Written Evidence on ‘Establishing a pro-innovation approach to regulating AI’

1. Executive Summary

This report has been prepared by Prof. Brent Mittelstadt, Prof. Sandra Wachter and Mr. Rory Gillis, drawing on prior work of the Oxford Internet Institute’s Governance of Emerging Technologies research programme. It is a response to the UK Government’s July 2022 policy paper on ‘Establishing a pro-innovation approach to regulating AI’, and was submitted in September 2022.

Based on our research in this area, we make a number of key recommendations regarding the government’s approach:

- Prioritise good everyday explanations over technical explanations when approaching the principle of transparency.
- Review existing definitions of fairness in UK law and ensure upcoming AI regulations harmonise with them.
- Consider adding privacy as a new cross-sectoral principle, defined in a manner that helps to protect individuals from unreasonable inferences and discrimination.

In addition, we make the following recommendations regarding implementation:

- Consider counterfactual explanations as a means of implementing the transparency cross-sectoral principle.
- Consider ‘Conditional Demographic Disparity’ as a means of implementing the fairness cross-sectoral principle.
- The fairness principle should ensure that AI systems actively help to tackle relevant inequalities.

2. Background

The Governance of Emerging Technologies (GET) research programme at the Oxford Internet Institute investigates the legal, ethical, and social aspects of AI, machine learning, and other emerging information technologies.[1] Our research projects address issues such as data protection and inferential analytics, algorithmic bias, fairness, diversity, and non-discrimination as well as explainable and accountable AI. In addition to analysing problems related to these technologies, the programme has also developed several solutions that have already been implemented by a variety of partners.

3. Approach

Do you agree that we should establish a set of cross-sectoral principles to guide our overall approach?

The establishment of cross-sectoral principles can be useful to guide the UK’s overall regulatory approach, but principles alone are not sufficient to ensure the ethical regulation of AI. Principles can be useful in focussing public debate and raising awareness of ethical

[1] More information about the programme can be found on its website here: https://www.oii.ox.ac.uk/research/projects/governance-of-emerging-technologies/.

challenges related to AI. They can also help to function as universal baseline standards that help to deal with ethical challenges in all sectors and applications.

However, in our previous research, we have also highlighted difficulties that mean that establishing cross-sectoral principles will be insufficient. These include the fact that it is hard to translate ethical principles into practice. This is why sectoral regulations can be helpful, for example in managing particular risks raised by the use of AI in healthcare or in criminal justice.[2]

Do the proposed cross-sectoral principles cover the common issues and risks posed by AI technologies?

The cross-sectoral principles will only help to cover AI's risks if they are made more concrete. This is because the principles use essentially contested concepts with many competing but valid definitions.[3] Fairness, for example, can be defined in many ways according to one's moral or political beliefs. Failing to define the principles more concretely will allow companies to satisfy regulation according to weak definitions, or to effectively 'shop' between different fairness definition or metrics for the one that presents their system or business practice in the best possible light. This will not require them to make meaningful changes to make their products safer, and therefore defeats the point of having principles in the first place. In addition, without added regulations that add clarity in particular cases, it can be hard to verify accountability or adherence to abstract principles.

Citizens often have strong views about how to define these principles, which should guide their definition in regulation. Regarding transparency, our research shows that people prefer good 'everyday explanations' of AI decisions rather than technical explanations of the underlying code.[4] These explanations are clearer and easier to understand. For example, a good everyday explanation of a decision made by an AI system could be that "You were denied parole because you had four prior arrests. If you had two prior arrests, you would have been granted parole." Everyday explanations can help users and citizens better understand how AI impacts their lives, the importance of which was highlighted by the government in their 2021 guidance on 'Ethics, Transparency and Accountability Framework for Automated Decision-Making'.

In defining fairness, it would be worthwhile to look at definitions found in past UK regulatory frameworks. It is not a new concept, and it would make sense for AI regulation to harmonise with existing definitions where possible. Looking at previous regulatory frameworks should reveal that fairness should not be understood as meeting a quantifiable or unchanging threshold, but as a contextual standard. Its definition is determined through various legal judgements and *instrumental* criteria, rather than through pre-existing *substantive* criteria.[5]

[2] See Sandra Wachter and Brent Mittelstadt (2018) 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI'. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248829.

[3] See Brent Mittelstadt (2019) 'Principles alone cannot guarantee ethical AI'. Available at https://arxiv.org/abs/1906.06668.

[4] See Brent Mittelstadt, Chris Russell and Sandra Wachter (2018) 'Explaining Explanations of AI'. Available at https://arxiv.org/abs/1811.01439. In addition, see Tim Miller (2018) 'Explanation in artificial intelligence: Insights from the social sciences'. Available at https://www.sciencedirect.com/science/article/abs/pii/S0004370218305988.

