to keep data separate, TRE staff will need to know how the data will be processed into rows. For example, if a row will represent individuals, then the TRE staff should set aside all rows corresponding to a subset of individuals. This is particularly important for episodic health data where each individual may have more than one row.

*We are currently investigating how the data should be selected. For example, should it be random and different for each project? We will add our recommendation on this into the next draft.*

**Responsibility:** TRE for implementation. **Usage & Understanding:** Researchers.

### 10.2.7 TECH 7: Identifiability Control: To assess the risk of a model containing identifiable data, TREs should run a range of checks on the model to be released

It should be noted that if the model will be released with additional controls on the queries of the model (e.g. hosted within a webservice rather than released openly) (see Section 10.2.16), these checks may not be required.

**Responsibility:** TRE for implementation; **Understanding:** Researchers & data governance teams.

#### 10.2.7.1 TECH 7.1: Identifiability Control: TREs should utilise a Risk Assessment Check-list

TREs often carry out risk assessments as part of their disclosure control process. These steps are additional steps to carry out on top of the existing risk assessments.

**Responsibility:** TRE for implementation. **Usage & Understanding:** Researchers & data governance teams.

#### 10.2.7.2 TECH 7.2: Identifiability Control: Researchers should provide sufficient detail (in a standardised manner) to TRE staff to enable model assessment

Researchers should provide:

1. The trained model in a standard file format.
2. The code that was used to train the model (the complete pipeline, including any data pre-processing).
3. Their assessment of model predictive performance.

Note that researchers must be made aware of this requirement at the start of the project as a condition of model disclosure (see ELA Section 11.2.2).

**Responsibility:** TRE & researchers; **Understanding:** Governance teams.

#### 10.2.7.3 TECH 7.3: Identifiability Control: TREs should run the model to be released against set-aside data to ensure that it provides the expected result.

TRE staff should assess the performance of the model on the data held out from the researchers. Poor performance relative to the performance quoted by the researchers is an indicator of over-fitting, which can indicate disclosive models. Performance that deviates significantly from that reported by the researcher should be reported back to the researcher and investigated before further analysis is undertaken.

**Responsibility:** TRE for implementation; **Understanding:** Researchers & data governance teams.

#### 10.2.7.4 TECH 7.4: Identifiability Control: TREs should check the size of the trained model

The model file size ought to be orders of magnitude smaller than the size of the training data (an exception is some large neural network models that can have more tuneable parameters than observations). If the model file size is of a similar order to the data size, researchers should be asked to justify this. It is worth noting that some ML packages by default include the training data within the saved model file (for reproducibility). This check will identify this issue.

**Responsibility:** TRE for implementation; **Understanding:** Researchers & data governance teams.

#### 10.2.7.5 TECH 7.5: Identifiability Control: TREs should use a tool that reads values from the data they provided and scans the released file for variables within the data

This checks for accidental incorporation of training data within the release.

**Responsibility:** TRE for implementation; **Understanding:** Researchers & data governance teams.

#### 10.2.7.6 TECH 7.6: Identifiability Control: TREs should eyeball the code used to train the model

TRE staff should have the access and expertise to inspect the code used to train the model to identify any (potentially accidental) poor practices (e.g. storing the data within the model). In addition, they should be able to check that the code is training the model in the manner described in the researcher's report.

**Responsibility:** TRE for implementation; **Understanding:** Researchers & data governance teams.

#### *10.2.7.7 TECH 7.7: Identifiability Control: TREs should check the file type of the model to be released to check that it is contained within a list of accepted file types for release*

TRE staff should only accept model files in a set of common formats that can be loaded with common tools and inspected. Proprietary formats or formats that only allow the model to be run and not inspected should not be permitted.

**Responsibility:** TRE for implementation; **Understanding:** Researchers & data governance teams.

#### *10.2.7.8 TECH 7.8: Identifiability Control: TRE staff should run attack simulations using the set-aside data*

Our results (Appendix A) show that model vulnerability cannot be prospectively determined for a model/dataset combination. Therefore, to have confidence that a model is safe, TRE staff need to empirically assess model vulnerability. We recommend that this is done by:

- Passing the training data through the trained model to obtain predictive probabilities.
- Passing the data held out by the TREs through the trained model to obtain predictive probabilities.
- Attempting to build a new model that can predict whether a particular example is in the training data or the held-out data. This simulates a Worst-Case membership inference scenario (see Appendix A on page 69), where the attacker has access to training and non-training data. Note that this is not necessarily a realistic attack scenario, but serves to provide a conservative assessment of the model's vulnerability.
- This will result in a series of metrics that describe the membership inference risk, that would be interpreted by the TRE staff.

An equivalent worst-case attack for attribute inference would be: Pass each training set record through the trained model with different values for that attribute, and note whether there is a unique value for that attribute that yields the highest predictive probability and whether that value is the 'true' attribute value for that record. Report the percentage of the training set for which there is a unique attribute value that gives the highest confidence and the accuracy of those inferences. Repeat for the hold-out data and report on the differences. This description applies to categorical variables. Similar attacks for continuous variables work based on the upper and lower bounds of the prediction attack.

The GRAIMATTER DARE sprint project has developed as a minimal viable product (MVP) as a suite of attack simulations that can be applied by TREs.

*Location of the MVP attack simulation suite – will be provided later*

**Responsibility:** TRE for implementation; **Understanding:** Researchers & data governance teams.

### 10.2.8 TECH 8: Identifiability Control: Attack simulations should be run following a set of principles

- Use the same standardised data dictionary format.
- Define a format in which raw-data-to-design-matrix transformations should be specified; ideally both human and automatically readable.
- Make open-source on GitHub to code to run attack simulations which invite others to use it. Regular update frequency etc.
- Ideally, provide confidence intervals.
- Deal with different risk appetites for TREs.
- Add to existing GitHub site – *GitHub organisation will be added*

**Responsibility:** TRE for implementation; **Understanding:** Researchers & data governance teams.

### 10.2.9 TECH 9: Identifiability Control: Researchers should provide a 'data dictionary' describing the inputs to their model in a standardised format.

For any model to be useable for making predictions on new data, it is necessary to know what data should be provided to them, and in what format. If TREs (or preferably some higher body) agree on a standard format for this, then it is possible to write automated scripts that "attack" the models and provide the TRE output checker with useful information about the vulnerability of the model.

Information is needed because, to run various inference attacks on models, it is necessary to know the fields that they trained on and how they categorised them.

The data dictionary should include (note: similarities with Python pipelines and PyTorch data loaders):

- Name of the original dataset
- Source code function that takes as input the original dataset and performs any pre-processing (feature selection, feature creation, normalisation, one-hot encoding, handling of missing values, etc.)
- Description of the outputs of the pre-processing function (i.e., the inputs to the model): feature names, encoding (e.g., one-hot categorial or continuous), and a list of indices that represent each feature (e.g., feature indexes [1,2,3] may correspond to a single one-hot encoded feature).
- How the data is split into training, validation, and testing; the number of samples used?
- Description of the model outputs: type (classification, regression, etc.), encoding of the outputs (e.g., softmax, linear, etc.), number of outputs (e.g., number of classes).
- Description of the model: list of libraries and models used, architecture, training algorithm and how the model is initialised.
- Random seeds used (for reproducibility).

Inputs

- Clear documentation of their model architecture, data used and how it is split
- A script that will enable the TRE to load the model and examine it
- A script that will enable the TRE to test the model

Furthermore, to help against inadvertently introducing the risk of backdoor or poisoning attacks, the TRE staff should check the transformation pipeline to ensure that:

1. The number of records is not increased unless specific data augmentation techniques have been agreed.

2. Transformations are not applied selectively to any subset of the records.

*Appendix F provides an example data dictionary template.*

**Responsibility:** Researchers to provide. **Understanding:** TRE & data governance teams.

### 10.2.10 TECH 10: Identifiability Control: To increase efficiency, researchers may wish to employ community developed Safe Wrappers

Trained models should not be allowed to be released from TREs if they were trained with unsafe hyperparameters, with unsafe variants (for example, not using differentially private versions where these are available) or have been manually altered after the training algorithm had been run. Safe wrappers are methods which inherit from existing machine learning libraries, providing the same functionality but enforcing limits on relevant hyperparameters, and providing some traceability.

A safe model version will check if the researchers have complied with a safe training practice without any malicious intent and a report is provided to the output checker.

While we can know that some hyper-parameter/algorithm variants typically lead to disclosive models, it is in the nature of logic that we cannot know the converse in advance for most algorithm-dataset combinations. Therefore, safe wrappers should also include functionality to run automated attacks on the specific trained model the researcher wishes to release.

The GRAIMATTER DARE sprint project has developed a set of minimal viable product (MVP) safe wrappers building on python scikitlearn/TensorFlow toolkits employing the safe parameters determined by experiments across a range of models, data types and hyperparameter settings (see Appendix A for more details). These are available to be used and tested by the community from *(github repository link will be provided).*

The community of TREs and researchers should build many more such safe wrappers to meet the needs of the research community. e.g. R library versions. As the wrappers themselves do not contain any personal data, they can be widely shared across TREs rather than having to be developed bespoke for each TRE.

TREs should encourage researchers to utilise Safe Wrappers. This will reduce the length of time a TRE will take to verify that a model can safely be released.

Furthermore, there is a distinction between a researcher knowing that *some* choices may lead to unsafe models (a concept that needs to be learned once), and the specific knowledge of *which* those choices are (a subject of ongoing research effort). If the definitions of 'unsafe' hyper-parameter combinations and algorithm variants are held in a central (human and machine-readable) file in the TRE, then this can be easily maintained and updated. The alternative is to rely on continuously updating and re-delivering the training provided to researchers.

The rationale for using wrappers:

1. We recognise that manually checking trained ML models will place a huge resourcing strain upon TREs and researchers. Without some form of support, this is likely to lead to lengthy delays in the release of models, or possible limited TRE-appetite for supporting the use of ML and for researchers to use TREs.
2. Incentivisation (e.g. faster approval of results) will help to overcome resistance from researchers.

3. Many researchers will happily accept suggestions of 'safe' hyper-parameter ranges or algorithm variants. While researchers should be free to override these suggestions (with the understanding that they will still need, in a way that will be approved, to make the model 'safe' for it to be allowed to be exported from the TRE), those are conscious design choices with attendant risks which should be reported to the TRE output checkers to make a principles-based decision.
4. There is a high degree of correlation between a trained model's vulnerability to privacy attacks, and the risk of it 'over-fitting' the training set and hence failing to generalise. Therefore, the use of wrapper models is intended to encourage good practice, and should not significantly impact the accuracy of the trained model.
5. Wrapper approaches form a natural way for triggering/embedding automated testing of the privacy risks associated with specifically trained models so that all the information is to hand when the TRE output checker comes to deal with a request for model release, rather than introducing further delays into the process.

**Responsibility:** Researchers to employ. **Understanding:** TRE staff & & data governance teams.

### 10.2.11 TECH 11: Identifiability Control: Safe wrappers should be developed following a set of principles

Clearly, there will be slight changes depending on the features supported by different languages. In the section below, we describe them with reference to python as this is the predominant language for Machine Learning development and libraries.

1. To ensure consistency of behaviour across different ML models or implementations, as much of the 'safe wrapper' behaviour as possible should be implemented within a super class. This class should include functionality for checking hyper-parameter values, making checkpoints of the models to guard against malicious tampering, saving models, and producing reports for TRE output-checking staff.
    a. For example, in our implementation, we have called this class SafeModel.
    b. Classes for specific model types should then use multiple inheritances so they contain as little code as possible. **The philosophy here is to augment the functionality of existing code, not to re-implement it**.
    c. For example, our class SafeDecisionTreeClassifier() inherit from both SafeModel and sklearn.trees.DecisionTreeClassifier() as super classes.
2. All constraints on hyper-parameter values needed to prevent excessive disclosure risk should be stored in a human and machine-readable file held centrally by the TRE. We suggest the use of a single file, with write access limited to the TRE administrators for security and consistency reasons.
    a. An example of part of such a file is provided in Appendix G.
    b. We will be publishing recommendations for constraints for different model types as part of our full report. Naturally, these will be subject to change, so an online version will be made available.
    c. TREs should put in place schedules for periodically checking their constraints files are up to date.
3. Whether there is a differentially private version of the optimisation algorithm used in training a model, such as for Support Vector Machines, and Artificial Neural Networks (e.g. Tensorflow privacy) this should always be used.
    a. This is one example of where the use of class inheritance naturally provides straightforward mechanisms for overriding default choices/optimisers. It would make sense for the differential privacy 'epsilon' factor to be defined in a single place in the constraints file to be consistent across different model types.

4. Safe wrappers should consider attributes or parameters of a model that researchers might unintentionally or maliciously change between 'fitting' a model, and the model is saved and release requested.
    a. One obvious example is a researcher training a model with unsafe parameters and then changing the parameters to 'safe' values but not retraining
    b. Snapshotting a copy of the fitted model after the automated training, and comparing this to the model the researcher requests to release is one useful mechanism.
    c. In our example implementation, we have found that although much of this process can take place in the super class, knowing exactly which model attributes to check and what format they take in python needs to be done for different implementations of algorithms.
5. Safe Wrappers should be made open source and shared with the community to maximise benefit and improve the ongoing maintenance
6. Safe Wrappers should make sure training data is not included in the model: for example, a Safe_KNearestNeighbour() class might just return a message saying this algorithm is not permitted since it inherently encodes the training data.
7. Where Safe wrappers are being developed for different languages, they should provide equivalent functionality in different languages.

**Responsibility:** TRE for implementation; **Understanding:** Researchers.

### 10.2.12 TECH 12: Identifiability Control: Instance based models should have additional controls such as training on aggregated data or using differentially private methods

Instance-based models are popular within ML (KNN, SVM, etc). However, they require special treatment as, in their default use, they need all (KNN) or some (SVM) of the original data rows to be able to make predictions. It is vital that TRE staff are cognisant of this, and can ensure that, for example, researchers are unable to release a standard Support Vector Classifier.

TRE staff require the expertise to be able to assess if the model under consideration falls within this category so that they can ensure that the researchers have taken the necessary steps to ensure disclosure.

Researchers should be informed at an early stage of project scoping that release of instance-based models (in their standard form) will not be permitted.

There are several options for additional technical controls for instance-based models:

- **Train on aggregated/binned data instead:** although unlikely to be practical in many cases, if the training data can be de-identified (by, e.g. achieving k-anonymity via aggregating individuals and binning disclosive features) then there is no danger with rows being disclosed in the trained model.
- **Use differentially private training algorithms or training algorithms with anti-overfitting:** Some instance-based models have been adapted to use differentially private training algorithms. These place provable bounds on the disclosure risk of the trained model, and ought to be used where available. For example, the GRAIMATTER DARE sprint project has developed as a set of minimal viable product (MVP) safe wrappers for SVNs and neural networks utilising differentially private versions of the libraries. An alternative to DP model training is model training with anti-overfitting techniques (regularisation and dropout), since [27] shows that anti-overfitting techniques may yield better accuracy-privacy trade-offs than DP,
  *Links to these will be provided once finalised*

- **Train on synthetic data generated by a differentially-private mechanism:** If neither of the previous steps is appropriate, a final option is generating synthetic data (in a private manner) and then training the model on this synthetic data. When trained in this way, we can think of instance-based models as no longer being instance based, since the 'instances' they contain should be synthetic data points only. However, care must be taken that the synthetic data generator is generating samples sufficiently different from those in the training data; see notes on synthetic data.

Employing any one of these 3 additional technical controls can reduce the accuracy of the model. Researchers should consider instead limiting the number of queries on the model as an alternative option.

Since DP mechanisms are necessarily random, the material released from the TRE using the same method on the same data will be different each time. Importantly, DP guarantees are violated if several such releases are compared and one is chosen. As an example, if researchers firstly fit a non-DP model to some data, then set up a DP algorithm, generate fifty potential releases using this algorithm, choose the one of these fifty that most closely resembles the non-DP model, and release this 'best' model, then the net procedure is no longer DP. It is thus imperative that the randomisation inherent to the DP algorithm be dependent on a random seed, and that the seed corresponding to the final release is chosen by TRE staff who are agnostic to how well the released model performs.

We recommend that the use of Differential Privacy alone is no substitute for following best practices, such as adhering to techniques to limit overfitting (e.g., data augmentation, regularization, dropout in ANNs, etc.) as well as performing attacks on trained models to empirically estimate their vulnerability. This recommendation stems from two main concerns with differential privacy in machine learning. First is the utility loss: differential privacy imposes an accuracy loss on the trained models, especially for small values of epsilon. This may not always be acceptable, especially in the biomedical sector. Second, the only available theoretical bounds on the privacy leakage of ML models are for epsilon values below 1 (at which point the accuracy loss is high), and these are not tight bounds. Values of epsilon above 1 still offer (empirical) protection, but since there is no theoretical bound on this level of protection, empirical risk assessment is still needed [27], [38].

**Responsibility:** TRE and researchers for implementation; **Understanding:** Data governance teams.

### 10.2.13 TECH 13: Identifiability Control: Ensemble methods should receive special treatment

Ensemble methods are ML methods in which several individual ML models are combined (e.g. taking the average of their outputs, or having them vote). TREs should take care to ensure that both the overall model is safe, and the base models that constitute it are safe. It is straightforward to make examples that show both that safe base models can create an unsafe ensemble, and that a safe ensemble can include unsafe base models. It is recommended that TREs do not assume that a safe ensemble implies safe base models or vice versa. This risk is only relevant in circumstances when an attacker has access to the inner workings of the model (white box attack), and not when all they can do is query the ensemble model.

**Responsibility:** TRE for implementation. **Understanding:** Researchers and data governance teams

### 10.2.14 TECH 14: Identifiability Control: When using Federated Learning the final models should be tested for vulnerability

Federated learning in this context describes an ML architecture where the training data is spread across multiple TREs and a 'local' model is trained in each TRE which is iteratively updated with learned information from all the other TREs. At each iteration, the local model is shared with a 'central' TRE which aggregates the multiple models into one which encompasses information learned over all the TREs. The aggregated model is shared back with the other TREs for further training there. This continues until training is complete and the aggregated model is the final trained model. Identifying if a model is disclosive needs to be done using the data on which it was trained, which is not possible for the final aggregated model as the central TRE does not have access to the other TREs' data. Therefore, it is recommended that within each TRE, the local models are tested for vulnerability. It can be assumed that the federated TRE network is trustworthy, following governance approval, and therefore only the final versions of the models are tested for disclosure by running attack simulations within all the TREs.  The final aggregated model can be released only if all attack simulations are completed (see Figure 6).

