Written evidence submitted by Dr Dan Stowell, Dr Emmanouil Benetos, Dr Bob Sturm, Dr Laurissa Tokarchuk, Centre for Intelligent Sensing (ALG0036)

Introduction

1. We are a group of academics based in the Centre for Intelligent Sensing at Queen Mary University of London (QMUL). Our work has a strong focus on artificial intelligence, machine learning and other algorithmic approaches which make inferences based on sensor data: primarily sound recordings as well as mobile phone sensor data (see Appendix for definitions of technical terms such as "machine learning"). Our work concerns the engineering of such "intelligent sensing" systems. We also work with commercial partners on data driven applications.

2. In this submission we make recommendations on accountability and bias in decisions made automatically by algorithms such as those driven by machine learning.

3. Algorithms have been used in decision making for many decades, although in the past human operators have been the ones implementing the calculation: e.g. bank loans, benefits entitlements. The current discussion arises partly because of the increasing use of data and fully-automated processes, which implies society placing a greater degree of trust in algorithms and less in human discretion; and partly because many of the modern algorithmic methods are increasingly inscrutable.

Algorithms and bias - the good news and the bad news:

- 4. Fully-automated processes may be used to replace the individual judgment of a frontline operator, and so can remove one potential source of subjectivity and bias. This is particularly useful in situations of decision-making involving sensitive personal information. Even well-intentioned people have *conscious and unconscious biases* that affect judgments, even aspects of physiology such as levels of hunger detrimentally impact human judgement in critical ways. Hence, algorithmic decision-making offers an opportunity to improve one aspect of impartiality.
- 5. However **algorithms are not unbiased.** Algorithms themselves enforce the biases present in the algorithm designer and/or the data used to "train" the algorithm. This can have effects as deleterious as the biases that they remove. As a simple example: if past data on employee recruitment is used to "train" an algorithm for making employment decisions, then in many cases it will reproduce the biases that were present in previous practice, unless deliberate effort is made to mitigate that. The removal of such bias is non-trivial and is an ongoing research field. Biases are difficult to uncover if the algorithm is complex; if the algorithm is simple and auditable - ideally, publicly auditable - then there is much greater chance that algorithm-based decisions can be fair. Likewise, if the data used to "train" the algorithm is auditable (and anonymised when appropriate), algorithmic biases could be identified.

UK legal context

6. Algorithmic decisions could apply to individual people, to families, to businesses, organisations, etc. For simplicity, and for legal context, here we will primarily consider the context for decisions made about individuals.

7. The Equalities Act 2010 specifies nine **protected characteristics**, a set of personal characteristics which one might wish to protect from bias. Those characteristics are not the be-all-and-end-all, but for the present purposes we use them as the primary concern: many decisions (not all) made by public and private bodies should be invariant to protected characteristics (PCs).

- 8. One idea is to forbid protected characteristics to be included in the input to an algorithm. However, this is ineffective in practice, since such characteristics are implicitly reflected in other characteristics (e.g. educational and career history). In other words, it is implausible to remove all “clues” about protected characteristics from the input. Instead, the focus must be on ensuring that the **outputs** - the decisions made - are **statistically independent of protected characteristics** (except when there are deliberate “positive biases” such as *affirmative action/positive discrimination* initiatives). This is analogous to being aware of your own "unconscious bias" and making sure that your behaviour is fair even so, rather than pretending that one’s biases do not exist.

9. In addition to protected characteristics, commercial providers increasingly have access to large collections of **behavioural data**, which might be gathered from GPS or accelerometer tracking (e.g. from phones or smart watches) or from online behaviour. These data often yield highly detailed and personal profiling, probably more than most citizens expect, and this has a sensitivity that goes beyond those addressed in the Equalities Act. There is the capacity for example for decisions to be made based on inferences such as “this person is often driving late at night”.

10. In algorithmic decision-making, there is little case law yet established. The Data Protection Act (DPA) gives citizens a right to a meaningful explanation about the logic used in decisions that have been made about them. In addition, the EU General Data Protection Regulation (GDPR) enables citizens to question and fight decisions that affect them that have been made on an algorithmic basis, although the “right to explanation” enabled to citizens provides limited information, and regulations offer limited rights towards “right to not be subject to automated decision-making”. The UK government should work towards issuing guidance on what these rights should mean in practice.

