AI and Ethics (2021) 1:301–310
https://doi.org/10.1007/s43681-021-00039-2

ORIGINAL RESEARCH

AI auditing and impact assessment: according to the UK information commissioner's office

Emre Kazim[1] · **Danielle Mendes Thame Denny**[2] · **Adriano Koshiyama**[1]

Received: 15 November 2020 / Accepted: 9 January 2021 / Published online: 7 February 2021

Abstract
As the use of data and artificial intelligence systems becomes crucial to core services and business, it increasingly demands a multi-stakeholder and complex governance approach. The Information Commissioner's Office's 'Guidance on the AI auditing framework: Draft guidance for consultation' is a move forward in AI governance. The aim of this initiative is toward producing guidance that encompasses both technical (e.g. system impact assessments) and non-engineering (e.g. human oversight) components to governance and represents a significant milestone in the movement towards standardising AI governance. This paper will summarise and critically evaluate the ICO effort and try to anticipate future debates and present some general recommendations.

Keywords Information commission's office · Artificial intelligence · Machine learning · Impact assessment · AI auditing · AI governance · Regulation · Standards

1 Introduction

As the documentary *Social Dilemma* [1] has been recently shown, the use of data by artificial intelligence (AI) systems can have nefarious effects and these effects are increasing at the forefront of the public mind. In the 'age of the algorithm' [2], the decisions that affect the daily lives of the people are being made not by humans, but by mathematical models that instead of bringing greater fairness–for example, by being transparent and eliminating bias—are often opaque, unregulated, and incontestable. These 'weapons of math destruction' [3] need to be curtailed and the efforts of the Information Commissioner's Office (ICO) is to orient an ethical use of AI.

The ICO publications are part of a growing literature concerning the governance of AI systems. Broadly, we can interpret the literature as addressing technical (e.g. system impact assessments) and non-technical (e.g. human oversight) components to governance [4]. The ICO's work leads this conversation by producing guidance that encompasses both components and represents a significant milestone in the movement towards standardising AI governance structures. It has sparked and stimulated a critical debate and is also likely to inform future legalisation in this area.

In addition to the authority and influence of the ICO, it is well placed given its standardisation of Data Protection Impact Assessments (DPIA) [5]. Indeed, in the longer term, we anticipate that that DPIA and the AI auditing and impact assessments will be integrated. The guidance seeks to provide "a solid methodology to audit AI applications and ensure they process personal data fairly, lawfully and transparently" [6, p. 5];

It is aimed at those concerned with compliance and to technology specialists, with risk evaluated in terms of rights and freedoms. The guidance is not a statutory code and is to be read as complementing existing ICO resources.

It is structured according to four questions, as summarised below:

- Part 1: *What are the accountability and governance implications of AI?*

  This draws upon the accountability principle, which is the responsibility to comply with the data protection principles and for demonstrating that compliance.

✉ Emre Kazim
e.kazim@ucl.ac.uk

[1] Computer Science, University College London, London, UK

[2] Technology, Media and Communications Law, Fundação Armando Alvares Penteado, São Paulo, Brazil

- Part 2: *What do we need to do to ensure lawfulness, fairness, and transparency in AI systems?*

  Where discussion is presented regarding the lawful bases for processing personal data in AI systems, assessing and improving AI system performance and mitigating potential discrimination to ensure fair processing.
- Part 3: *How should we assess security and data minimisation in AI?*

  Here, assessing security risks and compliance with the data minimisation principle (identifying the minimum amount of personal data needed to fulfil purpose, and to only process that information, and no more) is explored.
- Part 4: *How do we enable individual rights in our AI systems?*

  Understanding when and how individual rights apply to processed data and implementing effective mechanisms for individuals to exercise those rights.

  Welcoming the ICO's innovative, pertinent and crucial intervention, in this paper, we summarize each of these sections and offer recommendations in each part. We conclude with our *general* suggestion of what we believe will be the fulcrum of future debate.

2 Section summaries and recommendations

2.1 ICO Guidance

The ICO's seeks to provide 'a solid methodology to audit AI applications and ensure they process personal data fairly [6]. It comprises auditing tools and procedures that the ICO will use in its own audits and investigations, including indicative risk and control measures. The two principal audiences are: (i) those with a compliance focus like Data Protection Officers (DPOs), general counsel, risk managers and the ICO's own auditors; and (ii) technology specialists, such as machine learning developers and data scientists, software developers/engineers**,** and cybersecurity and IT risk managers. Risks are stated in terms of impact on rights and freedoms and measures are thereby offered to mitigate risks that AI systems may present within this value framework. Non-technical responsibilities are to structure governance measures, while technical responsibilities concern the 'controller' for the processing. It is clearly stated that the guidance is not a statutory code and is to be read as complementing existing ICO resources. The most relevant UK legislation is identified as the UK Data Protection Act 2018, and internationally, it is the General Data Protection Regulation ((EU) 2016/679 (General Data Protection Regulation) 2016).

