Consultation response: Explaining AI decisions

Date: 24 January 2020

Sent to: explain@ico.org.uk

UK Finance is the collective voice for the banking and finance industry.

Representing more than 250 firms across the industry, we act to enhance competitiveness, support customers and facilitate innovation.

Introduction

1. Data protection and privacy are increasingly important areas of regulation, with GDPR focusing minds among firms and growing public awareness. Furthermore, this particular guidance on explaining AI decisions from the ICO covers a new and important area; getting it right will be important to ensuring that socially beneficially innovation is encouraged in the UK economy, with individual rights protected and public trust maintained.
2. The ICO is to be commended for being a global leader and taking on this new and complex regulatory challenge.
3. However, the issue of explaining AI decisions is complex, as is the draft guidance, with very little earlier regulatory guidance to refer back to. It is of interest to individuals from a range of functions within firms; indeed, the guidance itself states that it is for compliance teams, DPOs, technical specialists and management. Furthermore, the fact that the ICO is one of the first regulatory authorities to issue such guidance means it has attracted interest from stakeholders from multiple countries within UK Finance's members.
4. As such, preparing an industry-level response is a lengthy process. Each firm needs to obtain input from numerous internal stakeholders (who do not necessarily specialise in regulatory compliance) and develop an in-house view. In turn, there then needs to be a similar process of discussion and iteration at industry level, discussing the guidance, identifying areas of concern or uncertainty, agreeing on proposed solutions and then having these validated by members.
5. Therefore, in order to ensure that industry is able to prepare well-considered and comprehensive feedback, we strongly recommend that the ICO use a much longer consultation period for future guidance. We note that in financial services, three months is the standard length for a consultation, with two to three months common in other industries such as telecommunications.
6. Context, audience and the materiality of the use case are the key factors that play into the content and delivery of AI decision explanations. Given the wide variety of use cases, industries, model types and the pace of technological change, guidance on AI explainability needs to allow flexibility to firms to tailor and adapt their explanations. Ultimately, explanations need to be useful and meaningful to the individuals they affect and need to be realistic for firms to provide. A principles-based approach, rather than a prescriptive approach, is therefore needed.
7. The guidance contains many positive points, in particular:
    a. The general approach of laying out a range of options and methods for providing explanations and suggesting a range of tools that can be used to increase

interpretability, while providing firms, who are best placed to do so, discretion to determine the approach that will work best.

b. The recognition that interpretability will often not be possible, allowing for other types of explanation that do not rely on interpretability, when appropriate contextual factors allow.

c. The emphasis on the importance of context when designing a suitable explanation.

8. Additionally, the "Principles to Follow" in Part 1 of the draft guidance are an important foundation for the guidance, in particular the following:
    a. "There is no one-size-fits-all approach to explaining AI-assisted decisions."
    b. "Don't just give any explanation to people about AI-enabled decisions - give them a truthful and meaningful explanation, written or presented appropriately, and delivered at the right time"
    c. "When planning on using AI to help make decisions about people, you should consider the setting in which you will do this, the potential impact of the decisions you make, and what an individual should know about a decision"

    These principles affirm the importance of context, audience and materiality as noted above, and that the guidance on explaining AI decisions should first and foremost be flexible because there are far too many variables for a one-size-fits-all approach to work.
9. However, we also wish to set out a number of areas where it is unclear how to interpret the guidance, with potential for it to be read as highly prescriptive or giving rise to unintended consequences. These issues are interconnected but we have broadly grouped them as:
    1. Regulatory coordination
    2. Greater clarity on the status and nature of the guidance;
    3. Applying the guidance to different types of AI 'decision';
    4. Distinguishing between explanations provided to data subjects and supervisors as opposed to processes for internal purposes.
10. The remainder of this response discusses some more detail-focused points, made at the end of this response in section 4, grouped according to the Part of the draft guidance they relate to.

