Summary of response to the consultation on ICO guidance on the AI auditing framework, with comments

Introduction

In February 2020, the ICO published draft guidance on the AI auditing framework, with an initial deadline of 1 April 2020 for comments. Due to the coronavirus pandemic, this deadline was extended until 1 May 2020.

Our survey asked for:

- feedback on how well pitched each section of the guidance was;
- views on the list of controls organisations could use to mitigate some of the risks AI poses to individual rights;
- practical examples that could further help our thinking, and
- provided an opportunity for respondents to make any further general comments.

The ICO would like to thank all those organisations and individuals who took the time to read the draft guidance and give us their views, and those who offered to work with us further.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 are especially grateful that even in times of a global pandemic, you made time to engage with us on this guidance. 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Overall, we received 65 responses to our draft guidance. 48 respondents followed the survey questions, and 17 sent a standalone document with their feedback to us directly.

Due to the way the survey was run, the largest proportion of responses received was from unknown sectors (30). Of the sources known, the private sector had the largest proportion. The distribution of responses is shown in Figure 1 below.

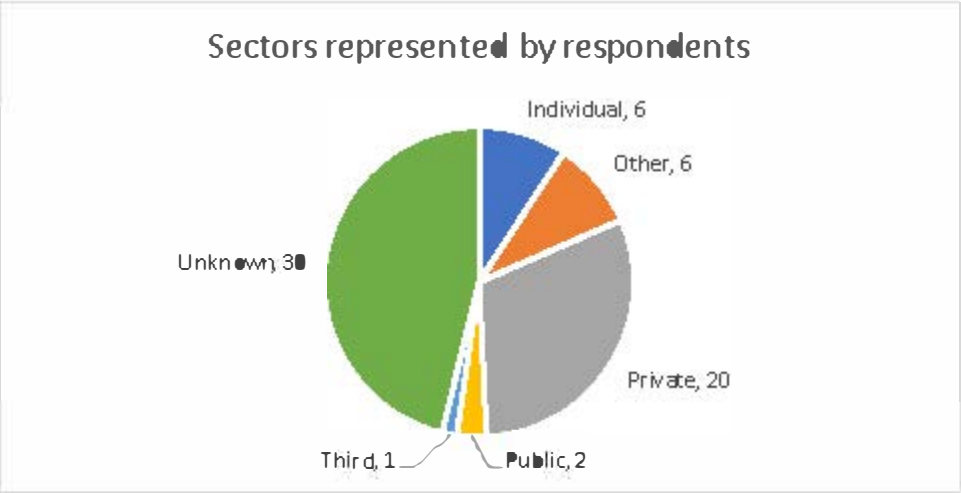

Figure 1: Sectors represented by those that responded to the consultation

Overall, the response to our consultation was generally positive.