Introducing an *executive summary*, and a glossary to compile many technical terms (e.g. privacy, fairness, etc.) outlined across the text would greatly *clarify and standardize* the discussion. Some disambiguation may also be needed (e.g. accuracy—statistical accuracy or accuracy principle). A thoroughly example of a glossary though can be found in the European Commission's 'Ethics Guidelines for Trustworthy AI' [7]. Additionally, the 'risk-based approach' [6, p. 9] also requires clarity; here, 'decision-makers' are referred to in terms of 'reconsidering risk appetite'. However, there is no note on who 'decision-makers' should be or specification of their duties. A *mapping of duties and risks* to responsibilities/roles could clarify this issue. Finally, it is stated that *freedom of information* is not considered in the guidance [6, p. 10]; elsewhere in the guidance, it notes various issues to do with 'rights', such as in the context of explainability or proprietary issues, which fall firmly under the umbrella of freedom of information. Clarifying the scope of the guidance would help in this case.

2.2 What are the accountability and governance implications of AI?

According to the guidance, *accountability* comprises: (i) compliance responsibility; (ii) assessment and mitigation of risks; and (iii) demonstrability of compliance and choices. Compliance responsibility cannot be delegated to data scientists, rather it is for the senior management to ensure this. Importantly, governance and risk management capabilities should be proportionate to the use of AI. The ICO is developing a general accountability toolkit, which can be built upon an approach to AI accountability.

Accessing risk involves:

i determining how to undertake DPIAs for AI systems.
ii identifying whether one is a 'controller' or 'processor' (The guidance does not offer a clear definition of these terms, however, in the context of DPIA, the ICO cites the definitions from GDPR: 'controller' means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data. 'Processor' means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller GDPR Regulation (EU) 2016/679 (General Data Protection Regulation) (n 9)).
iii assessment of risk to rights and freedoms of individuals in design and deployment.
iv method of demonstrating approach taken.

The guidance suggests that 'algorithm impact assessments' can be undertaken in a combination with DPIAs. It is suggested two versions of an assessment can be maintained: (i) a thorough technical description for specialists; and (ii) more high-level description of the processing, explaining how the outputs impact on individuals for broader readership

[6 p. 17]. Here, integration with data stewardship literature would have been useful [6, p.21], by drawing parallels and identifying differences with respect to controller/processor relationships in data protection. Furthermore, we recommend having a few worked-out reports or templates to help gauge the minimum requirements needed for both reports.

Necessity for processing is assessed to show that purposes could not be served in a less intrusive way, like in the DPIA case where demonstration of proportionality is required. Indeed, we welcome this emphasis on proportionality, drawing on lessons from DPIA, which we believe is crucial to ensuring that risk mitigation does not diminish benefits [8]. A score or level should be assigned to each risk measured against likelihood and severity of impact on individuals. Additional mitigating measures can be included along organisational measures already in place (e.g. training). The DPIA should conclude by noting:

i additional measures planned.
ii whether identified risk has been eliminated, reduced, or accepted.
iii overall level of 'residual risk' after measures.
iv whether the ICO needs to be consulted.

The document should be considered 'live' and thereby open to updates. Responsibilities are nuanced and special provision should be made where there are providers of AI development tools and AI prediction services. Also, responsibilities with respect to procuring AI models for local deployment must be cognisant that compliance with data protection law remains with the procuring organisation.

We welcome the clear three-step accountability framework, namely allocation of responsibility, risk assessment and mitigation, and demonstration of compliance, which provides an overarching framework; however, our reading of the guidance is such that fundamental data protection principles are paramount. AI and its trade-offs and competing interesting are secondary [6, p. 12, 14] — we recommend *including more on how these data protection principles can be translated and where appropriate expanded in the context of AI*.

A more general concern we have is what should be assessed in the DPIA. We note that the guidance expresses that considerations are best served if undertaken at the earliest stages of project development [6, p. 16]. The crucial ones regard: (i) "the intended outcomes for individuals or wider society, as well as for you"; and (ii) "an explanation of any relevant variation or margins of error in the performance of the system which may affect the fairness of the personal data processing". In relation to (i) having clear foresight at the earliest stages, before knowing the underlying aspects of the model and the actual environment and feedback from users, may make the technical assessment harder. Concerning, (ii), without having first processed the data by building a few models and trying different modelling pipelines, any discussion about margins of error in the model's performance would be technically challenging.

Management of AI-related trade-offs:

The guidance identifies several AI trade-offs. These are:

i Privacy vs statistical accuracy (here is important to mention that the term 'accuracy' refers to the accuracy principle of data protection law; and 'statistical accuracy' refers to the accuracy of an AI system itself).
ii Statistical accuracy vs discrimination.
iii Explainability vs statistical accuracy.
iv Privacy vs explainability.

With respect to (iii), it is noted that 'black box' models should only be used if: (i) thorough consideration of potential impacts and risk evaluation in advance has been conducted; and (ii) the system includes supplemental interpretability tools that provide domain-appropriate levels of explainability. Moreover, with respect to (iv), the risk of exposure of personal data, and commercial security (revealing of proprietary information) with respect to explainability is also explored: here, the guidance states strongly that their own 'research and stakeholder engagement' indicates that this risk is quite low and that data protection compliance cannot be 'traded away'. But the law can be restrictive, unclear and even paradoxical concerning when the explanation right can be triggered and to what extent computer scientists should be able to design explainable responses [9].

In general, trade-offs should be managed by first identifying them and then considering technical means, lines of accountability, and regular reviews to monitor and control the trade-offs. Documentation should be made available regarding the methodology for identifying and assessing the trade-offs, the Nuffield Foundation prepared a set of examples of practical guides to mediate trade-offs in using AI systems, mitigating the negative societal implications of algorithms, data, and AI [10].