Written evidence submitted by Dr Dan Stowell, Dr Emmanouil Benetos, Dr Bob Sturm, Dr Laurissa Tokarchuk, Centre for Intelligent Sensing (ALG0036)

Transparency: black boxes and white boxes

11. Black-box vs white-box: at various points this distinction is useful. “Black box” means that there is no feasible way to inspect the components of the algorithm to get a clear understanding of how the outputs were arrived at. This is true of much of the "deep learning" revolution of the past decade and more (see Appendix). By contrast, a competent technician can inspect a “white box” algorithm and then explain how a particular decision was arrived at.

12. There is no chance of banning black-box algorithms in commercial use, and this would not be desirable, even when processing personal data. In governmental/public decisionmaking, there is a standard of democratic accountability which implies black-box algorithms are undesirable, especially when dealing with sensitive topics.

Accountability

13. Accountability - by which we mean the ability for decisions made with the help of algorithms to be challenged (in court or otherwise), and for the processes that led to those decisions to be interrogated. Such decisions often involve people as well as computers; we assume here that the human side of things is well-developed in existing law, and focus on the accountability of the automated component.

14. One component of accountability is for the steps involved in the decision-making to be explained. The GDPR (General Data Protection Regulation) is widely seen to have mandated the ‘right to explanation’. While the principles behind such a mandate are just, the feasibility of such an approach is widely questioned. While some assume this right applies to an ‘independent expert’ many have interpreted it as being explainable to the layperson. Explainability to the layperson however is problematic and, as has already been shown in Austria and Belgium, results in the high-level data release and explanations that do not satisfactorily address why a particular decision was made about a particular individual. Moreover, even this unsatisfactory level of explainability is often only possible for white-box models and not for black-box models.

15. One component of accountability is, for a trained model (as in most machine learning methods), to specify which data were used to train the model, and which data were used to verify that the trained algorithm then behaves as intended. These aspects are particularly important in holding black-box models to account.

Written evidence submitted by Dr Dan Stowell, Dr Emmanouil Benetos, Dr Bob Sturm, Dr Laurissa Tokarchuk, Centre for Intelligent Sensing (ALG0036)

Some algorithms[1] are more accountable than others

- 16. Black-box models such as “deep learning”/”neural networks” (also “random forests”) are typically the least accountable because of their lack of transparency, and should not be used in situations where accountability is particularly important.
- 17. Some white-box models are relatively simple to analyse and to explain. Good examples are “decision tree”, “linear regression” and “logistic regression” methods.
- 18. Bayesian models (such as “Bayesian networks”) are particularly recommended, because (a) the assumptions encoded are made explicit and up-front, in the prior and in the model structure; and (b) they tend to be less prone to unexpected behaviour than many other algorithms.

How to shine a light on algorithm behaviours

Auditing of bias in decision-making algorithms:

- 19. Even black-box systems can be audited for how their decisions co-vary with protected characteristics. This can be done even for commercially-sensitive algorithmic workflows.
- 20. Well-known mathematical statistics can be used to test whether any given algorithm's decisions tend to co-vary with protected characteristics. There is not a clear consensus on the best measurement to make: standard mathematical measures such as correlation or mutual information may be relevant; “statistical parity” is a specific measure discussed in relation to algorithmic fairness. The wider point is that decision-making algorithms can be audited by inspection of their inputs and outputs using mathematical tools. Ideally such tests would be measured on all protected characteristics together rather than separately, because there may be niche biases that depend on multiple criteria (as an arbitrary example: exhibiting bias against women from a specific ethnic background, rather than against women in general).
- 21. Audits can be performed by probing the system with fictitious data generated by sampling from UK demographic data, or by a company's own anonymised customer data (e.g. by counterfactually varying the effects)

[1] See the Appendix for definitions of the machine learning algorithms we refer to.