1. Regulatory Coordination

11. As the private and public sector increase their understanding and adoption of AI, it is foreseeable that the issue of explainability will only attract greater attention over time. As more user cases are developed and enter production, there will be an increased understanding and appreciation of the "explainability needs" specific to different industries and use cases. In this eventuality, it is feasible that, given its broad mandate to promote data protection and privacy, the ICO fulfils the role of providing overarching principles and high level guidance that can be readily adapted by sectoral regulators seeking to provide more specific, context- and risk-based, guidance to the firms they oversee. As such, we strongly recommend that the ICO approach the development of its guidance in a way that allows it to coordinate closely with sectoral regulators that are also working on this area or that may seek to do so in the future. In particular, the FCA has announced that it is exploring the issue of transparency and explainability of AI in the financial sector in partnership with the Alan Turing Institute, and plans to publish the outcomes of its research shortly[1]. In order to ensure a coherent approach for UK firms, the work of sectoral regulators should dovetail cleanly with the ICO's overarching principles and guidance.

[1] https://www.fca.org.uk/news/speeches/future-regulation-ai-consumer-good

12. The UK government has taken sound steps towards its goal of establishing the UK as a world leader in AI, including its plan to establish rules and standards that can make the most of AI in a responsible way[2]. At the same time, many countries are vying to put themselves at the forefront of the data and AI revolution and are taking similar steps. For example, the latest action taken by the U.S. is to draft guidance to government agencies with a focus on fostering innovation and promoting leadership, by avoiding actions that could hamper innovation and promoting coordinated regulatory actions.[3] Of note, its principles to favour approaches that maximise net benefits to society and to pursue flexible approaches that can adapt to rapid changes are likely to help the U.S. to realize the economic benefits of technology innovations. It highlights the importance of providing suitable good practice guidance and 'helpful steers' while avoiding unnecessary prescription to help firms provide explanations to data subjects that are meaningful and implement effective internal governance while also being able to innovate.

2. Greater clarity on the status and nature of the guidance

Key points: It should be clarified that:

- *The core purpose of the guidance is to help firms think through potential approaches for complying with GDPR articles 13, 14, 15, 21 and 22 in relation to fully automated decisions using AI, that have a significant or legal effect on the data subject.*
- *The same level of explanation is not required under GDPR for decisions that have a 'human in the loop', but the guidance can nonetheless be used as a resource for firms when building such applications.*

Discussion

13. It is not clear how the ICO anticipates applying the guidance. On page 1 of each Part, the guidance states that it is not a 'statutory code' and sets out 'good practice'. However, in Part 1 there are numerous sections that draw a close connection between complying with the guidance and meeting GDPR requirements. Page 18 of Part 1 also refers to regulatory action, which suggests that ICO might enforce against firms for not following the guidance closely (noting, though, that the guidance also highlights other remedies available before resorting to enforcement).

14. We recognise that the ICO has been called on to develop guidance on explaining 'AI decisions', and that this concept does not fit cleanly within the GDPR framework, which instead refers only to automated decision-making (with significant / legal effects) and to other kinds of processing. This creates a challenge for the ICO in attempting to apply GDPR rules to 'AI decisions'. Given this somewhat 'awkward' fit, we recommend making the status of the guidance as being good practice (rather than statutory guidance) clearer, and also better differentiating between automated decisions caught by Article 22 and other types of AI decisions. Specific suggestions:
    - On pages 8 and 9 of Part 1, the guidance sets out the requirement to provide explanations to data subjects under Articles 13, 14, 15, 21 and 22 (plus Recital 71) for automated decision-making that has a legal or significant effect. Given the up-front status of the guidance as 'good practice', it should be clarified on page 9 that the guidance is intended to help firms think through how to explain AI decisions that are caught under Article 22, such as the provision of 'meaningful information' under Articles 13, 14 and 15.