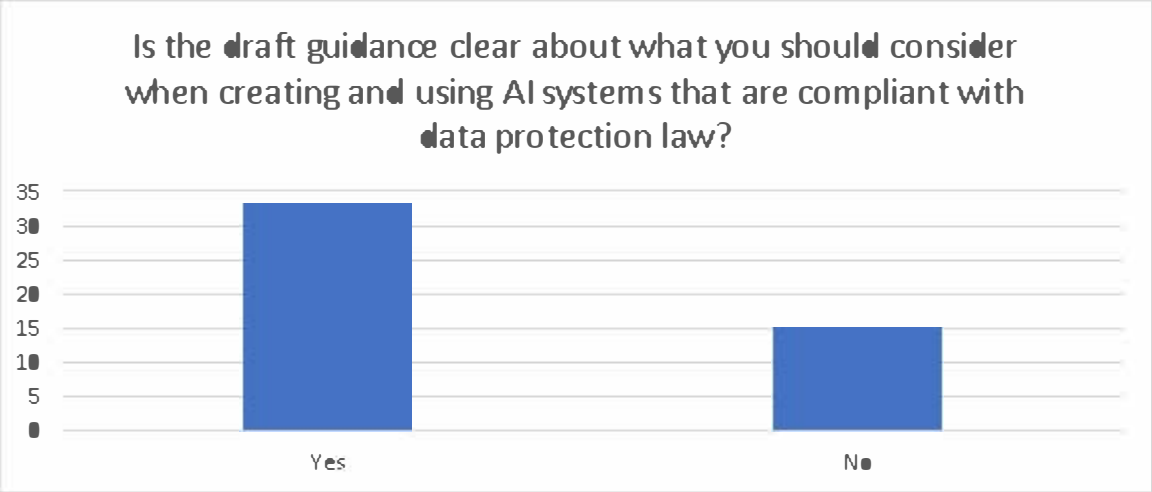

Figure 2: responses to whether the draft guidance was clear

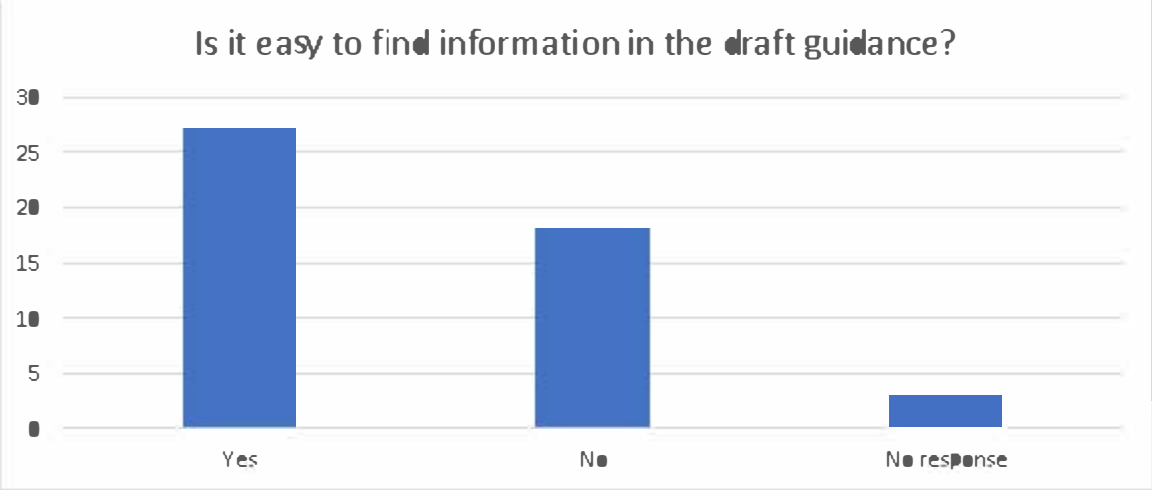

Figure 3: respondents on whether it was easy to find information in the draft guidance

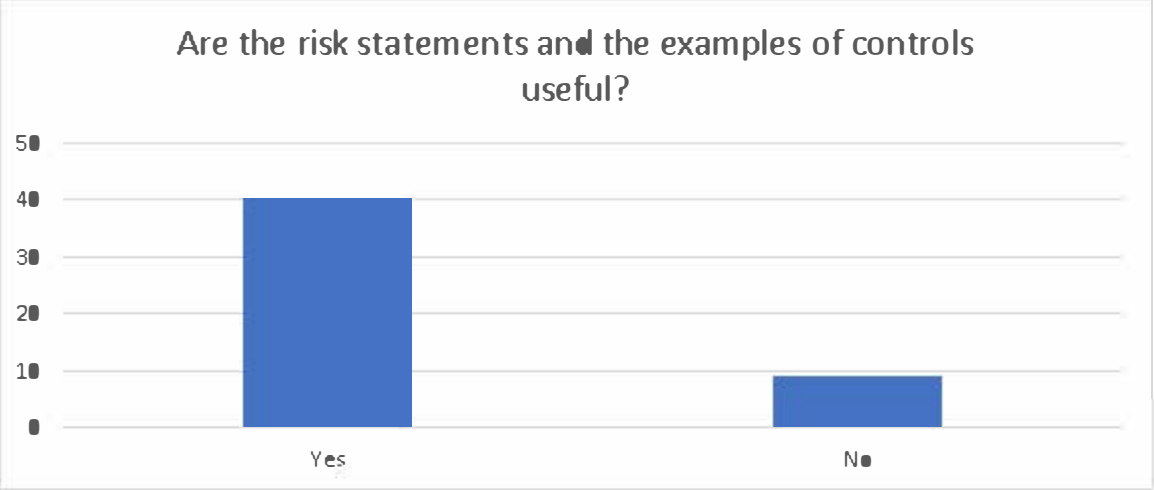

Figure 4: respondents on whether the risk statements and examples of controls were useful

Specific sections received a more mixed response. The quantitative responses to the sections are illustrated below.

