Summary of responses to the consultation on ICO/Turing draft guidance on Explaining AI decisions, with comments

Introduction

In December 2019, the ICO published draft guidance on Explaining decisions made with AI, jointly produced with The Alan Turing Institute (The Turing), with a deadline of 24 January 2020 for comments.

Our survey asked for feedback on key areas in relation to the overall proposed guidance, requested views on the usefulness of the proposed explanation types and steps, and provided an opportunity for respondents to make any further general comments.

The ICO and The Turing would like to thank all those organisations and individuals who took the time to read the draft guidance and give us their views, and those who offered to work with us further. We have carefully noted all your comments and these have been invaluable in shaping our thinking on this topic as we produced the final version of the guidance.

Quantitative summary

Overall, we received 42 responses to our online survey and a further 26 general responses via our ExplAIn inbox.

The largest proportion of responses received (33) was from the private sector, with smaller numbers of responses from those working in other sectors. Seven respondents declined to provide this information. The distribution of responses is shown in Figure 1 below.

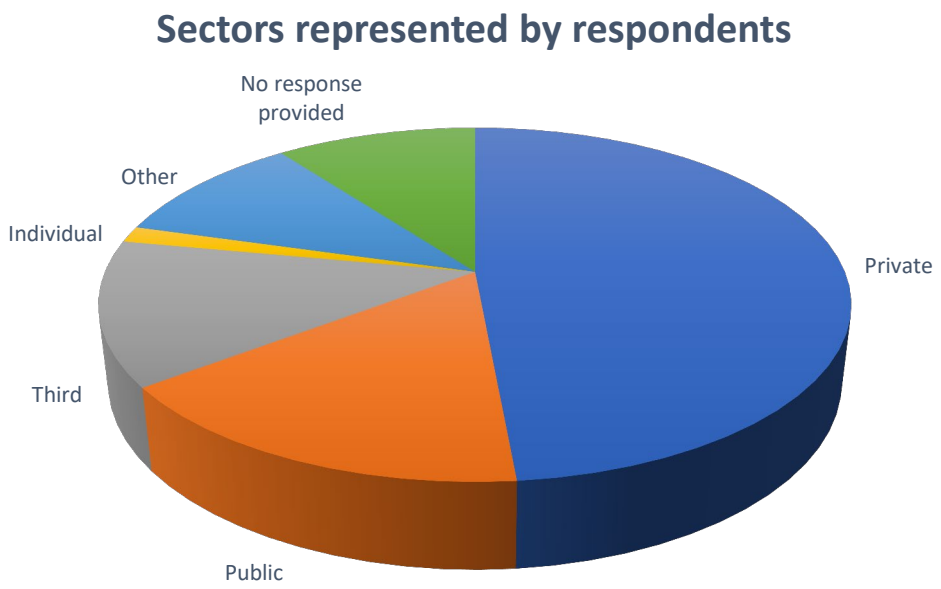

Figure 1: Sectors represented by those that responded to the consultation

Overall, the response to our consultation was generally positive. The majority of quantitative responses collated from the survey indicated that respondents understood what we were looking to explain within the guidance. A selection of these quantitative responses are summarised below. Please note that respondents were not required to answer every question, so the total number of respondents will not necessarily be reflected in the number of responses received to each question.