[2] https://www.weforum.org/agenda/2018/01/theresa-may-davos-address/

[3] https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf

- On a related point, it might also be helpful for the guidance to highlight that the principles set out could also assist firms in thinking through and preparing explanations of decisions based solely on automated processing that does *not* involve AI / ML, as a part of compliance with Articles 13, 14, 15, 21 and 22. What the customer needs / wants to know is likely to depend more on the context / use case (eg: approval for a loan) than on the exact technology used.
- Pages 10 and 11 suggest that there is an *implicit* obligation under GDPR on firms to also explain AI use-cases not caught by Article 22. In particular, from page 11: "So, whichever type of AI-assisted decision you make (involving the use of personal data), data protection law still expects you to explain it to the individuals affected." GDPR sets out specific transparency and explanation rules for processing caught by Article 22 (i.e. fully automated) because it recognized that there are materially different risks involved versus when there is human intervention. Therefore, it is questionable to argue that GDPR requires the same level of explanation for other processing not caught by Article 22.
- This is not to say that transparency and providing suitable explanations of AI- and AI-assisted decisions are not important. However, the discussion of detailed GDPR requirements on pages 8-11 should be amended to make clear that the guidance sets out potential good practice for AI-assisted decisions, in line with the statements on page 1.

3. Application of the guidance to different types of AI 'decision'

Key points:

- *The guidance should explain in more detail what is meant by a 'decision' in scope of the guidance.*
- *The guidance should make clear that firms need to decide whether explanations are best delivered before a decision, after a decision or on request only, according to the circumstances and type of decision.*

Discussion

15. The guidance is framed around the broad principle that decisions relating to an individual should be communicated to them, together with an explanation of that decision, if it produces legal effects or similarly significant effects to them. This is a principle that our members agree with on the whole, but we note that there are some nuances to this principle which should be covered more clearly in the guidance.
16. There are types of AI decisions (or what could arguably be called 'AI decisions') that utilise personal data but which we do not think the ICO intends to be within scope of this guidance. Additionally, it is not entirely clear whether the guidance anticipates firms providing an explanation proactively, or whether this can be provided to data subjects on request. It should be clarified that this should be determined case-by-case by firms, in line with the circumstances and type of decision.
17. For example, in financial services there are large numbers of processes which are becoming increasingly automated. At least for the time being, AI is used primarily in back-office functions, rather than in customer-facing products and services, as was found by a recent FCA survey[1]. In many of these scenarios we do not think there would be any benefit in providing an explanation for every decision, and indeed this could be counterproductive for data subject understanding. These scenarios could include:
    - when the decision is not directly about an individual,
    - when the system makes a high volume of decisions.
18. First, our members use AI tools for risk management and similar applications. Although these kinds of application use personal data to inform a 'decision', the decision affects the

business and macro policies, like the firm's overall risk appetite and the level of exposure to different asset types sought by the firm; the decisions are not in respect of any specific individual. We do not think that the ICO intends for explanations of such 'decisions' to be provided in the manner described in the guidance.

19. **We therefore recommend** that the discussion of 'AI decisions' in Part 1 be amended to clarify that such 'macro decisions' are not in scope. Similarly, it would be helpful to clarify that use of an individual's personal data for the purposes of model training / development does not amount to a 'decision'. (We note of course that GDPR would still require the firm to provide more general information about the personal data collection and processing in both cases).
20. Second, the guidance appears to be primarily aimed at AI decisions that will be made relatively infrequently in an individual's life, such as diagnoses of illnesses or loan applications in a finance context. However, AI systems can also make decisions continuously. For example, many of our members have systems that monitor customer transactions in order to stop fraudulent payments and identify money laundering / terrorist financing, etc. In this case, for every payment / transfer of funds the system 'decides' whether the payment is legitimate, with the potential to freeze the transaction or take other action if there are danger signs. The guidance gives the impression that, if these systems employ AI, the customer should receive an explanation after every payment. Similarly, where AI systems are used to authenticate customers (eg: when logging onto a banking app), a decision about the individual could be made several times a day. It is questionable whether many customers would want to receive a communication about each such decision.
21. **We therefore recommend** that the guidance be amended to make clear that *proactively* providing an explanation is not always required after a decision and that it can be appropriate, depending on the circumstances, to only provide an explanation at the time the customer is onboarded, on request in accordance with Article 22, or potentially only when certain 'unexpected' decisions are made. This clarification would work best in Part 1, page 20 and also Part 2, Steps 6 and 7.
22. (We highlight that this is distinct from the issue around 'gaming' of AI explanations, where explaining antifraud measures could compromise the system. This issue is very important to our members, so the guidance on page 17 of Part 1 is welcome, though we have some comments on this below in section 4 of our response.)