**Responsibility:** TREs and researchers for implementation; **Understanding:** Data governance teams.

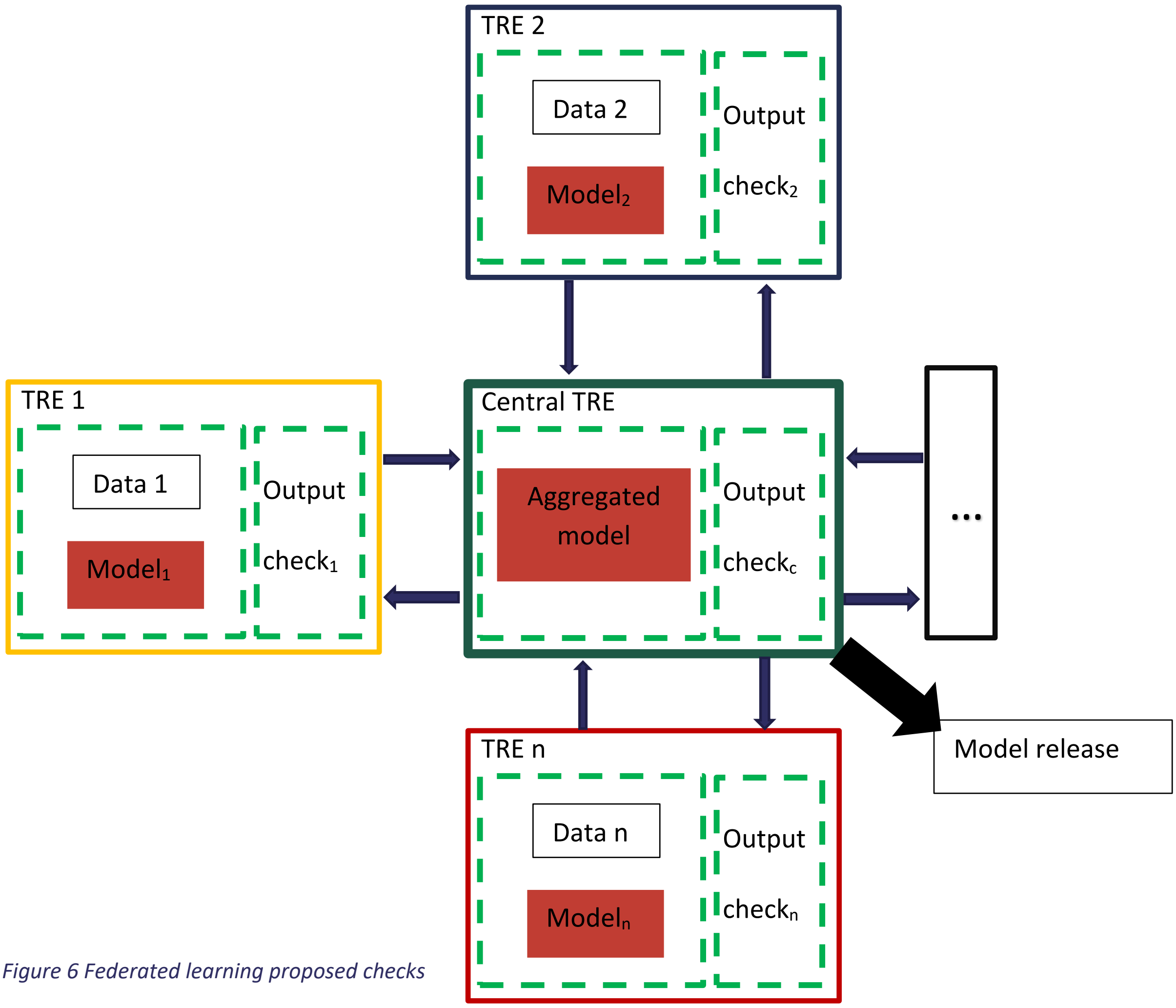


*Figure 6 Federated learning proposed checks*

### 10.2.15 TECH 15: Identifiability Control: Synthetic data should still be considered to be personal data

Synthetic data generators (SDGs) can often produce arbitrary amounts of data from a fixed sampling distribution, so should be thought of as specifying a *distribution* rather than a *dataset.* A classifier can be thought of as a conditional distribution of outcomes given covariate values, but an SDG specifies a joint distribution of covariates and outcomes.

We recommend that

- If an SDG is used exclusively inside a TRE, the aggregate process of generating synthetic data and fitting a classifier to this synthetic data should be considered as a classifier fitted to original data and tested as described above.
- If an SDG is to be released from the TRE, a provably differentially-private fitting mechanism should be used for the SDG. The level of differential privacy should be specified before beginning data analysis. Users of synthetic data generated by the released SDG should be made aware that the sampling distribution of the SDG may be different from that of the original data, and if models are fitted to the synthetic data we would expect attenuation in performance when used on non-synthetic data.

*If an SDG and a classifier are both to be released from the TRE, we are investigating if we should recommend that these be fitted to disjoint datasets.*

**Responsibility:** TRE and researchers for implementation. **Understanding:** Data governance teams

### 10.2.16 TECH 16: Model Query Control: Controls should be considered to limit the queries on a model once it is released from a TRE

Limiting the number of queries on a model is much safer than releasing it for external deployment as the usage can be tracked and any odd behaviour detected so that appropriate action can be taken.

There are several ways to limit the number of queries on a model. For example:

- **Secure Web Service:** the model could be held securely within a web service which receives query data and returns only the answer from the model. Such a webservice could be run by the TRE itself, a trusted third party or by the organisation of the researcher. Controls for a model deployed within the secure web service include:
  - restricting access to a small number of IPs that can query it, e.g. if the model is only used by NHS Scotland, the web service could be configured to reject all other incoming query traffic.
  - Maintaining a log of all queries received can help identify an attempted attack (e.g. if there are a high number of repetitive queries from the same IP this highlights an attempted attack).
  - Restricting the number of queries within a time period.
  - *Other controls will be added after more research*
- **Controls on a software program:** the software program which embeds the trained model should include controls around the number of times a particular end-user can query the model within a time period. However, there might be practical difficulties in ensuring that the program is used in this way, and can't be reverse-engineered. *More needed.*

Where there is a risk that the controls could be circumvented (for example, by user collusion), additional technical controls, such as those listed within Sections 10.2.3 to 10.2.12, should be considered.

*We are still investigating Model Query Controls*

**Responsibility:** TRE for implementation. **Understanding:** Researchers and governance teams.

### 10.2.17 TECH 17: Model Query Control: Controls placed on a model to limit the number of queries should be tested by an external party

In addition to internal security processes, TREs are regularly penetration-tested by an external party to check for security holes. If it is required that controls are implemented to limit the number of queries on the model. External penetration tests of these controls should be considered (consistent with model risk) to ensure that there are no security holes which could mean a hacker could circumvent the controls. For example, a webservice which is hosting the trained ML model should be externally penetration-tested.

**Responsibility:** Researchers. **Understanding:** Researchers & governance team.

## *10.3 Technical recommendations which require future investigation*

The GRAIMATTER project was an 8-month sprint project. As such, we did not have the time and resources to cover all areas. This is a non-exhaustive list of areas we would like to focus on as part of future work.

- *Transfer learning*
- *Imaging data*
- *Balancing the needs of researchers to have all the data vs enabling "worst-case" attacks – use of synthetic data as a proxy must be from the same distribution?*
- *Group disclosure: does this need to be handled separately.*
- *Collusion of users to mount disclosure attacks on TREs.*
- *Possibly refining the right metrics for risk influence: discussion with PPIE e.g. Risking disclosure of 1 person while benefitting 99 others. Acceptable or not?*
- *Combination of metrics to quantify disclosure: Explore and implement multiple criterion decision analysis.*
- *Quantification of risk appetite by TREs being involved in discussions around data privacy.*
- *Exploration of the risks in case of linking to a specific 'someone' in the dataset and other inferences about data.*
- *Homomorphic encryption.*
- *Impact of using safe wrappers on model accuracy.*
- *How TREs remain up to date? Mitigate new risks and challenges? Running new attack scenarios.*

# 11 Ethical and Legal Aspects (ELA)

## 11.1 ELA Background

We have investigated the legal and ethical issues accompanying ML model release from TREs. We identified and assessed how current UK legislation applies to TREs supporting trained ML model release and the extent to which the legislation addresses ethical issues pertinent to the release of ML models out of TREs, and what happens after that release, developed based on personal data, such as transparency, privacy, data protection and non-discrimination. We looked at the main legal obligations incumbent on the researchers and TREs, including protecting the confidentiality of the personal data held by the TRE and used for training ML models. We took into account the duty of data controllers to protect confidentiality but also share data in the public interest. In particular, given the aforementioned identification of data breaches as a major risk for ML model release from TREs, we investigated how current UK laws apportion responsibility for a personal data breach, misuse of private information, and breach of confidential information in research projects when releasing trained ML models from the TRE.

We drew from the field of AI ethics and governance from an international level (UNESCO Draft Recommendation on the Ethics of Artificial Intelligence [39]) and EU level (proposed Artificial Intelligence Act) to inform our analysis, especially in considering and guiding the reform of applicable UK frameworks. We considered whether a bespoke ethical impact assessment is needed before the release of a trained ML model, based on TREs data, as suggested in the UNESCO Draft Recommendation, and potentially also after the release through an ethics-based auditing system.

Our research on the security of the machine learning model considers the specific provisions of the Data Protection Act 2018 which aims to:

- demonstrate the need for a more detailed and particularised implementation of appropriate technical and organisational measures [7].
- acknowledge the requirements of the Data Protection Act 2018, that a level of security commensurate with and appropriate to the risks [7] arising from machine learning models is urgently required as these tools are increasingly used to process personal data.
- acknowledge the regulatory need for logging and, in particular, the identity of the person [7] who consulted the data, despite the lack of joined-up and interlinked registration systems to assist with the traceability of an ML model. As a result, these recommendations partially address the UNESCO call for traceability, human oversight and determination [39] for "any stage of the life cycle of AI systems".
- provide a stepping stone towards the roadmap to an 'effective AI assurance ecosystem'[40].
- incorporate the need for "greater algorithmic transparency and accountability"[41].
- incorporate the call for answerability, which can manifest itself in a continuous chain and designation of human responsibility [43], as proposed by Leslie in the Alan Turing Institute Guide for the Responsible Design and Implementation of AI systems in the public sector.

## 11.2 Ethical and Legal Recommendations:

The personal data used in TREs, which may be released in a trained ML model, will be governed by data protection law if it exists in a particular jurisdiction. Internationally, the European Union's General Data Protection Regulation (GDPR) [44] is the most prominent data protection framework, and the UK implemented it into domestic law before it left the EU as a Member State [45]in the Data Protection Act 2018. Given our focus on UK-based TREs, we proceed with some key points about their application.

The key challenge is responsibility for a possible data breach resulting from the release of a trained ML model, where the output checker does not have access to human-readable outputs. The aforementioned attacks (model inversion and membership inference) may lead to the identification of personal data, thereby rendering the model a personal dataset, to which data protection law would apply.

Understanding also that the disclosure of a trained ML model, which is not as human-readable as typical TRE releases such as the disclosure of standard statistics (e.g. graphs and tables etc.), raises certain risks and complications for TREs is important. The security of processing is a responsibility attributed to the controller and the processor, as outlined in Article 32 of the GDPR [44]. The controller and processor must take "state of the art", technical and organisational measures into account. These include "appropriate" measures commensurate with the risks, in terms of (a) pseudonymisation and encryption, (b) ongoing resilient systems and services, (c) ability to restore personal data (PD), and (d) regular testing for effectiveness of the security of the processing.

For a trained ML model, the chain of responsibility may be unclear where the initial data controller and TRE processor are no longer involved once the ML model has been transferred out of the TRE, and if the responsibility for any potential future personal data breach has not been identified nor allocated correctly in the terms of contractual agreement that a TRE operator has made with a researcher. These are issues we aim to address in these ELA recommendations.

Our recommendations here relate to the application of data protection legislation to trained ML models once they leave the TRE, the insertion of new contractual terms in existing contractual agreements or linked agreements which researchers sign when accessing TRE data, and the ethical review process for projects using TREs to produce ML models.  A simplified way to link the initial ethical approval to the potential risks involved with the post-export of an ML model from the TRE would be to incorporate contractual terms in existing contractual agreements governing the post-export of the ML model.

*We are also working on adding additional clauses to the legal requirements set out in the Human Rights Act 1998 and the Equality Act 2010 to ensure compliance with human rights laws and that there is no direct or indirect discrimination on the basis of protected characteristics.*

### 11.2.1 ELA 1: Data Controllers, TREs and researchers should consider that Data Protection legislation may apply to the trained model and data sharing agreements may be required

Data protection legislation (in the UK, the Data Protection Act 2018 which currently implements the EU GDPR) is generally considered not to apply to the anonymous aggregate level releases from classical statistical data analysis from a TRE as appropriate controls ensure that these releases do not contain personal data. A trained ML model, in contrast, may be considered to potentially contain pseudonymised personal data including special category personal data, therefore requiring specific technical and organisational measures to ensure the processing is compliant with data protection law. In particular, it needs to be considered whether appropriate data security measures have been adopted to reflect the risk of data breach.

Assessing the level of risk as to whether the trained ML model and/ or its output are inextricably linked to the personal data it was trained on will determine how the law categorises the model. We consider the following 3 categories:

1. A trained ML model can be considered to only contain anonymous data and therefore data protection law may not apply (as would be the case for aggregate level releases from classical statistical data analysis).
2. A trained ML model is considered to potentially contain pseudonymised personal data, therefore requiring specific technical and organisational measures to ensure the processing is data protection law compliant. In particular, whether appropriate data security measures have been adopted to reflect the risk of data breach.
3. A trained ML model is considered to carry more risk of including personal data given the risk of the deliberating ‘hiding’ of data or vulnerable parts of a model. Certain forms of attack (model inversion and membership inference) may render ML models as personal data(sets) [19]. In addition to differential privacy techniques and measures to tackle overfitting, Data Controllers (through the Data Governance Committees) may need to consider further legal protection in contractual terms to govern the transfer of responsibility, obligations and rights to a new data controller/processor associated with the released ML model, ensuring prior written authorisation of the controller, or indeed the retention of responsibility, obligations and rights by the original data controller in the contract. In this instance, new contractual

agreements and/or new terms in existing contractual agreements will be required between the data controller and researchers before the ML model is released from the TRE environment.

A data protection impact assessment (DPIA) will need to be carried out by the researchers/their organisation and submitted as part of the data governance and ethics approval processes.

*We plan to reach out to the ICO ourselves including on this aspect, although the consultation may not be completed before the end of this project.*

### 11.2.2 ELA 2: Existing Data Governance and Ethical approval processes should consider a range of risks and controls with each application providing sufficient details to support the review process

Both Data Governance and Ethical approval processes should consider the risks and controls associated with the training and release of ML models, particularly as regards the possibility that such models may contain personal data, as this is a different scenario from conventional TRE releases which would not typically contain personal data. In this section, we term these processes as approval processes rather than spelling them out each time.

As described in the Scoping section, both researchers and projects must go through both the ethical and data governance approval processes to access TRE data. In their applications, researchers need to articulate the wider benefits of their work when they wish to release an ML trained model and explain how they will secure the models. This is especially important when ML trained models may be released as full consideration of benefits may add weight to the value of the research versus additional risks from disclosure.

It should be the responsibility of the researchers to provide sufficient details of the likely risk and controls for the Data Governance Committees and Ethical Boards to review. These risks and controls should be revisited and finalised before the trained ML model is released from the TRE as a collaborative process between the researchers and the TRE. Any significant changes to the risk profile compared to what was outlined in the initial approvals should be notified to the approval bodies and new/amended applications for approval may be required in such circumstances.

*However, before the researchers should be obliged to take on this responsibility, the risk and controls surrounding the use of ML models trained on personal data should form a specific part of a high-level decision by the Information Commissioners Office acknowledging the importance of the security of processing personal data in the machine learning industry. In light of the importance of this work, we are currently investigating a prior consultation with the Information Commissioner for approval of the minimum baseline recommendations published in this Green Paper will form part of the work to be conducted in the aftermath of the publication of the final Green Paper. Receiving concrete guidance from the regulatory body, in an attempt at progressing the security of personal data in the machine learning industry, would recognise its importance as a fundamental building block. This, in turn, will help progress the work by the Central Digital and Data Office Algorithmic transparency template, the Information Commissioners Office AI and data protection risk toolkit and the Ada Lovelace Algorithmic Impact Assessment: AIA Template.*

**Responsibility:** Data Governance and Ethical approval teams; **Understanding:** TRE and researchers

#### 11.2.2.1 ELA 2.1: Approval processes should consider the additional disclosure risks and the controls used to protect data confidentiality and the specific legal and ethical implications

The following should be considered by the approval process:

- Information on the release of a trained model and any safe channels to be used to facilitate the release/deployment/license/transfer.
- The risks, the risk spectrum, controls and benefits for the ML model release:
  - How the technical recommendations provided in the Technical Section will be implemented/observed e.g.
    - TRE staff will run attack simulations and the researchers will use safe wrappers to reduce the risk that the trained ML model contains identifiable data (Identifiability Controls)
    - Or, the risk the trained ML model contains identifiable data is significant and therefore controls will be placed on the queries of the model by hosting the release within a secure webservice (Model Query Controls).
  - The appropriateness of releasing the ML model considering the risks
  - The purpose of the release
  - The public benefits of the release
  - The benefits of the mode of the release e.g.
    - The trained ML model is shared openly, after being subject to appropriate disclosure controls, supporting peer review, scientific reproducibility and re-use.
    - Or, the trained model will be hosted within a secure webservice. As the queries on the model will be controlled, privacy-preserving methods (required to safely release the model openly) will not be required resulting in a higher degree of model accuracy.
- Mechanisms for the traceability of released ML models, as ways to address issues of trust and trustworthiness.
- Whether the organisations involved are considered ‘safe’ e.g.:
  - The organisation responsible for applying any controls to limit the queries on a model
  - The TRE
  - The organisation the researchers and any collaborating researchers are from (academic, public sector, industry)
- Potential controls to stop the model from being used for another purpose than listed within the approval documentation e.g.
  - If the model is released completely openly with no conditions then such controls would be added to a user agreement (ELA 6)
  - If the model is embedded within a software or hosted by a trusted third party, controls should be included to necessitate an amendment or a new approval process for an alternative purpose.
  - The consequence of legal sanctions in form of breach of contract if the model is used for other purposes as this would breach the Data Security Measures as per the suggested clauses to be added to the user agreements etc.