Figure 5: responses to how well pitched was the 'About this guidance' section

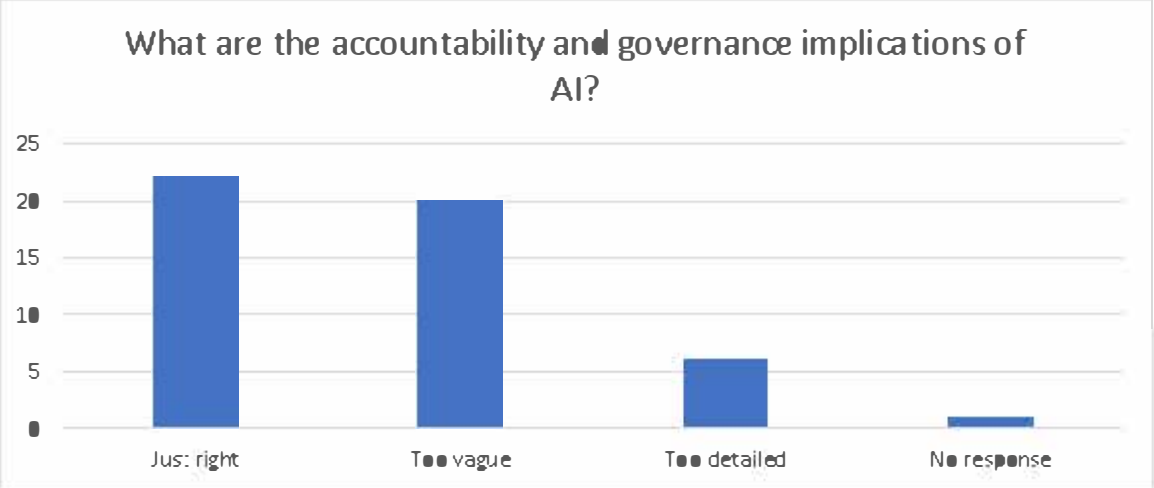

Figure 6: responses to how well pitched was the section on accountability and governance implications of AI

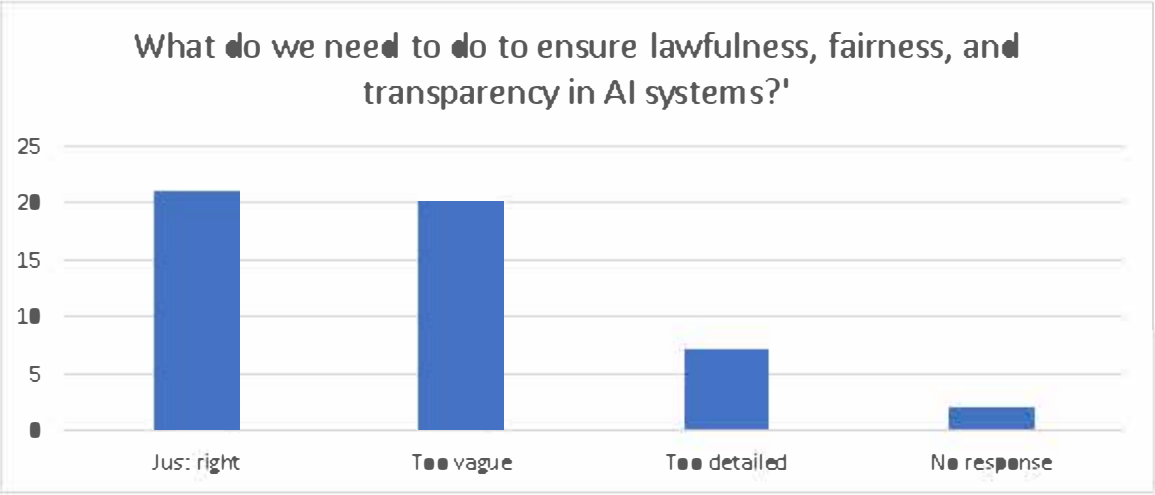

Figure 7: responses to how well pitched was the section on lawfulness, fairness, and transparency in AI systems