4. Distinguishing between explanations provided to data subjects and supervisors as opposed to processes for internal purposes.

Key points: The guidance should be revised to make it clear that firms can determine what explanation and documentation is appropriate for each use case, and for each key explanation purpose:

- *Communications to customers (or other data subjects) about decisions*
- *Internal use: for model development, governance and controls*
- *Explanations to regulatory authorities by requirement*

The guidance should make clear that firms need to determine what is appropriate for each purpose, in accordance with the GDPR's accountability principle, with the guidance intended to help firms approach this task in a considered and effective manner.

Discussion:

23. As we understand it, the guidance is intended primarily to help firms prepare explanations aimed at data subjects. The guidance also refers to using the explanation process to improve internal governance and documentation (eg: page 11 of Part 1, page 4 of Part 2), but on our reading of the guidance this is secondary, with the *focus* being on providing effective explanations to data subjects. **This interpretation informs our following comments:**

Length of explanations and consumer engagement –

24. Explanations under Step 2 of Part 2 in the guidance risk being extremely long and often not meaningful to data subjects. Much of the content seems to have been prepared with specific scenarios in mind, such as medical diagnosis or use in the judicial system. This is no doubt helpful in those contexts, but some content is not so relevant to other sectors such as financial services. Many of our members are therefore unsure how to apply some sections of the guidance.
25. We understand from the overarching commentary in Part 1 that the guidance is intended to provide options to firms and to allow flexibility as to how best to implement AI explanations, according to the circumstances. However, this is not clear; discussions with members reveal different interpretations of the level of flexibility versus prescription implied by the guidance. Indeed, sections of Part 2 seem to strongly push firms towards providing exhaustive explanations covering all of the potential subject matter for every explanation type, even when there is little relevance or where the information would not be meaningful to individuals.
26. There is helpful guidance in Part 2 to the effect that firms should consider context when deciding which types of explanation need to be prioritised (eg: page 15 and Step 6). However, it is not clear that firms should consider context when deciding which *exact information* should be included in an explanation. The checklist on page 15 of Part 2 states that firms should consider the depth of the explanation but there do not appear to be other statements to this effect in Step 2. Furthermore, the language in Step 2 for each explanation type suggests that an exhaustive approach could be needed. The subheadings are 'What you need to show' and 'Information that goes into this explanation' and are followed by long lists of bullet points. This suggests that explanations ought to cover all of these bullets by default.
27. We are not certain that this is the ICO's intention but, in our view, the draft guidance can be read in this way.
28. Similarly, it should be clarified that firms do not need to prepare / provide explanation types that are not relevant to the use case. For example, providing a safety or fairness explanation might not be relevant to a use case with only limited impact on the data subject. Page 4 of Part 2 seems to say that this is the intended approach at the top of the page, but at the bottom of the page then seems to say that all explanations should nonetheless be provided, just in case. This ambiguity exists throughout the guidance; for example, the first and third bullets of the checklist on page 85 seem to imply that *all* explanations are to be provided.
29. If firms take an exhaustive approach this is likely to lead to very long explanations, which data subjects are unlikely to engage with effectively and are unlikely to find meaningful (even if provided in layers).

