Written evidence submitted by Dr Janet Bastiman (ALG0029)

Executive Summary

- Algorithms used in decision-making can be too complex to describe in clear English
- Data used in algorithms can go through multiple levels of abstraction such that it is impossible to determine the original input
- Bias in algorithms arises from poor scientific practices used in industry and through reuse of academic publications only designed to cover a very limited scenario.
- Intellectual property of enterprises using decision making algorithms is contained in the weightings of the algorithm and the data used to create them
- Clear transparency of the algorithm is not achievable, instead institutions should publish clarity on accuracy of the algorithm and rigor in its creation including removal of disparate impact in accordance with UK law.
- All personal data covered by UK and international laws must be observed when used to create algorithms although once abstracted into algorithmic weightings there should be no requirement to recreate the algorithm after withdrawal of permission for use.

Introduction

I have 14 years’ experience in the private sector, the past 6 of these as Chief Technical Officer or equivalent title, where my responsibilities have included ensuring any algorithms developed and deployed are appropriate and robust. For the past two years, this has involved specific applications of deep learning alongside preparation of a company for the EU General Data Protection Regulation Act 2016\(^1\) (EU GDPR). As such, I have a significant insight into the needs, problems and practicalities of how businesses may be affected by any future legislation and wish to ensure that these issues are considered as part of this inquiry.

Evidence

1. Algorithms have been used to determine results from a series of inputs since the start of the digital era. They are fundamentally instructions on how to combine inputs to produce a specific output, and each algorithm in isolation can be converted to a very clear English definition consisting of how important each input is in achieving a specific output (known as the input weighting). As systems have become more complex, algorithms have also become more complex. That which we call an algorithm may consist of multiple sub algorithms connected in multiple ways. While this can be explained in terms of the structure of the system, the inputs and outputs at each stage rapidly become more abstracted, thus explaining the algorithm in plain English becomes significantly more difficult.

2. To ensure transparency in algorithms we would need to ensure that all stages of the algorithm could be explained in a manner that a non-expert individual could understand.

3. Algorithms can be created “by hand” that is: a human designs, tests and reworks the algorithm until it achieves the desired result. Algorithms may also be created through machine learned processes. For machine learned processes\(^2\) the key differential is that the system itself adjusts the algorithm’s input weightings. For complex systems, many hundreds or thousands of weightings are being adjusted in parallel. The resulting systems can be

---


\(^2\) There are many different terms for machine learned processes, which I shall combine for the purposes of this submission. These include but are not limited to: artificial intelligence, artificial neural networks, machine learning, deep learning, adversarial networks, convenets and resnets
explained mathematically, however the inputs for such systems are abstracted from the raw data to an extent where the numbers are practically meaningless to any outside observer.

4. While hand-designed algorithms can be held to scrutiny through clear definition without much difficulty, the complexity in the mathematics and concepts of machine learned algorithms has resulted in a concept of “mathwashing”\(^3\) where we believe that algorithms are objective and unbiased because they are too complex to describe to the general public.

5. Individuals within the technology industry understand just that since programs are only as good as their programmers, even machine learned algorithms are only as good as the data used to generate them. Any decisions and biases used in collecting the data will be reflected in the final algorithm even though it will not be clear from looking at the structure of the algorithms.

6. Bias in algorithms can be benign (e.g., the posts the Facebook chooses to show from our friends in our top stories), concerning (filtering our input to show only news that agrees with our existing biases), offensive (identification of non-Caucasian faces as animals in image recognition technology) to life-changing (denial of employment or finance, legal risk). While all of these biases require resolution, institutions dealing with life-changing decisions must be held to the highest scrutiny as quickly as possible.

7. A significant paper by a multi organisational group was published in 2015\(^4\) on ensuring that bias can be identified in algorithms. In this paper, the authors assert that if an algorithm makes a prediction over 80% of the time correlated with a protected attribute (race, gender, age etc) then there is disparate impact. The protected attributes may vary between algorithms, but these can be clearly defined in the initial definition of the algorithm purpose and can also be published by the institution creating the algorithm.

8. With machine learned algorithms, even if protected attributes are removed from training data (race, gender, age etc.) these attributes can still be inferred by the system from non-protected attributes leading to a bias in the final algorithm. For example, with a full postcode there will be high correlations with individuals of specific race, religion, and affluence within large parts of the United Kingdom. This shows the need for testing of a final algorithm and not assuming lack of bias because of the removal of key fields within the data.

9. Bias in algorithms can be introduced in multiple ways, all of which can be overcome with good scientific and engineering practises, which inherently avoid bias from assumed results and seeks to prove, to high significance, the accuracy of the algorithm. This can be achieved by the institution:

   1. Defining the purpose of the algorithm in terms of the problem and not predefining the solution e.g “determine the factors that contribute to an individual being at low risk of mortgage defaults” instead of “identify individuals with poor credit history who are more likely to default on a mortgage”. When suggesting a correlation, data collected to train and test the algorithm are more likely to support the conclusion rather than ensure that the input data is broad and unbiased.