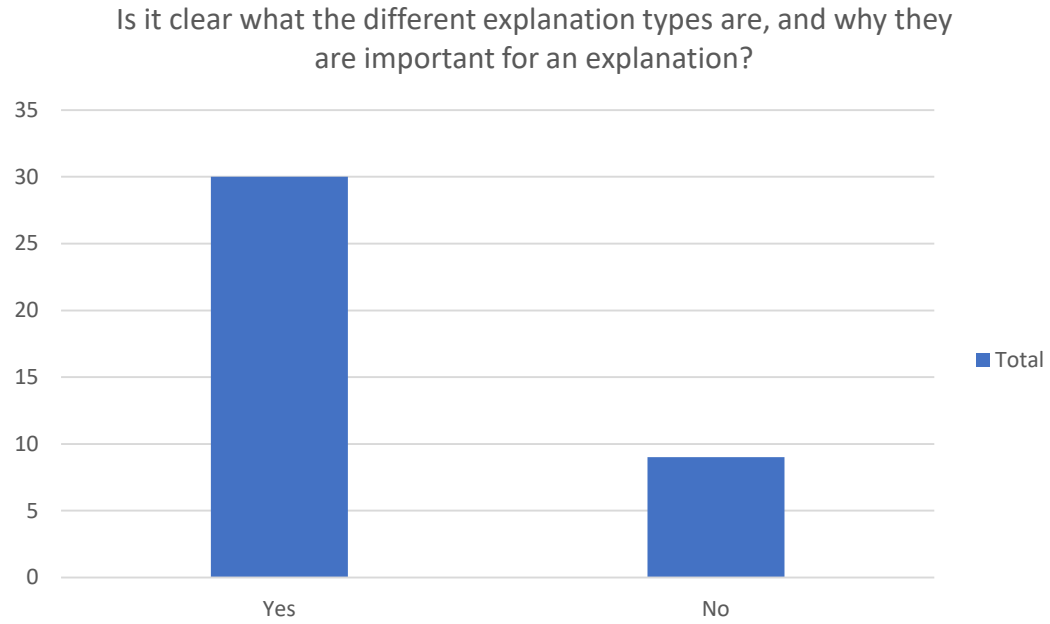

Figure 2: Q6 – approximately three quarters of respondents (30 of 39) agreed that the different explanation types are clear

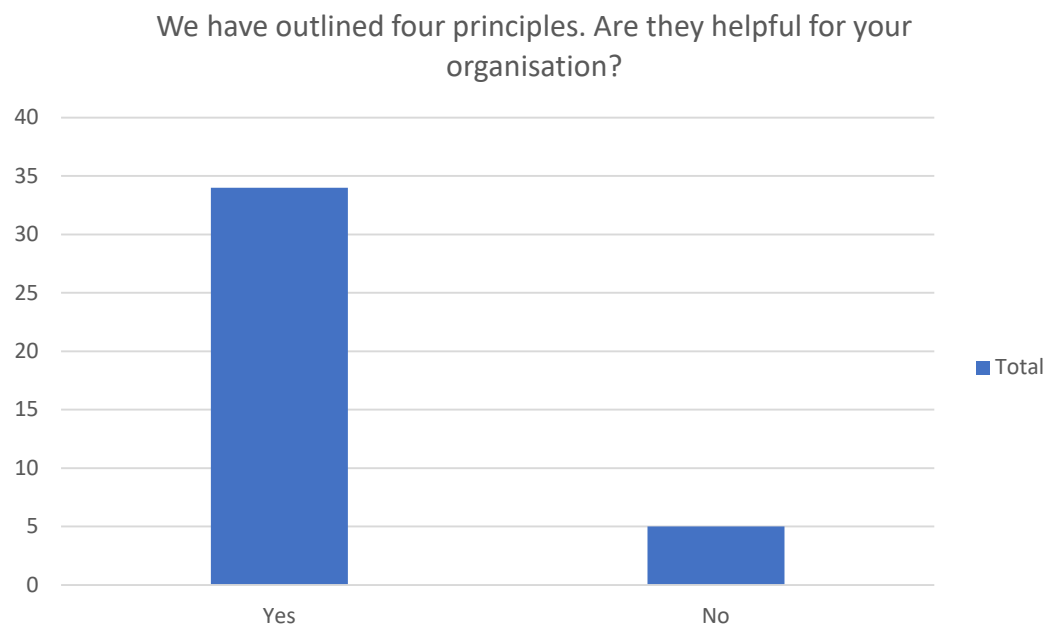

Figure 3: Q8 – a large majority of respondents (34 of 39) agreed that the principles are helpful

