A Review of the UK-ICO's Draft Guidance on the AI Auditing Framework

Emre Kazim [*1] **Adriano Koshiyama** [*1]

Abstract

The Information Commissioner's Office (ICO) timely call for consultation regarding their 'Guidance on the AI auditing framework: Draft guidance for consultation' (February 2020) is part of a growing literature concerning the governance of Artificial Intelligence (AI) systems. The ICO's draft leads the UK's national conversation by producing guidance that encompasses both technical (ex. system impact assessments) and non-technical (ex. human oversight) components to governance and represents a significant milestone in the movement towards standardising AI governance. Welcoming this crucial intervention, we summarise and critically evaluated each section of the draft guidance, offering feed-back in line with the call for consultation. We conclude with a note on what we anticipate will be future debates and by presenting our general recommendations.

1. Introduction

The United Kingdom Information Commissioner's Office (ICO) timely publication 'Guidance on the AI auditing framework: Draft guidance for consultation' (February 2020) (UK-ICO, 2020a) is part of a growing literature concerning the governance of Artificial Intelligence (AI) systems. Broadly, we can interpret the literature as addressing technical (ex. system impact assessments) and non-technical (ex. human oversight) components to governance (Ada Lovelace & DataKind UK, 2020). The ICO's draft guidance leads the national conversation by producing guidance that encompasses both components and represents a significant milestone in the movement towards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) (Bieker et al., 2016). Indeed, in the longer term we anticipate that that DPIA and AI impact assessments will be integrated.

The guidance seeks to provide 'a solid methodology to audit [...] and ensure they process personal data fairly'. 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. The draft guidance is structured according to four questions, each of which we comment upon below. We conclude with our general recommendations and areas we believe will be the fulcrum of future debate.

2. Section Summaries and Recommendations

2.1. ICO Guidance

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 thorough example of a glossary is found in (HLEG AI, 2019). Additionally, the 'risk-based approach' (UK-ICO, 2020a; 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 (UK-ICO, 2020a; p. 10); else-where the guidance 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.

2.2. Part 1: What are the accountability and governance implications of AI?

According to the guidance, governance and risk management should be proportionate to the use of AI. We welcome emphasis on proportionality, drawing on lessons from DPIA, which we believe is crucial to ensuring that risk mitigation does not diminish benefits (Floridi & Cowls, 2019).

It is suggested two versions of an assessment:

[*]Equal contribution [1]Department of Computer Science, University Collehe London, London, UK. Correspondence to: Emre Kazim <ekazim@cs.ucl.ac.uk>.

• a **thorough technical description** for specialists; and

• more **high-level description of the processing** (UK-ICO, 2020a; p. 17). Here integration with data stewardship literature would have been useful (UK-ICO, 2020a; p. 21). Furthermore, we recommend having a few worked out reports or templates to help gauge the minimum requirements needed for both reports.

We welcome the clear three step accountability framework, namely allocation of responsibility, risk assessment and mitigation, and demonstration of compliance; 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 (UK-ICO, 2020a; p. 12, p.14) – we recommend including more on how data protection principles can be translated 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 (UK-ICO, 2020a; p. 16). The crucial ones regard:

i. "the in-tended 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: In general, trade-offs should be managed by first identifying them and then considering technical means, lines of accountability, and regular reviews for monitoring and control. Documentation should be made available regarding the methodology for identifying and assessing trade-offs (for examples, see (Whittlestone et al., 2019). Further clarification is needed regarding how a non-mathematical/engineering intervention might ensure certain parameters are respected and/or trade-offs assessed. Moreover, within this context of assessing trade-offs we found the worked 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 SME environment (UK-ICO, 2020a; p. 13). One suggestion might be to have data scientists and other stakeholders accredited/associated (e.g. Royal Statistical Society accreditation (UK-RSS) to a trade/professional association like medical doctors and lawyers in some countries.