   2. Ensuring rigorous testing of the algorithms is in place consisting of:
      1. data never seen by the algorithm
      2. data from a broad range of sources and of varying quality

---


Written evidence submitted by Dr Janet Bastiman (ALG0029)

3. Data from edge cases and unexpected sources presented deliberately to “break” the algorithm and highlight errors.
4. Validation of the algorithm to ensure it passes the bias tests proposed by Feldman et al described in paragraph 7.

10. Research and progress within the field of machine learned algorithms is progressing rapidly as key private sector companies are allowing their employees to publish their results and open source the frameworks to create new algorithms. As such, technology companies are incentivised to gain publicity from publishing rather than patenting their own innovations.

What companies do protect are:

1. Detailed structures of their networks – companies will generally show an example network to demonstrate the innovation without giving away the precise structure used in their income-generating technology.
2. Weights of the network – even where the structure is published in detail, the final weights to achieve the results are not given in detail.
3. The data itself – usually the data set is rights limited (e.g. medical data) and cannot be published, so the data will either be described or the academic paper will be written to use a freely accessible data set such as ImageNet or MNIST for images or equivalent for text, speech or music.

11. Since the intellectual property in machine learned systems is encapsulated in the structure, weighting, and input data that comprise the final algorithm, any legislation requiring clear transparency of the algorithm itself could have negative impact on the commercial viability of private sector institutions using this technology.

12. Given the complex nature of these decision-making algorithms, even if the full structure, weighting, and training data were published for an end user, it is unlikely that they would be able to understand and challenge the output from the algorithm.

13. Transparency in complex machine-learned algorithms is almost impossible to define. The multiple levels of abstraction involved reduce the data into a form that cannot be returned to the original form. Some appreciation of aspects of the data can be determined from the algorithms via visualisation from the filter layers and not all activation features correspond to any identifiable parts. In this respect, while the algorithm could be visualised, this would be unhelpful to any party trying to determine fairness or bias.

14. Where personal data is concerned, all institutions should already be fully aware of the forthcoming EU GDPR legislation, the UK’s Data Protect Act (DPA), as well as the rights and responsibilities handling data from third parties (either directly or indirectly). Institutions must be able to state the source and restrictions on any data used in the creation of algorithms to an extent where they are able to comply with requests to remove such data. Guidance should be provided to institutions on how this legislation affects their use of data. It is my belief that where an institution has permission to use the data (either directly as a

---

5 For example: https://www.techemergence.com/top-tech-companies-open-source-their-ai-secrets/ accessed 16th April 2017
6 ImageNet provides a list of images for training networks on object recognition http://www.image-net.org/ accessed 16th April 2017
7 MNIST is a library of handwritten numbers for digit recognition http://yann.lecun.com/exdb/mnist/ accessed 16th April 2017
8 As listed in e.g. http://deeplearning.net/datasets/ accessed 16th April 2017
Written evidence submitted by Dr Janet Bastiman (ALG0029)

data owner, or indirectly as a data processor) then it may be used to create algorithms as the institution requires.

1. Where the permission to use data is withdrawn then that data must not be used for any algorithm creation or improvement from the point of notification onwards, that is, it must be removed from all test and training data sets to prevent use in any future algorithm creation. Furthermore, if it is possible to extract the data from the algorithm then the algorithm itself must also be updated.

2. Where data has been abstracted to the extent that it is impossible to determine any personally identifiable data from the algorithm, system or any components, then there should be no requirement to force an institution to recreate or retrain such a system following the removal of data from the data set except where the inclusion of the data has been shown to introduce unacceptable bias.

15. I make the following recommendations:

1. Best practise for algorithms be defined that would be mandatory for government and voluntary for private sector. Adherence to this best practise would provide confidence to the users of the services provided by these institutions that the algorithms are fair and unbiased. This best practise should include:
   1. Clear definition of the problem space for the algorithm; its inclusions and exclusions, and protected attributes tested to avoid disparate impact.
   2. Clear definition of the type of data gathered to test and (if applicable) train the algorithm to ensure that bias in the data and testing process is removed through transparency.
   3. Publication of the accuracy of each algorithm in a plain English form

2. Institutions using data covered by the EU GDPR and UK DPA must maintain records of permission grants to use that data and also be able to remove data from systems as required by this legislation. Data subject to additional international legislation must be recorded in a manner such that the institution can adhere to any further restrictions imposed by permissions withdrawal outlined in such legislation. Where the data has been used to train algorithmic systems, either machine learned or human adjusted, and has become abstracted as described in paragraph 13, there should be no need to retrain these systems. The data must be removed from the training and testing data sets so that the personal data is unavailable for use in future evolutions of the algorithms.

April 2017

11 http://www.plainenglish.co.uk/ accessed 15th April 2017