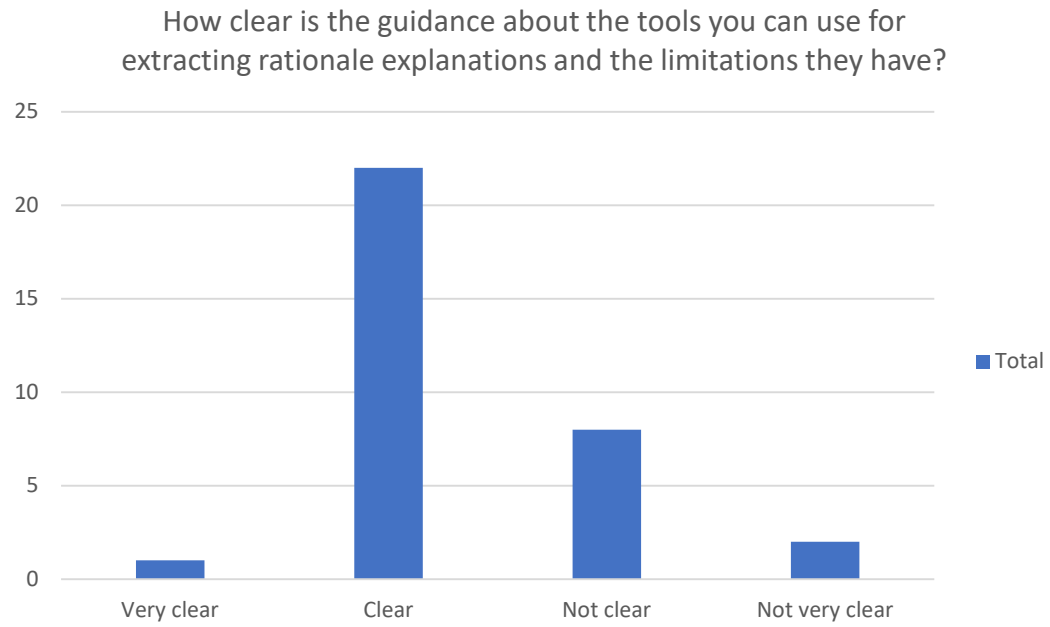

Figure 4: Q15 – over two thirds of respondents (23 of 33) felt that the guidance on tools for extracting rationale explanations is "clear" or "very clear"

