When it comes to explaining AI decisions, context matters



Alex Hubbard, Senior Policy Officer at the ICO, looks at some of the key themes identified in the ICO and The Alan Turing Institute's interim report about explanations of AI decisions.

03 June 2019

Explainability of AI decisions is a key area of the AI auditing framework. This guidance from Project ExplAIn will inform our assessment methodology.

If an Artificial Intelligence (AI) system makes a decision about an individual, should that person be given an explanation of how the decision was made?

Should they get the same information about a decision regarding criminal justice as they would about a decision concerning healthcare?

These are just two of the issues we have been exploring with public and industry engagement groups over the last few months.

In 2018, the Government tasked the ICO and The Alan Turing Institute (The Turing) to produce practical guidance for organisations, to assist them with explaining AI decisions to the individuals affected. This work has been titled 'Project ExplAIn'.

The ICO and The Turing conducted research, including citizens' juries and industry roundtables, to gather views from a range of stakeholders with various, and sometimes competing, interests in the subject.

Today, we published the findings of this research in a Project ExplAIn interim report.

Key themes

The report identifies three key themes that emerged from the research:

1. the importance of context in explaining AI decisions;
2. the need for education and awareness around AI; and
3. the various challenges to providing explanations.

1. Context

The strongest message from the research is that context is key. The importance of providing explanations to individuals, and the reasons for wanting them, change dramatically depending on what the decision is about.

People who took part in the citizen juries (jurors) felt explaining an AI decision to the individual affected was more important in areas such as recruitment and criminal justice than in healthcare. Jurors wanted

explanations of AI decisions made in recruitment or criminal justice settings in order to challenge them, learn from them, and check they have been treated fairly. In healthcare settings, however, jurors preferred to know a decision was accurate rather than why it was made.

Jurors said they only expected an explanation of an AI decision if they would also expect a human to explain a decision they had made. They also wanted explanations of AI decisions to be similar to human explanations. But the industry roundtables questioned whether AI decisions should be held to higher standards, because human explanations could sometimes misrepresent the truth in order to benefit the explainer or to appease the individual.

2. Education

The findings showed a need to engage and inform the public in the use, benefits and risks of AI decision-making, through education and raising awareness. Although there wasn't a clear message on who should be responsible for this, delivering a balanced message is important, suggesting a need for a diverse range of voices.

3. Challenges

Industry roundtable participants generally felt confident they could technically explain the decisions made by AI. However, they raised other challenges to 'explainability' including cost, commercial sensitivities (eg infringing intellectual property) and the potential for 'gaming' or abuse of systems.

The lack of a standard approach to establishing internal accountability for explainable AI decision systems also emerged as a challenge.

What's next?

The findings set out in the Project ExplAIn interim report will feed directly into our guidance for organisations. This will go out for public consultation over the summer and will be published in full in the autumn.

The ICO has said many times that data protection is not a barrier to the use of innovative and data-driven technologies. But these opportunities cannot be taken at the expense of being transparent and open with individuals about the use of their personal data.

The guidance will help organisations to comply with data protection law but will not be limited to this. It will also promote best practice, helping organisations to foster individuals' trust, understanding, and confidence in AI decisions.

As well as benefiting the ICO and The Turing on Project ExplAIn, it is hoped that these findings will help inform others in their own thinking, research and development of explainable AI decisions.

All materials and reports generated from the citizens' juries are freely available to access. The project ExplAIn guidance will also inform the ICO's AI auditing framework, which is currently being consulted on and which is due to be published in 2020.



Alex Hubbard is a Senior Policy Officer at the ICO

Next blog