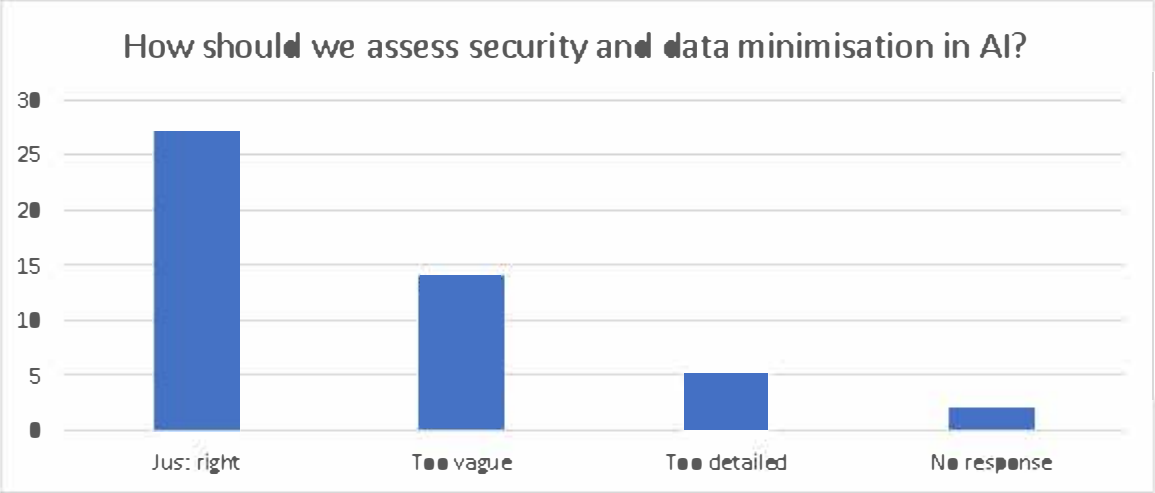

Figure 8: respondents on how well pitched the section on security and data minimisation in AI was

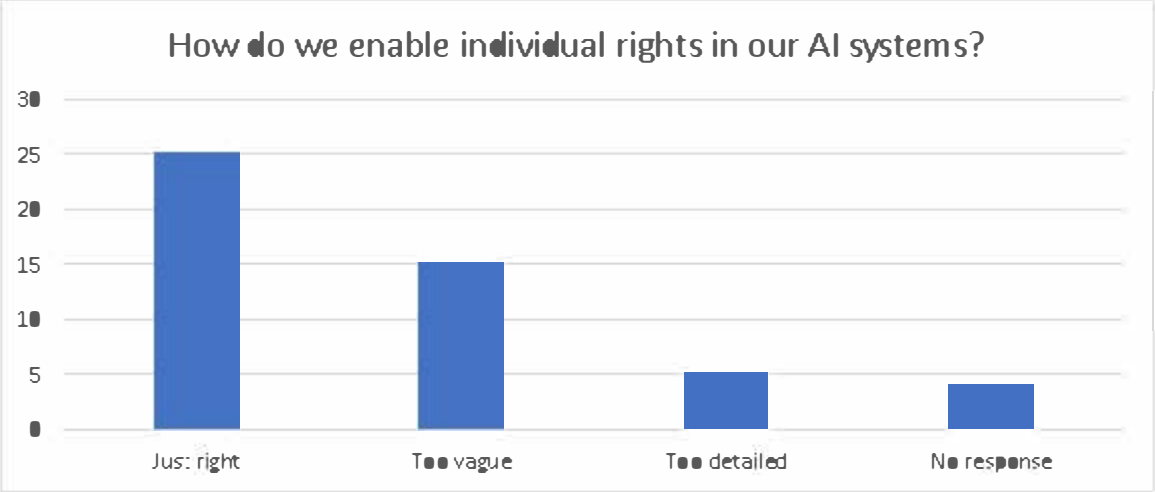

Figure 9: responses to how well pitched the section on individual rights and AI systems was

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

About this guidance

In the draft guidance, we stated that the AI auditing framework will provide tools and resources to assist our investigation and assurance teams when assessing the compliance of organisations using AI. Some respondents queried whether we have the power to 'audit algorithms' themselves.