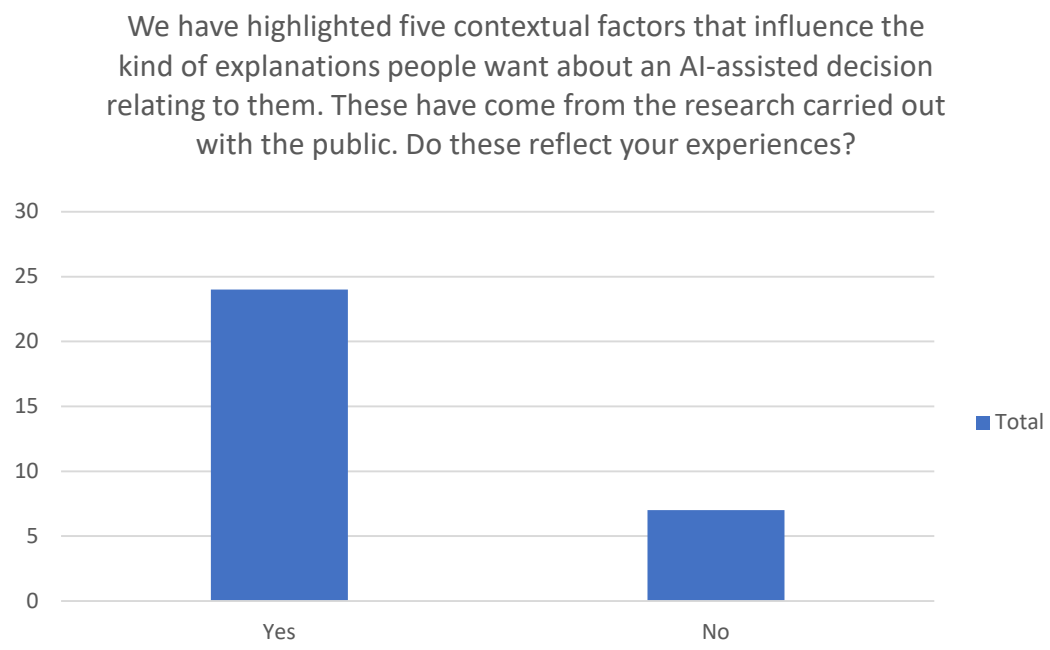

Figure 5: Q19 – 24 out of 31 respondents agreed that the five contextual factors outlined reflected their experiences

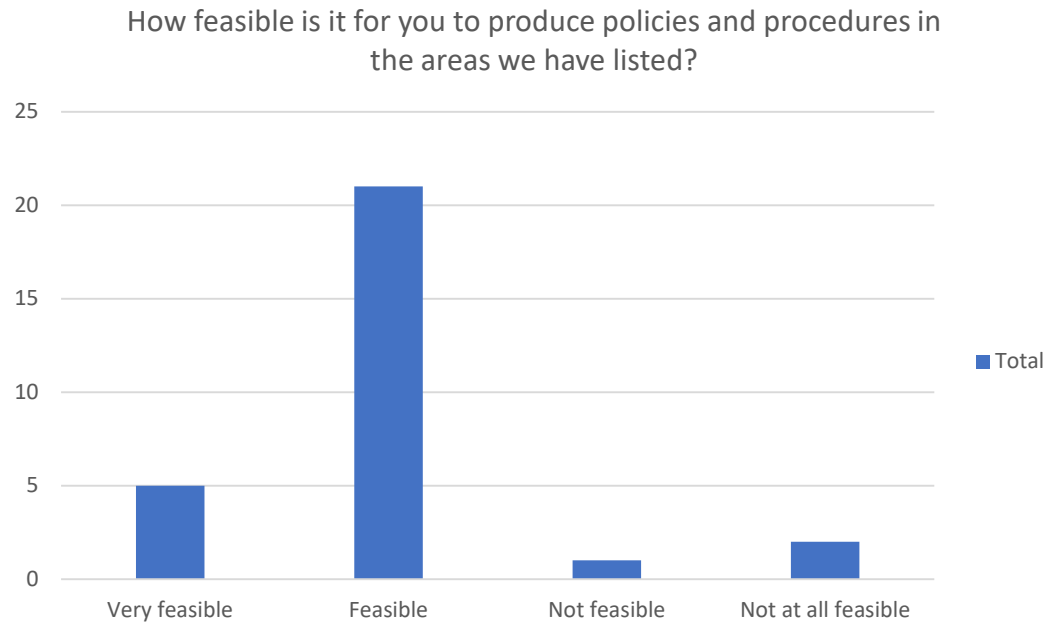

Figure 6: Q27 – nearly all (26 of 29) respondents felt that it would be "feasible" or "very feasible" to produce the policies and procedures outlined in the guidance

However, there were a couple of questions that left the respondents more divided. These responses are shown below, and the qualitative responses are discussed in more detail later on in this summary.