Information for internal use as opposed to explanations to data subjects –

30. The lists of information to include for each explanation type in Step 2 are very detailed. We observe that much of the content sets out important questions for firms to consider when they design new processing activities, complete Data Protection Impact Assessments (DPIA) and draft internal documentation as a part of oversight and governance, and to share with regulatory authorities if required.
31. The guidance would also be useful for the construction of explanations in a business-to-business context, where one firm is providing an AI tool to another firm to deploy. This would help the recipient firm to better understand the tool and design data subject-facing explanations. (See also further comments on this in section 4 of our response).
32. However, building on the previous comment, we do not think that it will necessarily be useful to provide all of this information to data subjects. Depending on the use case, we doubt that many individuals will want to read much of this information, particularly the 'process-based' explanations and the information under 'what you need to show' for each explanation type in Step 2. For example:
    - Information to include in a 'rationale explanation' (page 18, Part 2) – "Explain how the procedures you have set up help you provide meaningful explanations of the underlying logic of your AI model's results." Clearly it is important to have confidence in the internal processes that are used to prepare explanations, but a description of this process is unlikely to be meaningful or of interest to data subjects.
    - 'Responsibility explanation' (pages 9-10 and 19-20 of Part 2) – It is clearly very important for firms to understand as a part of their internal governance who is responsible for oversight of the various components of a product or service so that they can ensure appropriate human accountability. However, the guidance seems to suggest a level of individual liability towards data subjects that is not appropriate in many corporate settings. It is the firm that is ultimately responsible for how it treats customers, employees, or other data subjects, not individuals working for the firm. Beyond providing data subjects with information about who they can take a query or challenge to (if applicable), it is unlikely to be appropriate to provide data subjects with detailed information about staff roles and responsibilities within the firm.
    - 'What you need to show' for a 'fairness explanation' (pages 21-23, Part 2) – These are important considerations when designing and implementing an AI tool but it is not clear how these are supposed to be reflected in explanations to data subjects (or indeed why data subjects would want an explanation of internal compliance processes).
    - 'Safety and performance explanation' (pages 24-25, Part 2) – Though again important for internal control we doubt that data subjects want an explanation of information security measures, as is suggested by the 'what you need to show' and 'process' lists. (We would also highlight that the 'gaming' concerns noted at the end of Part 1 are particularly relevant here).

33. It will no doubt be useful for firms to document many of the items and steps set out in the guidance for internal use as a part of compliance and control, and for potential conversations with auditors or supervisors. However, such documents would not normally be written with data subjects as the intended audience. This risks essentially duplicating existing internal policies and documentation.
34. More importantly, there is a risk that data subjects will be confused by overly long explanations, even if provided in layers. There is some recognition on page 16 of Part 1 that excessively long explanations could reduce trust by confusing data subjects, but this section then nonetheless pushes for extensive explanations.

35. Given the potential for complexity in explaining AI decisions and systems, *homing in on the key information* will be vital to ensure data subject understanding and engagement.
36. It is not clear to us whether the ICO actually wants firms to provide all of this information by default, but this should be made clearer.

Emphasis on rationale explanations

37. Page 27 of Part 2 rightly recognises that in some instances it is not possible to prepare a rationale explanation, particularly where the domain the firm is using the AI model in requires the use of unstructured, high-dimensional data. However, throughout much of the rest of Part 2, there is a heavy emphasis on giving a rationale explanation, with little recognition of the limitations to doing so.

In light of the above, we recommend:

