Trade-offs

Reuben Binns, our Research Fellow in Artificial Intelligence (AI), and Valeria Gallo, Technology Policy adviser, discuss how using AI can require trade-offs between data protection principles, and what organisations can do to assess and balance them.

25 July 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.

AI systems must satisfy a number data protection principles and

requirements, which may at times pull organisations in different directions. For example, while more data can make AI systems more accurate, collecting more personal information has implications in terms of erosion of privacy.

Organisations using AI need to identify and assess such trade-offs, and strike an appropriate balance between competing requirements.

The right balance in any particular trade-off will depend on the specific sectoral and social context an organisation operates in, and the impact on data subjects. However, in this blog we discuss some considerations for assessing and mitigating trade-offs that are relevant across use cases.

We start off with a short overview of the most notable trade-offs that organisations are likely to face when designing or procuring AI systems.

Notable trade-offs

Privacy vs accuracy

Machine Learning (ML) uses statistical models to predict behaviour or classify people. In general terms, the more data is used to train and run the ML model, the more likely it is for the latter to capture any underlying, statistically meaningful relationships between the features in the datasets.

For instance, a model for predicting future purchases based on customers' purchase history will tend to be more accurate the more customers are included in the training data. And any new features added to an existing dataset may be relevant to what the model is trying to predict; for instance, purchase histories augmented with additional demographic data might further improve the predictive accuracy of the model.

However, collecting additional personal data can have an adverse impact on the privacy. The more individuals included in the dataset, the more

information collected about each person, the greater the impact.

Accuracy and Fairness

As we discussed recently, the use of AI systems can lead to biased or discriminatory outcomes. Organisations can put in place various technical measures to mitigate this risk, but most of these techniques also tend to reduce the accuracy of the AI outputs.For example, if an anti-classification definition of fairness is applied to an AI credit risk model, any protected characteristics, as well as known proxies (eg postcode as a proxy for race) would need to be excluded from consideration by the model. This may help prevent discriminatory outcomes, but it may also result in a less accurate credit risk measurement. This is because the postcode may also have been a proxy for legitimate credit risk features, for example job security, which would have increased the model's accuracy.There may not always be a trade-off between accuracy and fairness. For example, if discriminatory outcomes in the model are driven by a relative lack of data on a minority population, then both fairness and accuracy could be increased by collecting more relevant data.However, in that case, the organisation would face a different trade-off, between privacy and fairness.

Privacy and Fairness

Privacy and fairness might conflict in two ways. Firstly, as described above, an organisation may find that its system is unfair, due in part to a relative lack of data on a minority population. In such cases, it may want to collect data on more people from such groups so that its system is more accurate on them.Secondly, in order to test whether an AI system is discriminatory, it would normally be necessary to collect data on protected characteristics. For instance, to measure whether a statistical model has substantially different error rates between individuals with different protected characteristics, it will need to be tested with data that contains labels for those characteristics. If this data needs to be collected in order to perform the testing, then the organisation faces a trade-off between privacy (not

collecting those characteristics) and fairness (using them to test the system and make it fairer).

Explainability and Accuracy

As discussed in the interim report of our ExplAIn Project, the trade-off between the explainability and accuracy of AI decisions may often be a false dichotomy.Very simple AI systems can be highly explainable. In simple and relatively small decision trees, for example, it is relatively easy to understand how inputs relate to outputs. And although it is more challenging, there are also ways to explain more complicated AI decision-making systems.Nevertheless very complex systems, such as those based on deep learning, can make it hard to follow the logic of the system. In such cases, there may be a trade-off between accuracy and explainability, which will be also considered in greater depth in the ExplAIn project final guidance.

Explainability and Security

Providing data subjects with explanations about the logic of an AI system can potentially increase the risk of inadvertently disclosing private information in the process.Recent research has demonstrated how some proposed methods to make ML models explainable can unintentionally make it easier to infer private information about the individuals whose personal data the model was trained on. This is a topic which we will cover in a future upcoming blog on privacy attacks on ML models.Some literature also highlights the risk that in the course of providing an explanation to data subjects, organisations may reveal proprietary information about how an AI model works. Our research and stakeholder engagement so far indicate this risk is quite low. However, in theory at least, there may be cases where a trade-off will need to be struck regarding the right of individuals to receive an explanation, and the right of organisations to maintain trade secrets.Both of these risks are active areas of research, and it is not yet known how likely and severe they are likely to become. Organisations should monitor the latest research and consider realistic threat models in their given context.

How should organisations manage trade-offs?