In some cases, approvers might decide that the privacy risk associated with the disclosure may vary between (types of) attributes within the same dataset. In such cases, it may be appropriate to provide researchers and output checking staff with a ‘risk appetite’ on a feature-by-feature basis. It may be appropriate for some ML models not to be released from a TRE environment as they are too risky.

Experience shows that actively seeking to engage with data providers (rather than just making material available) tends to produce more positive outcomes. Data Governance Committees include representatives from the data providers who can approve projects. We recommend that researchers engage with these data provider representatives prior to application submission, particularly if they view the project as high risk.

**Responsibility:** Data governance teams; researchers for drafting **Understanding:** TRE

#### *11.2.2.2 ELA 2.2: Data controllers/data governance committees should be able to mandate a time limit for the use of a model, after the expiry of which the researcher needs to seek new approvals from the TRE*

Where a model is not openly shared on release, it may be possible that controls are placed on the time the model can be used without requesting fresh approval. Approval processes should consider whether a time limit is appropriate and if so, the controls that would be put in place to ensure that the model is not used after this time period without fresh approvals being granted.

**Responsibility:** Data governance teams; **Understanding:** TRE and researcher

#### *11.2.2.3 ELA 2.3: Approval processes should mandate necessary requirements to keep the data and the pipeline available to meet legal requirements*

There may be regulatory or legal requirements to keep a copy of the data and pipeline used to generate the trained ML for audit purposes for several years (e.g. if the model is used in a medical device such information may have to be kept for 15 years). The approval process should consider this requirement and the feasibility of meeting this need. The approval process should consider the implication for the persistency of all of the organisations involved to meet these requirements, i.e. what would happen should the TRE or another relevant party no longer exist within the time period.

**Responsibility:** Data governance teams; **Understanding:** TRE and researchers

#### *11.2.2.4 ELA 2.4: Consideration should be taken regarding the correct language to describe the process*

Determining the correct language for different uses of the trained model is important. The following terms could be used to help explain how a trained model can travel from a TRE to another TRE or elsewhere: release, deployed, used, licensed and transferred. Language is also important for determining who retains ownership, and/or control of the trained model, and how the technical and organisational measures are logged under existing Data Protection Act 2018 rules, in addition to the traceability obligations in terms of the life-cycle of an AI system. These are key issues that need to be addressed by the data governance and ethical application processes.

**Responsibility:** Data Governance teams; **Understanding:** TRE

#### *11.2.2.5 ELA 2.5: Researchers should complete DPIAs and the ICO AI and Data protection risk toolkit and provide these as inputs to the Data Governance and Ethical approvals process*

Researchers should complete Data Protection Impact Assessments (DPIAs) and the Information Commissioners Office AI and data protection risk toolkit [46] and provide these as inputs to the Data Governance and Ethical Approval process.

**Responsibility:** TRE; **Understanding:** Data Governance team

#### *11.2.2.6 ELA 2.6: Data controllers should have the option of refusing to release a model or recalling a released model if it is deemed that the risks of release/continued release are too high*

After these processes, if the TRE (acting on behalf of the Data Controller) deems the model to be too risky to release, it should be able to decide not to release the model and order the researchers and their organisation to cease working or using the model. The TRE should also be able to stop the continued use of a model it has previously agreed to release if after release it is shown to pose excessive/unacceptable risks to data security, privacy or other human rights (this may not be technically possible if the model has been openly released after identifiability controls have been applied i.e. there are no model query controls applied).

**Responsibility:** Data Governance team and TRE; **Understanding:** Researchers

### 11.2.3 ELA 3: Legal contracts/contractual terms should be considered to cover the responsibilities of each party if controls on the model are required after release from the TRE

If controls on the model are required after release from the TRE then the Data Governance Committees and Ethical Boards should consider the requirement for legal contracts/ the insertion of new contractual terms in existing contracts, e.g. a data user agreement, to ensure compliance with the responsibilities of each party to implement and maintain such controls. It is the responsibility of the researchers to ensure that these requirements are met in any subsequent agreements they enter into e.g. with a web service and in the form, such agreements may take e.g. a software licence. The researchers will be in breach of relevant contracts with the data controllers if the user does not fulfil these responsibilities and may be liable for damages for breach.

For example,

- The researchers must implement these controls vis-a-vis the model and the researchers/their organisation is liable for any loss or damage if these controls are not implemented.
- A company running a web service could be contractually required to ensure that the controls to limit the number of queries are in place.
- If a TRE is responsible for running the web service, the researchers who trained the ML model may wish to add an availability requirement.
- Exported ML models should not be allowed to be used for another purpose than the one listed in the approval within the approval documentation. This should be controlled by a contractual agreement (e.g. software licence between the researchers and subsequent users, and also other contractual agreements between the data controller and researchers) which forbids such activity. It would be the responsibility of the researchers who trained the ML model to add such a clause to such a contractual agreement (e.g. software licence) with which they make the ML model available to other users.

**Responsibility:** Data Governance & Researchers; **Understanding:** TRE staff

### 11.2.4 ELA 4: Data Use Registers should be extended to include information on trained models and detail about their release and controls

It is recognised as good practice for TREs to keep a record of all of the research projects they support and which datasets were used. Health Data Research UK drafted a white paper providing recommendations for a data use register standard [3], supporting TREs to make such data publicly available within a standardised format.
We recommend that the data use registry standard is enhanced/extended to specifically include:

- A category of release is a trained model
- Details covering what the model was trained to do and the datasets it was trained on
- A new entry for each time a newly trained model is released
- A record for the attribution of responsibility between controllers/processors, processors/sub processors and how this changes over time to enable it to be added to any contractual agreement in the future
- The intended and actual uses of the trained ML model
- Information on any regulatory approvals of the model and associated dates e.g. medical device regulation or CE mark
- The controls are utilised to reduce the risk of personal data being released to an acceptable level.

The use of such a data use register could be extended by requiring the compulsory registration of the ML model in order to facilitate its auditing progress. This could be facilitated easily by modifying and increasing the information collected by the HDRUK data use register, as described above. This information could include updated registration numbers which are continually logged on the Information Commissioner's Office register of controllers and new processors. This information could include up-to-date certified audits (reference number(s)) by the Central Data and Digital Office for high-risk machine learning models (note the algorithmic template). This information could include a certified audit (reference number) by the Information Commissioners' Office for the use of a high-risk machine learning model (note the AI risk tool). This information can then be linked to a contractual agreement. This information can then be linked to a contractual agreement.

**Responsibility:** TRE; **Understanding:** Data governance teams and researchers

### 11.2.5 ELA 5: Researchers, TRE staff and Data Governance/Ethics Committees should be required to complete ML model training courses

Completing these courses (see Section 13 on Training) should be an obligation that researchers must fulfil before accessing TRE data and releasing an ML model. TRE staff and those who sit on data governance and ethics committees should also complete the courses targeted to their specific audience.

The training courses of researchers, TRE staff and Data Governance/Ethical Boards should include legal and ethical aspects of ML model disclosure (as described within the Training Section).

**Responsibility:** Data Governance teams, TRE staff & researchers

### 11.2.6 ELA 6: Clauses should be added to the terms of use of any resulting trained ML model

If the resulting model is embedded within software, there should be legal terms added to the software user licence which prevent hacking of the model to determine any personal data within the software e.g. if a trained model is embedded within a medical device, those using that medical device should be prohibited from hacking the trained model as per the terms of use they sign up to. Transfer learning should also be disallowed. There should be explicit legal terms which stipulate that appropriate data security measures must be taken to ensure the security of any personal data and to mitigate against the risk of its disclosure. Hacking is illegal under the computer misuse act but these additional clauses provide additional clarification.

A legal term outlining the "permitted purpose" of the trained ML model should be added. This would restrict future use outside the "permitted purpose" assigned. To cater for the potential for a future change to the "permitted purpose", a condition should be added that prior approval and consent from the controller of the ML model should be sought prior to such use and change of purpose. In the event that the end-user proposes to use open-source software when developing the trained ML model, we propose that, that any software user license should be subject to the provisions of an additional ethical open-source license, which should form part of the legal clauses (*note the drafting of such an ethical open-source license should be worked upon, in conjunction with the legal team at the Information Commissioners Office and as part of a potential run-on GRAIMATTER project)* (see Appendix E).

**Responsibility:** Data governance teams; **Understanding:** TRE staff, researchers & end-users

### 11.2.7 ELA 7: Additional clauses should be added to researcher declaration forms

Researchers often must read and sign researcher declaration forms to analyse data within TREs. Such forms explain the researcher's responsibilities and behaviours to which they must adhere. We recommend additional clauses are added to researcher declaration forms to cover the training and release a ML model, including:

- Researchers must abide by the controls which have been approved by data governance and ethical committees.
- Researchers agree that details of the project will be recorded on the data use register
- Researchers are required to submit a new or updated Ethical and Data Governance approval documentation with any changes to the use of the trained ML model once it leaves the environment i.e. if it is incorporated into a product with a CE mark and is being sold commercially. This may involve a new or modified DPIA to be undertaken.

For instances where the data governance and ethical approvals require that models are assessed by the TRE to ensure there is the removal of personal data from the trained ML model (i.e. other controls such as limiting the queries on the model are not being utilised), these new clauses should be added:

- There is a responsibility on the researchers/their company or organisation to carry out due diligence of disclosive personal data within the trained model and in such cases must report it to the TRE and follow rules (whether those already in existence or augmented) about how to address such a situation and e.g. de-identify the personal data/otherwise mitigate the data breach
- Researchers agree that the TRE staff can run attack simulations on their model
- Researchers must be aware that to disclose a trained model they will need to provide information on their data-to-variable transformation to facilitate attack simulations by TRE staff in both human- and machine-readable ways.

*Draft clauses will be provided in an Appendix*

**Responsibility:** TREs; **Understanding:** Data governance teams & researchers

### 11.2.8 ELA 8: Approval processes may wish to consider a risk based approch vis-a-vis personal data

It may be necessary for users to conduct a Data Protection Impact Assessment (DPIA) to be included with their Ethics and Data Governance applications according to the Data Protection Act 2018/GDPR. We recommend a risk-based approach is taken to all ML model research/exports from TREs in the form of DPIAs by users and a risk assessment to be carried out by the Data Governance and Ethics Boards and TRE in their decision-making in granting access to the TRE data in the first place and then this risk assessment must be revisited and re-assessed before the ML model is exported.

These risk assessments should incorporate an appraisal of data protection and security risks and ethical risks and could also incorporate the Information Commissioner's Office AI and data protection risk toolkit [46] and the Ada Lovelace Institute's algorithmic impact assessment [47].

See also the Tech recommendation on risk assessment 10.2.7.1 for a technical perspective on this issue.

*We are currently working on this idea. A high-level idea vis-à-vis the risk of personal data identification (there may be other grounds on which a model should be restricted / not be released which are out of the scope of these recommendations) is provided in the table below*

| | **Model openly available and shared openly** | **Model embedded within the software with legal controls around hacking** | **Model embedded within the software with legal controls around hacking and limits on the number of queries** | **Model only accessible via queries to a secure web-server – limiting controls on the number of queries and users** |
|---|---|---|---|---|
| **Passes attack simulations to check for identifiable data and TRE checks** | 🟨 | 🟩 | 🟩 | 🟩 |
| **Passes attack simulations to check for identifiable data and TRE checks and use of** Safe Wrappers | 🟨 | 🟩 | 🟩 | 🟩 |
| **Fails attack simulations to check for identifiable data and TRE checks** | 🟥 | 🟥 | 🟨 | 🟩 |
| Differential Privacy Inference based models | 🟩 | 🟩 | 🟩 | 🟩 |
| Instance-based models<br>e.g. SVM | 🟥 | 🟥 | 🟥 | 🟩 |

**Responsibility:** Data Governance teams and researchers: **Understanding:** TRE & researchers

### 11.2.9 ELA 9: Template text should be available from TREs covering the range of standard controls and processes which could be applied to streamline the approval process for researchers and Data Governance Committees and Ethical Boards

Many TREs provide Standard Operating Procedures (SOPs) or template text which covers the controls that they apply to ensure personal data is protected. Researchers can utilise such information within application forms, for example by directly referencing the SOP rather than listing the controls in their own words. Data Governance Committees and Ethical Boards who review many applications which refer to such SOPs are not required to re-review the SOPs for every new application as they are already familiar with the text. This reduces the effort on behalf of the researchers and reviewers as well as reduces the number of re-submissions of applications due to missing information.

We recommend that the additional controls required for processing trained ML models are also provided within SOPs or template documents. This will help to streamline the Data Governance and Ethical application processes.

**Responsibility:** Data Governance teams & TRE; **Understanding:** TRE staff and researchers

### 11.2.10 ELA 10: TREs should agree to confidentiality agreements should the researchers have concerns re their IP

As per recommendation TECH 9, researchers should share details on their inputs to their model and details of their training method to support TREs to efficiently run attack simulations. If researchers have concerns re the confidentiality of the TRE relating to their IP, then confidentiality agreements should be considered between the TRE and the researchers.

**Responsibility:** Data governance teams & TRE; **Understanding:** TRE staff and researchers

### 11.2.11 ELA 11: The Data Controller should approve the processes TREs will apply for identifiability controls

This information could just be included as links to SOPs within the data governance applications (as is often the case for non ML projects) or could be pre-agreed as a set of approved processes.

**Responsibility:** Data Governance teams; **Understanding:** TRE and researchers

## 11.3 Additional recommendations

So far, our recommendations only focus on the issues specifically associated with risks and controls for protecting personal data (essentially the role of the TREs). In this section, we identify other aspects of ethical or legal aspects where we feel additional recommendations are required but are out of this narrower scope:

- Researchers should consider the issue of group privacy (and not only individual privacy) in the context of ML model release from TREs.
- TREs should carefully consider what additional technical and ethical/legal controls will be needed for updating existing ML models. TREs should carefully consider the checks, risks and possible controls needed if there is the possibility of allowing (knowingly or not) ML models initially trained on external data into the TRE (transfer learning using pre-trained models).

  - Researchers should be made aware of their legal responsibilities in addition to the Data Protection Act 2018 and UK GDPR e.g. Computer Misuse Act 1990, Fraud Act 2006, National Security and Investment Act 2021, Human Rights Act 1998, Equality Act 2010. (Appendix D)
- A discussion should be had on whether trade secrets are permitted to be developed, in the TRE on the basis that the controller and the public will not know what is happening to personal data, for example during the process of anonymisation and subsequent use of the trade secret for a different purpose [48].

# 12 Costing

## 12.1 Costing Background

TREs generally work using a cost recovery model. They may charge for a range of different services such as

- the work required by data analysts to extract and pseudonymise relevant data for the specific research project (this is often the case if the TRE is also the same organisation that owns the data)
- providing a linkage service as a trusted third party
- providing expertise on the data sets
- providing the secure hardware and software infrastructure for researchers to analyse the data without being able to be released without going through disclosure Identifiability Control and limiting access to the internet
- providing a disclosure control service where trained members of the TRE team assess files which researchers would like to release out of the environment for personal data.

Providing support for research projects training ML models will increase the support requirements above that of a 'normal' research project and as such these costs need to be considered.

## 12.2 Costing Recommendations:

### 12.2.1 C 1: TREs should charge for the additional work to undertake disclosure control of trained ML models and run attack simulations

TREs should estimate the additional effort to support the controls required to ensure personal data is not encoded within models to be released and to run attack simulations on models. It may be that for the first few projects that TREs support these additional costs are considerable (see additional funding requirement below, but over time this will become routine practice and result in efficiencies expected of mature processes.

**Responsibility:** TRE; **Understanding:** Data governance teams & researchers

### 12.2.2 C 2: Additional funding should be made available to support TREs to develop the tools and frameworks to support ML training

The design and development of new processes to support the recommendations listed within this document will take time and funding. We recommend that there are infrastructural funds made available for TREs to seek to support them to provide this new functionality to the community.

**Responsibility:** Funding bodies

### 12.2.3 C 3: The costs for limiting queries on the model should be considered

TREs could provide a service, such as a secure webservice, to host the trained model and provide the security required for querying the model based upon the required controls. Such controls may include limiting the IP addresses which could submit a query, limiting the number of queries over a period of time, and guarding against denial-of-service attacks. Service level contractual agreements may be required.

Such a service could also be provided by a trusted third party or the organisation of the researcher. Whether the service is provided by a TRE, a trusted third-party or the researcher organisation, the costs of providing the service would need to be covered by the researcher. The costs of designing, implementing and running the service should be considered.

**Responsibility:** Researchers; **Understanding:** Data governance teams & TREs

### 12.2.4 C 4: TREs should consider the additional costs of maintaining the data and development pipeline to support legal requirements e.g. for a certified medical device

There may be regulatory or legal requirements to keep a copy of the data and pipeline used to generate the trained ML for audit purposes for several years e.g. for medical devices such information may have to be kept for 15. TREs may have to consider the cost of supporting these requirements.

**Responsibility:** TRE for implementation. Researchers for the cost.

### 12.2.5 C 5: TREs should consider outsourcing some of the highly technical work

Although highly trained experts are needed to support research projects involving ML model training, substantial expertise at each TRE is not necessarily required. The expectation is that ML releases are likely to be relatively rare (compared to traditional statistical analysis which generates multiple release requests daily). It may be efficient for TREs, particularly those who do not have many ML users, to share ML checking expertise, calling it in when needed. Some TREs already provide "peer review" access to projects, which could be a mechanism for a pool of trained checkers to share expertise. However, even in the case of outsourced checking, it would be advisable for TRE staff to have a basic conceptual understanding of ML modelling.