Some respondents were unsure what the status of the guidance was. They did not know whether it was guidance on best practice, or guidance on interpreting data protection law (or both). There was additional confusion about the guidance being portrayed as best practice but also being used as a standard for our investigation and assurance teams to assess compliance.

Some respondents criticised our definition of 'artificial intelligence'. Some felt it was too broad and included systems that would not ordinarily be referred to as AI.

Finally, there were responses that suggested that more could be done to signpost which sections would require assistance from technology experts to interpret what we say.

Our response

The Data Protection Act 2018 (DPA 18) gives us the power to carry out various auditing and investigation activities. We believe that, in some cases, this includes the recovery and analysis of evidence, including the AI systems themselves.

The guidance informs you about what we think constitutes best practice for data protection-compliant AI as well as how we interpret data protection law as it applies to AI systems that process personal data. We take on board that this could have been made clearer and have taken steps to remedy this in the guidance. For example, we have advised that where we use 'should', you should interpret this as best practice, but where we use 'must', you should interpret this as a legal obligation under data protection law.

We acknowledge that our initial definition of AI could be interpreted broadly. However, neither the General Data Protection Regulation (GDPR) nor the DPA 18 defines 'artificial intelligence', so none of your legal obligations depend on exactly how it is defined. We think it is useful to understand broadly what we mean by AI in the context of the guidance. We have included two definitions in the guidance: one used within the AI research community and one used in the data protection context. We use the umbrella term 'AI' because it has become a standard industry term for a range of technologies, including machine learning.

We take on board that we could have been clearer to signpost sections which would likely need assistance from a technology expert to interpret what we mean and have taken steps to correct this. For example, we have included a subsection at the start of each part where we describe who this is for. We have also added a glossary of some of the technical terms used for reference.

What are the accountability and governance implications of AI?

There were several comments on our guidance about when to carry out a data protection impact assessment (DPIA). Responses indicated that it is not a legal requirement for organisations to carry out a DPIA when AI is being used to process personal data and that the draft guidance misinterpreted the law when it said it is.

Some respondents suggested that some activities we interpreted as typically being carried out by controllers could be interpreted as processor activities.

We received several requests from respondents to include more 'trade-offs' that individuals in compliance focus roles are likely to face. Some felt that we had underplayed the risks associated with commercial sensitivity when discussing the trade-off, it may have with explainability.

Finally, the mathematical solutions to assessing trade-offs that were included in the draft guidance was questioned by several respondents about how practical they would be in a real-word environment.

Our response

We take on board the point that not all applications of AI that processes personal data is likely to result in high risk and therefore trigger the legal requirement to undertake a DPIA. However, we think that in the vast majority of AI systems that process personal data, a DPIA will need to be undertaken. In any case, it is good practice to carry out a DPIA for a new project even if you are not legally required to do so. We recognise that your assessment of whether a DPIA needs to be undertaken will be made on a case by case bases. In those cases where your assessment reaches the conclusion that an AI system is not likely to result in high risk processing, you will still need to document how you came to that decision.

We have also more closely aligned this section with our existing GDPR guidance on determining whether you are a controller or a processor to ensure our approach remains consistent.

Our work has identified that, when AI systems involve a number of organisations in the processing of personal data, assigning the roles of controller and processor can become complex– for instance, when some of the processing happens in the cloud. This raises questions of policy, and we plan to work with Government to explore these areas, with a view to addressing these issues in more detail when we revise our Cloud Computing

Guidance in 2021. This Guidance will also be subject to external stakeholder consultation prior to its finalisation.

We understand that there are likely to be several competing interests that you will have to consider when designing and developing an AI system. Our thinking has developed on this and we now think that 'trade-offs' is a misleading term when talking about these competing interests in a data protection context, as it could imply that you can trade away one legal obligation for another when this is not the case. Instead, we think that it is about striking the right balance between competing interests while ensuring you comply with your obligation under data protection law. In the latest guidance, we have reframed 'trade-offs' as 'competing interests' when talking about them in a data protection context. In addition, we have removed the sections which discuss specific competing interests and have, where appropriate, embedded them in the relevant sections of the guidance.

We agree that the mathematical solutions that we described in the draft guidance were not as helpful as we initially thought. Therefore, we have decided to remove the worked example from the latest guidance to avoid confusion. We have kept a short discussion on mathematical approaches but note that they can be difficult to meaningfully put them into practice.

