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

GRAIMATTER Public Summary

September 2022

Summary

Artificial intelligence and machine learning have the potential to provide substantial benefits to public health services. However, these tools are often trained on sensitive personal data, which could create privacy risks. This project studied the risks and developed methods and tools to provide reassurance that the confidentiality of the data is maintained.

Throughout this project members of the public were invited to share their thoughts and concerns which have helped the research team in shaping the recommendations.

This work was funded by UK Research and Innovation as part of the DARE UK (Data and Analytics Research Environments UK) programme. The specific project was Guidelines and Resources for AI Model Access from TrusTEd Research environments (GRAIMATTER).

Glossary of key terms

Machine learning (ML)	Training a computer or machine to perform complex tasks in a way that is similar to how humans solve problems. ML is just one type of Artificial Intelligence.
Trusted Research Environment (TRE)	Trusted Research Environments (TREs) are highly secure computing environments that provide (remote) access to data for approved research.
Output checking	This is the process by which TRE staff check files which researchers would like to export from the TRE to ensure that they do not contain any potentially identifiable data.
Statistical disclosure control	Steps taken with data to eliminate (or reduce) the risk of disclosing information about a person from the data.

Context: Machine Learning and Secure Environments

Artificial intelligence and machine learning in health

Artificial intelligence is increasingly being used to help and support a range of public health operations and outcomes, for example in helping doctors make diagnoses. This can improve the efficiency of health services as well as deliver new services.

Many Artificial intelligence solutions involve 'Machine Learning' (ML): training a computer or machine to perform complex tasks in a way that is similar to how humans solve problems (see Box 1). The resulting 'trained ML model' can make predictions when provided with a new example or new situation. 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.

Because ML models 'learn' from examples, they need a very large amount of data to learn how to make effective predictions and reduce the number of situations they don't 'recognise'. In many scenarios, such as health care, the training data is likely to be personal and sensitive, and the best practice is to hold and analyse this data within a 'TRE'.

Box 1: How machine learning works

ML is different from the traditional model of computing, where a programmer decides what needs to be done and writes the code to achieve it. In ML, the computer is given a large amount of data, some general rules and then 'learns' how to make decisions, with limited human direction.

Consider the images above. An ML programme would be fed with many similar images labelled 'dog' or 'muffin' and would develop its own rules for deciding which is which. The trained model could then make its own predictions:

Dog or Muffin?

(Image credit @teenybiscuit)

Trusted Research environments (TREs) and output checking

A 'trusted research environment' allows researchers to work on and develop ML models using highly sensitive data, confident that the data never leaves this secure environment in an uncontrolled manner (see Box 2). As part of the TRE process, all statistical results undergo 'output checking': manually reviewing releases to ensure that there is no possibility of any confidential data being accidentally disclosed – this is 'disclosure control' and is well-understood in traditional statistical fields.

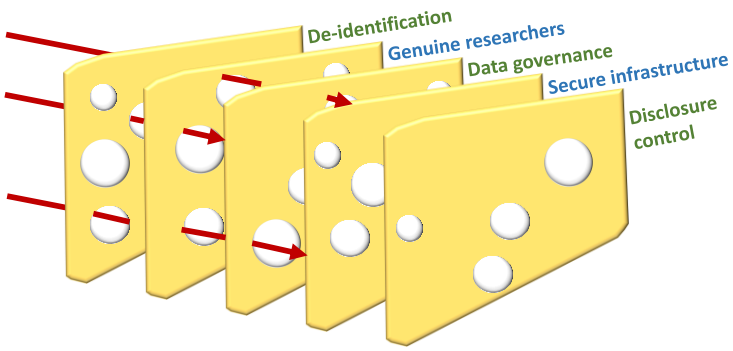
ML models also need to be checked when released from TREs, but they create substantial challenges for the traditional output-checking ML model. Unlike statistics, ML models

- are too complex (even the 'simple' ones) for humans to read and understand
- cannot be classified as 'safe to release' or 'unsafe' simply by looking at the output
- offer more ways for attackers to re-engineer the ML models and reveal personal data
- may allow researchers to include confidential information deliberately or accidentally in the output

The combination of growing demand, growing potential benefits, and significant confidentiality challenges create a need to develop output-checking solutions specifically targeted at ML models.

Box 2: Trusted Research Environments

TREs have multiple controls to ensure that data use is secure and ethical. One way to think about it is the 'Swiss cheese' model developed by the Scottish Safe Havens:



The TRE reduces detail in the data (de-identification), ensures that researchers are trained and trustworthy, has governance procedures to ensure use is ethical, operates a restricted computer facility, and checks all outputs before release. No control is perfect, but together they provide a secure and reliable research environment.

Research shows that some TREs do not allow the use of ML models in their secure environment because they are unsure how to handle them.