**Responsibility:** TRE.

# 13 Training

## 13.1 Training Background

Previous work [49] showed that uncertainties amongst TRE staff limited the uptake of ML modelling, including an understanding of how to carry out disclosure control on trained ML models. This project aimed to provide the tools and clear guidance to allow TRE staff to have confidence in releasing models. Use of those tools/guidance requires a good understanding of ML modelling, and so there is a need for training TRE staff in the specifics of assessment.

## 13.2 Training Recommendations:

### 13.2.1 TR 1: Training courses and documentation should be developed for TRE staff on ML, how to run attack simulations and the risks of disclosive data within trained models

These courses should provide sufficient detail for the TRE staff to both carry-out tests and advise researchers on how to use the tools and guidance. The trained TRE staff should be able to engage with researchers as equals in ML. In line with current output checker training, this course should also cover how to engage with the researcher to build positive communications. As noted above (Costing Section), not all TRE staff need to go through such a course (and possibly none if output checking is outsourced).

### 13.2.2 TR 2: Introductory training courses and documentation should be developed for TRE staff on ML

All staff should be aware at an introductory level of the issues around ML models, how these are resolved, and who to contact for further advice (i.e. the specialist ML output checker). The purpose of this is to ensure that all TRE staff know how to process ML model outputs, even if they are not capable of running the tests themselves.

Such training should include:

- The risks of disclosure of personal data from trained models.
- The controls available to them to mitigate these risks such as
  - Attack simulations run by the TRE to check for unsafe practice
  - Use of safe wrappers
  - How to limit queries to the model
- Legal and ethical implications

**Usage:** TRE staff

### 13.2.3 TR 3: Training courses and documentation should be developed for researchers on the risk of disclosure control when training models and should consider controls and legal and ethical components

Such training should be a requirement for access to the data for AI model training projects. Again, the courses should follow good practice in engaging researchers in the 'why' of output checking, to build a positive community of interest. Output checking is most efficient [37] when both the researchers and the checkers understand and agree on the basic questions: what is to be checked, how will it be checked, and why is it being checked? Researchers are unlikely to have considered the disclosure risk of models or be aware of tools developed for checking. Therefore, efficient output checking of ML requires the training of researchers.

Researchers should be made aware of the different types of contractual agreements and use that could apply to the deployment or release of a trained model, in terms of a user agreement, a material transfer agreement, a license, or a trusted user agreement. This is important as in the future there may be different categories of contracts required for different transactions (research/commercial). There may also be 'linked agreements' within these agreements, the best example here is the International Data Transfer Agreement published by the Information Commissioners Office. This is also important in terms of the types of open-source licenses attached to the development of the trained models. Note this clarification in terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) [50]. When creating linked agreements as separate agreements / supplemental agreements in terms of "dual licensing", it will be important to understand that applying for different permissions or waiving them, will result in the original CC license not applying.

For commercial researchers, recognising and understanding that contractual agreements (including any linked agreements) must contain fair, reasonable and non-discriminatory (FRAND) [51] clauses is crucial. Understanding if the terms are FRAND compliant, in terms of the security of processing personal data by a machine learning model, will require a clear understanding of the baseline recommendations of this Green Paper which are intended to be submitted to the Information Commissioners Office under the prior consultation function in the Data Protection Act 2018. This clear understanding will help the interaction between technologists and lawyers in the drafting of these linked contractual agreements navigating machine learning model practices, which in turn, will help prevent or settle any potential disputes that may arise from unfair contract terms included in a subsequent contractual agreement (e.g. licenses, medical device agreements, patent licenses and trade secrets). Such training should include:

- The risks of disclosure of personal data from trained models.
- The controls available to them to mitigate these risks such as
  - Attack simulations run by the TRE to check for unsafe practice
  - Use of safe wrappers
  - How to limit queries to the model
- Legal and ethical implications

This training will help researchers to provide relevant information on data governance and ethical applications for such boards/committees to assess.

**Usage:** Researchers

### 13.2.4 TR 4: Training courses and documentation should be developed for Data Governance Committee members and ethical boards to assess applications which include the training of ML models

Such training should include:

- How to assess the risks and controls of disclosure of personal data from trained models
- How to assess the risks and controls on the limitations on queries to the model
- Legal and ethical implications

Developing such training in combination with the target groups will help to ensure that the training is relevant and useful, as well as highlight issues which the training team might not have considered.

As Data Governance Committee members include data controllers who are responsible for approving the use of their data, such training should help to identify and forestall any concerns that data controllers may have.

**Usage:** Data governance and ethical process teams.

# 14 PPIE

## 14.1 PPIE Background

We established a PPIE Group, with 8 people with an interest in health data research. We worked to ensure that across the PPIE members there is diversity including, but not limited to gender, ethnicity, age and geography. The PPIE group met virtually 5 times throughout the project to:

1. Understand health data and what is machine learning

2. Present the legal challenges of machine learning and AI in health datasets

3. Understand how the legal challenges of machine learning and AI could be addressed

4. Checking the PPIE group's understanding of the legal challenges and building on the project insights to date

5. Review the outputs and confirm PPIE approval of the recommendations

Across the workshops, several tools were used to drive engagement with the group, including Menti, word clouds, video and graphics.

The group have input into this draft green paper from a public perspective. The lay co-applicants were actively included in all workstreams and consulted on all recommendations.

### 14.2 PPIE Recommendations

*Populate after workshops*

#### 14.2.1 P 1: Public representatives should be involved in data governance and the ethical approvals process

**Responsibility**: Data governance and ethical process teams

#### 14.2.2 P 2: The use of data for training ML models and the controls on the models should be visible to the public through searching data use registers

**Responsibility**: Data governance and ethical process teams & TRE

## 15 Future Research and Development

### 15.1 Research and Development Background

*We have only been able to open this can of worms during this sprint project. There is still lots of research left to do before this becomes a mature field. There is still lot of work required by the community to develop mature processes. People need to be trained. Here are a few of the many outstanding questions*

### 15.2 Research and Development Recommendations

*More research is required in the following areas*

1. *Training materials*
2. *More wrappers – tool kit*
3. *Development of more informative and more easily understandable risk metrics that balance information about the 'global' risk (e.g. mean accuracy of inference attacks) vs. the 'local' risk specific subgroups of people.*
4. *Difference between AUC when using safe wrappers verses not*

5. *Developing an improved understanding of membership inference risks, (and how to measure them), for unstructured data such as images and text etc., where it is inappropriate to focus specifically on exact combinations of feature values.*
    - *In other words, under current definitions, a 'perfect' attack would say that someone's data was not in the training set if we changed one pixel in an image, or one decimal place in the calculation of BMI, both of which change continuously and are subject to measurement uncertainty. So we need a new understanding of membership that accounts for the inherent uncertainty in feature gathering.*
6. *PPIE*
7. *Public education*
8. *Trust and public benefit*
9. *Commercial components*
10. *Health economics – costs to support AI within a TRE. What is feasible for a researcher to pay/industry*
11. *What does all of this cost*
12. *Exemplar projects running AI in different TREs*
13. *Benefit sharing*
14. *Participatory data stewardship*
15. *Group disclosure – from a technical and ethical perspective*
16. *Methods for stratifying/analysing disclosure risks for different sub-group, and then approach to making sure that certain groups are not more vulnerable to disclosure of their data.*
17. *Genetics data*
18. *Best practice sharing – impactful rather than research*
19. *Non-TREs*
20. *Consented datasets – pay for the model*
21. *HIC to set up the same model as the US –*
22. *Updating of models –*
23. *Trusted party to host models – semi-disclosive model*
24. *Public Summaries*
25. *Legal Research: Consider the creation of a new regulated profession with categories for different industries, e.g. one could be the accredited NHS researcher, (see Goldacre Review)*
26. *Legal Research: A review of the disclaimer and warranty clauses of open-source licenses of the commonly used tools in Github Libraries / chosen commercial tools.*
27. *Legal Research: Consider crafting a codified "SAFE TRE Practice" or "SAFE Model Practice" with a view to it becoming a statutory framework for future legislation in AI or a sub-set of AI regulation called, Machine Learning, for example.*
28. *Legal Research: The lack of mention of "anonymous" practices in the current legislative provision (Data Protection Act 2018), albeit in a work-in-practice Code format provided by the Information Commissioners Office, needs to be addressed.*
29. *Legal Research: The necessity for an acceptable Open-Source License to be added to the MIT License including restrictive clauses on the lifecycle of the ML Model*
30. *Defining 'disclosure'*
31. *Water marking*
32. *Cryptography*
33. *Encryption of the model and key management*

34. *Further work on high-dimensional personal data, particularly genomics, and privacy considerations specific to this setting*
35. *More dynamic models e.g. ethics audits for data governance and ethics approval processes which follow a project's lifetime (and beyond) for AI-related projects without posing disproportionate additional burdens on researchers, TREs, Ethical Boards and Data Governance Committees.*

# 16 Appendix A: Risk Assessment of AI Models and Hyperparameters

The evaluation of risks associated with releasing trained machine learning (ML) models from TRE, is part of the Factor Analysis of Information Risk (FAIR) assessment. It considers the potential privacy risks and negative users (attackers and malicious) of the model. It assesses the potential vulnerability to a data leak in releasing the model.

ML models have several components as shown in Figure 7:

- *Architecture:* one can imagine it as a skeleton. It's how the model is structured, in the same way as there are many types of animals, and they all have unique skeletons, the same with models.
- *Training data*: it can be thought of as the filling of the skeleton: muscles, skin and so on.
- *Hyper-parameters*: can be thought of as the clothes that this person (or model) we are creating is the right size and style for the person/model. An overfitted model would be when the clothes are made to measure and only fits the specific individual and no one else, in which case the model generated is no good. Also, some parameters may not work well with some data, just as some styles of clothes are simply not appropriate for a certain type of event.
- *Distribution of training data*: it can be thought of as the characteristics of the data, the maximum, minimum, average, if there are many outliers, etc.

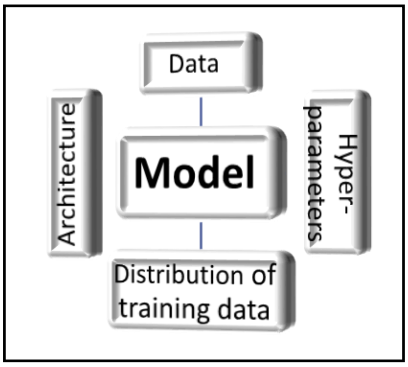
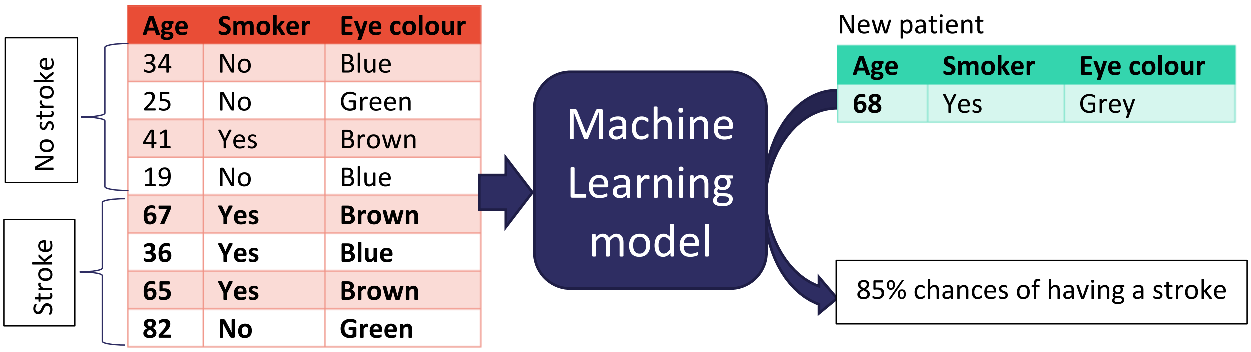


*Figure 7 Components of a Machine Learning model*

Let's create an example. Imagine we want to predict whether patients will have a stroke or not. We can take health records of patients who did not have a stroke and patients who did and train the model with the selected data. Each row in the table corresponds to one patient. In a machine learning context, the columns are referred to as features. In the example (Figure 8) below a model with 3 features is constructed (age, smoker, eye colour). Some features will be more relevant to the prediction than others. In the example below we can see that age and smoker might be informative for the prediction, but that eye colour probably isn't.



| | Age | Smoker | Eye colour |
|---|---|---|---|
| No stroke | 34 | No | Blue |
| | 25 | No | Green |
| | 41 | Yes | Brown |
| | 19 | No | Blue |
| Stroke | **67** | **Yes** | **Brown** |
| | **36** | **Yes** | **Blue** |
| | **65** | **Yes** | **Brown** |
| | **82** | **No** | **Green** |

New patient

| Age | Smoker | Eye colour |
|---|---|---|
| **68** | Yes | Grey |

85% chances of having a stroke

*Figure 8 Example of a machine learning model*

To query the model, it is necessary to have a row with the same type of data (same features) used to train the model which, in this case, would be age, smoker and eye colour. The model will be able to predict whether this new patient will suffer a stroke or not.

The minimum required for an attack is that the attacker or adversary needs to be able to query the model. The attacker will send rows of input data (age, smoker, eye colour) and the model will respond with either a predicted class (stroke / non-stroke) or scores (probabilities) associated with each class. The adversary will typically try and use this process (possibly repeated multiple times) to learn something about the personal data that was used to build the model.

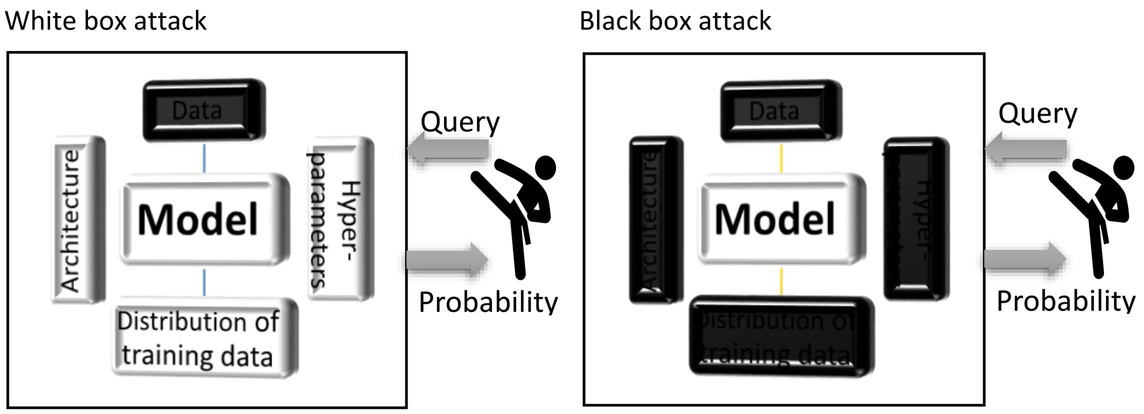


*Figure 9 Type of attacks*

The situation in which the attacker can only query the model is known as a *black box* attack. Sometimes an attacker might have more access and be able to examine the model itself. This is known as a *white box* attack. The kind of information available in white- and black-box attacks are depicted in Figure 9 above.

We focus on Membership Inference Attacks (MIA). In such events, an adversary or attacker is trying to determine whether a set of input values (they have access to, or have generated) is part of the original training dataset of the target model (referred to as target train). Our objective in assessing a model is to measure the highest potential MIA attack accuracy and, establish criteria in which models with identifiable data are safe.

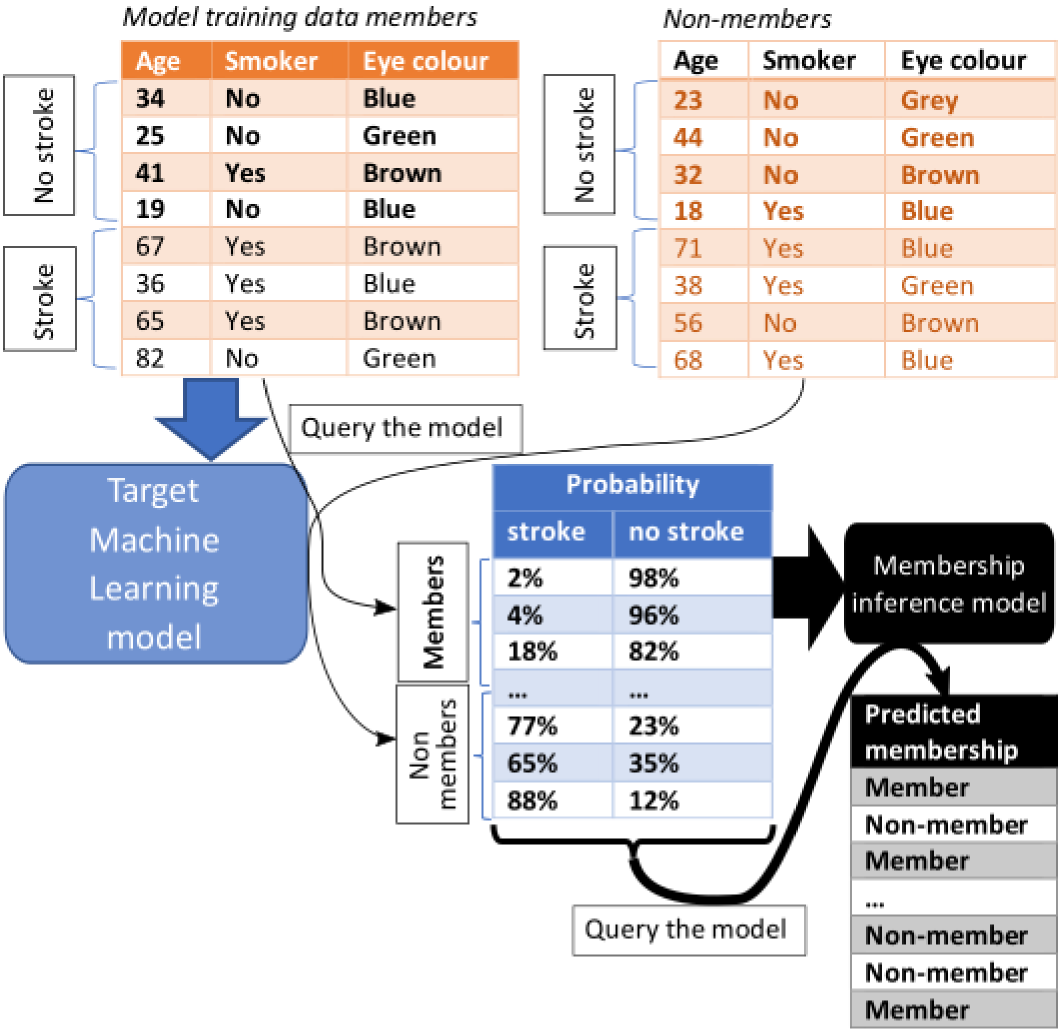
Our simulated attack experiments proceed as follows. After pre-processing, the target dataset is split into 3 equal parts: train, shadow and test. The split is repeated 5 times, varying the rows included in each part. For each classifier of interest, a set of values of hyperparameters to be explored is defined. Five target models are created for each combination of classifier hyperparameters, one for each data split. So far, all of our experiments have been on health record data.

