Consultation on the ICO’s ExplAIn draft guidance

The ICO and The Alan Turing Institute are consulting on our co-badged ExplAIn guidance. This guidance aims to give organisations practical advice to help explain the processes, services and decisions delivered by AI, to the individuals affected by them.

We are looking for a wide range of views from organisations across all sectors and sizes.

The guidance is comprised of three parts. Depending on your level of expertise, and the make-up of your organisation, some parts may be more relevant to you than others. You can pick and choose the parts that are most useful.

You can answer as many or as few of the questions as you want to. You can also save your progress and return at a later date.

If you would like further information about the consultation, please email explain@ico.org.uk.

Please send us your response by 17:00 on Friday 24 January 2020. You can email it to explain@ico.org.uk or complete the online version of this survey.

Privacy statement

Please note, your responses to this survey will be used to help us with our work on explainability only. The information will not be used to consider any regulatory action, and you may respond anonymously should you wish. For more information about what we do with personal data see our privacy notice.
Q1. Does the guidance provide what your organisation needs when considering how to explain AI-enabled decision to individuals?
☐ Yes
☒ No
Please provide further detail:

Content is good but it is difficult to access. Key points are highlighted but it is poorly structured. An executive summary which could be provided to a board is what is needed. Organisations have many different needs. Some sort of precedent models, or sample documents would be useful. There are examples mostly in the second section but sometimes these are not relevant in specific domains. Document two in particular is too long. An interactive document would be useful.

There needs to be training material to support this. A suite of documents ought to be provided for relevant industries. Maybe even decision trees, diagrams or workflows might be appropriate.

While the content is interesting it is hard to know how it applies in a particular situation. Different people need different explanations.

As technologists we found the legal explanations not very clear. Data protection officers might find these explanations clearer.

The following questions relate to ‘Part one: The basics of explaining AI’:

Q2. What other definitions, if any, should we cover to help inform the guidance?

This seemed very skewed to supervised learning rather than unsupervised. There are many different types of AI.
Q3. How clear are the summaries of the relevant legislation and how they apply?

<table>
<thead>
<tr>
<th></th>
<th>Very clear</th>
<th>Clear</th>
<th>Not clear</th>
<th>Not at all clear</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDPR</td>
<td>☐</td>
<td>☐</td>
<td>☒</td>
<td>☐</td>
</tr>
<tr>
<td>Equality Act 2010</td>
<td>☐</td>
<td>☒</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Judicial review</td>
<td>☐</td>
<td>☐</td>
<td>☒</td>
<td>☐</td>
</tr>
</tbody>
</table>

Q4. What other legislation, if any, should we cover?

Regulation around advertising and marketing needs to be included. This can impact the conduct of elections. For instance, the controversy around real-time bidding illustrates this. These documents are also very UK centric.

There are also issues with existing legislation. For instance, GDPR does not help when the data has been anonymised. This is problematic as many have shown that it is often very easy to determine who the individuals in a data set are from cues in their records eg. Alexa - German magazine; New York Times (NYT) location data showing daily routine from home to work etc etc. post code and relate information can be put together.

Q5. In your experience, which of the benefits and risks we have outlined are most relevant in explaining (or not explaining) AI-assisted decisions to individuals?

This was a useful section.
Q6. Is it clear what the different explanation types are, and why they are important for an explanation?
   ☑ Yes
   ☐ No
   Please provide further detail:
   
   This was good but maybe the inclusion of a section containing definitions, or a glossary would be beneficial.

Q7. What other explanation types, if any could we include?

Sometimes explanations are difficult (e.g. if using a deep learning model in a medical application). If it is possible to validate that a human would reach the same conclusion, do we always need an explanation?

Q8. We have outlined four principles:

   Be accountable
   Be transparent
   Consider context
   Reflect on impacts

   Are they helpful for your organisation?
   ☑ Yes
   ☐ No
   Please provide further detail:

   Should we consider trade-offs? For instance, there may be trade-off between explainability and accuracy. Businesses should be considering this type of thing in any case. One important principle is to reflect on impacts. This is a crucially important design step in any decision algorithm. This requires considerably more thought to prevent it simply being a box ticking exercise. How can this be policed?
In terms of allocating responsibility, impact assessment etc we need to make a distinction between the roles and responsibilities of the manufacturers (designers/programmers/marketers) and those using the algorithms e.g. companies using and algorithm for job selection or insurance. How can we ensure that those employing an algorithm are employing it correctly for the purposes that it was designed for? This division is also unclear in the second document although in the case studies it appears to be on the users rather than the manufacturer.

