Professor Sandra Wachter and Dr Brent Mittelstadt of the Oxford Internet Institute (OII), University of Oxford — Written evidence (NTL0058)

This response has been prepared by Prof. Sandra Wachter and Dr. Brent Mittelstadt of the Oxford Internet Institute (OII), University of Oxford. It reports on prior work undertaken by Prof. Wachter, Dr. Brent Mittelstadt and Dr. Chris Russell as part of the Governance of Emerging Technologies research programme at the OII. We would like to thank the Committee for the opportunity to submit this evidence and we are available for any further queries.

Our response focuses on our prior work on "counterfactual explanations" and a bias test called "Conditional Demographic Disparity." These methods are complementary and can be used in any algorithmic or AI system in criminal justice. Both have been implemented by major technology companies and are thus readily available for use (details below).

The best way to think about transparency of algorithmic systems and artificial intelligence in criminal justice is at two levels. To have trustworthy and reliable AI oversight on the (1) individual and (2) system levels is required.

Individual level transparency

The first aim of transparency is to shed light on how and why an algorithm made a particular decision, recommendation, or other output in a particular case. This individual level, or case-specific form of transparency, will be interesting for judges, the police and the person affected by the decision. A method we have recently developed called "counterfactual explanations" can help all parties in criminal justice understand how the algorithm arrived at the conclusion. This is a technical method for explaining the behaviours of algorithmic systems. It can be applied in principle to any form of automated decision-making system used in criminal justice, meaning it can be embedded in the software itself during development or implemented as an external auditing package to test production systems. We and others (such as Tim Miller) also showed that individuals prefer counterfactual explanations in their daily lives and see them as good "everyday explanations."

Counterfactual explanations are a very simple premise that can be used with algorithmic decision-making, machine learning, and AI systems regardless of their complexity; for example, they can be used in neural networks. Counterfactual explanations take the form of a statement such as:

> "You were denied parole because you had 4 prior arrests. If you had 2 prior arrests, you would have been granted parole."

The idea is to give an explanation that describes a small possible change to a case, or to the world, that would have led to a different outcome. These changes can be determined with certainty by testing the system that actually made the original decision. In this hypothetical case, the testing, or search for valid counterfactual explanations, revealed that the individual would have received a different outcome (i.e., being granted parole) if they had only been arrested two times previously, rather than the four times they had been arrested in reality.

One clear benefit of counterfactual explanations is that they can provide advice or a roadmap for an individual to change their situation and receive a preferred outcome. Since only limited information is released, these explanations are less likely to reveal trade secrets and intellectual property rights. Yet, counterfactual explanations are of course not a substitute for transparency and do not sanction opacity.

But this is not their only advantage. In many cases, counterfactuals can act as an early warning that the system has a problem. Imagine the following counterfactual explanation:

> "You were denied parole because you are Black. If you had been white, you would have been granted parole."

Such an explanation would suggest a racial bias exists in the system. Further testing could then be carried out to determine the extent and harm of such a bias (see below).

System level transparency and fairness

One limitation of individual-level transparency and counterfactual explanations is that they only provide information about individual cases or recommendations. The method cannot, at least on its own, assess whether a system is unbiased. They also cannot detect proxy discrimination (e.g. race via postcodes). For that, system-level transparency is required.

AI uses many data points from often untraditional sources, meaning it will often be impossible to rely on human judgement alone to assess if the decisions and the data are unfair. In our paper "Bias Preservation in Machine Learning" we showed that the majority of the bias tests (13/20) that are offered at the moment clash with UK and EU law. We therefore created a bias test, called "Conditional Demographic Disparity" or CDD, that lives up to the standards of these laws.

To evaluate if a system is holistically fair, unbiased, or trustworthy, our bias test can be helpful. An algorithmic system used for parole recommendation: based on the decision criteria for parole or sentencing (e.g., prior arrests, charges, age), 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 with the following type of statement:

> "Your current decision system is not granting Black people parole at a comparable rate to other groups. Is this on purpose?"