Indeed, such trade-off analysis is likely to be become a critical and vibrant area of academic, policy, and law activity. Trade-off analysis is inherently practical, as such the principal dimensions of analysis will be related to sector specific and use-case specificity. Facial recognition can be taken as an example of this, where fairness (i.e. how the system performs with respect to varying demographics — such as race and gender) may require a trade-off with the ethical concern of privacy (i.e. the anonymisation of demographic data). Here, the logic would be that to track the performance in terms of potential negative biases across demographics, a more 'invasive' (with respect to privacy) approach may be appropriate. Similar trade-off can be explored when comparing, for example, fairness and system accuracy, etc.

Crucially, as noted above, the practical nature of these problems — insofar as they are not theoretical but very much emended in the world with real-world consequences to people — means that best practice is likely to emerge through experience. Notwithstanding this, the argument is not being made that systems should simply be deployed and then retrospectively lessons are to be learned due to the harms, rather the argument is that specificity of sectors and use-cases will necessarily evolve nuances in trade-off analysis that any trade-off schematic could demarcate.

One way in which the trade-offs can be communicated is through visualisation where choices can be plotted on a graph. The guidance notes that this can reveal the 'production-possibility frontier' that may help decision-makers 'understand how system design decisions may impact the balance between different values'. In addition to communication strategies, there are mathematical approaches to minimisation of trade-offs, like Constrained Optimisation and Differential Privacy; however, the guide then asserts that 'it can be hard to meaningfully put them into practice' [6, p. 34].

More clarification is needed regarding how a non-mathematical/engineering intervention might ensure certain parameters are respected and/or trade-offs assessed. The guidance could clarify this point. Moreover, within this context of assessing trade-offs, we found the analysed example on p. 31 unclear. The X and Y axes present numbers that could be difficult to technically calculate, and the charts could be described further.

We believe that good governance of AI systems will require new skill sets and interdisciplinary expertise, as such the call for upskilling and diversity is commendable. However, we note that this may be challenging in a start-up or small-to-medium enterprise — SME environment [6, p. 13]. One suggestion might be to have data scientists and other stakeholders accredited (e.g. Royal Statistical Society accreditation [11]) or associated to a trade/professional association like medical doctors and lawyers in some countries.

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

Three key concepts are introduced and explored in this section: lawfulness, fairness and transparency. In the main text, these concepts are defined as:

- *Lawfulness* – defined in terms of a requirement to 'identifying the purpose of a system and how this relates to law'.
- *Fairness* – defined in terms of the requirement to avoid 'discrimination and consideration of the impact of individuals' reasonable expectations.
- Contrastingly, while the importance of *Transparency* is asserted, it is not immediately defined [6, p. 36].

Lawfulness

With respect to lawfulness, before data processing is performed, it is crucial to know which laws may be applicable depending on the context—e.g. specific laws for recruitment, financial services, etc. The decision should be documented, lawful bases cannot be swapped later (without good reason) and lawful basis should also be included in privacy notices. Also, processing of special categories requires both a lawful basis and an additional condition for processing.

It also distinguishes the purposes between AI development and deployment. This is applied to cases where a system was trained for a specific purpose but then deployed in a different context or repurposed. This is more acute when a system is implemented by a third party, and when processing is undertaken for a different purpose than the one in mind when designed.

We welcome the recommendation that legality be evaluated before and at the start of the development—this will map the legal framework for developers to be cognizant of when they engineer and, thus move more into a legal/ethical-by-design approach. Further, by distinguishing (in terms of legality) development and deployment, the benefit had is research and can be more flexible (typically conducted in the development phase), whereas impact, i.e. deployment, can be more concrete in terms of the mitigation of harms. This way, while risk is mitigated, research and innovation will not be stifled. The European Commission's White Paper on Artificial Intelligence reflects this balance by calling for an ecosystem of excellence and innovation that is trustworthy and safe [12].

The guideline then *outlines some AI-related considerations for each of the GDPR's lawful bases*, noting:

i *Consent*–freely given, unambiguous, and able to withdraw (See, GDPR see Articles 4(11), 6(1)(a) 7, 8, 9(2)(a) and Recitals 32, 38, 40, 42, 43, 171).
ii *Performance of a contract*–using AI is objectively necessary to deliver a contractual service.
iii *Legal obligation, Public task or Vital interests*—such as for detection of a crime, public interest exercised by public authority or emergency diagnosis of patients' unable of providing consent.
iv *Legitimate interests*–requires identification of a legitimate interest (the 'purpose test'); showing the processing is necessary to achieve it (the 'necessity test'); and, balancing against the individual's interests, rights and freedoms (the 'balancing test').
v *Automated decision-making (including profiling)*–consideration of the impact of Article 22 of the GDPR. Requires necessity, legal authorisation and explicit con-

sent, and that individuals are given information about the processing, establishing mechanisms of human intervention or challenging of a decision and carrying out regular checks.

Regarding statistical accuracy, it is noted that an AI system need not be 100% statistically accurate to comply with the accuracy principle. Rather, alternative measures should be used to assess how good a system is. The guidance briefly discusses 'false positive/type I error' and 'false negative/type II error' and argues that a balance needs to be struck between these two sources of errors. Key takeaways are that those responsible are adequately trained, that prediction is clearly labelled as inferences and is not claimed to be factual, and that reasonable expectations are managed.