The following questions relate to ‘Part two: Explaining AI in practice’

Q9. Are there any steps missing in the summary steps?
☐ Yes
☐ No
What are the missing steps?

Different roles require different explanations. There are 108 pages of Part two. To be practically useful someone (probably the chief technology officer) will have to summarise this and provide more readable explanations. The content is however interesting. Step three for instance was broken down well. This has to be live document. Filter out the information to different levels of knowledge.

Each of the different methods are susceptible to the possibility of introducing bias inadvertently.

Should there be some attempt to access when bias might be introduced? In Canada they have an algorithm impact assessment. This sounds useful.

We need an audit system in place and ideas such as pre-certification processes to ensure that people are following standards. The present system looks somewhat like self-certification.

Once an automated process is in place sampling etc. needs to be instigated.
Q10. Is it clear what information should go into the explanations we have described?
☑ Yes
☐ No
Which explanation is unclear and why?

Q11. What other elements of the data collection and pre-processing, that contribute to explainability, if any, should we include?

This is the area in which bias can tend to creep in. Algorithms in themselves are not biased, the background data is.

Data integrity and quality are the building blocks of AI and this is vital. Data gaps (i.e. the correct representation of people in data is key). The “invisibility of women” might occur accidently. Ethnicity is another obvious point of weakness.

Q12. In step two, is it clear how you should choose your priority explanation types?
☐ Yes
☒ No
Please provide further detail:

This does not help in deciding priority.

In healthcare applications process- based explanations are important. Outcome explanations are more relevant for individuals.

Sometimes if you cannot explain your process you might not be able to explain your outcome.

In a “black box” situation if you cannot do an outcome-based process you might need a process one.

Q13. Are the examples for choosing suitable explanation types clear?
☐ Yes
☒ No
Please provide further detail:
Examples need to be tailored to specific industries. The examples are too high level.

This is a very fast changing area although while the technology changes fast, data problems are more static.

Q14. After reading the guidance about selecting an appropriately explainable model in step four, how helpful do you feel this will be for your work?

- Very clear ☐
- Clear ☒
- Not clear ☐
- Not at all clear ☐

Q15. How clear is the guidance about the tools you can use for extracting rationale explanations and the limitations they have?

- Very clear ☐
- Clear ☒
- Not clear ☐
- Not at all clear ☐

Q16. What other rationale explanation extraction tools, if any, could we include?

These examples are clear but incomplete. While they provide a starting point, this is a dynamic area. Highlighting tools at this point is not that useful.

Q17. Is it clear how you should take the statistical output of the AI system and translate it into meaningful explanation?

- Yes ☐
- No ☒

Please provide further detail:

This is an area of expertise. It is a skillset in itself that is not necessarily explainable in a document.
You cannot necessarily just pull something off the shelf. One aspect which is not mentioned is how the AI system is actually used.

The big players need to get together to develop appropriate training systems.

Q18. Step five discusses how to train staff, who implement your AI system, to interpret the outputs and apply them to the circumstances of an individual. After reading this, do you feel confident about applying this training in your organisation?
☐ Yes
☒ No
Please provide further detail:

This section was very short. It was good that it was included, however the recommendation that “you have to ensure that” is very strong language for what is supposed to be a guidance document. Within the medical sphere, post market surveillance is as important as explanation to an individual.

We have highlighted five contextual factors that influence the kind of explanations people want about an AI-assisted decision relating to them. These have come from the research carried out with the public.

Q19. Do these reflect you experiences?
☒ Yes
☐ No
Please provide further detail:

Q20. What other contextual factors, if any, could we include?

Perhaps equality is something which might be included.
Contextual factors such as the diversity of development teams are also an issue.

Q21. Do the types of explanation we have suggested for each contextual factors make sense to you?
☒ Yes
☐ No

This was felt to be among the most useful and clear parts of the documentation.

Q22. How likely are you to implement the detailed proactive engagement measures in your organisation?

Very likely ☐
Likely ☐
Unlikely ☒
Very unlikely ☐

Q23. What other measures, if any, should we include?

This is a domain specific issue. AI applications in areas such as document comparison are radically different from healthcare applications. Each domain has their own set of regulations as well. In addition, regulation systems in other jurisdictions need to be taken into account.

However, this does show those specific articles within GDPR which are relevant to AI.