Several attack scenarios have been defined to determine the risk of personal data leak from ML models: Worst Case, Salem 1, Salem-synth and, Salem 2. The table below contains a summary of their main characteristics.

*Table 1 Attack scenarios main characteristics.*

| **Scenario** | *Salem 1* | *Salem-synth* | *Salem 2* | *Worst case* |
|---|---|---|---|---|
| **Shadow model** | Same classifier as target model, same hyperparameters. | | | NA |
| **Shadow data** | Split of the target data (held out). | Simulated set from target data. | Unrelated to the target (breast cancer). | NA |
| **MIA model** | Random Forest Classifier | | | |
| **MIA data** | Shadow model predicted the probabilities of the shadow data. | | | Target model predicted the probabilities of the target train and target test data. |

The *Worst-Case* is a white box scenario, and it does not need any shadow model. It is described in Rezaei [52] as the easiest possible for the attacker. To perform a MI attack, a new binary set of data (member, non-member of target train data) is created containing the predicted probabilities of the training and test data by the target model. A Random Forest Classifier is fitted with half of this new set. This scenario is not supposed to simulate a realistic attack (if the attacker has access to the data, they do not need to attack) but instead to assess whether there are potentially vulnerabilities in the model that could potentially be leveraged by an attacker. This can give us a good estimation of the maximum capability of an attacker to succeed. In some cases, the risk of data leakage could be overestimated, but it does guarantee (as much as possible) that any ML model allowed out of a TRE is safe. At the same time, it's easy to implement (see Figure 10).



| Age | Smoker | Eye colour |
|---|---|---|
| 34 | No | Blue |
| 25 | No | Green |
| 41 | Yes | Brown |
| 19 | No | Blue |
| 67 | Yes | Brown |
| 36 | Yes | Blue |
| 65 | Yes | Brown |
| 82 | No | Green |

| Age | Smoker | Eye colour |
|---|---|---|
| 23 | No | Grey |
| 44 | No | Green |
| 32 | No | Brown |
| 18 | Yes | Blue |
| 71 | Yes | Blue |
| 38 | Yes | Green |
| 56 | No | Brown |
| 68 | Yes | Blue |

| Probability | |
|---|---|
| stroke | no stroke |
| 2% | 98% |
| 4% | 96% |
| 18% | 82% |
| … | … |
| 77% | 23% |
| 65% | 35% |
| 88% | 12% |

| Predicted membership |
|---|
| Member |
| Non-member |
| Member |
| … |
| Non-member |
| Non-member |
| Member |

*Figure 10 Worst case scenario diagram*

The *Salem 1, Salem-synth* and *Salem 2* scenarios are based on adversary attacks described on Salem [34]. The three scenarios work in the same way and need a shadow model, represented in Figure 11. The attacker does not have access to the data used for training and testing the ML model. The attacker can only query the model and obtain probabilities, and it needs to create or find data for the "shadow model" to perform the attack. The same combination of the target classifier with identical hyperparameters as the target model is used to generate the shadow model with the shadow data. In *Salem 1* the data to train and test the shadow model is the shadow data split from the target dataset; in *Salem-synth* the shadow data is synthesised from the target dataset (still data from the same distribution); and *Salem 2* uses an unrelated set of different distribution, which is breast cancer data available in the python package scikit-learn. In all cases, the shadow data is split into two equal parts, train, and test shadow. Half of the predicted probabilities by the shadow model of the target train and test shadow data are used as a new set to train a binary classifier (Random Forest).

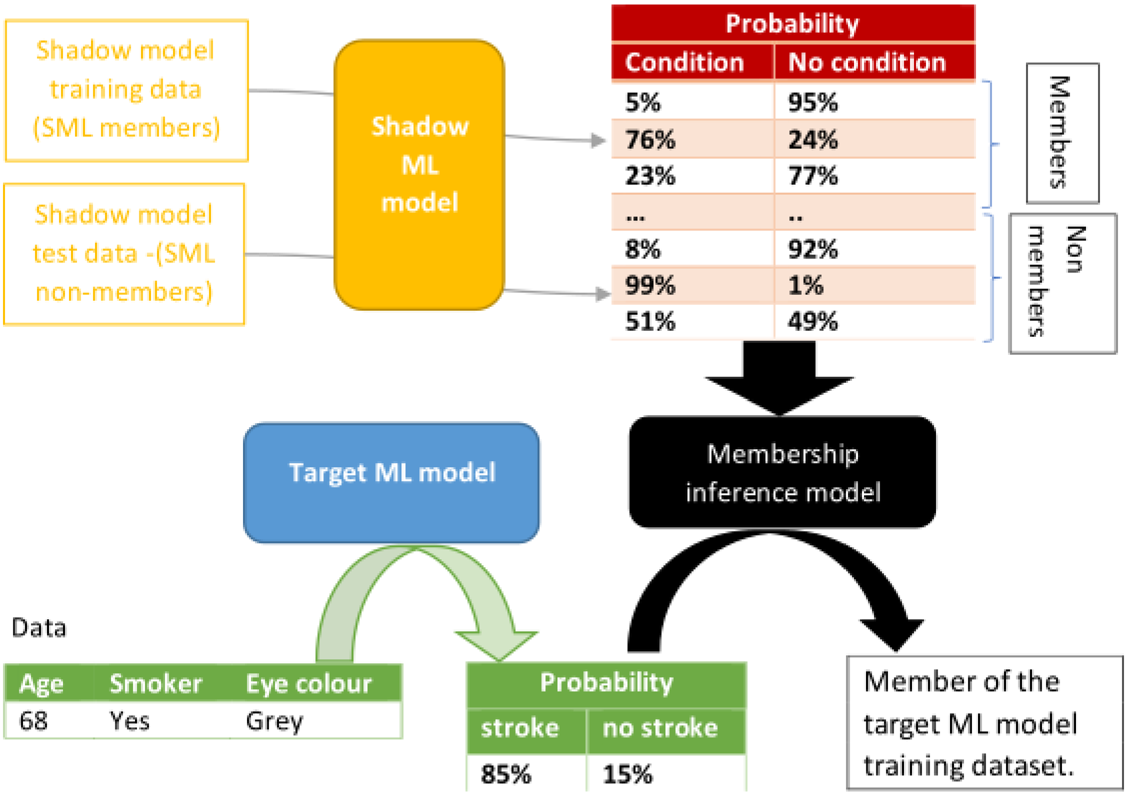


*Figure 11 Salem scenarios diagram*

In all the scenarios (Figure 11), the MIA model predicts whether a data point belongs to the target model training data and is validated with the test MIA set.

The predictions obtained from the target, shadow and MIA models are used to generate a confusion matrix and subsequently calculate the metrics of True Positive Rate (TPR), False Positive Rate (FPR), False Alarm Rate (FAR, also known as False Discovery Rate - FDR), True Negative Rate (TNR), Positive Predictive Value (PPV), Negative Predictive Value (NPV), False Negative Rate (FNR), Accuracy (ACC), F1Score, Advantage – which measures the capability of the attacker to distinguish if a data point belongs to the target training set [53] and is measured as the absolute value of the difference between TPR and FPR, and their corresponding probabilities are used to calculate AUC accordingly.

The results presented here correspond to 7 target classifiers with a hyperparameter combination range from 12 to 3840 per classifier. The figures below show the metrics' true positive rate of the Random Forest classifier for the parameter n_estimators and the mean AUC per target classifier (see Figure 12 and Figure 13). Our results highlight variability depending on the classifier and also the target dataset (data not shown).

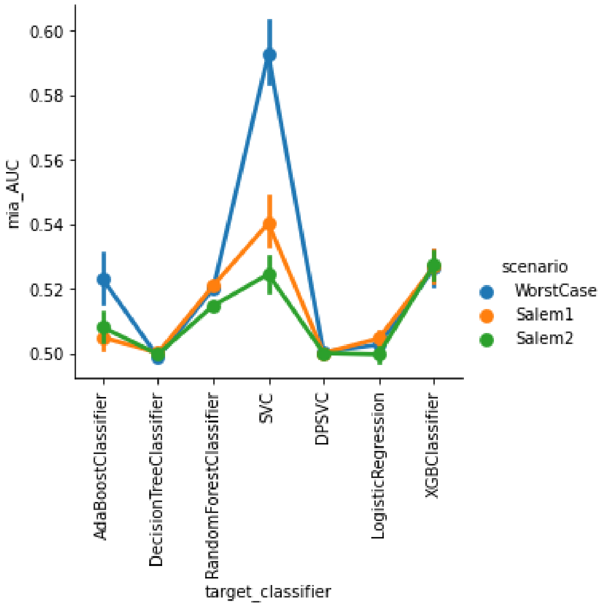


*Figure 12. AUC by target classifier and scenario.*

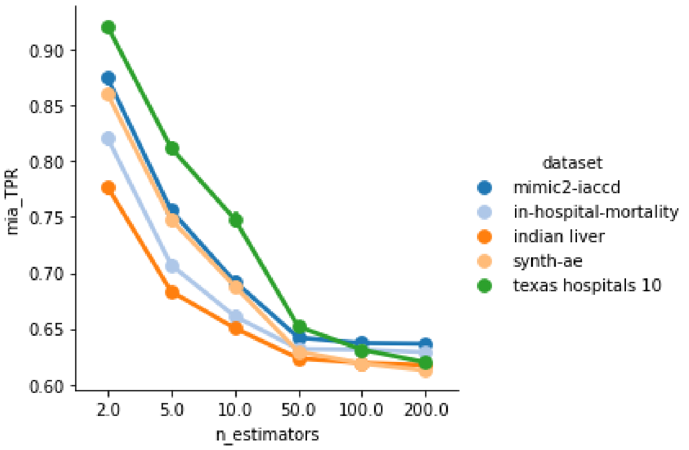


*Figure 13. True positive rate by n_estimators for Random Forest Classifier by dataset.*

## 16.1 Experiment 1: Scenario comparison

A comparison of the scenarios was performed to determine which scenarios ought to be included in an attack suite for the TRE staff. In particular, the scenarios were compared to find the scenario which gave the most conservative risk assessments.

The comparison was performed over several open tabular health datasets. For each dataset, the scenarios were compared with a range of target classification models (non-neural network models so far). In each case, the membership inference attack classifier was a Random Forest with default sklearn parameters. Target model hyper-parameters were varied over a large range and various attack metrics were calculated. In all, a total of ~500,000 experiments were performed.

The experiments showed that, on average, the Worst Case scenario had the most conservative risk estimates. A summary of the results is shown below. We show just Worst-Case versus Salem 1 (Salem 2 had consistently lower attack performance than Salem 1, see Figure 14). Here we show three metrics: the membership inference area under the ROC curve (mia_AUC, see Figure 15), the membership inference advantage and the membership inference F-score (Figure 16). In all cases, a higher score means more attack success. Each metric is represented by one pair of plots. The left-hand plot shows a histogram of the difference in the metric between the Worst Case and Salem 1 scenarios. A general positive trend suggests the Worst Case scenario consistently gives higher risk values. The right-hand plot shows the metric for the Worst Case scenario (x) versus the metric for Salem 1 (y). The key point from this plot is the absence of points in the upper-left quadrant (where Salem 1 would indicate riskiness and Worst-Case would not).

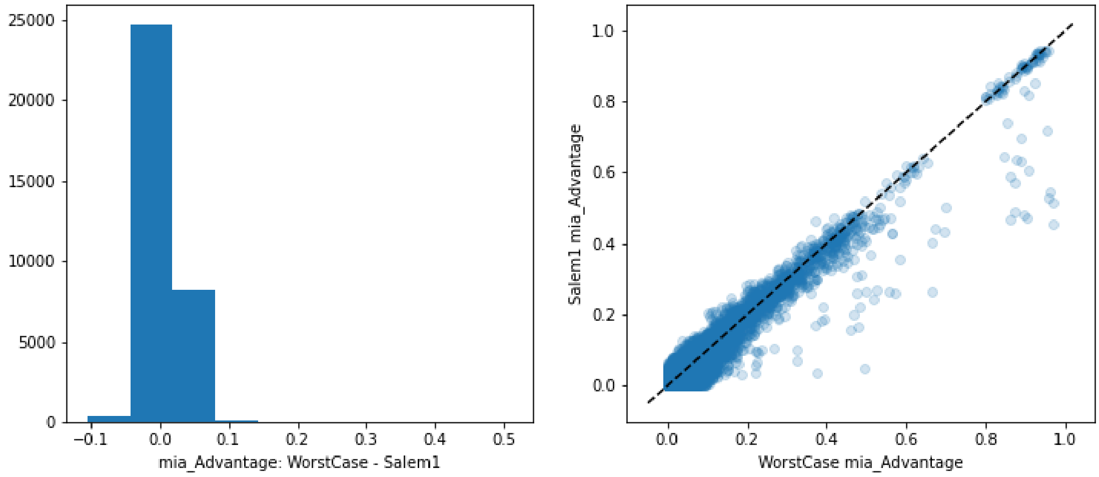


*Figure 14. Advantage comparison of the worst case versus Salem 1 scenarios.*

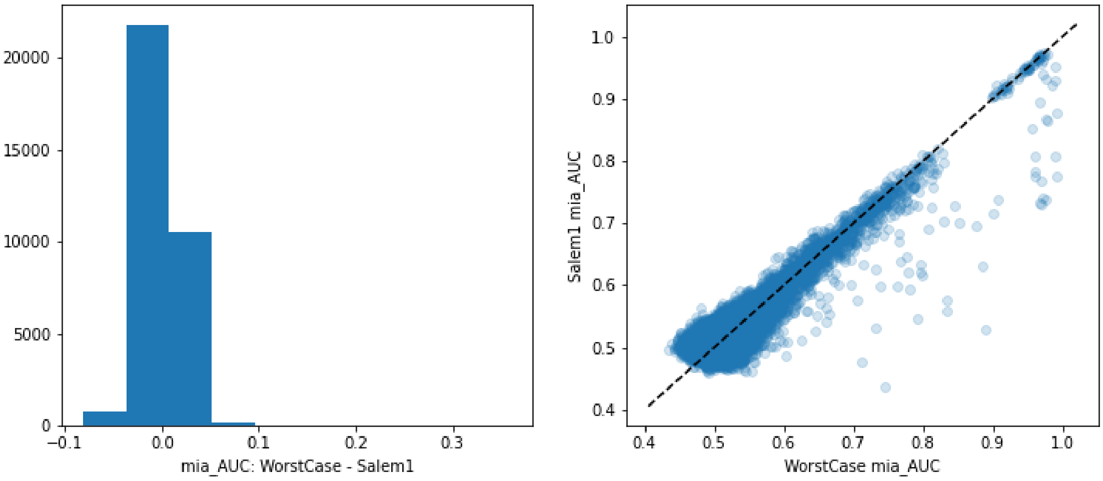


*Figure 15. AUC comparison of the worst case versus Salem 1 scenarios.*

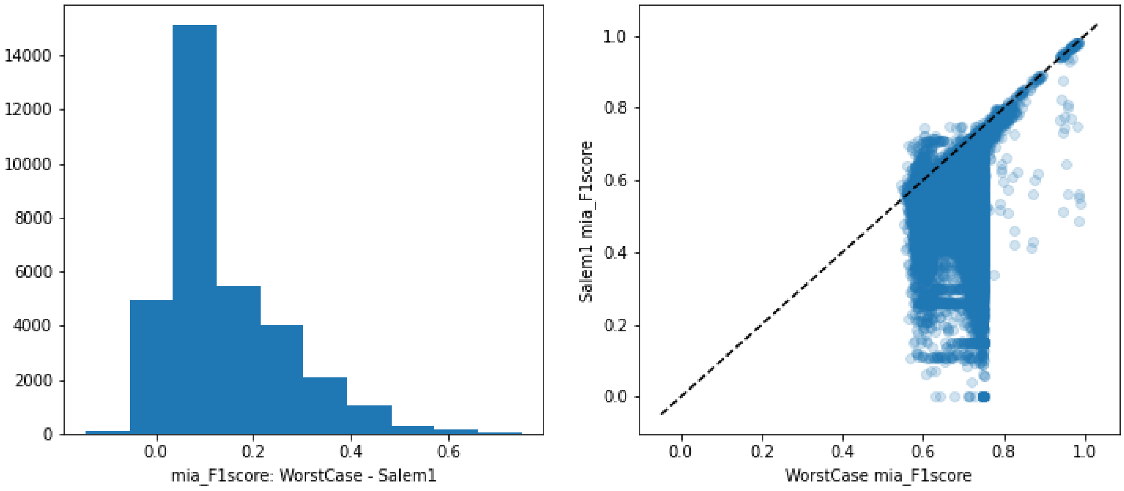


*Figure 16 F1 score comparison of the worst case versus Salem 1 scenarios*

The following additional conclusions can be drawn from these results:

Even classical ML algorithms (decision trees, random forests, Support Vector Machines) can be vulnerable to membership inference attacks, as can be seen by points in the upper right quadrants of the right-hand plots.
There are many situations in which the Worst Case scenario identifies risk when Salem 1 does not. It would be dangerous to release such models as they clearly have a degree of vulnerability, even if current attack methods would be unable to expose them.

## 16.2 Experiment 2:

In our second suite of experiments, we investigated the extent to which the riskiness of a set of hyper-parameters is generalised over datasets. This is an important consideration for TRE staff. If riskiness does generalise across datasets, then one could be confident that a model was safe based on experiments with that model on previous datasets, reducing the burden on TREs to perform their experiments.