Our project (named GRAIMATTER) carried out the first detailed review of the topic. The goal was to develop recommendations that operators of TREs, researchers, data governance and ethics committees can follow. We considered technical, ethical and legal concerns, and also had a public engagement panel to help us consider how the wider public views our findings.

Whilst we focused on TREs, our recommendations also apply to non-TRE environments developing ML models on sensitive data. Our recommendations can be seen as 'best practice'.

We assumed throughout that all the other elements of 'best practice' within TREs are being followed, so our analysis and recommendations are i*n addition to* regular TRE operating practice.

The detailed recommendations green paper can be found accessed here: 10.5281/zenodo.7089491

This document provides a lay summary of the challenge and the recommendations.

What are the risks associated with machine learning?

We assessed the risks by assuming that there is an 'attacker' who wishes to extract confidential personal information from the ML model. While this is not a likely attack scenario in practice, it is useful for us to consider this as a 'worst case' scenario – someone deliberately setting out to 'break' the ML model. Successful protection against a well-prepared, malicious attacker provides some confidence that our checks and adjustments to the ML model are effective.

Box 3: Attack types

A 'black box' attack is where the attacker can only send queries to the model and hope to infer something from the answers. For example, the attacker might ask "Has Mr AA Hancock aged 32 from 27 Railway Cuttings been treated for stroke?" and try to infer from that whether Mr Hancock is in the survey.

A 'white box' attack uses extra information on the way the model was setup to refine the attack. For example, in the above case the attacker might know that the model was set up to emphasise the detection of strokes. A white box attack should have a much higher chance of success.

We considered 'risk scenarios'; situations which might lead to a successful attack. We identified three risk scenarios:

- A naïve researcher mistakenly includes inappropriate information in the ML model, unaware of the possibility of an attack
- A malicious researcher deliberately tries to hide data or other information in the ML model
- An external attacker outside the TRE uses the ML model (and possibly published information about the ML model) to attack it

There are two main types of attacks (see Box 3): a 'white box' attack where the attacker has detailed information about the ML model, and a 'black box' attack where the attacker can just ask questions about the ML model and guess whether this relates to real people.

Finally, we considered two types of disclosure risk:

- Was a specific person in the dataset used to train the ML model ('membership inference')?
- If a specific person was in the dataset, can we find out some additional information ('attribute inference')?

Box 4: Query versus identity controls

We broadly assumed that the ML models would be released from the TRE, and hence we need to have 'identity controls' applied to ensure personal data can't be re-engineered. The model itself is checked to ensure it is anonymous.

However, another possibility is not to release the model at all, but just to allow users to send in queries. This 'query control' stops white box attacks, and can limit the chance of black box attacks, as the query service can check for suspicious activity (such as unusual characteristics being requested, or multiple requests with very similar data). In this scenario the results returned from the model are anonymous.

We focused on membership attacks as these are key. Knowing that someone was in a dataset may be a problem in itself (if, for example, the dataset concerned people with an embarrassing or stigmatising illness); but it also allows for inference attacks ("We know Mr Hancock is in the dataset; what else can we find out or infer about him?"). Without a successful membership attack, you need to rely upon other information to carry out an inference attack (for example, knowing that the data includes every patient in Dundee hospital in June 2021, and your uncle was an inpatient then).

So, to build an attack ML model we considered:

- What scenarios might lead to an attack risk?
- What attacks might be feasible in this scenario?
- What might such an attack uncover?

There are many different types of ML models, and many different types of data – pictures are very different from clinical records, for example. To try to keep this manageable, we focused on popular ML models and data types, and then considered the different scenario-attack-disclosure combinations applicable in these cases. This gave us a measure of the risks involved - in a wide but not exhaustive set of cases, but enough for us to feel confident about our findings. We then considered ways that the ML model might be adjusted to protect against any risks identified.

What did we find?

We considered four aspects of the risk problem

- Can we provide assessments of the risk? That is, given an ML model, we want to be able to run tests on it and conclude something like "There is a 40% chance that *this* sort of attack on

that sort of ML model in *these* circumstances would lead to a successful re-identification of someone in the training data"

- Can we suggest ways to make the ML model safer? For example, can some values be omitted from the ML model in a way which doesn't affect the predictive power but does protect the source data?
- Can we suggest ways that the ML model could be built differently to make it inherently safe?
- Does the ML model need to be released at all? What happens if you only allow users to ask queries about the ML model, but don't share it? See Box 4.

For the first, the answer was yes, but safe thresholds vary depending on the specific data set and the type of model trained. The tests are sensitive to the ML models, risk and datasets under review so that the tests require expert interpretation – and multiple tests are needed to come to a judgement whether a given ML model is safe or not. This is nevertheless a great step forward, as we now have the tests and know how to apply them – and as we run tests on more general cases, it might be that general rules will start to emerge. TRE operators can run these tests using data set aside from the data used to train the ML model. We have shared the tests developed by GRAIMATTER for TREs to adopt and extend for their needs.