[5] For more on this topic, see Sandra Wachter, Brent Mittelstadt and Chris Russell (2020) 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI'. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547922.

What, if anything, is missing?

Privacy is an important cross-sectoral principle that should be considered regarding the risks from AI. AI systems increasingly use large and diverse datasets for training and decision-making purposes, and individuals have a moral right to know which data is used, and to have control over such uses. Research from the GET programme has revealed two important considerations that the government should make in the framing of any cross-sectoral principle. As with fairness, one possible starting point for considering the cross-sectoral principle of privacy is to review previous definitions in UK legislation, such as in the Data Protection Act 2018.

The unique way in which AI operates demands privacy legislation that can account for inferences and classifications usually not made by individuals.[6] Current data protection and non-discrimination laws fail to protect privacy in the required manner. The use of large and diverse data by AI systems has allowed them to begin to make inferences about individuals without their knowledge or permission. For example, technology companies can seemingly predict whether a person has a particular disease from search engine interactions. These inferences can be invasive, counterintuitive, and are often unverified. It is therefore important that people have their 'right to reasonable inferences' protected, through greater consideration of the purposes of data processing and by offering individuals a right for an explanation of why a particular inference is justified.

In addition, without proper privacy protections, AI can engage in active discrimination against individuals.[7] This can happen when an AI uses people's data to group them in decision-making, without their knowledge or consent. This can affect already disadvantaged communities and can also harm novel groups of people. For example, a modern AI may group users of a particular web browser together and show them higher prices on a website. This is because the relevant group of users of such browsers are not protected in law.

These types of discrimination are wrong because they hinder a person's access to important goods and compromise people's right to self-determination. They are becoming increasingly more concerning as AI is being used more often to make life-changing decisions about people. These decisions can be made without people's knowledge. Complaint-based mechanisms are, therefore, often insufficient, as people do not even know they are being discriminated against. Self-determination is a foundational concern of UK non-discrimination law, but the law has not yet been updated to reflect this. Comprehensively safeguarding an individual's right to privacy when interacting with AI would help to deal with this problem.

3. Implementation

Do you have any early views on how we best implement our approach?

Our research offers a technically feasible solution to implement the transparency requirement in complex AIsystems. Counterfactual explanations are a technical solution that explains why

[6] See Sandra Wachter and Brent Mittelstadt (2018) 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI'. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248829.

[7] See Sandra Wachter (2022) 'The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law'. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4099100.

an AI has made a certain decision. For example, in finance, a counterfactual explanation might say that 'you were denied a loan because you previously failed to pay back your credit card debt. If you hadn't have done this, you would have been successful in your application'.

These can be useful explanations because they offer a roadmap for showing how an individual can change their behaviour. Also, as they only require the release of a smaller amount of information, they are easy to understand and less likely to infringe on trade secrets and IP rights. In addition, a biased counterfactual explanation (such as 'you would have been given a loan if you had been white') can reveal the need to rectify the algorithm in question. Though we think these are a useful regulatory tool, they are not a substitute for a completely transparent AI system.

Our research also offers a statistical measure of automated discrimination detection that could help to implement the fairness requirement, called 'Conditional Demographic Disparity' (CDD). CDD works by comparing the disparity of outcomes of protected groups and can be helpful in determining if a system is holistically fair, unbiased, or trustworthy.[8] For example, in an algorithmic system used for parole recommendations, our bias test would show how the algorithm affects certain communities. It would act as a 'watchdog' or alarm system that would inform the user that the current decision system is not granting Black people parole at a comparable rate to other groups. The user can then decide if this was on purpose.

Counterfactual explanations and CDD are highly flexible methods that can be quickly adapted and implemented by any researchers and developers to work with a variety of systems and case types. The former has already been adopted by companies including Google and IBM, and the latter by Amazon. In addition, the relevant research is free and open access, so can be easily viewed by interested parties.

In your view, what are some of the key practical considerations?

A key practical consideration in the implementation of the fairness cross-sectoral principle is 'bias preservation'.[9] Most existing technical measures of AI fairness, which have been developed in the United States, do not live up to the aims of UK non-discrimination law. This is because UK non-discrimination law aims to reduce discrimination by helping to 'level the playing field' to achieve substantive (rather than merely formal) equality. Given this, it is an important practical consideration to ensure that an appropriate measure of AI fairness is used in the implementation of the cross-sectoral principles, such as CDD.

What will the regulatory system need to deliver on our approach?

No response.

How can we best streamline and coordinate guidance on AI from regulators?

No response.

[8] For more on this topic, see Sandra Wachter, Brent Mittelstadt and Chris Russell (2020) 'Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI'. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547922.

[9] See also Sandra Wachter, Brent Mittelstadt and Chris Russell (2021) 'Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law'. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772.