We investigated a wide range of hyper-parameter configurations across a range of popular ML models for each of our datasets and then computed the minimum and maximum value of a range of attack metrics across the datasets. This resulted in, for each metric, minimum and maximum attack metrics across datasets and we were interested in identifying if there were hyperparameter values which were safe in one dataset and then very risky in others. Note that all datasets were of a similar type – tabular data.

The results showed that for popular model choices, there exist many hyperparameter values that appear safe for one dataset, but highly unsafe for others, suggesting that it would not be sufficient for TRE staff to rely upon previous analysis to determine if a new model was safe or not.

Example results for the Random Forest Classifier are shown below (Figure 17, Figure 18, and Figure 19). In each plot, the minimum value of the metric is plotted on the x-axis and the maximum on the y. Each point is a hyper-parameter configuration. Configurations with risk performance that generalises between datasets would have very similar minimum and maximum values and therefore exist on the y=x diagonal. We can see that although this is the case for some, there are many for which it is not. In particular, there are many cases where the attack AUC ranges from 0.5 (unsuccessful) to 1.0 (totally successful).

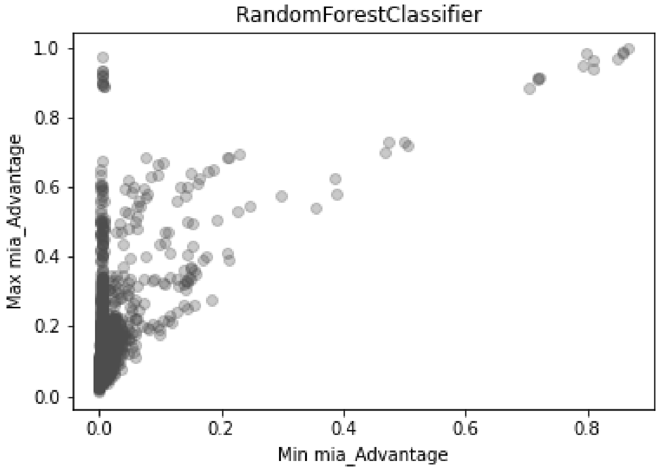


*Figure 17. Maximum versus minimum attack advantage*

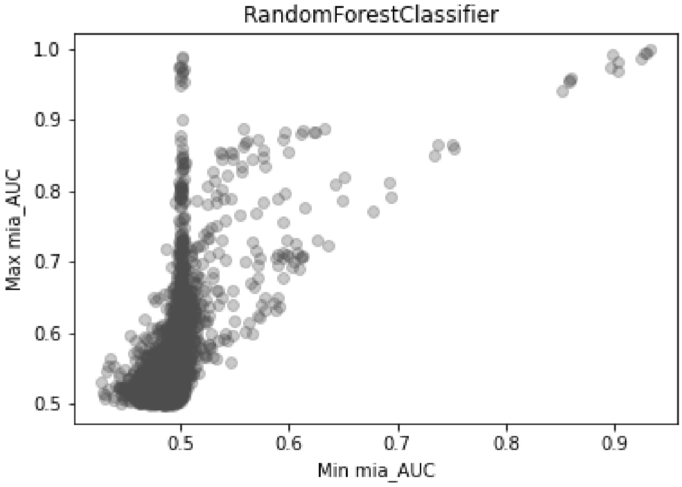


*Figure 18. Maximum versus minimum attack AUC*

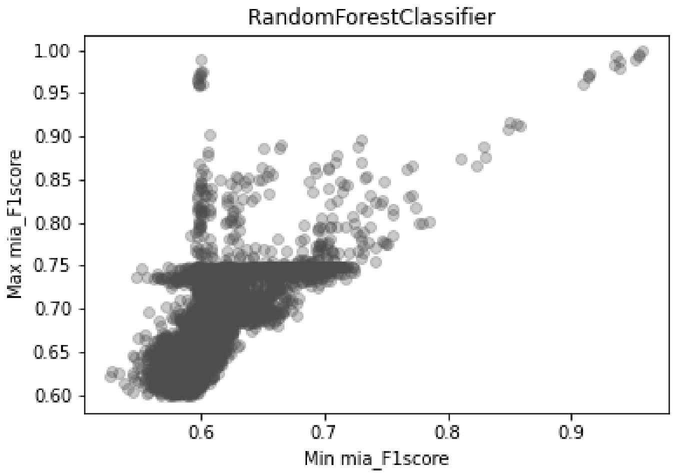


*Figure 19. Maximum versus minimum attack F1 score*

It is worth noting that although we can see that riskiness of hyper-parameter configurations does not generalise overall, there appear to be many examples that are either always risky within the context of our experiments (top-right), or never risky (bottom-left). Parameter configurations identified as being risky in at least one experiment can be used to form the ranges to be used by the Safe Wrapper classes (Section 10.2.10).

## 16.3 Attribute Inference attacks

The experiments above refer to a framework for investigating Membership Inference attacks. A similar investigation is now underway to explore the vulnerability of algorithm-hyperparameter-dataset combinations to Attribute Inference Attacks and investigate the relationship between these two different types of risk.

Informed by Experiment 1 above, we have designed a 'worst case' attribute inference attack that assumes an attacker has access to a data record and the 'true' label, but that the value of one attribute for that record is missing.

- If the attribute is categorical, we then complete copies of the record with each possible value, and note the predicted label, and the confidence (predicted probability for that label). If there is a unique attribute value that leads to a prediction with the highest confidence, the attack model outputs that label as its prediction, otherwise, it does not make a prediction.
- If the attribute is continuous, then we estimate the lower and upper bounds on the prediction that an attacker could make as follows. First, we take *n* (e.g. 100) samples from across the range and use the target model to make predictions using those samples in place of the missing attribute value. Next, we record the lowest and highest values where the trained model has confidence equal to its maximum. Finally, we mark the record as 'at_risk' if the upper and lower bounds are both within k% of the true value for that attribute.
- Finally, for each attribute, we can then construct risk measures, and provide comparisons of these for the training set, and a previously unseen test set.

The plots below, Figure 20, show typical results from a random forest classifier trained on the Texas Hospital Mortality Data. Top plot shows for each continuous variable, the proportion of records that can be estimated within +/- 10% of their true value, separated by training and test data.

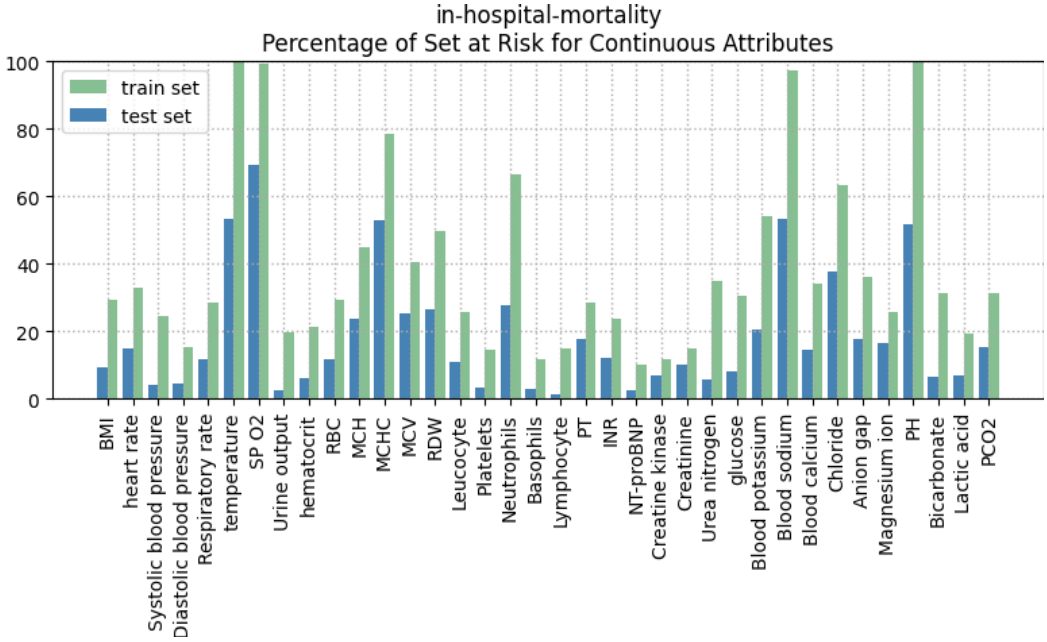


*Figure 20. Percentage of risk by continuous attributes*

The next plot, Figure 21, shows, for categorical variables, the proportion of records where the attack model would predict the missing value of an attribute.

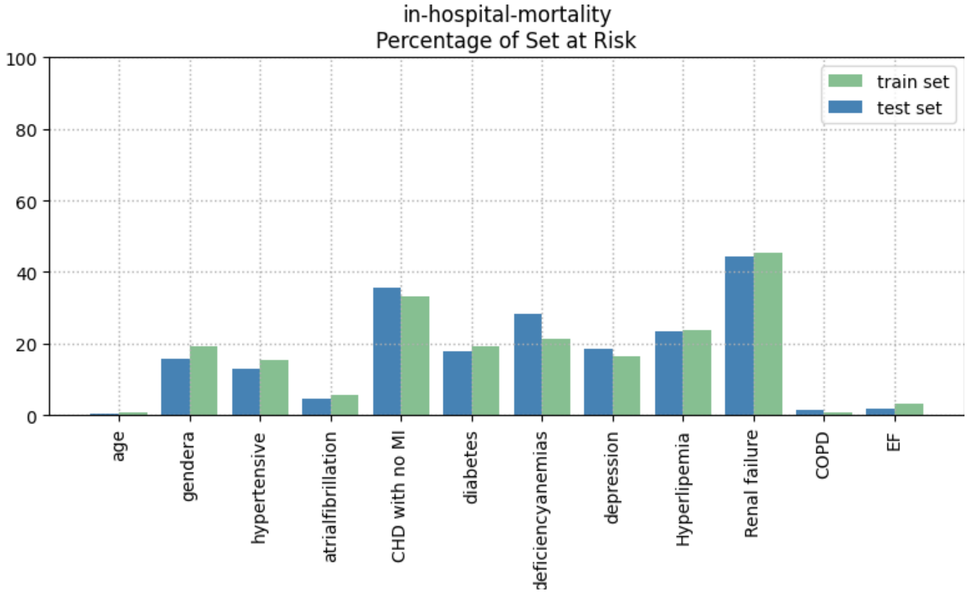


*Figure 21. Percentage of risk by categorical attributes*

The final plot, Figure 22, shows the improvement over a ‘most frequent value’ estimator achieved by the attack model, in the cases where it does make a prediction.

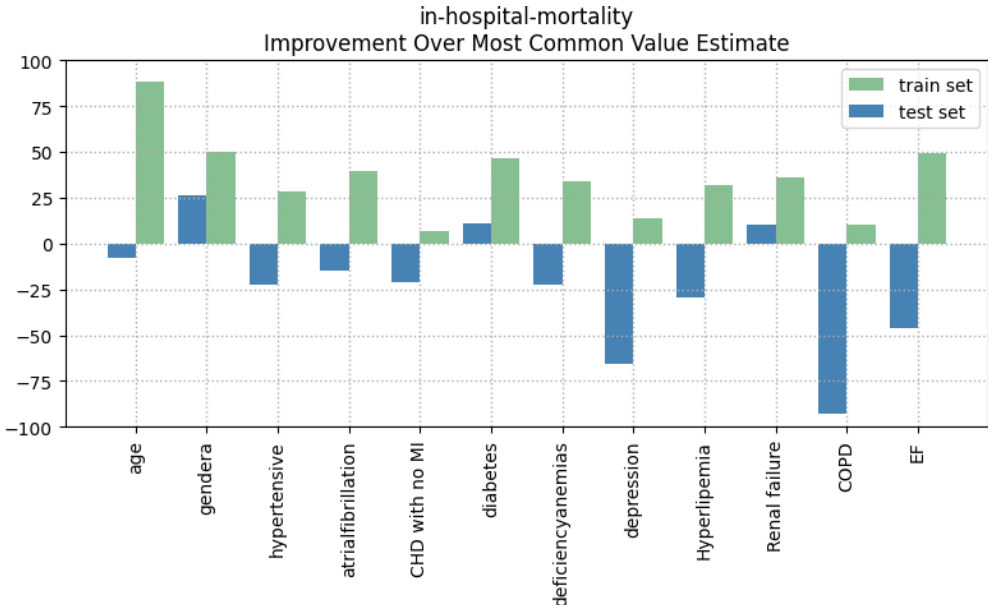


*Figure 22. Attack model improvement over most common values estimate*

As can be seen, these initial results demonstrate that Attribute Inference Attacks can pose a significant threat to some trained models.

The next phase is to explore the effect of different hyper-parameter and dataset choices as per the MI work.

## 17 Appendix B: Attack simulation tool kit for TREs staff to use

*Detailed information about how the tools were developed (which should in part be covered by Appendix A) and where to find the tool kit*

*A SafeModel wrapper class has been developed in python, and versions produced for the following ML model types:*

| *Model type* | *Base Implementation/Library* | *Notes* |
|---|---|---|
| *Decision Tree Classifier* | *Sklearn.trees.DecisionTreeClassifier* | *Model also reports k-anonymity value for tree* |
| *Random Forest* | *Sklearn.ensembles.RandomForestClassifier* | *Model also reports k-anonymity value for forest* |
| *k-Nearest Neighbours* | *N/A* | *Safe model rejects attempts to use this algorithm* |
| *Support Vector Classifier* | *Sklearn + bespoke* | *Implements Differential Privacy version on top of sklearn* |
| *Artificial Neural Networks* | *Tensorflow Keras functional model* | *Enforces choice of TensorFlow differentially private optimiser* |

*Considerable effort has been put into the additional functionality and making this transparent to researchers.*

*From a researcher's point of view, the principal differences are two extra functions: save_model() and request_release(). The former is self-explanatory, the latter is the most important, and is the function they call to trigger the TRE output checking process.*

*'Behind the scenes, the superclass does the following:*

- *The object constructor init() function checks whether the parameters supplied by the researcher (or the base implementation defaults) match the constraints held in the TRE's rules.json file.*
  - *If they do not, it changes them where possible*
  - *And warns the researcher that these changes have been made, and any other changes they need to make*
- *The fit() method calls the base implementation's fit() method to train the model. It then saves a 'snapshot' copy of the model's attributes at that stage*
  - *Where appropriate (e.g. Decision Tree and Random Forest Classifiers it also calculates the k-anonymity of the model*

*The request_release function does the following checks:*

- *Do the model's current parameters meet the constraints in the TRE rules.json?*
- *Have the model's parameters been altered (maliciously or otherwise) between training and a request for release (e.g., to hide an unsafe training regime)?*
- *Have any other critical parts of the mode been changed since the fit() method was called? ( for example the trees in a Decision Tree or Random forest, the support vectors in an SVM or the weight in a neural network).*
- *(if appropriate) Was the DP-variant of the optimiser used?*
- *(if appropriate) what are the differential privacy guarantees (e.g. epsilon values)*
- *(To be added): What is the result of running a worst-case Membership inference attack on this model?*
- *(To be added): What is the result of running a worst-case Attributeinference attack on this model?*
-

*The results of these tests are then collated into a human and machine-readable output file along with some recommendations. Below are some typical excerpts from respectively a 'safe' model, a model trained with unsafe hyper-parameters, and one trained unsafely then maliciously changed to look superficially 'safe':*

```
{
 "researcher": "j4-smith",
 "model_type": "RandomForestClassifier",
 "model_save_file": "testSaveRF.pkl",
 "details": "Model parameters are within recommended ranges.\n",
 "recommendation": "Run file testSaveRF.pkl through next step of checking procedure"
}
{
"researcher": "j4-smith",
 "model_type": "RandomForestClassifier",
 "model_save_file": "unsafe1.pkl",
 "details": "WARNING: model parameters may present a disclosure risk:\n- parameter bootstrap = False identified as different than the recommended fixed value of True.",
 "recommendation": "Do not allow release",
 "reason": "WARNING: model parameters may present a disclosure risk:\n- parameter bootstrap = False identified as different than the recommended fixed value of True."
}
```

```
{

 "researcher": "j4-smith",

 "model_type": "RandomForestClassifier",

 "model_save_file": "unsafe-malicious.pkl",

 "details": "Model parameters are within recommended ranges.\n",

 "recommendation": "Do not allow release",

 "reason": "Model parameters are within recommended ranges.\WARNING: basic parameters differ in 2 places:\nparameter bootstrap changed from False to True after the model was fitted\nparameter min_samples_leaf changed from 2 to 10 after the model was fitted\n"

}
```

*Reference a manuscript which will describe the method and results for an AI expert audience*

## 18 Appendix C: Legal and Ethical Background

Legal and Ethical Framework

## 19 Appendix D: Template wording for TRE Researcher Declaration Forms

*Will be incorporated in next iteration*

## 20 Appendix E: Template wording for end-users of trained models

*e.g. could be added to a software licence for a medical device*

## 21 Appendix F: Example data dictionary template

*This work is currently in progress.*

The following is an example of how information about the encoding of different features could be provided:

```
data:{

 0: {'name': 'parents', 'indices': [0, 1, 2], 'encoding': 'onehot'},
1: {'name': 'has_nurs', 'indices': [3, 4, 5, 6, 7], 'encoding': 'onehot'},

2: {'name': 'assessment_completed', 'indices': [8, 9, 10, 11], 'encoding': 'onehot'},

3: {'name': 'children', 'indices': [12, 13, 14, 15], 'encoding': 'onehot'},

4: {'name': 'atrialfibrillation', 'indices': [16], 'encoding': 'int64'}

...

}
```

## 22 Appendix G: Example constraints file

Below is a snippet from a file rules.json that contains examples of how the 'safe envelope' for algorithm parameters can be stored centrally as a set of constraints within a human and machine-readable file.{

```
 "DecisionTreeClassifier": {
  "rules": [
   {
    "keyword": "min_samples_leaf",
    "operator": "is_type",
    "value": "int"
   },
   {
    "keyword": "min_samples_leaf",
    "operator": "min",
    "value": 5
   }
  ]
 },
 "RandomForestClassifier": {
  "rules": [
   {
```

```
   "operator": "and",
   "subexpr": [
    {
     "keyword": "bootstrap",
     "operator": "equals",
     "value": true
    },
    {
     "keyword": "min_samples_leaf",
     "operator": "min",
     "value": 5
    }
   ]
  }
 ]
},
"SVC": {
 "rules": [
  {
   "keyword": "dhat",
   "operator": "min",
   "value": 1000
  },
  {
   "keyword": "C",
    "operator": "min",
    "value": 1
   },
   {
    "keyword": "eps",
    "operator": "min",
    "value": 10
   },
   {
    "keyword": "gamma",
    "operator": "min",
    "value": 0.1
   }
  ]
 },
....
```

# 23 Appendix H: Case studies

This section includes a few basic examples to illustrate some of the main risks that machine learning models can have regarding data privacy. Most of them use invented data, except for the hospital survival one that is based on a real anonymised publicly released data set. The data sets used are relatively small in most cases to keep things as simple as possible to explain the principle. The risks presented here apply to both small and large datasets. The three first examples are cases of disclosure that potentially could happen before the measures proposed in this document have been implemented. Whereas the last one, the proposed measures have been put in place and disclosure risk is identified and prevented.