This alarm would allow judges and the police to investigate the decision criteria, design decisions, or unintended biases that have caused the inequality, and make a decision about whether the disparity is justified and why. The question of justification is of course a complex one, but it inevitably relates to underlying legal frameworks and the ways in which fairness and biases are measured in AI systems.

Fairness and equality are by no means new concepts in UK law, so it is sensible to first look at how they have been treated in past regulatory frameworks and jurisprudence when proposing new regulations for AI. In "Why fairness cannot be automated" we looked at exactly this question: how are fairness and equality operationalised in EU and UK law? What we found is that fairness and equality are fundamentally contextual concepts in the EU and the UK. Equality is not achieved by meeting a specific, quantifiable, unchanging threshold, for example a specific ratio of outcomes between protected groups. Rather, the meaning of fairness and equality are determined on a case-by-case basis according to Member State laws and judicial interpretation. There is not a specific *substantive* measure of fairness prescribed by the law. What we found, however, is that there are certain *procedural* requirements in how fairness is measured that can be thought of as a 'gold standard' for comparing outcomes between groups, and thus measuring fairness in practice. We proposed a new fairness metric (CDD) to capture this procedural gold standard and explained how it can be used as a consistent evidential baseline for measuring AI fairness across different cases.

There are compelling reasons to require the usage of CDD to measure fairness in AI systems in the context of criminal justice. Non-discrimination law in the EU and UK aims at substantive equality. This means simply treating different protected groups equally going forward (i.e. 'formal equality') is not enough; rather, the law also aims at 'levelling the playing field' for groups that have been historically disadvantaged. In "Bias preservation in machine learning" we proposed a classification scheme for ways of measuring fairness in algorithmic systems (i.e., 'fairness metrics') that reflects the distinction between formal and substantive equality. We distinguish between 'bias preserving' and 'bias transforming' fairness metrics.

Metrics that are 'bias preserving' treat the status quo as a neutral starting point to measure inequality. In effect, these metrics take the acceptability of existing inequalities for granted. This is a problem if we want to use AI not simply to uphold the status quo, but to actively make society fairer by rectifying existing social, economic, and other inequalities. And it likewise clashes with the aims of non-discrimination law to achieve substantive equality. In contrast, 'bias transforming' metrics do not take the status quo for granted, but rather actively question what existing inequalities and biases are appropriate to teach a model or AI system.

To help achieve system-level transparency and fairness in practice, we recommend requiring organisations using AI to make important decisions in the context of criminal justice to (1) at a minimum, use bias transforming metrics to measure fairness and make fair decisions and (2) ideally, publish summary statistics based on our recommended bias transforming metric, Conditional Demographic Disparity. Doing so will ensure all parties involved in a case of potential discrimination or unfair automated decision-making can have access to a common set of statistical evidence, which can then be used to decide what is actually substantively 'fair' in their specific case.

Application and use of the recommended approaches

Google has acknowledged our work explicitly and implemented counterfactual explanations in two major products between 2018 and 2020 to help users understand the outputs of AI systems: Google Cloud, and the "What-If Tool" for

TensorFlow. Other businesses that have implemented counterfactual explanations in their products include IBM, Accenture, Vodafone South Africa (in its "Just 4U" payment plans, and the drone insurance company Flock. In January 2021, Amazon released an AI accountability toolkit for "bias detection": SageMaker Clarify**.** Amazon implemented our bias test CDD as one of the statistical measurements that allows testing of fairness in automated systems.

Our research describing how counterfactual explanations and CDD work is open and freely available for everyone. The methods are highly flexible and can be quickly adapted and implemented by any researchers and developers to work with a variety of systems and case types.

References

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI." *Proceedings of the conference on fairness, accountability, and transparency*. 2019.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law." *W. Va. L. Rev.* 123 (2020): 735.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." *Harv. JL & Tech.* 31 (2017): 841.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI." *Computer Law & Security Review* 41 (2021): 105567.

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.

5 January 2022