Figure 7: Q9 – 16 of 33 respondents felt that there are missing summary steps in the guidance. We discuss this in more detail on pages 9 and 10 of this summary document

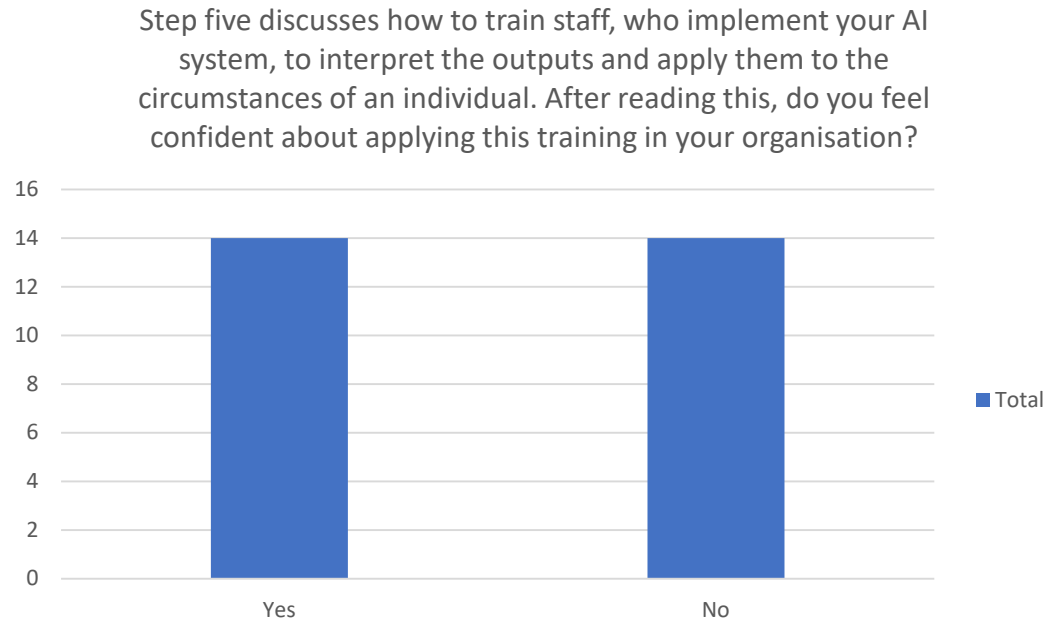

Figure 8: Q18 – half the respondents felt that the guidance did not give them confidence about training staff that implement AI to interpret the outputs from the systems. This is discussed in more detail on page 9 and 10 of this summary document

While we cannot respond individually to each contribution, we have provided an overview below of the key themes that have become apparent and some comments on our emerging thinking from each area of the consultation as we finalised our guidance.

Key themes

General points about the guidance overall

There appeared to be some confusion about the status of this guidance, with several respondents indicating that they viewed this as guidance for complying with data protection legislation, and not a more general "best practice" guidance document. Part of this stemmed from the language used within the guidance, which occasionally stated that organisations were required to take certain actions (eg "*you will need to produce all the documentation prepared and testing undertaken*").

Several respondents were also concerned that the guidance has been written for organisations providing the explanation, and that there was no guidance written from the point of view of the decision recipient.

There were requests from respondents for templates to aid them in providing explanations to decision recipients. This included requests for spreadsheet templates to help with following the steps outlined in Part 2, and a Data Protection Impact Assessment (DPIA) style template to help with documentation of the process.

A few respondents questioned why we had focussed solely on machine learning techniques and not other methods, such as ontology-based or symbolic AI.

We received some responses that indicated that more guidance should be given to help organisations explain decisions to children and other vulnerable groups. These comments included the suggestion that organisations consider the Python ELI5 (Explain Like I'm Five) package, or follow recommendations given to health organisations which state that all explanations should be provided at a level that can be understood by an 11 year old.

Another common theme amongst the responses was that the guidance focussed too much on organisations that developed and deployed AI systems in-house. However, a significant number of organisations procure ready-built systems that they then deploy. Similarly, organisations may have several different systems performing different tasks that have been developed by different teams, both internally and externally. These respondents therefore recommend that additional content is added to aid organisations in the procurement process.

Finally, there were several comments that related to the technical focus of the guidance. These respondents would like to see an increased focus on practical steps they can take to create an understandable explanation,

and less detail about the technical approach to extracting explanations from AI systems.