The code for all of the examples below can be found in separate attached files.

## 23.1 Finding out unknown information about a famous person

Researchers from a hospital have created a model that is capable to predict the risk of cancer for people with multiple other diseases. Early diagnoses of cancer save lives. As it is normal for scientists, their methods and description of the data are published in a specialised journal. The model is publicly available and anyone with the right technical background can easily access it (see Figure 23).

When a doctor, for instance, asks the model about a new patient, it responds with a percentage representing the chances that the individual will suffer from cancer in the future. For example, a patient for whom the response is 24% is very unlikely to suffer from cancer, however, if a patient's response is much higher, e.g., 95%, the doctor can identify the individual as high risk.

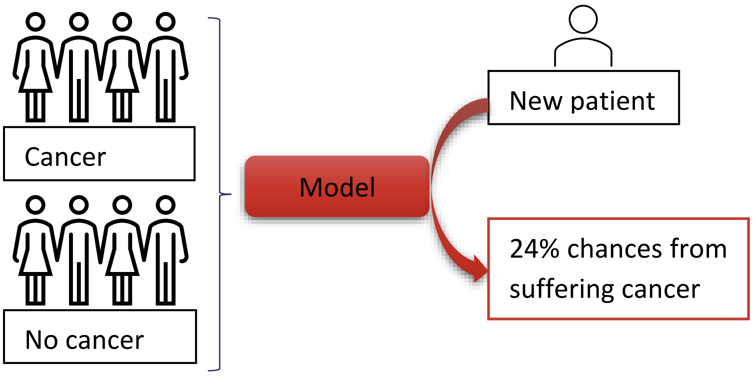


*Figure 23 Model to predict risk to suffer from cancer*

The data used to create the model included 200 patients, half of which suffered from cancer. Their names and address were removed from the data to ensure the preservation of their privacy. The data employed to create the model was kept in a safe environment and never made available to the public as it is illegal to do so. Table 1 shows example data.

The researchers didn't realise that the hospital where the data was collected is a hospital local to a famous Member of Parliament (MP). This MP is well known, and a quick search reveals a lot of his health information: you can easily discover that he **suffered from cancer, is diabetic, asthmatic, smokes,** and **is 62 years old**.

*Table 1 Data used to build the model for predicting risk to suffer from cancer*

| Diabetes | Asthma | Weight | Blood pressure | Smoker | Age |
|---|---|---|---|---|---|
| Yes | No | Obese | High | Yes | 72 |
| Yes | Yes | Normal | High | No | 83 |
| No | No | Obese | High | Yes | 63 |
| Yes | Yes | Obese | High | No | 77 |
| **Yes** | **Yes** | **Overweight** | **Slightly high** | **Yes** | **62** |
| Yes | No | Underweight | Normal | No | 50 |
| No | Yes | Normal | Slightly high | Yes | 66 |
| Yes | Yes | Normal | Slightly high | No | 44 |
| No | No | Normal | Slightly high | Yes | 61 |
| No | No | Normal | Slightly high | No | 56 |

Cancer

Famous MP

No cancer

Let's imagine you are a highly skilled technical person, who realised that potentially the MP's data could have been used to train the model and therefore it might be possible to *attack* the model to find out more about their health. First, you already discovered a good deal about the MP by doing an online search. You also know the possible values from data that you don't have: blood pressure (low, slightly high, high, very high, extremely high) and weight (underweight, normal, overweight, obese).

Based upon this information, you create all the possible combinations of the known and unknown data and come up with 20 possible options (based upon the fact that blood pressure can take on 5 different values, and weight 4). Next, you ask the model about each possible combination to get the model's response confidence.

It is these confidence values that form the basis of the attack. In particular, it can be the case that a model has higher confidence for examples that were in the training set, than examples that were not. Let's imagine the 5 responses with the highest confidence you got are presented in Table 2.

*Table 2 Model response with unknown variables*

| Weight | Blood pressure | Response confidence |
|---|---|---|
| **Overweight** | **High** | **93,8%** |
| Normal | Slightly high | 57,7% |
| Normal | High | 54,4% |
| Overweight | High | 54,2% |
| Obese | Slightly high | 54,2% |

You quickly realise that the first row has a much higher confidence response. This makes it very likely that these are indeed the true values for the MP and therefore you conclude that the MP BMI group is 3 (overweight) and the blood pressure group is 2 (slight high).

The researchers could have prevented this from happening had they used a different configuration of the model when they were training it.

## 23.2 Finding private information from publicly available data

Imagine you are part of our small research group working with diabetes. Our latest project is to generate a machine learning model capable to predict whether a given person suffers from type 2 diabetes, which is life-style related (not genetic). We are recruiting participants for our study, but unfortunately, we have a limited budget, and we can only pay for 20 individuals to take part. All the participants wish to keep their type-2-diabetes-status private.

We record their birthdays, whether they smoke or not, their diabetes status, and their HbA1C (a biochemical measure). HbA1C above 50 indicates the person suffers from type-2-diabetes. We put their data in a safe environment out of the public reach. Their data is as shown in Table 3.

*Table 3 Data to create the diabetes model*

| Name | Smoker | Birthday | HbA1C | Diabetes |
|---|---|---|---|---|
| Jose | Yes | 07-01-2001 | 83 | True |
| Nicholas | Yes | 10-06-1978 | 44 | False |
| Jesse | No | 03-10-1992 | 38 | False |
| William | No | 24-03-2000 | 17 | False |
| Joshua | No | 20-10-1975 | 58 | True |
| Kristen | Yes | 12-03-1972 | 47 | False |

We decide to build a very simple model, and set up a threshold for our diabetes indicator (HbA1C). The people in charge of checking the safety of the model do not detect anything risky in it, and we get all thumbs up. We decide we want to make it public as it is expected to improve lives.

After the model's release, it becomes obvious there's something odd about it. The model contains a coefficient to aid diabetes prediction. And when we look at it and put all the information we have together, we realise that each birthday corresponds to a coefficient. Also, this coefficient turns out to be the same as HbA1C values minus 50. The Table 4 below illustrates it:

*Table 4 Odd coefficient-HbA1c-birthday relationship*

| Name | Birthday | HbA1C | HbA1c – 50 | Coefficient |
|---|---|---|---|---|
| Kristen | 12-03-1972 | 47 | -3 | -3 |
| Adam | 01-07-1975 | 63 | 13 | 13 |
| Joshua | 20-10-1975 | 58 | 8 | 8 |
| Patrick | 10-11-1977 | 85 | 35 | 35 |
| Nicholas | 10-06-1978 | 44 | -6 | -6 |
| Kimberly | 03-01-1980 | 76 | 26 | 26 |

Therefore we now find out (too late) that anyone who knows a birthday can find out right away whether the person has diabetes or not and infer their exact HbA1C levels, both information is confidential. This is a violation of the privacy of the individuals in the study.

This might imply an investigation from the institution your research group belongs to, for determining if all the procedures were followed, and who is responsible for the breach. Also, it is possible to have legal consequences.

## 23.3 Identifying if someone famous has suffered from cancer

Scientists have created a model that predicts whether patients will respond well to a specific drug used to treat cancer for people with multiple diseases. Administering the appropriate treatment for cancer patients is crucial for recovery and the model will save lives. As it is normal for scientists, their methods and description of the data are published in a specialised journal. The model is publicly available and anyone with the right technical background can easily access it.

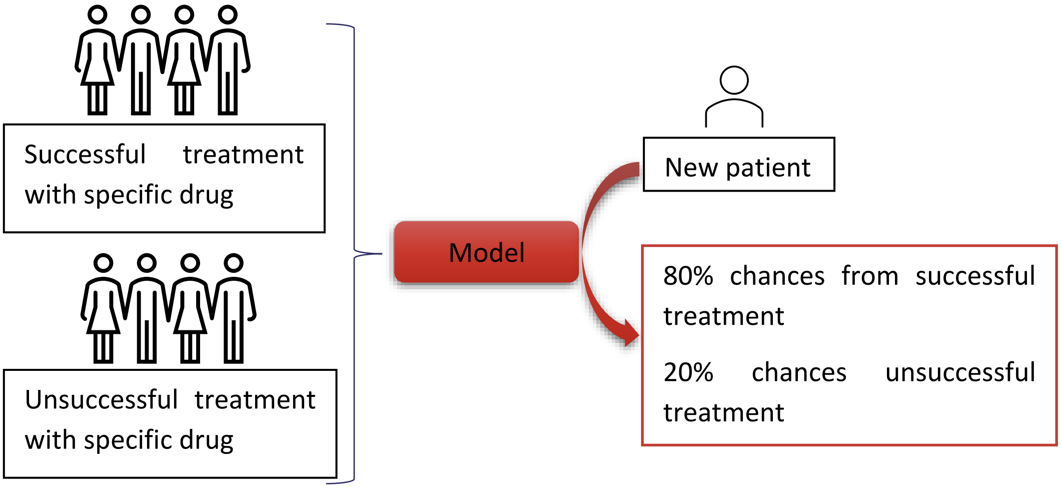

Successful treatment with specific drug
Unsuccessful treatment with specific drug
Model
New patient
80% chances from successful treatment
20% chances unsuccessful treatment

*Figure 24 Machine learning model to predict cancer treatment success*

When a doctor, for instance, asks the model about a new patient, it responds with two percentages indicating the chances of treatment with the specific drug being successful and not successful accordingly (see Figure 24). For example, a patient for whom the model response is 80% success and 20% non-success suggests it is highly likely that the specific drug will be successful, and so the patient can start treatment immediately. If the response is 10% success and 90% non-success it means that probably the treatment with the specific drug will not be successful, so doctors need to find another treatment that is likely to succeed better. This process is called personalised medicine.

*Table 5 Data to build the machine learning model to predict cancer treatment success*

| Diabetes | Asthma | Weight | Blood pressure | Smoker | Age |
|---|---|---|---|---|---|
| **Yes** | No | Obese | High | Yes | 72 |
| **Yes** | Yes | Normal | High | No | 83 |
| **No** | No | Obese | High | Yes | 63 |
| **Yes** | Yes | Obese | High | No | 77 |
| **Yes** | **Yes** | **Overweight** | **Slightly high** | **Yes** | **62** |
| **Yes** | No | Underweight | Normal | No | 50 |
| **No** | Yes | Normal | Slightly high | Yes | 66 |
| **Yes** | Yes | Normal | Slightly high | No | 44 |
| **No** | No | Normal | Slightly high | Yes | 61 |
| **No** | No | Normal | Slightly high | No | 56 |

Successful treatment

Famous MP

Unsuccessful treatment

The researchers didn’t realise that the hospital where the data was collected is the hospital used by a famous Member of Parliament (MP), as shown in Table 5. This MP is well known, and a quick only search reveals a lot of his health information: you discovered that is diabetic, asthmatic, overweight, has slightly high blood pressure, smokes, and is 62-year-old. What you don’t know is whether the MP suffered from cancer or not. All people in the study suffered from cancer, so if you were able to establish that the famous MP took part in the research study it would mean that he did indeed suffer from cancer – information that should not have been released.

Let's imagine you are a highly skilled technical person, who realised that potentially the MP data could be in the model. You also have read and understood the methods employed in their publication. There are several steps you need to follow to attempt to establish if the MP is in the model. First, you need to create some fake information about fictional patients. Obviously, it will not be identical to the original set, but it does not matter for your purpose. Next, you generate your drug success prediction model, as described in the publication, which is the same as the original, except that it is trained on some of your fake data.  Using this fake model, you make a series of predictions for treatment response using your fake data – some examples of which were used to train your model, and some not. These treatment response values for examples that were and were not used to train **your** model lets you see if your model gives different response values for examples that it saw in training compared to those it didn't. You see that it does, and conclude that, given that the model is very similar to the original, the original will too. Because you are a machine learning expert, you decide to use these responses to build another model that can predict, from these responses, whether an example was used to train your model or not. The process is illustrated in Figure 25.

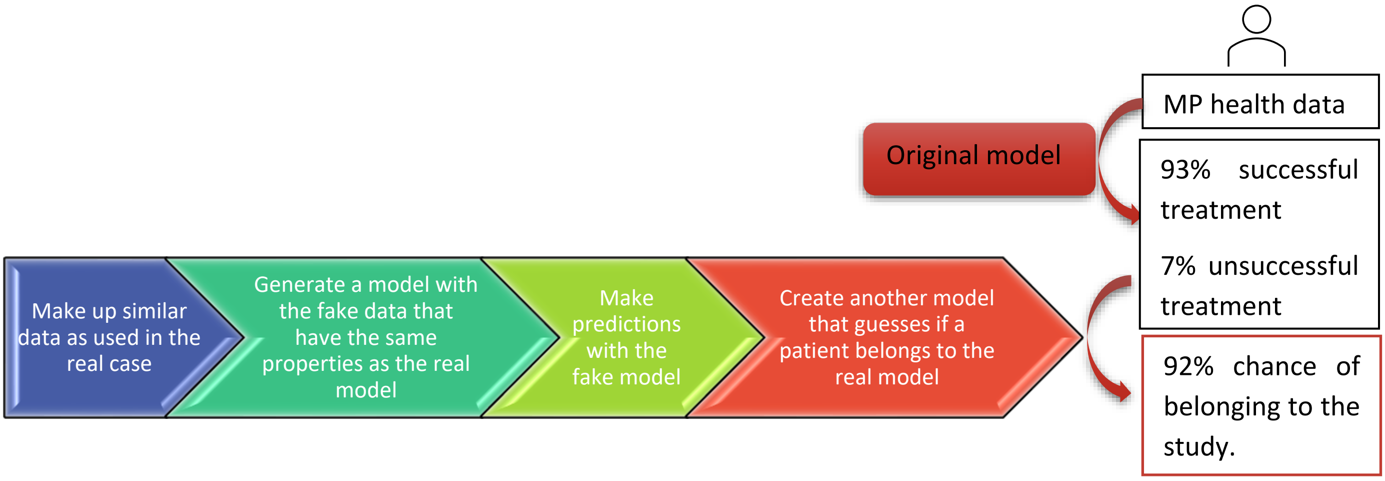

MP health data
Original model
93% successful treatment
7% unsuccessful treatment
Make up similar data as used in the real case
Generate a model with the fake data that have the same properties as the real model
Make predictions with the fake model
Create another model that guesses if a patient belongs to the real model
92% chance of belonging to the study.

*Figure 25 Process that an attacker needs to follow during an attack*

You are now ready to determine if your MP was in the data or not (and therefore learn if they had cancer or not). You use the original published model to make a prediction about the treatment response for the famous MP (from which you found online all the information you need) and ask the second model how likely this response is to belong to data used for training or not. If the answer is high, you are confident that the famous MP was in the training set and therefore suffered from cancer, but if it is low, you can only say he was not part of the study, but still don't know whether he suffered cancer or not.

## 23.4 Successful candidates in a job interview

Imagine you are a senior manager in a factory. You are currently recruiting people for the summer period, which is your busiest time of the year. In the past, you had bad experiences with some of the workers you had, and this time you want to ensure you recruit the best possible candidates. For the process, you rely on an agency to help you. They assure you that in the past they had very positive experiences from using a machine learning model that is capable to guess whether an interview candidate is a drug user (although this wasn't the intended primary objective of the model), and therefore you can discard them immediately.

So, what is this model that the recruiting agency is talking about? A while ago there was some research done to help drug users who were in serious financial difficulties. Their main goal was to improve the quality of life for drug users. 50 people took part in this study with the promise of their data being kept anonymous. The data recorded look like as shown in Table 6.

*Table 6 Data used to create a machine learning model to predict insolvency for drug users*

| Name | Age | Education | Sex | Number of previous convictions | Housing | Number of previous rehabilitations | Solvent |
|---|---|---|---|---|---|---|---|
| Michelle | 34 | Some secondary | Female | 2 | Other | 0 | No |
| Lindsay | 35 | Secondary | Female | 0 | Supported | 1 | No |
| James | 40 | University | Male | 2 | Other | 4 | No |
| Kenneth | 36 | University | Male | 3 | Other | 1 | No |
| Emily | 32 | None | Female | 1 | No fixed | 2 | Yes |
| Rebecca | 22 | Some secondary | Female | 0 | Supported | 0 | Yes |

Once the model was built it was decided to make it public and work with charities, as they are often reaching these local communities well and have expertise in helping them prevent extreme financial problems (see Figure 26). At some point afterwards, some of the participants made public that they were part of the study. Some people, including the recruiting agency that is helping you, found out. Also, they know other people who were not part of the study. Obviously, it is not acceptable to ask job interview candidates whether they are drug users or not. Instead, the agency cleverly uses all of the information they gathered previously.

