Enabling access, erasure, and rectification rights in AI systems

Reuben Binns, our Research Fellow in Artificial Intelligence (AI), discusses the challenges organisations may face in implementing mechanisms in AI systems that allow data subjects to exercise their rights of access, rectification and erasure.

15 October 2019

This post is part of our ongoing Call for Input on developing the ICO framework for auditing AI. We encourage you to share your views by emailing us at AIAuditingFramework@ico.org.uk.

Under the General Data Protection Regulation (GDPR) individuals have a number of rights relating to their personal data. These rights apply to personal data used at the various points in the development and deployment lifecycle of an AI system, including personal data:

- contained in the training data;
- used to make a prediction during deployment; or
- that might be contained in the model itself.

This blog post describes the considerations organisations may encounter when attempting to comply with three specific rights – access, rectification and erasure - in relation to AI systems, and where exemptions may apply.

Rights relating to training data

Organisations that create machine learning (ML) models will invariably need to obtain data to train those models.

For instance, a retailer creating a model to predict consumer purchases based on past transactions will need a large dataset of customer transactions on which to train the model.

A potential challenge for fulfilling individuals' rights is the difficulty involved in identifying the individuals the training data relates to.

Right of access

Typically, training data only includes information relevant to predictions, such as past transactions, demographics, or location, but not contact details or unique customer identifiers. Training data is also typically subjected to various 'pre-processing' measures to make it more amenable to ML algorithms.

For instance, a detailed timeline of a customer's purchases might be transformed into a summary of peaks and troughs in their transaction history.

This means training data can be much harder to link to a particular individual. However, in relation to data protection law this cannot be considered in itself to be pseudonymisation or anonymization, and the data must still be considered when responding to individuals' requests under the GDPR.

However, even if it lacks associated identifiers or contact details, and has been transformed through pre-processing, training data may still be considered personal data, because it can be used to 'single out' the individual it relates to, on its own or in combination with other data (even if it cannot be associated with a customer's name).

For instance, the training data in a purchase prediction model might include a pattern of purchases unique to one customer.

In this example, if a customer were to provide a list of their recent purchases as part of their request, the organisation may be able to identify the portion of the training data that relates to that individual.

In these kinds of circumstances, the organisation is obliged to respond to a data subject's request, assuming they have taken reasonable measures to verify the identity of the data subject, and no other exceptions apply.

Requests for access, rectification or erasure of training data should not be regarded as manifestly unfounded or excessive just because they may be harder to fulfil or the motivation for requesting them may be unclear in comparison to other access requests an organisation typically receives.

Organisations do not have to collect or maintain additional personal data to enable identification of data subjects in training data for the sole purposes of complying with the regulation (as per Article 11 of the GDPR). There may be times, therefore, when the organisation is not able to identify the data subject in the training data (and the data subject cannot provide additional information that would enable their identification), and therefore cannot fulfil a request.

Right to rectification

The right to correct inaccurate data may also apply to training data. However, the purpose of training data is to train models based on general patterns in large datasets, so individual inaccuracies are less likely to have any direct effect on an individual data subject. Organisations should therefore prioritise the rectification of personal data that might be used to take action in relation to an individual, over training data whose accuracy at an individual level is less likely to affect the individual.

Returning to our example, it may be more important to rectify an incorrectly recorded customer delivery address than to rectify the same incorrect address in training data. This is because the former could result in a failed delivery but the latter would barely affect the overall accuracy of the model.

Right to erasure

Organisations may also receive requests for erasure of training data. Organisations must respond to requests for erasure, unless a relevant exemption applies and provided the data subject has appropriate grounds. For example, if the training data is no longer needed because the ML model has already been trained, the organisation must fulfil the request. However in some cases, where the development of the system is ongoing, it may still be necessary to retain training data for the purposes of re-training, refining and evaluating an AI system. In this case, the organisation should take a case-by-case approach to determining whether it can fulfil requests.

Complying with a request to delete training data would not entail erasing any ML models based on such data, unless the models themselves contain that data or can be used to infer it (situations which we will cover in the section below).

Rights relating to personal data involved in AI systems during deployment

There are only a few differences between the considerations that apply to training data and the personal data involved during model deployment.

Typically, once deployed, the outputs of an AI system will be stored in a profile of an individual and used to take some action in relation to them.

For instance, the product offers a customer sees on a website might be driven by the output of the predictive model stored in their profile. Where such data constitutes personal data, it would be subject to the rights of access, rectification, and erasure. Whereas individual inaccuracies in training data will usually have only a negligible effect, an inaccurate output of a model could directly affect the data subject.

Requests for rectification of model outputs (or the personal data inputs on which they are based) are therefore more likely to be made, and should be treated with a higher priority, than requests for rectification of training data.

Rights relating to the model itself

In addition to being used in the inputs and outputs of a model, personal data might also be contained in a model itself. As explained in a previous blog post Privacy attacks on AI models, this could happen for two reasons; by design or by accident.

Fulfilling requests about data contained by design

When personal data is included in models by design, it is because certain types of models, such as Support Vector Machines (SVMs), contain some key examples from the training data in order to help distinguish between new examples during deployment. In such cases, a small set of individual examples will be contained somewhere in the internal logic of the model.

The training set would typically contain hundreds of thousands of examples, and only a very small percentage of them would end up being used directly in the model. Therefore, the chances that one of the relevant data subjects makes a request are very small; but it is possible.

Depending on the particular programming library in which the ML model is implemented, there may be a built-in function to easily retrieve these examples. In such cases, it might be practically possible for an organisation to respond to a data subject’s request. If the request is for access to the data, this could be fulfilled without altering the model. If the request is for rectification or erasure of the data, this would not be possible to achieve without having to re-train the model (either with the rectified data, or without the erased data), or deleting the model altogether.

Fulfilling requests about data contained by accident

Aside from SVMs and other models that contain examples from the training data by design, some models might ‘leak’ personal data by accident. In such cases, unauthorised parties may be able to recover elements of the training data or infer who was in it by analysing the way the model behaves.

The rights of access, rectification, and erasure may be difficult or impossible to exercise and fulfil in these scenarios. Unless the data subject presents evidence that their personal data could be inferred from the model, the organisation may not be able to determine whether personal data can be inferred and therefore whether the request has any basis.

Organisations should regularly and proactively evaluate the likelihood of the possibility of personal data being inferred from models in light of the state-of-the-art technology, so that the risk of accidental disclosure is minimised.

Your feedback

We would like to hear your views on this topic and genuinely welcome any feedback on our current thinking.

Please share your views by leaving a comment below or by emailing us at AIAuditingFramework@ico.org.uk.

Dr Reuben Binns, a researcher working on AI and data protection, joined the ICO on a fixed term fellowship in December 2018. During his two-year term, Dr Binns will research and investigate a framework for auditing algorithms and conduct further in-depth research activities in AI and machine learning.

Next blog