2.3. Part 2: 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.

Lawfulness: Before data processing is performed it is crucial to know which laws are applicable. The guidance distinguishes the purposes between development and deployment (an issue particularly acute when a system is implemented by a third party). We welcome mapping legality at the start of the development in order for developers to move more into a legal/ethical-by-design approach (European-Commission, 2020; Kazim & Koshiyama, 2020). The guideline then outlines some AI-related considerations for each of the GDPR's lawful bases. Within the context of the statistical accuracy measures mentioned (UK-ICO, 2020a; 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 note that within this section there is an inconsistent statement (UK-ICO, 2020a; p. 39) regarding when testing should be done: clarification is needed (it is hard to develop a statistically accurate system without first training, mainly if it is ML-based) (Norvig & Russell, 2002).

Fairness: There are several reasons why an AI system may lead to discrimination and technical means to mitigate such discrimination currently exist (Corbett-Davies & Goel, 2018; Mehrabi et al., 2019; Pleiss et al., 2017). The surveying of technical approaches 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 (Oneto & Chiappa, 2020). However, an additional supplement examining 'Proxy variables' (UK-ICO, 2020a; p. 54), which is cursorily raised, would be particularly helpful (Warner & Sloan, 2019), along with more guidance/further references on 'concept drift' (UK-

ICO, 2020a; p. 49, 17, 18).

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 (UK-ICO, 2020a; p.55-56). The guidance could 'rank' these or provide best practices (Lerner, 1979; US Equal Employment Opportunity Commission, 2002; Bernardin et al., 1980; Barrett, 1998). 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 (Brundage et al., 2020; Gebru et al., 2018; Chouldechova, 2016).

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' p.60) is made clear with respect to why the system is being developed and deployed. 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. Part 3: How should we assess security and data minimisation in AI?

Security requirements are not one size-fits-all but should be directed by specific risks.

Data minimisation and privacy-preserving techniques: There is sound technical guidance on how to perform data minimisation in the con-text of AI systems (UK-ICO, 2020b; Dwork et al., 2014) We suggest some clarification on this point since the guidance hints that these steps should take place before running some experiments internally (UK-ICO, 2020a; 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: There are several in-stances (UK-ICO, 2020a; 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) (Friedman et al., 2001; Efron & Hastie, 2016). A clear guidance concerning the limitations of these might be welcome given the ongoing scientific effort to research such models.

From this section we welcome that the guidance notes that AI introduces its own risks, as the drawing heavily on data protection provisions may mean overlooking risks particular to AI systems (Russell & Bohannon, 2015; Bostrom, 2013).

2.5. Part 4: How do we enable individual rights in our AI systems?

As personal data is 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. The following are described ways that personal data is contained in models and procedures to mitigate risks:

• **By design (e.g. SVMs)** – models should be implemented in ways that allows for identification and easy retrieval of such personal data

• **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 automated decisions with legal or similar effect can be enabled by including the right to:

• **obtain human intervention**;

• **express their point of view**; and

• **contest decisions and obtain explanations**.

These safe-guards 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;

• an **outlier case**, where the circumstances are substantially different from those considered in the training data, or

• 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 (Lukowicz, 2019).

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. To

maintain human oversight, auxiliary systems can be used as decision-support to aide 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 sup-port 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 (Cranor, 2008). More generally the guidance notes that controls should be in place to keep risks within targets, with processes to swiftly act and assess compliance.

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.

ii. **Lack of interpretability** – difficult for human reviewer to interpret the decisions being automatically made.

Distinguishing solely from non-solely auto-mated 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 ii. there is a concern regarding how far an input needs to be explained (UK-ICO, 2020a; 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?

3. Summary

Additional specific critical comments/recommendations, our general recommendations are:

• **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 or whether it needs to be adapted/amended. As a corollary to this, in Part 4, 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.

• **Case studies**: templates to help DPOs, etc. with their reporting would be beneficial, where this could be achieved by discussion through an open forum for relevant stakeholders helping to build up a repository.

• **Risks of other Machine Learning (ML) systems**: in addition to the focused guidance on Supervised Learning (Friedman et al., 2001) further guidance would be welcome about the unique issues and risks presented in other forms of ML, like Reinforcement (Sutton & Barto, 2018) and Unsupervised learning (Ghahramani, 2003).