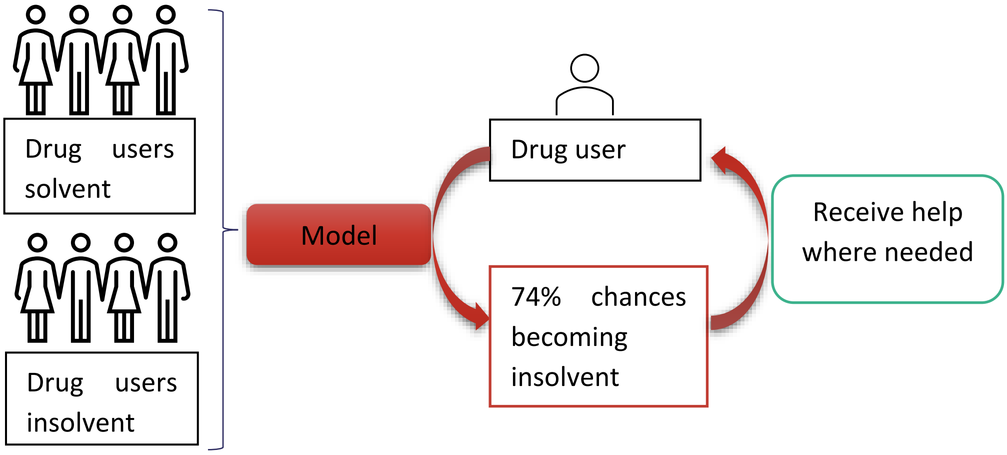


*Figure 26 Diagram of the machine learning model to predict insolvency for drug users*

They start by using the model to predict how likely the people whose information they have (the people who made it known they were in the study, and those that they know not to be in the study) are to have serious economic issues. Let's have a look at what they obtained (see Table 7).

*Table 7 Example of response obtained by the model to predict insolvency for drug users*

| Name | Probability insolvency | Part of the study |
|---|---|---|
| John | 75.0% | Yes |
| Ashley | 74.9% | Yes |
| Elizabeth | 25.0% | Yes |
| Angela | 75.0% | Yes |
| Tyler | 74.9% | Yes |
| Jason | 24.9% | Yes |
| Christina | 50.0% | No |
| Jose | 50.0% | No |
| Thomas | 50.1% | No |
| Brittany | 49.9% | No |
| Nicholas | 49.8% | No |
| Megan | 50.8% | No |

The recruiting agency immediately spot a pattern: participants in the study have values of either around 75% or 25%, whereas non-participants have values of around 50%. The agency can use this rule to screen candidates, and not call for a job interview any candidates who the model gives a probability of around 75% or 25% as they are likely to have been in the study, and are therefore likely to be drug users, see diagram of the selection process on Figure 27.

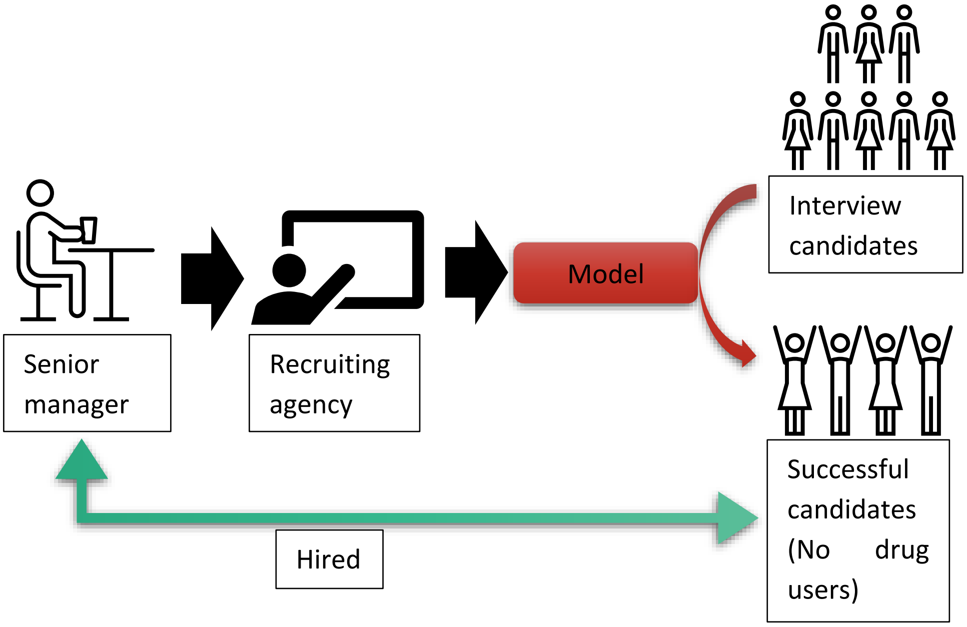


*Figure 27 Hiring process using a machine learning model*

In this case, the model violates the privacy of the people who volunteered to be part of it and did not want their information to be made public, and consequently, they are not managing to get jobs.

## 23.5 Hospital admission survival

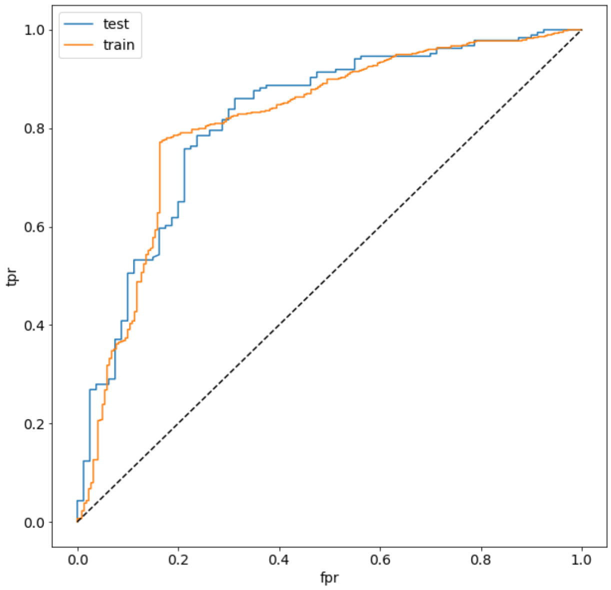


*Figure 28. ROC curve for hospital admission survival model*

Hospitals want to improve the care for admitted patients, by predicting the probability they have to survive. To do so a model was created from patients using a type of model called Super Vector Machine (SVM). This type of model belongs to "instance based" type of models, which are extremely popular in machine learning. They are powerful and capable of making very accurate predictions. Figure 28 shows the ROC curve.

However, to function, they need access to training data during training and when they are making predictions. The performance obtained is pretty good, as shown in the figure below. TPR is the True Positive Rate, which represents the proportion of correctly identified survival of patients versus the False Positive Rate (FPR), which represents the proportion of incorrectly identified survival. The further the curves deviate from the diagonal black line, the better. The level of performance here suggests that the model could be extremely useful.

This model was created in a safe environment out of public reach to ensure patient privacy. The creators of the model want to make it public as it would clearly benefit the population. To do so, the creators need to be able to take the model out of the safe environment.

The people in charge of the safe environment, check the model carefully to decide whether they will allow the model to be made public or not.

It was found that the model contained 441 exact copies of the original patient information out of the 798 included. They spoke with the creators and asked them to remove this private patient data from the model. The researchers explained that it is not possible as the model will not work without it. The safety staff gave them three options: to create another model using non-instance-based methods (if they still want it to be public), not to release it (possibly deploying it within the safe environment), or re-training the model using a different training algorithm that allows patient privacy to be preserved.

## 24 References

[1] M. Veale, R. Binns, and L. Edwards, ‘Algorithms that remember: model inversion attacks and data protection law’, *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 376, no. 2133, p. 20180083, Nov. 2018, doi: 10.1098/rsta.2018.0083.

[2] F. Ritchie, ‘The “Five Safes”: a framework for planning, designing and evaluating data access solutions’, Zenodo, Sep. 2017. doi: 10.5281/zenodo.897821.

[3] ‘Health Data Research UK publishes “Recommendations for Data Standards in Health Research”’, *HDR UK*. https://www.hdruk.ac.uk/news/health-data-research-uk-publishes-recommendations-for-data-standards-in-health-research/ (accessed Jun. 07, 2022).

[4] T. Hubbard, G. Reilly, S. Varma, and D. Seymour, ‘Trusted Research Environments (TRE) Green Paper’, Zenodo, Jul. 2020. doi: 10.5281/zenodo.4594704.

[5] Scottish Government, ‘Charter for Safe Havens in Scotland: Handling Unconsented Data from National Health Service Patient Records to Support Research and Statistics.’ http://www.gov.scot/publications/charter-safe-havens-scotland-handling-unconsented-data-national-health-service-patient-records-support-research-statistics/ (accessed Nov. 08, 2021).

[6] M. Sperrin, D. Jenkins, G. P. Martin, and N. Peek, ‘Explicit causal reasoning is needed to prevent prognostic models being victims of their own success’, *J. Am. Med. Inform. Assoc.*, vol. 26, no. 12, pp. 1675–1676, Dec. 2019, doi: 10.1093/jamia/ocz197.

[7] ‘Data Protection Act 2018’. https://www.legislation.gov.uk/ukpga/2018/12/section/66/enacted (accessed Jun. 01, 2022).

[8] D. Bzdok, N. Altman, and M. Krzywinski, ‘Statistics versus machine learning’, *Nat. Methods*, vol. 15, no. 4, Art. no. 4, Apr. 2018, doi: 10.1038/nmeth.4642.

[9] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft, ‘Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning’. arXiv, Nov. 30, 2009. Accessed: May 21, 2022. [Online]. Available: http://arxiv.org/abs/0911.5708

[10] L. Breiman, ‘Random Forests’, *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[11] T. Chen and C. Guestrin, ‘XGBoost: A Scalable Tree Boosting System’, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[12] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, ‘Super learner’, *Stat. Appl. Genet. Mol. Biol.*, vol. 6, p. Article25, 2007, doi: 10.2202/1544-6115.1309.

[13] C. Dwork and A. Roth, *The algorithmic foundations of differential privacy*. Boston, MA Delft: Now, 2014.

[14] ‘End-User License’, *Responsible AI Licenses (RAIL)*. https://www.licenses.ai/enduser-license (accessed Jul. 21, 2022).

[15] ‘Design choices for productive, secure, data-intensive research at scale in the cloud’, *The Alan Turing Institute.* https://www.turing.ac.uk/research/publications/design-choices-productive-secure-data-intensive-research-scale-cloud (accessed Nov. 10, 2021).

[16] S. Kavianpour, J. Sutherland, E. Mansouri-Benssassi, N. Coull, and E. Jefferson, ‘A review of Trusted Research Environments to support next generation capabilities based on interview analysis’, *J. Med. Internet Res.*, May 2022.

[17] ‘Survey: Leakage and Privacy at Inference Time’, *DeepAI*, Jul. 04, 2021. https://deepai.org/publication/survey-leakage-and-privacy-at-inference-time (accessed Jun. 10, 2022).

[18] Mansouri-Benssassi E., Krueger S., Ritchie F., Smith J, ‘Work Session on Statistical Data Confidentiality 2021 - Work Session on Statistical Data Confidentiality 2021 - UNECE Statswiki’. https://statswiki.unece.org/display/confid/Work+Session+on+Statistical+Data+Confidentiality+2021?preview=/314934659/330368628/SDC%20for%20ML%20-%20paper%20outline_UNECE%20format%20v2.pdf (accessed Nov. 11, 2021).

[19] E. Mansouri-Benssassi *et al.*, ‘Machine Learning Models Disclosure from Trusted Research Environments (TRE), Challenges and Opportunities’, arXiv, arXiv:2111.05628, Nov. 2021. doi: 10.48550/arXiv.2111.05628.

[20] ‘How should we assess security and data minimisation in AI?’, Nov. 22, 2021. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/ (accessed Jul. 21, 2022).

[21] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, and L. Sweeney, ‘Privacy Preserving Synthetic Data Release Using Deep Learning’, in *Machine Learning and Knowledge Discovery in Databases*, Cham, 2019, pp. 510–526. doi: 10.1007/978-3-030-10925-7_31.

[22] K. E. Emam, L. Mosquera, and J. Bass, ‘Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation’, *J. Med. Internet Res.*, vol. 22, no. 11, p. e23139, Nov. 2020, doi: 10.2196/23139.

[23] Liley, J. Bohner, G. Emerson, S *et al.*, ‘Development and assessment of a machine learning tool for predicting emergency admission in Scotland’, *medRxiv*, Aug. 2021, doi: 10.1101/2021.08.06.21261593.

[24] J. Jordon, J. Yoon, and M. van der Schaar, ‘PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees’, *Int. Conf. Learn. Represent.*, Sep. 2018, Accessed: Nov. 08, 2021. [Online]. Available: https://openreview.net/forum?id=S1zk9iRqF7

[25] E. De Cristofaro, ‘A Critical Overview of Privacy in Machine Learning’. https://emilianodc.com/PAPERS/IEEESP2021.pdf (accessed Nov. 08, 2021).

[26] J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia, ‘The limits of differential privacy (and its misuse in data release and machine learning)’, *Commun. ACM*, vol. 64, no. 7, pp. 33–35, Jun. 2021, doi: 10.1145/3433638.

[27] A. Blanco-Justicia, D. Sánchez, J. Domingo-Ferrer, and K. Muralidhar, ‘A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning’, *ACM Comput. Surv.*, Jul. 2022, doi: 10.1145/3547139.

[28] S. Gambs, F. Ladouceur, A. Laurent, and A. Roy-Gaumond, ‘Growing synthetic data through differentially-private vine copulas’, *Proc. Priv. Enhancing Technol.*, vol. 2021, no. 3, pp. 122–141, Jul. 2021, doi: 10.2478/popets-2021-0040.

[29] T. Chanyaswad, C. Liu, and P. Mittal, ‘RON-Gauss: Enhancing Utility in Non-Interactive Private Data Release’, arXiv, arXiv:1709.00054, Oct. 2018. doi: 10.48550/arXiv.1709.00054.

[30] M. E. Gursoy, A. Inan, M. E. Nergiz, and Y. Saygin, ‘Differentially private nearest neighbor classification’, *Data Min. Knowl. Discov.*, vol. 31, no. 5, pp. 1544–1575, Sep. 2017, doi: 10.1007/s10618-017-0532-z.

[31] E. M. Weitzenboeck, P. Lison, M. Cyndecka, and M. Langford, ‘The GDPR and unstructured data: is anonymization possible?’, *Int. Data Priv. Law*, p. ipac008, Mar. 2022, doi: 10.1093/idpl/ipac008.

[32] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, ‘Inverting Gradients - How easy is it to break privacy in federated learning?’, in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 16937–16947. Accessed: Jul. 21, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html

[33] A. Salem, A. Bhattacharyya, M. Backes, M. Fritz, and Y. Zhang, *Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning*. 2019.

[34] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, ‘ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models’, arXiv, arXiv:1806.01246, Dec. 2018. doi: 10.48550/arXiv.1806.01246.

[35] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, ‘Membership Inference Attacks against Machine Learning Models’, arXiv, arXiv:1610.05820, Mar. 2017. doi: 10.48550/arXiv.1610.05820.

[36] S. Kavianpour, J. Sutherland, E. Mansouri-Benssassi, N. Coull, and E. R. Jefferson, ‘A Review of Trusted Research Environments to Support Next Generation Capabilities based on Interview Analysis’, *JMIR Prepr.*, Sep. 2021, doi: https://doi.org/10.2196/preprints.33720.

[37] K. Alves and F. Ritchie, ‘Runners, repeaters, strangers and aliens: Operationalising efficient output disclosure control’, *Stat. J. IAOS*, vol. 36, pp. 1–13, Sep. 2020, doi: 10.3233/SJI-200661.

[38] B. Jayaraman and D. Evans, ‘Evaluating differentially private machine learning in practice’, in *Proceedings of the 28th USENIX Conference on Security Symposium*, USA, Aug. 2019, pp. 1895–1912.

[39] UNESCO, ‘Draft Recommendation on the Ethics of Artificial Intelligence - UNESCO Digital Library’. https://unesdoc.unesco.org/ark:/48223/pf0000378931?posInSet=13&queryId=f4082765-2f1f-4710-a706-047db14472d1-draft-data-297 (accessed Nov. 08, 2021).

[40] 'The roadmap to an effective AI assurance ecosystem', *GOV.UK*. https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem (accessed Jun. 01, 2022).

[41] 'What is our new Algorithmic Transparency Standard? - Data in government'. https://dataingovernment.blog.gov.uk/2021/11/29/what-is-our-new-algorithmic-transparency-standard/ (accessed Jun. 01, 2022).

[42] 'Algorithmic transparency template', *GOV.UK*. https://www.gov.uk/government/publications/algorithmic-transparency-template (accessed Jun. 01, 2022).

[43] 'Understanding artificial intelligence ethics and safety', *The Alan Turing Institute*. https://www.turing.ac.uk/research/publications/understanding-artificial-intelligence-ethics-and-safety (accessed Jun. 01, 2022).

[44] 'EUR-Lex - 32016R0679 - EN - EUR-Lex'. https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed Jun. 06, 2022).

[45] 'European Union (Withdrawal) Act 2018'. https://www.legislation.gov.uk/ukpga/2018/16/section/1/enacted (accessed Jun. 09, 2022).

[46] 'AI and data protection risk toolkit', May 26, 2022. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/ (accessed Jul. 21, 2022).

[47] 'Algorithmic impact assessment: user guide'. https://www.adalovelaceinstitute.org/resource/aia-user-guide/ (accessed Jul. 21, 2022).

[48] R. C. : 881 N.W.2d 749, 'State v. Loomis'. https://harvardlawreview.org/2017/03/state-v-loomis/ (accessed Jul. 21, 2022).

[49] S. Kavianpour, J. Sutherland, E. Mansouri-Benssassi, N. Coull, and E. Jefferson, 'A review of trusted research environments to support next generation capabilities based on interview analysis', *Rev. Trust. Res. Environ. Support Gener. Capab. Based Interview Anal.*, Sep. 2021, Accessed: May 21, 2022. [Online]. Available: https://preprints.jmir.org/preprint/33720

[50] 'Creative Commons — Attribution 4.0 International — CC BY 4.0'. https://creativecommons.org/licenses/by/4.0/ (accessed Jun. 09, 2022).

[51] T. S. Court, 'Unwired Planet International Ltd and another (Respondents) v Huawei Technologies (UK) Co Ltd and another (Appellants) - The Supreme Court'. https://www.supremecourt.uk/cases/uksc-2018-0214.html (accessed Jun. 13, 2022).

[52] S. Rezaei and X. Liu, 'On the Difficulty of Membership Inference Attacks', arXiv, arXiv:2005.13702, Mar. 2021. doi: 10.48550/arXiv.2005.13702.

[53] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, 'Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting', arXiv, arXiv:1709.01604, May 2018. doi: 10.48550/arXiv.1709.01604.