Within the context of the statistical accuracy measures mentioned [6, p. 48], we recommend that the guidance should go beyond metrics only applicable for classification problems, and could include notes related to common practice used to reliably estimate 'statistical accuracy', such as cross-validation or covariance-penalty methods. We also point that there is an *inconsistent statement* [6, p. 39]. The guidance states that 'even if it is appropriate ground for the use of the system, this may not be an appropriate ground for processing personal data to train an AI system'. Some clarification may be needed, because it is hard to develop a statistically accurate AI system without having to train it first, mainly if it is ML-based [13].

Fairness (bias and discrimination)

Within the UK context, the key legislation around anti-discrimination, is the UK Equality Act 2010 that sits alongside data protection law, applies to a wide range of organisations and gives individuals protection from direct and indirect discrimination, whether generated by a human or an automated decision-making. There are several reasons why an AI system may lead to discrimination, like when building models based on imbalanced training data or when training data reflect past discrimination. But demonstrating that an AI system is not unlawfully discriminatory under the EA2010 is a complex task, and is a separate and additional obligation relating to discrimination under data protection law. The ICO guidelines outline technical means to mitigate such discrimination, relating only to data protection law:

i *Anti-classification*–fairness through exclusion of protected characteristics when making a classification or prediction [14, p. 25].
ii *Outcome/error parity*–fairness through monitoring how members of protected groups are treated comparatively by a model.
iii *Equal calibration*–fairness through ensuring equal calibration between model prediction and the actual likelihood of the event happening.

The guidance explores processing 'special category data', including race, religion or belief, and sexual orientation, to assess and address discrimination in AI systems. In cases where protected characteristics are used to assess and improve potentially discriminatory AI, the following should be considered: (i) whether the new purpose is compatible with the original purpose; and (ii) how new data can be obtained (e.g. issues of consent). Finally, the guidance asserts that profiling with the 'intention of inferring special category data' should be dealt with as a 'special category data irrespective of whether the inferences are incorrect' [6, p. 60].

This surveying of a number of technical approaches to fairness is commended because it is likely that bias and the mitigation of it, will require significant engineering interventions, rather than solely legalistic and broader governance [17]. However, an additional supplement examining 'Proxy variables' [6, p. 54], which is cursorily raised, would be particularly helpful [18], along with more guidance and further references on 'concept drift' [6, p. 17, 18, 49], given the importance these topics.

More concretely, *clarification is needed over technical explanations on bias and discrimination.* The guidance discusses measures to assess (e.g. statistical parity) and mitigate bias (e.g. anti-classification), but later it informs the reader that such metrics conflict with one another [6, p. 55, 56]. The guidance could 'rank' these or provide best practices and current legislation that back some of the metrics sectors, like 4/5 rule in recruitment [19–21]. Another way to present these forms would be to reference academic papers and technical reports where methodologies are outlined, and clear guidance is provided on how to implement these different methodologies [22]. For instance, the incompatibility problem of fairness definitions is a well-documented issue, so an outline followed by a reference list may have been a simpler manner to deal with this issue [23].

Transparency

Building upon a discussion of special category data and discrimination, the guidance explores mitigation of such risks. The overarching theme here is the need for transparency, which is that the purposes 'intention' [6, p. 60] is made clear with respect to why the system is being developed and deployed [24]. This should be done right from the beginning (design phase), with clear policies and good practice regarding procurement and lawful processing of data, including robust testing of any anti-discriminatory measures and monitoring of performance, and with senior management being responsible for signing off the chosen approach. The guidance does not provide a full section on the topic of transparency, as done for lawfulness and bias and discrimination. It may be beneficial to develop this topic in future updates of the guidance.

2.4 How should we assess security and data minimisation in AI?

Two security risks that AI can increase are:

i *Data*–loss or misuse of the large amounts of personal data (unauthorised or unlawful processing, accidental loss, destruction or damage).
ii *Software*–potential for software vulnerabilities introduced as a result of the introduction of new AI-related code and infrastructure.

In this vein, the security requirements are not one size-fits-all but should be directed by specific risks. Context is crucial. *Security in AI differs from 'traditional' technologies* depending on:

i the way the technology is built and deployed.
ii the complexity of the organisation deploying the system.
iii the strength and maturity of the existing risk management capabilities.
iv the nature, scope, context and purposes of the processing.

Actions to be taken include recording and documenting personal data migration (auditing), deleting when the data are no longer needed, applying privacy-enhancing technologies (PETs), and incorporating appropriate security measures when sharing data. Furthermore, mitigation of security risks associated with third-party code by separating, where possible, the ML development environment from the rest of the IT infrastructure; this can be achieved by 'virtual machines' or 'containers' or by deploying coding languages and libraries that are more 'secure'.