• **Regression and Forecasting**: beyond addressing 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.

• **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.

Future research will cover these areas:

• **Legal Status of Algorithms**: we anticipate that the legal status of algorithms will increase in important over the coming years. Themes such as the nature of legal culpability and even questions of agency and personhood will be debates (Treleaven et al., 2019);

• **Sector Specific Standards**: we anticipate that best practice and particularities of sectors will emerge within the literature and wider calls for AI auditing; and

• **Integration of data protection and AI**: as noted, 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 emerging as a contentious issue within the literature.

References

Ada Lovelace and DataKind UK. Examining the black box: Tools for assessing algorithmic systems. Technical report, AdaLovelace Institute, https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-

auditing-framework-draft-for-consultation.pdf, 2020.

Barrett, R. S. *Challenging the myths of fair employment practices.* Quorum Books/Greenwood Publishing Group, 1998.

Bernardin, H. J., Beatty, R. W., and Jensen JR, W. The new uniform guidelines on employee selection procedures in the context of university personnel decisions. *Personnel Psychology*, 33(2):301–316, 1980.

Bieker, F., Friedewald, M., Hansen, M., Obersteller, H., and Rost, M. A process for data protection impact assessment under the european general data protection regulation. In *Annual Privacy Forum*, pp. 21–37. Springer, 2016.

Bostrom, N. Existential risk prevention as global priority. *Global Policy*, 4(1):15–31, 2013.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., et al. Toward Trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, arxiv pre-print. 2016.

Corbett-Davies, S. and Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

Cranor, L. F. A framework for reasoning about the human in the loop. 2008.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Efron, B. and Hastie, T. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

European-Commission. White paper on artificial intelligence–a european approach to excellence and trust. 2020.

Floridi, L. and Cowls, J. A unified framework of five principles for ai in society. *Harvard Data Science Review*, 2019.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., and Crawford, K. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

Ghahramani, Z. Unsupervised learning. pp. 72–112, 2003.

HLEG AI. Ethics guidelines for Trustworthy AI. *B-1049 Brussels*, 2019.

Kazim, E. and Koshiyama, A. Lack of vision: A comment on the EU's White Paper on Artificial Intelligence. *Available at SSRN 3558279*, 2020.

Lerner, B. Employment discrimination: Adverse impact, validity, and equality. *The Supreme Court Review*, 1979: 17–49, 1979.

Lukowicz, P. The challenge of Human Centric AI, 2019.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Norvig, P. and Russell, S. *Artificial Intelligence: a modern approach*. Prentice Hall, 2002.

Oneto, L. and Chiappa, S. Fairness in machine learning. In *Recent Trends in Learning From Data*, pp. 155–196. Springer, 2020.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689, 2017.

Russell, S. and Bohannon, J. Artificial Intelligence: Fears of an AI pioneer. *Science (New York, NY)*, 349(6245): 252–252, 2015.

Sutton, R. S. and Barto, A. G. (eds.). *Reinforcement learning: An introduction.* MIT Press, Palo Alto, CA, 2018.

Treleaven, P., Barnett, J., and Koshiyama, A. Algorithms: law and regulation. *Computer*, 52(2):32–40, 2019.

UK-ICO. Guidance on the ai auditing framework: Draft guidance for consultation. https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf. Technical report, Information Commissioner's Office, United Kingdom, 2020a.

UK-ICO. Data minimisation and privacy-preserving techniques in ai systems (accessed: 07/06/2020). https://ico.org.uk/about-the-ico/news-and-events/ai-blog-data-minimisation-and-privacy-preserving-techniques-in-ai-systems/, 2020b.

UK-RSS. Accreditation scheme (accessed: 07.06.2020). https://rss.org.uk/membership/professional-development/accreditation-scheme/.

US Equal Employment Opportunity Commission. Enforcement guidance: Reasonable accommodation and undue hardship under the americans with disabilities act. 2002.

Warner, R. and Sloan, R. H. The proxy problem: Fairness and artificial intelligence. *Available at SSRN 3441888*, 2019.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., and Cave, S. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. *London: Nuffield Foundation*, 2019.