22. Specifying algorithm requirements: public bodies or their subcontractors commissioning some system that provides algorithmic decision-making should include in the **technical specification, required characteristics of the decisionmaking process**. These might include:

- the level of accuracy required (e.g. with respect to sample data which the body does *not* reveal to the implementor), quantified through a measure such as *error rate* when compared against the decisions of experts;
- the kinds of bias that are not permitted (e.g. correlation of decisions with protected characteristics);
- if the algorithm is “trained”, then the allowable sources of data that can be used for training;
- if the algorithm is “trained”, then the extent to which a transparent record is required of the data sources that were in the end used for training;
- the extent to which the algorithm must be transparently understandable by an independent auditor (some contexts may be too sensitive to allow “black box” models, while for some it is acceptable).

23. A related suggestion is, irrespective of how any given algorithm was created and by whom, to establish **a standard for documenting** (in outline) the characteristics of an algorithm used in decision-making, listing:

- which assumptions have been encoded (ideally, details of the Bayesian prior)
- which data (if any) have been used for training the algorithm
- What model/algorithm has been used (this could be quite explicit for algorithms in public life; for commercial algorithms this is often considered a trade secret)

Conclusion: summary of recommendations

A. Where algorithmic decisions are required to be fair or unbalanced with respect to personal characteristics, the focus must not be on whether those characteristics were used as inputs, but how the algorithm’s outputs co-vary with the characteristics.

B. The UK government should work towards issuing guidance on what citizens’ DPA rights mean in practice, with respect to algorithmic decision making.

C. In situations where accountability is particularly important, black-box algorithms (including “deep learning”, “neural networks, “random forests”) should not be used. However, in many situations they can be used, providing that sufficient accountability is maintained (e.g. via our other recommendations).

D. A technical committee of the UK government should design a general auditing procedure for algorithmic decision-making systems in the public and private sectors. Such a procedure can and should be applicable even to black-box and commercially sensitive scenarios, e.g. with a mechanism for auditors to access sensitive data on a carefully-managed basis.

E. A technical committee of the UK government should consider what guidance or legislation might be needed to set boundaries on the use of behavioural data (online and physical) in algorithmic processes.

F. A technical committee of the UK government should consider the desirability of the specification and documentation of the properties of algorithms used in decision-making, to provide an audit trail of aspects such as assumptions and training data.

G. A technical committee of the UK government should consider the formation of trusted independent body of experts who would serve in auditing capacity when examining 'algorithmic explanations'. The committee could also consider contracting for a length of service/non disclosure clauses to allow for more detailed examination of commercial algorithms without breaching company trade secrets.

H. Maintain and expand legislation based on UK DPA and EU GDPR in order to effectively enable citizens’ rights to explanation, information, and to opt out of algorithmic decision-making.

April 2017

Written evidence submitted by Dr Dan Stowell, Dr Emmanouil Benetos, Dr Bob Sturm, Dr Laurissa Tokarchuk, Centre for Intelligent Sensing (ALG0036)

Appendix - Definitions

Artificial intelligence (AI): The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

Machine learning (ML): Branch of AI which involves the creation of algorithms that can automatically learn from data.

ML algorithms mentioned in the text:

- **Neural net (artificial neural network)**: Class of ML algorithms which process information via a large collection of simple units, originally inspired by analogies with collections of neurons in the brain.
- **Deep learning**: Class of ML algorithms that employ neural nets with very many layers. These often require large amounts of data and computing power to be used effectively, but are responsible for many impressive results in recent years.
- **Decision trees**: Machine learning and decision support method that uses a tree-like model of yes-no decisions (e.g. “Is the person aged 42 or above?”).
- **Random forests**: Method for classification and regression that operates by constructing a multitude of decision trees.
- **Linear regression**: Approach for modeling the relationship between variables as a direct linear formula (e.g. “each cigarette takes 11 minutes off your life expectancy”).
- **Logistic regression**: A relatively simple regression model widely used to predict categorical variables.
- **Bayesian networks**: Also called *probabilistic graphical models* or *belief networks*, they are a type of statistical model that represents random variables and their dependencies. Unlike many other methods, Bayesian networks explicitly model our prior beliefs and uncertainties, and use them to infer the certainty of their outputs.