The guidance identifies the following types of privacy attacks: (i) model inversion attacks; (ii) membership inference attacks; (iii) black-box and white-box attacks; and (iv) 'invading' models that include training data by design. In response to these, the following needs to be noted:

i *Personal data*–assessing whether the training data contains identifiable personal data.
ii *Skills*–staying up to date with the state of the art in both methods of attack and mitigation.
iii *Overfitting*–addressing cases where a model pays too much attention to the details of the training data, effectively almost remembering examples from the training data rather than just the general patterns.
iv *APIs*–Application Programming Interface (API) would ensure system is not be subject to white-box attacks because the API's users would not have direct access to the model itself.
v *Procurement*–procurement policy should have enough information sharing between each party to perform respective assessments as necessary.
vi *Defence strategies*–robustness methods need to be put in place to protect systems against adversarial cases, where examples are fed into a model to create misclassification and pose risk to rights and freedoms (e.g. facial recognition systems). Real-time monitoring techniques (e.g. rate-limiting) are feasible forms to mitigate security risks.

Implementing these suggestions [6, p. 74] may demand a substantial effort to ensure that the personal data used to train a model are not exposed to a client. This may be harder to implement for bigger companies, and possibly for start-ups and SMEs with small support and teams, resulting in the stifling of innovation in smaller-scale ventures.

Data minimisation and privacy-preserving techniques available for Supervised Learning ML systems

The minimum amount of personal data needed to fulfil purpose should be identified, and 'to only process that information, and no more' (for relevant provisions in the legislation see, Article 5(1)(c) and Recital 39, and Article 16 (right to rectification) and Article 17 (right to erasure) of the GDPR). This means that the first step is to understand and map out all the processes in which personal data may be used, including differentiation between the training phase and the inference phase. There is sound technical guidance on how to perform *data minimisation* in the context of AI systems [25, 26]. We suggest some clarification on this point since the guidance hints that these steps should take place before running some experiments internally [6, p. 77]. Without such knowledge on which features to use and data points to consider, it is going to be challenging to diagnose which parts of the dataset can be excluded.

Minimisation of personal data in the training stage involves:

i Using a variety of *standard feature selection methods* to decide useful attributes for modelling.
ii Privacy-enhancing methods like perturbation, adding 'noise' and federated learning.
iii Differential privacy is mentioned but noted as being challenging to meaningfully implement.

There are several instances [6, p. 73–74, 88–89, 91] where the guidance mentions the deployment of models that include training data by design (e.g. SVMs, KNNs, etc.). As it stands, it could be harder to use these when compared to ones that are only 'parameter-based' (e.g. Neural Networks, Random Forest) [27, 28]. A clear guidance concerning the limitations of these might be welcome given the ongoing scientific effort to research such models.

Minimisation of personal data in the inference stage includes:

i Converting personal data into less 'human readable' formats.
ii Making inferences locally.
iii Privacy-preserving query approaches.

Finally, *anonymisation is considered and data protection law still applied, however, contrastingly* 'anonymous information' is no longer personal data; hence, data protection law does not apply to it.

From this section, we welcome that the guidance notes that AI introduces its own risks. As the guidance draws heavily on data protection provisions, there is a risk that the particular risk introduced by AI systems is overlooked. The emphasis on how AI systems introduce new problems is critical [29, 30].

2.5 How do we enable individual rights in our AI systems?

As personal data are contained in the training data and, in some situations, in the model itself or as an inference from it, the individual rights of information, access, rectification, erasure, and to restriction of processing, data portability, objecting, are applicable at different stages of the AI lifecycle. Following are described ways that personal data are contained in models and procedures to mitigate risks:

i *By design* (e.g. SVMs)–models should be implemented in ways that allows for identification and easy retrieval of such personal data.
ii *Accident* (e.g. leaking)–a regular and proactive evaluation of the possibility of personal data being inferred from models should be followed.

Individual rights relating to solely automated decisions with legal or similar effect can be enabled by including the right to (i) obtain human intervention; (ii) express their point of view; and (iii) contest decisions and obtain explanations. These safeguards cannot be token gestures; for a system to qualify as not solely automated meaningful human intervention is required in every decision. However, it is the case that errors may not be easy for a human oversight to identify, understand and fix.

When recourse is sought, the overturning of a decision may be as a result of (i) an outlier case, where the circumstances are substantially different from those considered in the training data, or (ii) the underlying design assumption are not fit for purpose. *Key steps in facilitating meaningful human review* are (i) considering it in the design phase, like interpretability requirements and user-interface design; and (ii) providing appropriate training and support for human reviewers. The emphasis on the importance of human oversight aligns the guidance with broader calls within the literature for human-centric AI [31].

However, individuals have a right to meaningful information about decision-making, and there may be cases where the system is too complex to explain and thereby contest [32]. To avoid this, and to maintain human oversight, auxiliary systems can be used as decision-support to aid human decision-makers. This contrasts with 'automated decision-making', where the systems make decisions automatically. The guidance notes that there is a GDPR constraint restricting fully automated decisions to a limited lawful basis, while there is a broader scope when systems are used to support decisions.

Importantly, human overview must be active and participatory, i.e. it cannot be a 'rubber-stamping' exercise–indeed the regularity of agreement should be monitored. We welcome this discussion of how a system with human oversight can become effectively solely automated when the human-in-the-loop becomes simply a rubber-stamping exercise. To our knowledge, this is a risk that is not well explored in the literature [33]. More generally, the guidance notes that controls should be in place to keep risks within targets, with processes to swiftly assess compliance and take actions.