Our comments

Our guidance has been drafted in response to the commitment in the Government's AI Sector Deal, but it is not a statutory code of practice under the Data Protection Act 2018. It has been written by the ICO and The Turing, as we were tasked to do this in the government's AI Sector Deal, published in 2018. We acknowledge that this could have been made clearer within the guidance, and have modified the language used to reflect this.

It should be noted that the purpose of this document is to aid organisations using artificial intelligence (AI) systems to explain decisions to individuals. As such, it is not intended as a guide for individuals as to what explanations they can obtain. Although we have not written this guidance from the point of view of the decision recipient, we conducted research, including Citizens' Juries, to ensure the views of the public were represented in the guidance. The explanation types, for example, were based on responses received during the Citizens' Jury exercises.

We are considering whether the provision of templates in relation to explaining AI decisions would be feasible. Once we have explored the options, we will publish any templates we create.

On a wide interpretation of the definition, other forms of AI could also be included. However, most AI-assisted decisions that use personal data will rely on machine learning, rather than symbolic AI. Rules-based systems such as symbolic AI systems are also much easier to interpret/derive explanations from, precisely because they produce deductions. They are therefore not our major concern in this guidance.

We have added some additional information about how to provide explanations to different types of audience.

We agree that the draft guidance tended to focus on systems that are both developed and deployed by the same organisation. We acknowledge that the draft guidance did not specifically address the position of organisations that procure systems from third party developers, particularly small and medium enterprises (SMEs). We have therefore included further guidance in relation to this.

We have modified the guidance to increase focus on "how" to produce explanations from AI systems using less technical language. We welcome input from organisations that work in this area as to how we

could improve this, so we can incorporate these comments if we update the guidance in the future.

Part 1 – The basics of explaining AI

We received several suggestions in response to our question about other definitions that should be included in the guidance. These included: some that disputed the way we defined artificial intelligence; definitions relating to algorithms and machine learning; and definitions relating to terminology from GDPR/DPA 2018.

We also requested views on other legislation that respondents felt should have been included within the guidance. We received several suggestions, a large number of which could be considered sector specific (Consumer Credit Act, Medical Devices Regulations, advertising and marketing legislation), as well as more general suggestions (Human Rights Act).

In response to our question about additional explanation types, we received several additional suggestions. Some of these appeared to relate to the explanation types already included. For example, a response called for a "comparability explanation" – comparing a decision made by AI with a decision made by a human. This could fall within the "safety and performance" explanation outlined in the guidance. There were also requests for explanations that would relate to the auditing of systems, such as explainability to regulators. This is likely to fall under the remit of the AI auditing framework being developed by the ICO[1], rather than this guidance, as this guidance is focussed mainly on explaining decisions to decision recipients.

Finally, we also received requests that we discuss the potential trade-offs that exist between accuracy and explainability of AI systems.

Our comments

We note that there are several terms that could be defined more clearly in the guidance. We have added further definitions to Part 1 of the guidance, so that organisations are clear on what we mean. Although there are several ways of defining some terms, such as artificial intelligence, we are using the definitions outlined at the start of the guidance.

We are reluctant to include numerous references to sector-specific

[1] https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations/

legislation within this guidance, as it is meant to provide general advice across all sectors. We are also unable to provide a significant amount of advice in relation to legislation which the ICO does not regulate. We have therefore included a list of further legislation, but recommend that organisations talk to their sector's regulator if they require further guidance in relation to their legislation.

Although there are concerns about the trade-off between explainability of an AI system and its accuracy, this is not a topic for this guidance. Instead, the AI Auditing framework will provide details on this topic. We do, however, discuss "black box" systems and supplementary models in Part 2 of our guidance.