- Clarify in Part 1 that the purpose of the guidance is to help firms prepare explanations for data subjects, and also as an input for internal governance and documentation.
- Clarify in Steps 1, 2, 6 and 7 that firms should consider the context in order to identify the *key information* that is most relevant to data subjects and the *depth* of explanation that will be most useful, and should prioritise clearly communicating this (ie: context should factor into not just which explanations to prioritise, but also what points to focus on for each explanation). We note that firms' ongoing experience of dealing with customers / data subjects will help them refine what information is more helpful and meaningful to them.
- Clarify on page 4 of Part 2 that firms do not need to provide explanations that have limited relevance to the use case / context.
- Clarify that a rationale explanation is not necessarily required if the context makes it less relevant or if the use case requires the use of data that does not make this feasible (eg; requires high-dimensional, unstructured data, as per page 27). This is particularly relevant at: page 17, page 4, page 5 and page 9 of Part 2.
- Clarify on pages 4 and 15-17 of Part 2:
    - that the Step 2 descriptions of different explanation types are intended to set out *considerations* for firms to review when preparing their explanations, and do not to set out minimum / exact explanation content, and
    - that *firms can determine the length and depth* of explanations provided, as appropriate for the circumstances.
- Clarify that the points in Step 2 are an input not only for explanations to data subjects but also for internal governance and documentation such as DPIAs, legitimate interest assessments, or assessments of fairness, as appropriate.
- Amend headings throughout Step 2 to reframe the guidance as points for firms to *consider* case by case, rather than an apparent list of *requirements*. For example,
    - Rename the 'What you need to show' headings as 'Consider what data subjects and other stakeholders might need to know, for example...'
    - Rename 'What information goes into this explanation' as 'Consider what information might be needed for this explanation, for example...'
- Consider moving content more relevant to risk assessments, controls and governance than to explanations into the AI Audit guidance, particularly much of the content under the 'What you need to show' and 'process explanation' subheadings, and the 'Responsibility explanation'.
- Add to Step 7 a recognition that firms will need to consider the level of understanding of likely recipients and can provide explanations that do not amount to an exhaustive description when this will better engage data subjects. (For some audiences, providing a small amount of simple, clear information could be preferable to a more complete explanation).

38. 40. We think that these changes will help clarify that firms should consider the context of their explanations and seek to provide the most helpful information to data subjects, while also encouraging good practice in the preparation of internal governance and control documentation.

5. Additional specific comments

Overarching commentary, clarity and readability

- We note that some of the models used as examples of AI in Part 2 are not what would typically be considered 'AI', particularly linear and logistic regression models. It might be preferable to re-title the guidance to be about algorithmic and AI decision-making.
- There seems to be over-use of different terms for similar concepts e.g. unsupervised, fully automated, solely automated, human oversight, supervised, AI-assisted, etc. It is not always clear whether there is an important difference between such similar terms. It would be useful to standardise the language to only use one term for each concept throughout the guidance and explain any differences between similar terms that need to be used.
- Unlike in most ICO guidance, the language in the guidance is in some places quite convoluted or complex (eg: "…because this kind of explanation concedes the opacity of the algorithmic model outright…"). Given the intention for this guidance to be used by individuals with a range of roles within firms (data protection, compliance, model designers, computer scientists, etc) it would be helpful to review the final guidance for readability.
- Trimming out duplicative content could also help readability and facilitate use of the guidance by multiple teams across firms cooperating on AI explainability.
- The guidance calls for a weighing of costs and benefits of less vs more explainable AI. Today firms often weigh the costs and benefits of AI versus the best available non-ML proxy, which can include regressions required for model risk management. It should be clarified that firms do not need to consider all possible AI model types when making these decisions. Ultimately this process would be highly burdensome and inconsistent with model development practices.
- We note that there is necessarily a close connection between explaining decisions and the right to human review in Article 22. We presume that this will be covered in the forthcoming ICO guidance on AI Audit; it will be important to ensure that these two sets of guidance dovetail effectively.
- It is implicit that the guidance is for data controllers and not for data processors, but it would be helpful to make this clear. See also comments below in relation to Part 3.

Relating to Part 1 of the Guidance

- Pages 7 and 8 – Following on from comments made above in section 2 of our response, our understanding is that the 'trigger' for the guidance is not the *use* of AI or the production of an AI output per se, but rather the taking of a decision that is (partly) based on an AI output. It would be helpful to make this more explicit.
- The discussion of 'gaming' risks is particularly important to the financial services sector. We suggest expanding this section to also mention the legitimacy of limiting explanations that risk undermining system / cyber security.
- The discussion of DPIAs should more directly acknowledge that, although a DPIA will often be needed for implementation of an AI tool, this will not always be the case. This would align with the broader ICO guidance on this topic.
- Following on from our comments above under '*3. The scope and purpose of explanations'*: The 'reflect on impacts' principle in Part 1, pages 30-31 goes well beyond explaining the impact of decisions. Although the impact is relevant to the type of explanation, the guidance should seek to limit itself to this, anything else relating to the broader design of models should be in the AI audit guidance, guidance on legitimate interest assessments, etc. We also

highlight that consideration of wide societal impacts is not clearly related to GDPR obligations so guidance on this point should be circumspect.