Additional risk factors in AI systems

i *Automation bias*–routine reliance on output generated by a decision-support system (effectively rubber-stamping). This risk can be mitigated by training and monitoring of human oversight and design choices (for an acute example of this danger see [34].
ii *Lack of interpretability*–difficult for human reviewer to interpret the decisions being automatically made. Distinguishing solely from non-solely automated AI systems will require senior management review and sign-off. This risk can be mitigated by considering interpretability from the design phase and ensuring human review. More specifically, this involves predicting how outputs change if given different inputs, identifying the most important inputs contributing to outputs, and identifying when the output may be wrong. There are several methods addressing low interpretability, such as 'local' explanations (e.g. Local Interpretable Model-agnostic Explanation), providing an explanation of a specific output rather than the model, and ascribing confidence scores.

Regarding the lack of interpretability, there is the concern about how far an input needs to be explained [6, p. 100–102]. If an input of a model is the prediction coming from another AI system (like using multiple AI 'experts' for a diagnosis or credit checking), should we just provide a general overview

on how it is computed and refer to the technical document about it, or do a thorough presentation of it in the documentation? If there are multiples of it, and they themselves are composed of other predictions, could this high opacity constitute an offense to the right of explanation?

Staff training for human overview should cover:

i how an AI system works and its *limitations.*
ii *anticipating* when a system may be misleading or wrong and why.
iii fostering *scepticism* towards the system's outputs [35]
iv understanding of how their own *expertise should complement the system.*
v *meaningful explanations* for either rejecting or accepting the AI system's output.
vi how that human reviewers could *override* the output.
vii model *monitoring and training* to keep it fit for purpose and in check.

3 Recommendations and future research

In this paper, we summarised each section of the ICO's guidance and offered critical comments and recommendations specific to each section, in short, our general recommendations are:

i *Data and AI:* explicit discussion of the relationship between data protection and the relevant regulatory/standards associated with it and how this translates into auditing of AI systems, i.e. whether the data protection framework is merely applied to AI systems or whether it needs to be adapted/amended. The *GDPR framework of rights is transferred to AI impact*–the concern with this is that it is unclear whether such a framework is necessarily suitable, i.e. do data protection rights and AI impact-related rights parallel one another? An assessment of this would improve the guidance greatly.
ii *Case studies:* templates to help DPOs, and others who have auditing responsibilities to better complete their reporting would be beneficial, this could be achieved by discussion through an open forum for relevant stakeholders helping to build up a good practices repository.
iii *Risks of other machine learning systems:* in addition to the focused guidance on Supervised Learning [27, p. 50], further guidance would be welcome about the unique issues and risks presented in other forms of ML, like Reinforcement [36] and Unsupervised learning [37].
iv *Regression and Forecasting:* beyond addressing the classification problems, the guidance should discuss the metrics and methods used when an AI system is used to tackle a Regression or a Forecasting problem.
v *Target audience:* should be better specified (Data Scientists, DPOs, etc.) or a framework could be created where each group is targeted within a structure that integrates their respective duties.

Notwithstanding, we believe the following will become areas of future research:

i *Legal Status of Algorithms*: we anticipate that the legal status of algorithms will increase in important over the coming years. The guidance is indicative of this emerging debate and its importance, insofar as it distinguishes between responsibilities and obligations of those developing and deploying AI. Themes, such as the nature of legal culpability and even questions of agency and personhood, will be debatable [38].
ii *Sector-specific Standards*: we anticipate that best practice and particularities of sectors will emerge within the literature and wider calls for AI auditing. Metrics and risk assessments will likely be context-dependent and develop auditing frameworks that are path-dependent upon specific sectors.
iii *Integration of data protection and AI*: as noted in this paper, the ICO's guidance draws heavily from data protection measures and frameworks. It is likely that the relationship between data protection and AI ethics will emerge as a contentious issue within the literature. This is because privacy and fairness can be thought of as in direct conflict with one another [39] and therefore demand an integration of data protection and AI impact assessments through a framework of priorities, not by conforming AI to data protection provisions but by implementing a negotiated approach that recognizes the fundamental tension between these two values.

Finally, given that the academic research, regulatory approach, and industry best practice, with respect to the auditing of AI is a nascent area, we believe that the tracking of and comparison of various jurisdictions would be useful. In this article, we have primarily drawn upon the UK context and secondarily upon activity in the European Commission, however, the seemingly ubiquitous adoption of algorithmic systems has resulted in a vibrant debate regarding appropriate mitigation with respect to potential harms in many jurisdictions. Within both the national (UK) and international context, we have witnessed a movement away from abstract statements of ethical and governance principles [40] to more detailed frameworks and practical guides [41]—the quest to achieve trustworthy AI [7] has matured to the point where appropriate governance [42], regulation [12], impact assessment [4] and auditability standards are being proposed and formulated.

Outside of the EU and UK, the Canadian context is particularly active and relatively mature [43]. Indeed, the Canadian Government has published an AI Impact Assessment Toolkit, which offers a systematised approach to AI Auditing through a question flow chart [44]. Although not as advanced as their Canadian counterparts, the United States Congress is similarly debating a proposed 'Algorithmic Accountability Act of 2019′ (116th Congress (2019–2020) [45], and the Australian government's research and survey mention AI Audit within discussions of Australia's AI ethics framework [46].