Part 2 – Explaining AI in practice

A large number of responses we received suggest that there is a widespread belief that all the information outlined in the guidance should be provided to the decision recipient whenever a decision is made. There were concerns that this may lead to excessive information about the algorithm and other intellectual property being disclosed, or allow individuals to game the system.

Feedback also indicated that Part 2 may be too long and technical, especially for SMEs and organisations that procure systems from external vendors. There were suggestions that some information could be moved to an annexe.

Several respondents called for additional steps to be included in the process of building an explanation (see figure 7 above). Some of these, such as "transaction-level monitoring" appear to be sector-specific. Other steps refer to auditing requirements, which will be covered in the Auditing framework. There were also comments that the steps were not in the correct order for some organisations.

Some responses indicated that the examples provided through the guidance were not necessarily helpful to their organisation. We also received requests for more examples, with one suggestion that we outline some case studies that run through the guidance to show organisations how building an explanation would develop through the steps.

It was also noted that Step 5 appeared to be very short when compared to other sections (see figure 8 above). Several responders have suggested additional content that could help organisations complete this task.

There were some comments that noted that the contextual factors are important through the whole process, not just one step of seven. It was felt that these factors should be introduced in Part 1, alongside the explanation types and principles.

Finally, we had some suggestions that an additional step could be added to advise organisations on suitable methods for monitoring model drift, and assessing if the decision outputs provided by the guidance are still accurate whilst the model is operational. It was also suggested that we provide guidance on the steps to take if the model does fail.

Our comments

We acknowledge that the phrasing could have been improved to clarify that the information listed should be considered, but is not necessarily required in every explanation. We have ensured this is clarified in the final guidance.

We do appreciate that Part 2 of the guidance is very technical, but we feel some technical detail is necessary to ensure some of the explanations outlined are built and understood in sufficient detail. However, some of the information provided may only be relevant to those looking to build their own AI systems, so to aid clarity we have moved some of this information to an annexe.

We note that different organisations and different sectors may require different processes, whether that be a different number of steps or steps in a different order, to provide an adequate explanation. The steps and order of steps outlined in the guidance were drafted as a reasonable starting point to encourage thought about the explanation to be provided. We have clarified this in the final guidance and renamed the "steps" as "tasks" to make it clearer that they do not necessarily need to be completed in the order listed.

We have tried to include a variety of examples in the text that would demonstrate how decisions made by commonly used AI systems can be explained using the methods outlined in the guidance. We expect that good practice examples will emerge over time as the techniques outlined in the guidance are implemented by organisations.

We appreciate that the guidance in Step 5 could be more detailed. We have expanded this section in the final draft of this guidance to provide additional details on how to train implementers.

We have modified the guidance to include further information about how to mitigate against model drift.

Part 3 – What explaining AI means for your organisation

Several respondents provided suggestions for additional roles that could be included in the guidance. These suggestions included sector-specific roles, such as Caldicott Guardian, which would only be applicable in health and care organisations. Several other suggestions are roles that are covered by the general groups already described. For example "testers" and "model validation team" fit within the "AI development team" role.

There were some concerns raised about the amount of documentation listed within the guidance. Some respondents felt this may lead to "explanation fatigue". However we also received several suggestions for additional documents and policies that may be required. The majority of these suggestions were in reference to auditing requirements.

Our comments

As in other sections, we do not wish to include roles that are sector specific within the guidance, and organisations should contact their sector's regulator if guidance is required in relation to these roles. We are therefore not minded to change the list present within the guidance at this time. We have, however, added more details to the roles to clarify where several teams or individuals may be categorised within one of the roles listed.

The aim of the guidance offered in relation to documentation is not that organisations are required to produce all the documentation listed in every case as it may not all be relevant. We are trying to cover a wide range of issues that might be relevant, so that organisations can select the documentation that would be useful to them.

As the ICO is producing separate guidance in relation to auditing of AI systems, it would not be appropriate to include documents or policies and procedures that may overlap with our auditing guidance. We therefore won't include those suggested additions within the guidance.