There is no hierarchy of data protection principles and striking the right balance across one or multiple trade-offs will in most cases be a matter of judgement, specific to the AI use case and the context it is meant to be deployed in.

Whatever choices organisations make, they will need to be accountable for them. From this perspective, there are a number of basic steps that organisations will be expected to undertake. Their efforts should be proportional to the data protections risks associated with the AI system. These steps are:

1. Identify and assess any existing or potential trade-offs, when designing or procuring an AI systems, and assess the impact they may have on the data subjects.
2. Consider available technical approaches to minimise the need for any trade-offs. Organisations should consider any techniques which can be implemented with a reasonable level of investment and effort.
3. Have clear criteria and lines of accountability in relation to the final trade-off decisions. This should include a robust, risk-based and independent approval process.
4. Where appropriate, take steps to explain any trade-offs to data subjects or any human tasked with reviewing AI outputs.
5. Review trade-offs on a regular basis, taking into account, among other things, the views of data subjects (or their representatives) and any emerging techniques or best practices to reduce them.

These processes, and their outcomes, should be documented to an auditable standard, and should be captured in the Data Protection Impact Assessment (DPIA), if required, with an appropriate level of detail.

Some of the things we would expect organisations to document include:

- consideration of the risks to the individuals that are having their personal data processed;
- the methodology for identifying and assessing the trade-offs in scope; the reasons for adopting or rejecting particular technical approaches (if relevant);
- the prioritisation criteria and rationale for the final decision; and
- how the final decision fits within the organisation overall risk appetite.

Organisations should also be ready to halt the deployment of any AI systems, if it is not possible to achieve an appropriate trade-off between two or multiple data protection requirements.

Outsourcing and Third Party AI systems

When AI solutions are either bought from or outsourced to third parties, careful consideration and independent evaluation of any trade-offs should be part of the due diligence process. Organisations should ensure that they have the authority to modify the systems before deployment, either directly or via the third party provider, so that they align with what they consider to be the appropriate trade-offs.

For instance, a vendor may offer a CV screening tool which is very accurate in selecting job candidates, but requires a lot of sensitive data in order to work. An organisation procuring such a system will need to consider whether they can justify collecting so much sensitive data from candidates, or tolerate a lower accuracy rate.

Culture, diversity and engagement with stakeholders

Organisations will need to make significant judgement calls when determining the appropriate trade-offs. While effective risk management processes are essential, the culture of an organisation will also play a fundamental role.

Undertaking this kind of exercise will require collaboration between different teams within the organisation. Diversity, incentives to work collaboratively, as well as an environment in which staff feel encouraged to voice concerns and propose alternative approaches will all be important.

Our understanding of the social acceptability of AI in different contexts, and of best practices in relation to trade-offs, are still developing. Whether or not the requirement is triggered by a DPIA, consultation with stakeholders outside the organisation can help organisations understand the value they should place on different criteria.

Assessing trade-offs: a worked example

In many cases trade-offs will not be precisely quantifiable, but this should not lead to arbitrary decisions. Organisations should perform contextual assessments, documenting and justifying their assumptions about the relative value of different requirements for specific AI use cases.

One possible approach to help organisations to identify the best possible trade-offs is to use a visual representation.

The 'production-possibility frontier' approach

Possible choices about how a system could be designed can be plotted on a graph, with the criteria – for example accuracy and privacy – on the X and Y axis. This is known as the 'production-possibility frontier' and is one way to help decision makers understand how system design decisions may impact different data protection values.

We have used this method in Figure 1 to visualise what a trade-off between privacy and accuracy might look like.

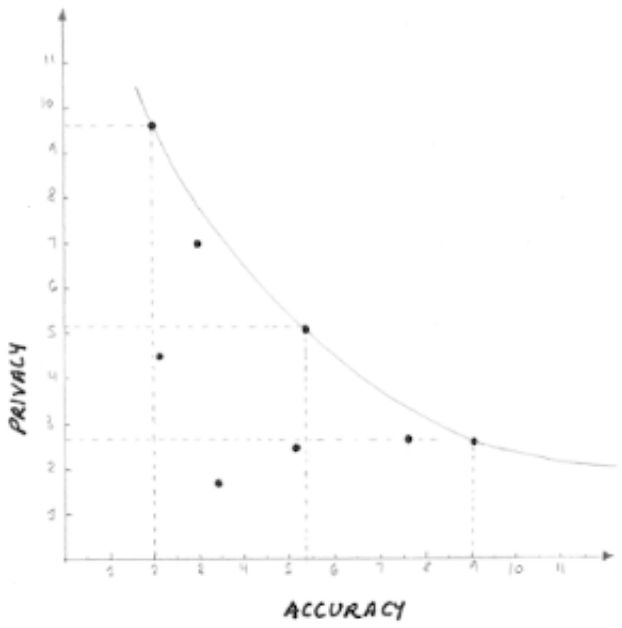

Figure 1

The data points in the graph represent the different possible technical configurations of the AI system, resulting from different design choices, ML models, and the amount and types of data used.