What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?

We were asked to provide an example of an unfair AI system when discussing the fairness principle.

Some respondents requested for further detail on when legal obligation is likely to be an appropriate lawful basis for processing personal data in the context of an AI system.

We received a suggestion to talk about the appropriateness of the legitimate interests lawful basis in the context of the initial research and development phase of an AI system, where purposes may be quite broad.

There was some useful constructive criticism about our discussion of the fairness principle and where it overlaps with the UK's anti-discrimination legal framework, notably the UK Equality Act 2010 (EA2010). It was noted by some that what we say organisations must or should do to comply with the fairness principle would not necessarily mean that they have complied with their obligations under the EA2010. It was pointed out to us that not all disparities in a data protection context would be instances of discrimination under equalities law. We were made aware that not all cases of discrimination will constitute unlawful discrimination under equalities law.

Some respondents provided further causes of potentially discriminatory AI systems that were not discussed in the draft guidance.

We were made aware of the conflict between algorithmic fairness and relevant non-discrimination law in the UK. For example, some of the techniques we discussed in the context of algorithmic fairness would not mitigate the risks of non-compliance with non-discrimination law.

Several respondents mentioned competing interests when assessing discrimination in an AI system. For example, collecting more data on a minority population to improve the statistical accuracy of the AI system and the risk of non-compliance with the data minimisation principle. Another example provided was where you may have other sector-specific regulatory obligations regarding statistical accuracy or discrimination which need to be considered alongside your data protection obligations.

Finally, some respondents wanted greater clarification about when to identify a lawful basis under Article 9 or 10 and Article 6 in cases where an AI system may infer special category data or criminal convictions data (or both).

Our response

We acknowledge the request for an example of an unfair AI system and have included some in the guidance.

We recognise that that your organisation may be required to audit your AI systems to ensure they are compliant with various legislation (including but not limited to data protection), and this may involve processing of personal data, for instance to test how the system performs on different kinds of people. Such processing could rely on legal obligation as a basis, but this would only cover the auditing and testing of the system, not any other use of that data. You must be able to identify the obligation in question, either by reference to the specific legal provision or else by pointing to an appropriate source of advice or guidance that sets it out clearly. We believe, though, that it is unlikely that it will be necessary to use AI to carry out this obligation.

We take on board the suggestion about legitimate interests as a lawful basis where initial research and development is taking place with broad purposes. We believe that legitimate interests may be an appropriate lawful basis depending on the circumstances, and recommend that, in some cases, as more specific purposes are identified, you review your legitimate interests assessment accordingly (or identify a different lawful basis).

We appreciate the comments and suggestions we received about the UK's anti-discrimination legal framework and how it relates to the fairness principle in data protection law. We have attempted to make it clearer that the guidance is only directed at how to comply with the fairness principle and what best practice looks like. Where it has relevance to the UK's anti-discrimination legal framework, we have noted it. This includes where what we say would not necessarily mean you will comply with anti-discrimination law. We have also made it clearer what our interpretation of the fairness

principle is. This will make clearer where the principle will overlap with anti-discrimination law and where it will diverge.

We also appreciate the suggestions of further causes of potentially discriminatory AI systems and have included some in the guidance.

We agree with the points made about techniques used to ensure algorithmic fairness and the potential conflict with equalities law. We have clarified in the guidance that these techniques may not be suitable to comply with equalities law.

We note the possible competing interests when addressing discrimination in an AI system and have included some words on some of these in the guidance.

We recognise that it may be difficult to identify whether an AI system has accidentally inferred special category data or criminal offence data (or both). If it is unclear whether or not your system may be inferring such data, you may want to identify a condition to cover that possibility and reduce your compliance risk, although this is not a legal requirement. Whether there intent to infer special category data or whether there is a reasonable degree of certainty that you haven't inferred special category data (or both) is also relevant.

How should we assess security and data minimisation in AI?

Several respondents commented that they did not think the guidance on security and AI systems was specific enough to AI. They suggested more discussion about specific security risks that AI creates and exacerbates.

Some respondents suggested that organisations using AI to process personal data should take a more holistic approach when it comes to assessing the security of AI systems. They highlighted that AI systems are just one component of a larger chain of software components, data flows, organisational workflows and business processes.