- The guidance rightly acknowledges that firms might need to limit some elements of their explanations in order to protect their intellectual property. The guidance also states, however, that this is unlikely to often arise, as it is not necessary to reveal source code or algorithmic trade secrets. We highlight that competitors can also engineer a model if they have sufficient information about its logic and features, even if the algorithm / code is not known. The guidance should more concretely recognise that firms will need to manage this risk.

Relating to Part 2 of the Guidance

- Given the length of Part 2, it might help navigation to add a contents page and also to add more links between related sections. Similarly, including the current section title in the header of each page would aid navigation.
- Before starting the 'seven steps', it could be helpful for firms to map out the relevant stakeholders to engage and how they are involved in the process.
- The expectation of firms to provide "reliable and accurate" representations of the behaviour of "black box" models (pages 5 and 31 of Part 2) risks imposing a standard that is difficult to achieve. Such models are so called "black boxes" precisely because they are less interpretable or their inner workings are relatively more opaque due to a high degree of complexity or dimensionality. Given how the supplementary explanation techniques function, it is therefore paradoxical for the expectation to be constrained to "reliable and accurate". We agree that the supplementary explanation techniques are an important method of increasing the interpretability of complex models and can provide a lot of useful information to model developers and reviewers, as well as towards providing a rationale explanation. However, we also agree that there are limitations to these techniques, as confirmed by the statement on page 53 of the draft guidance that states that "these strategies operate as imperfect approximations or as simpler surrogate models". Instead, the language on page 33 that reads "we have prioritised the need for it to provide a reliable, accurate and close approximation of the logic behind our AI system's behaviour" is a much more suitable and achievable standard. We, therefore, recommend that all instances in the guidance that refer to this expectation, including on pages 5 and 31 of Part 2, are revised to include the "close approximation" wording. Page 5, section 3 ('Build your rationale explanation…') should acknowledge that the rationale explanation can include a 'reliable and accurate but not complete representation of the system's behaviour'. This is a key point, which should be recognised in this introductory section, as it is on page 33.
- Page 10, bullet 2 arguably suggests that the choice of model should be based on whether a rationale explanation is required. This should be amended to make clear that there will be a range of points that the firm needs to consider for the use case in question, such as the importance of accuracy.
- Page 16, bullet 3, and page 32 checklist – mitigating risks of discrimination is important of course but removing unjustified biases from the data is not the only way of achieving this. For example, adversarial learning can be used to compensate.[4] Similarly, the important consideration for firms is whether *actions taken* are discriminatory, unfair or unjustified. Business processes can compensate for bias in a dataset, provided it is understood and accounted for. (As per other comments, managing bias issues and ensuring fairness is probably best covered in the AI Audit guidance, rather than guidance on explainability, though the two are connected.)
- Similarly, bullet 4 on page 16 should be amended to recognise that sometimes it can be important to include sensitive personal information in order to test counterfactuals (noting GDPR constraints around the use of special category data). Similarly, page 29 (and

[4] See for example: https://arxiv.org/abs/1801.07593 .

elsewhere) should acknowledge more directly that firms would need to take care to manage the risk that a model could 'triangulate' sensitive characteristics.