Q24. Do you have any suggested changes to the healthcare example we have included in Annex 1?
It was felt that the healthcare section was good but could be developed further. Worked examples would be useful, although we are aware that this might result in some people adopting such as example as a template. Some detailed guidance alongside the document might be useful. What might the output look like in different situations?

The following questions relate to 'Part 3: What explaining AI means for your organisation':

**Q25. How accurate is the characterisation of the following roles and responsibilities?**

<table>
<thead>
<tr>
<th>Role</th>
<th>Very accurate</th>
<th>Accurate</th>
<th>Not accurate</th>
<th>Very inaccurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Product manager</td>
<td>☒</td>
<td></td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>b. AI development team</td>
<td>☒</td>
<td>☒</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>c. Implementer</td>
<td>☒</td>
<td></td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>d. DPO and compliance team</td>
<td>☒</td>
<td></td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>e. Senior management</td>
<td>☒</td>
<td></td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

**Q26. What other roles and responsibilities, if any, should we include?**

Perhaps there needs to be a “maintenance of documentation” role

How do you ensure that explainability remains relevant? The people who initially provide the explanations may not be the same people who are asked to maintain the system.

**Q27. How feasible is it for you to produce policies and procedures in the areas we have listed?**
Please comment on any specific areas.

For some organisation (e.g. in healthcare) this system may not be just feasible but inherently necessary. In highly regulated domains, there is client pressure to provide explanations however for some organisations the suggested system may not be feasible. Much work is done from clients on their sites. All of this work requires time and cost – will the client pay for this?

Q28. What other policies and procedures, if any, should we include?

Training is obviously necessary. Maybe this is an emerging professional discipline. The ability to explain an AI system to a wider variety of audiences is a very unusual skillset. Teams of individuals are probably required.

Q29. Is it clear what types of information you need to document?
   ☒ Yes
   ☐ No

Q30. What other types of information, if any, should we include?

The bullet points listed on page 20 of document three is good.
Q31. Do you have any other comments on this guidance?

It is much too long. It would benefit from diagrams and workflows.

It needs to be restricted for specific roles and domains.

Worked examples would be ideal.

Expand on the checklists and the appendices.

It is a good step in the right direction.

More work needs to be done. It needs to be made clear that it is guidance and not regulation.

It is too focussed on machine learning. A more “end to end” approach would be better.

Some general comments:

The UN special rapporteur on extreme poverty recently wrote a report to the UN Human Rights Council about digital unfairness to poor and vulnerable section of populations internationally (including the UK). Many of those affected are the lesser educated and often illiterate. How can we ensure that they receive the same justice as others and that the explanations they receive are truly transparent to them and just. You mention complaining to persons responsible within companies but what if that is unsatisfactory. They need a route to complain such as an ombudsman that a company can point them to.

The first document on explainability seems not to require a detailed explanation of how the algorithm made the decisions (i.e. what feature did it use and what rules were programmed). We know that the numerical matrices produced by various machine learning models are opaque despite many research attempts make them explicable. Of-course if they were explicable, then a symbolic version of the algorithm could be used to make the process more transparent. Note that explaining the data used to train a machine does not give an explanation of how it uses that data to make decisions. It may not help with the explanation at all. Other questions may be where and how was the data labelled. We know that it is often sent to large data farms with poorer population working for a
pittance. Has this labelling been responsible or does it lead to certain biases?

In relation to the second document, it looks like either the responsibility falls on the user rather than the manufacturer. The document appears to conflate them.) How was this marketed to the user? Was the information that the user received accurate. Maybe the user is not technically capable of understand AI or investigating the tool that they have been given. How can they ensure that it has been tested adequately? Without sufficient knowledge their explanation may well be what they have been told rather that what they actually are. Smaller companies may be employing a decision algorithm for economic reasons and cannot afford to hire a technical advisor or tester.

Q32. What sector do you work in? Please tick all that apply:
☑ Private
☑ Public
☐ Third

Q33. What industry do you work in eg finance, health?

Our group represents private sector IT firms who work in finance health and government.

Q34. Where did you hear about this consultation?
☐ ICO Twitter account
☐ ICO Facebook account
☐ ICO LinkedIn account
☐ ICO newsletter
☐ ICO blog
☐ ICO staff member
☐ Colleague
☐ Twitter
☐ Facebook
☐ LinkedIn
☐ Other
If other please specify:

We heard about it from our professional body.