Indeed, it is clear that satisfaction of a particular standard—e.g. certification, auditability, etc.—will become mandatory (see both international [12] and national signalling by regulatory and standard bodies [47]). As noted above, in this respect, we anticipate that *standards will be both general and sector-specific,* where the former encompasses dimensions, such as privacy, explainability, safety and fairness, and these will be set by institutions and bodies with broader remits (for example, see the German Data Ethics Commissions 'risk-levels' approach [48]); and the latter will address idiosyncrasies of application (e.g. the UK's Financial Conduct Authority will lead standards of AI systems in financial services) [49], see also the Bank of France's review document of AI Governance in the Financial Sector [50]).

4 Conclusion

The ICO guidance here examined in detail is important when AI involves the use of personal data, and therefore it falls inside the remit of data protection law. Very often, AI uses or creates personal data, for instance, when vast amounts of personal data are used to train and test AI models, and upon deployment, when these data are fed through the model to make decisions about individuals, that even if they are only prediction or inferences are themselves personal data.

Besides that, there is an explicit discussion of the relationship between data protection and how it translates into auditing of AI systems. There is a pressing need for case studies to serve as templates development produced through discussion in an open forum for relevant stakeholders. Not only supervised learning present risks, other machine learning systems bring similar and different problems, so there is a need for additional guidance to the deployment of such tools.

And regression and forecasting techniques need to be better assessed in terms of metrics and methods used to analyse potential problems. And target audience need to have a better specification (Data Scientists, DPOs, consumers, or whomever they are) or a framework should be created where each group is targeted within a structure that integrates their respective duties.

Acknowledgements We would like to thank Prof. Philip Treleaven for his valuable supervision and comments on this work.

References

1. Jeff Orlowski.: The Social Dilemma (Exposure Labs, Argent Pictures, The Space Program 2020).
2. Serge Abiteboul., Gilles Dowek.: The age of algorithms (K Rae Nelson tr, 1a Edição, Cambridge University Press 2020).
3. Cathy O'Neil.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (1 edition, Crown 2016).
4. Ada Lovelace Institute and DataKind UK, 'Examining the Black Box: Tools for Assessing Algorithmic Systems.' (Ada Lovelace Institute and DataKind UK 2020) https://www.adalovelaceinstitute.org/examining-the-black-box-tools-for-assessing-algorithmic-systems/ accessed 8 June 2020.
5. Stefan Schiffner., and others (eds).: Privacy Technologies and Policy: 4th Annual Privacy Forum, APF 2016, Frankfurt/Main, Germany, September 7–8, 2016, Proceedings, **9857** (Springer International Publishing 2016) https://doi.org/10.1007/978-3-319-44760-5 accessed 8 June 2020.
6. Information Commissioner's Office ICO, 'Guidance on the AI Auditing Framework. ICO Draft Guidance for Consultation' (ICO Information Commissioner's Office 2020) https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf
7. European Commission, 'Ethics Guidelines for Trustworthy AI' (Shaping Europe's digital future - European Commission, 8 April 2019) https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai accessed 9 June 2020.
8. Luciano Floridi., Josh Cowls.: 'A Unified Framework of Five Principles for AI in Society' (2019) 1 Harvard Data Science Review https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/6 accessed 9 June 2020.
9. Lilian Edwards., Michael Veale.: 'Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 18.
10. Jess Whittlestone., and others.: 'Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research' 59.
11. Royal Statistical Society, 'Accreditation Scheme' (2020) https://rss.org.uk/membership/professional-development/accreditation-scheme/ accessed 12 June 2020.
12. Adriano Koshiyama., Emre Kazim.: 'Lack of Vision: A Comment on the EU's White Paper on Artificial Intelligence' (ResearchGate, 2019) https://www.researchgate.net/publication/340714178_Lack_of_Vision_A_Comment_on_the_EU's_White_Paper_on_Artificial_Intelligence accessed 12 June 2020; European

Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' (European Commission - European Commission, 2020) https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en accessed 12 June 2020.
13. Stuart Russell., Peter Norvig.: Artificial Intelligence: A Modern Approach (Pearson 2020) http://aima.cs.berkeley.edu/index.html.
14. Sam Corbett-Davies., Sharad Goel.: 'The measure and mismeasure of fairness: a critical review of fair machine learning'.
15. Ninareh Mehrabi., others.: 'A Survey on Bias and Fairness in Machine Learning' [2019]] http://arxiv.org/abs/1908.09635 accessed 12 June 2020.
16. Geoff Pleiss., others.: 'On Fairness and Calibration' (2017) http://arxiv.org/abs/1709.02012 accessed 12 June 2020.
17. Luca Oneto., Silvia Chiappa.: 'Fairness in Machine Learning' in Luca Oneto and others (eds), Recent Trends in Learning From Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019) (Springer International Publishing 2020) https://doi.org/10.1007/978-3-030-43883-8_7 accessed 12 June 2020.
18. Robert H Sloan., Richard Warner.: 'The Proxy Problem: Fairness and Artificial Intelligence' (Social Science Research Network 2019) SSRN Scholarly Paper ID 3441888 https://papers.ssrn.com/abstract=3441888 accessed 12 June 2020.
19. 'U.S. Equal Employment Opportunity Commission' (1978) https://www.eeoc.gov/ accessed 12 June 2020;
20. Barbara Lerner, 'Employment Discrimination: Adverse Impact, Validity, and Equality' (1979) 1979 The Supreme Court Review 17.
21. Richard S Barrett.: Challenging the Myths of Fair Employment Practices (Quorum 1998); Biddle, 'Uniform Employee Selection Guidelines on Employee Selection Procedures' (2018) http://www.uniformguidelines.com/ accessed 12 June 2020.
22. Timnit Gebru., and others.: 'Datasheets for Datasets' https://arxiv.org/abs/1803.09010v7 accessed 12 June 2020; Miles Brundage and others, 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims' https://arxiv.org/abs/2004.07213v2 accessed 12 June 2020.
23. Alexandra Chouldechova.: 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments' (2017) 5 Big Data 153.
24. Patrick Van Esch., and others.: AI-enabled biometrics in recruiting: Insights from marketers for managers, Australasian Marketing Journal (2020).
25. Cynthia Dwork., Aaron Roth.: 'The algorithmic foundations of differential privacy' (2014) https://www.cis.upenn.edu/~aaroth/privacybook.html accessed 12 June 2020
26. Information Commissioner's Office ICO, 'Data Minimisation and Privacy-Preserving Techniques in AI Systems' (1 November 2019) https://ico.org.uk/about-the-ico/news-and-events/ai-blog-data-minimisation-and-privacy-preserving-techniques-in-ai-systems/ accessed 12 June 2020.
27. Trevor Hastie., Robert Tibshirani., Jerome Friedman.: The elements of statistical learning: data mining, inference, and prediction, second edition (2nd edition, Springer 2016);
28. Trevor Hastie Bradley Efron.: Computer Age Statistical Inference (Cambridge University Press 2016).
29. Nick Bostrom.: 'Existential Risk Prevention as Global Priority' (2013) 4 Global Policy 15.
30. Stuart Russell., John Bohannon.: 'Artificial Intelligence. Fears of an AI Pioneer' (2015) 349 Science (New York, N.Y.) 252.
31. Paul Lukowicz.: 'The Challenge of Human Centric AI' (2019) 3 Digitale Welt 9.
32. Information Commissioner's Office ICO and The Alan Turing Institute, 'Explaining Decisions Made with AI' (ICO 2020) https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence/ accessed 2 October 2020.
33. Lorrie Faith Cranor.: 'A Framework for Reasoning about the Human in the Loop', Proceedings of the 1st Conference on Usability, Psychology, and Security (USENIX Association 2008).
34. Patrick Van Esch and others: Autonomous weapon systems: is a space warfare manual required? Computer Law & Security Review **33**(3), 382–389 (2017)
35. Patrick Van Esch., and others.: Job candidates' reactions to AI-enabled job application processes. AI and Ethics (2020)
36. Richard S Sutton., Andrew G Barto., Francis Bach.: Reinforcement Learning: An Introduction (second edition edition, MIT Press 2018).
37. Zoubin Ghahramani.:'Unsupervised Learning' in Olivier Bousquet, Ulrike von Luxburg and Gunnar Rätsch (eds), Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures (Springer 2004) https://doi.org/10.1007/978-3-540-28650-9_5 accessed 12 June 2020.
38. Philip Treleaven., Jeremy Barnett., Adriano Koshiyama.: 'Algorithms: Law and Regulation' (2019) 52 Computer 32.
39. Emre Kazim., Adriano Koshiyama.: "The interrelation between data and AI ethics in the context of impact assessments." AI and Ethics 1–7 (2020).
40. Anna Jobin., and others.: The global landscape of AI ethics guidelines. Nature Machine Intelligence, **1**(9), 389-399 (2019)
41. Leslie, D.: Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute (2019). https://doi.org/10.5281/zenodo.3240529
42. AI Governance Database, Nesta https://www.nesta.org.uk/data-visualisation-and-interactive/ai-governance-database/ (2020)
43 Yves Genest.: Artificial Intelligence and Auditing: Overview of Potential Impact on Public Sector Auditors (October 18, 2019), https://caaf-fcar.ca/en/performance-audit/research-and-methodology/research-highlights/3455-research-highlights-3 accessed 5 January 2021
44. Government of Canada, Algorithmic Impact Assessment, https://canada-ca.github.io/aia-eia-js/ accessed 5 January 2021
45. H.R.2231 - Algorithmic Accountability Act of 2019, 116th Congress (2019–2020) Sponsor: Rep. Clarke, Yvette D. [D-NY-9] (Introduced 04/10/2019). Committees: House - Energy and Commerce
46. Dave Dawson., and others.: Artificial Intelligence: Australia's Ethics Framework. Data61 CSIRO, Australia (2019).
47. Review into bias in algorithmic Decision-making, CDEI (2020). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/939109/CDEI_review_into_bias_in_algorithmic_decision-making.pdf.
48. Opinion of the Data Ethics Commission, Daten Ethik Kommission (Germany) (2020). https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.html
49. AI transparency in financial services – why, what, who and when?, FCA (2020) https://www.fca.org.uk/insight/ai-transparency-financial-services-why-what-who-and-when
50. Laurent Dupont., Olivier Fliche., Su Yang.: Gouvernance des algorithms d'intelligence artificielle dans le secteur financier: Document de réflexion (June 2020), Pôle Fintech-Innovation, ACPR.