- Responsibility explanation (page 20): Challenging a decision – The Guidance states that Individuals in receipt of other explanations, such as rationale or fairness, may wish to challenge the AI decision based on the information provided to them. The responsibility explanation helps by directing the individual to the person or team responsible for carrying a human review of a decision. However, the guidance seems to suggest that the firm should make sure that the team that reviews AI decisions when requested by data subjects includes an individual that helped make the *original* AI-assisted decision. We question whether this is always necessary. For example, if an AI tool holds up a transaction due to concerns that the customer's debit card has been compromised, the customer is unlikely to need to discuss the actual process by which the transaction was held up, provided there is an efficient means for the customer to advise the bank that the transaction was in fact authorised by the customer and was not fraudulent.
- The discussion of trade-offs between accuracy and explainability on pages 35-36 should acknowledge that the assessment of trade-offs will be industry and use-case dependent. (We also note that judging accuracy can be complex and context dependent, for example as relates to the tolerance for 'false positives' versus 'false negatives'; we understand that this issue will be explored in the forthcoming AI Audit guidance).
- The table starting on page 40 should acknowledge that some model types' interpretability depends not just on the type of model but also on whether the data is sparse, the nature of variable interactions, existence of splines, etc.
- 'Good' interpretability would be a useful area of future ICO work, in due course.
- On page 65 the discussion of counterfactual explanations should mention other limitations of this approach. First, the interaction between features, eg where there are non-linear and non-monotonic behaviours, is also a limiting factor, not just the number of variables. Furthermore, firms' ability to explain to individuals how they could change their behaviour to secure a different decision in the future will be constrained by such factors as the preferences of the individual and costs to the individual of taking alternative actions, which will not be known to the firm.
- The 'Implementer training' section on page 77 of Part 2 should be amended to recognise that individuals using AI recommendation tools also risk under-utilising them due to distrust of AI. This point is helpfully made on page 23 of Part 2.
- The descriptions of how to determine feature importance from pages 49 – 70 should acknowledge that the methods discussed can help identify what features are important but not how / why they are important, which will impact the rationale explanation in particular.
- The example on page 75 of Part 2 would be more instructive if the doctor had to disagree with the relevance of certain factors identified by the AI tool.
- Following on from points made above in section 3 of our response, Steps 6 and 7 of the guidance should include the *purpose of the explanation* as a part of the context that will inform the design of the explanation. For example, an explanation to be used in internal control and monitoring would not be designed or delivered in the same manner as an explanation to data subjects or regulatory authorities.
- The guidance refers to using demographic or 'protected characteristic' data several times as a part of system design and governance (eg: pages 16, 23 and 29 or Part 2 and page 18 of Part 3). It would be helpful for the guidance to recognise that GDPR constrains firms' ability to collect or use data in these ways, as it will *often* be 'special category data' under GDPR Article 9. It would also be useful for the forthcoming AI Audit guidance to discuss the use of such data not to make decisions, but to test the accuracy and fairness of models.
- The link at page 68 of the draft Guidance (Part 2) does not work ("Further readings on supplementary techniques").

Relating to Part 3 of the Guidance

- Following on from comments made above in relation to the unclear level of prescription versus flexibility, the Guidance seems to be very prescriptive in terms of what policies should cover, going into a lot of detail. However, in practice firms will have different approaches and policy frameworks, which might not cleanly align with the level of detail in the guidance. The guidance should be amended to recognise directly in Part 3 that firms have flexibility to tailor policies to their own internal governance. This will help ensure that firms can incorporate the relevant points from the guidance without needing to rewrite their existing policies
- We agree that there are different levels of explainability of outputs, depending on different audiences. However, it is not clear whether the Guidance is requiring that those in the AI development teams to also be directly responsible for providing an explanation to individuals. Although their input might be valuable, these teams would not necessarily be *responsible* for explanations.
- Part 3, page 7, provides some guidance on sourcing AI systems from third parties and the ultimate responsibility of the data controller. There are complexities in this area, which could helpfully be explored in more detail; it is likely that there would need to be a degree of shared responsibility for explanations. It would be useful for this section to acknowledge that firms outsourcing in this way may need to rely on representations from the external provider, when appropriate. However, external providers will not always provide a full explanation of their operation of their products, due to IP concerns (this is distinct from the issue of the IP concerns of the data controller that is doing the outsourcing, which are helpfully discussed in Part 1). There are also complexities when the firm using a bought-in AI tool applies it in a manner not originally covered by the design intentions or original technical specifications.

If you have any questions relating to this response, please contact

Walter McCahon
Manager: Data Policy
Walter.mccahon@ukfinance.org.uk

ENDS