We explored ways that ML models could be built inherently safely. For example, a technique called 'differential privacy' could be used strategically to protect the data by design. The team also developed a set of 'safe wrappers' – instructions that researchers incorporate into their own code so that inherently safe ML models are produced. However, these solutions can't guarantee that a ML model will be safe, and so the tests described above still have to be run (albeit with a higher chance of being passed).

We have therefore shown that it is possible to generate useful measures of the riskiness of ML models, as well as showing how researchers/TREs can develop safer ML models or working practices to embed privacy by design into ML modelling. However, at present, this does need expert input in the disclosure checking process. Because of this, we have not made strict recommendations about which control measures TREs should apply; we expect that TREs will find the combinations of checks and controls that best meet their needs and skill sets, at least for now.

What about the ethical and legal implications?

ML models are likely to have a more complicated ethical approval process than traditional analyses: the potential uses of the trained ML model are wide, and the ML model may also be transferred to other parties. For example, imagine a researcher being commissioned to develop a triaging ML model for use in A&E departments. The expectation is that once the ML model has been developed, the health board would apply it in its facilities, and possibly share it with other health boards.

These onward uses need to be considered at the ethical approval stage, as does the release mechanism: the ML model may be distributed per se after the identity controls have been applied, or it might be released only through a query server; or it could be that the ML model feeds into other TREs in a federated system, in which case the ML model may not need to undergo identity controls until it is finally released into the wild. These legal and ethical questions need to be raised and decided upon at the approvals stage, and we reflected on how such approvals processes may need to change to reflect ML modelling; see Box 5.

Modelling in a TRE can provide a helpful additional review point. When anything is released from a TRE, it goes through a formal output checking process. Along the technical checks described above, this checking process could include, for example, a review to ensure that the intended use is consistent

Box 5: Who has an interest in this?

Many different groups are involved in the effective and ethical use of ML models:

- *Data holders* want their data to be used efficiently and securely
- *TREs/research organisations* want to be certain that their processes support safe use
- *Ethical review committees* want to ensure that specific research projects are lawful
- *Researchers* want to be able to explore important issues
- *Members of the general public* want data to be used for the public good

And all of these want data use to be ethical, balancing public benefit against privacy risks.

with the use originally approved. The team has analysed the options and suggested ways that TREs could usefully exploit this review stage to ensure that projects are consistent with their design goals. We have drafted some template forms that could be used by TREs and researchers.

Throughout the project, we have worked with a public panel. This has helped us to understand what was important to the public, their thoughts on questions such as "what is the acceptable level of risk?" (as all judgments on release will have an element of subjectivity), and "how should we balance risks with benefits?".

What about training and costs?

Our technical and ethical/legal recommendations have a lot of indirect consequences: for example, we make several specific recommendations about training for ethical committees, researchers, and TRE staff. Our recommendations also have cost implications: for example, the assessment of an output as suitable for release needs someone with substantial experience in ML, and so the TRE needs to either have this experience in-house or a contract to buy-in the expertise as necessary. Output checking on ML models will be more expensive than for traditional analysis, for the foreseeable future, and we have tried to indicate to TREs where they should be planning for these additional costs.

What happens next?

The GRAIMATTER project was an 8-month activity. It generated significant evidence on how to reduce the risks of personal data leaks from trained ML models, supporting TREs to develop the capability to safely support such ML projects. However, this project just "scratched the surface" of this new field.

For these recommendations to be widely adopted, significant additional research and community engagement are required. In the future, we would like to work across many TREs and ML projects to test the recommendations in practice. We would like to work with the industry and the wider academic community to further understand their requirements. Such testing and community engagement will help us find general rules across real-world projects, and understand how the recommendations might be made more efficient in practice to refine and update the recommendations.

We would like to develop the training materials to upskill researchers, TRE staff, data governance and ethics committees with the required expertise. We have identified many technical areas which require further investigations such as imaging data, genetic data, transfer learning and federated learning. We would also like to expand the software we have developed to support researchers and TREs to utilise safe wrappers and run attack simulation tests.

We would like to work with public representatives to further understand their opinions on what they consider to be disclosive as well as develop a public education and outreach programme. There is significant additional research on the legal and ethical implications which we would like to undertake including the drafting of legal templates to support a range of use cases.

ML has the potential to transform processes making them more efficient and safer, benefiting society. However, the training and deployment of ML models is in its relative infancy and access to sensitive data for training has been limited. GRAIMATTER has developed recommendations which will support TREs to develop the capability to securely support ML projects, forging the pathway to enable the training of ML models at scale on relevant data.