Accuracy can be precisely defined, for example, in terms of precision and recall. Measures for privacy are likely to be less exact and more indicative in nature. However, there are several privacy indicators that organisations may choose to use as heuristics. These include the amount of personal data required, the sensitivity of this data, the extent to which it might uniquely identify the individual, the risk of harm the data processing may present to individuals, and the number of data subjects the AI systems will be applied to.

In the scenario in Figure 1, no system can achieve both high accuracy and high privacy, and any the trade-off between privacy and accuracy is significant.

For a different use case, the trade-offs may look very different, as visualised in Figure 2.

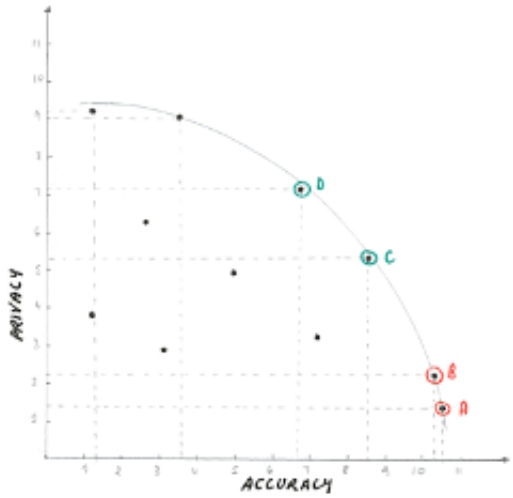

Figure 2

In this scenario, it may be easier to achieve a reasonable trade-off between accuracy and privacy. The graph also shows that the cost of sacrificing either privacy or accuracy is lower for those AI systems in the middle of the curve (C and D), than for those at the edge (A and B).

In this example, there are diminishing returns on accuracy for the possible systems at the bottom right. To justify choosing system A over B, an organisation would have to place a very high value on accuracy relative to privacy.

Visual representation of trade-offs can also include lower limits for either variable below which the organisation is not willing to go (Figure 3).

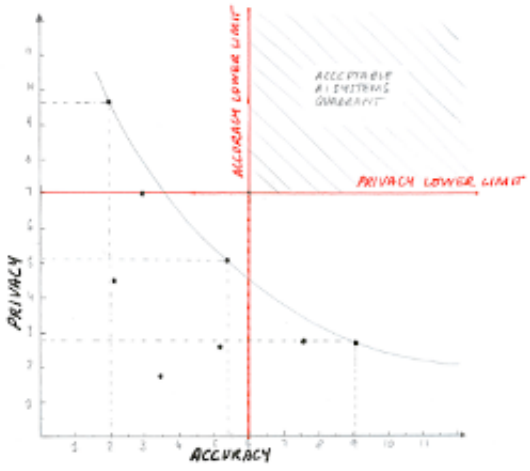

Figure 3

In the scenario in Figure 3, there is no possible system which meets both the lower limits for accuracy and for privacy, so the organisation should halt the project.

Mathematical approaches to minimise trade-offs

There may be cases in which some elements of the trade-offs can be precisely quantified.

In these cases, there are a number of mathematical and computer science techniques known as 'constrained optimisation' that aim to find the optimal solutions for minimising trade-offs. For instance, the theory of differential privacy provides a framework for quantifying and minimising trade-offs between the knowledge that can be gained from a dataset or statistical model, and the privacy of the people in it. Similarly, various methods exist to create ML models which optimise accuracy within the constraint of fairness, if the latter can be mathematically defined.

However, not all aspects of issues like privacy and fairness can be fully quantified. For example, differential privacy can measure the likelihood of an individual being identified with a particular piece of information, but not the sensitivity of that information. Therefore these methods should always be supplemented with a more holistic approach.

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.

Valeria Gallo is currently seconded to the ICO as a Technology Policy Adviser. She works with Reuben Binns, our Artificial Intelligence (AI) Research Fellow, on the development of the ICO Auditing Framework for AI. Prior to her secondment, Valeria was responsible for analysing and developing thought leadership on the impact of technological innovation on regulation and supervision of financial services firms.

Next blog