We were advised to discuss overfitting as being a possible reason you might cite to justify having more data points.

Several responses suggested that it is sometimes difficult to guarantee that no personal data is inadvertently shared. Our draft guidance stated that you remain responsible for ensuring personal data is not exposed. It was felt by some that this was an unreasonable expectation of controllers.

Respondents pointed out that mathematical methods such as differential privacy are not sufficiently mature enough to deploy in a real-life context.

It was suggested to us that the guidance includes information about synthetic data and how it could be used to help compliance with the data protection principle.

Finally, we received some responses that suggested that 'faceprints', in the context of facial recognition models, are not personal data because they are only identifiable to a specific model.

Our response

We appreciate that the section on security could have been more specific to AI. We have removed some more general parts. However, we believe that, overall, this section was applicable enough to AI even if it was also applicable to non-AI systems.

We agree that a holistic approach to security is worthwhile and effective. We have included a line suggesting that organisations follow this approach.

We acknowledge the point about overfitting. We have suggested that overfitting can happen where there are too many features included or where there are too few examples in the training data or both.

We agree with the point made about the difficulties of ensuring that personal data used to train your models is not exposed because of the way your clients have deployed the model. We have rephrased our expectations here to you being responsible for assessing and mitigating the risk that personal data used to train your models may be exposed by your clients deploying the model. By doing this, we remove the expectation of you mitigating the risk completely and instead opt for a more proportionate risk-based approach.

We have caveated the discussion about differential privacy, stating that it may not be appropriate or sufficiently mature to deploy in your particular context. We will continue to monitor developments and update the guidance accordingly.

We have included some information about synthetic data and how it could be used, in some cases, to help you comply with the data minimisation principle. Although, we note that there are risks associated with synthetic data being de-identified and the data not being useful for your purposes. Further guidance on synthetic data will be published in our anonymisation work.

We disagree that 'faceprints', in the context of facial recognition models, are not personal data. We think they are very much identifiable within the context of the specific facial recognition models that they are created for. When used for the purposes of uniquely identifying an individual, they would be special category data under data protection law.

How do we enable individual rights in our AI systems?

We received some feedback that suggested that it may be impossible for controllers to facilitate individual rights because their outsourced services do not get the information or functionality they need.

Several respondents questioned why we had not included guidance on specific individual rights.

We were asked for some clarification about whether individuals needed to be informed if their data was going to be processed to train an AI system if the controller did not know that they were going to use if for this purpose when they first collected it.

Some respondents disagreed that retraining and redeploying a model should not be prohibitively costly. This was in the context of fulfilling requests like erasure or rectification in models that contain personal data by design.

We were asked whether individuals have the right to meaningful information about the logic involved in an AI-driven decision where those decisions are not solely automated or where they do not produce legal or similarly significant effects or both.

Our response

We recognise that if you outsource an AI service to another organisation, this could make the process of responding to rights requests more complicated when the personal data involved is processed by them rather than you. When procuring an AI service, you must choose one which enables individual rights to be protected to meet your obligations as a controller. If your chosen service is not designed to easily facilitate such rights, this does not remove or otherwise change those obligations. If you are operating as a controller, your contract with the processor must stipulate that the processor assist you in responding to rights requests. If you are operating an AI service as a joint controller, you need to decide with your fellow controller(s) who will carry out which obligations.

We acknowledge that we did not discuss all individual rights. These rights are still important, and you must enable them (where applicable). However, we decided to only include rights where AI creates or exacerbates the risk of not enabling them. For guidance on individual rights not included in this guidance, read our Guide to the GDPR.

You must inform individuals if their personal data is going to be used for the purposes of training an AI system, to ensure that processing is fair and transparent. This information should be provided at the point of collection. If the data were initially processed for a different purpose, and you later decide to use it for the separate purpose of training an AI system, you will need to inform the individuals concerned (as well as ensuring the new purpose is compatible with the previous one).

We acknowledge that in some cases it could be prohibitively costly to retrain and redeploy a model that contains data by design. We believe that it will be less costly to retrain and redeploy your AI models accordingly if you have a well-organised model management system and deployment pipeline. However, we recognise that this is may not be a legal requirement.

We have clarified that individuals only have the right to meaningful information about the logic involved in an AI-driven decision where that decision is solely automated and has legal or significantly similar effects.