Written evidence submitted by the Institute of Mathematics and its Applications (ADM0008)

This submission updates the response submitted by the Institute of Mathematics and its Applications in April 2017, to the original call for evidence.

The Institute of Mathematics and its Applications (IMA) is the professional and learned society for mathematics in both industry (including Government) and research.

Background description

1. The Oxford English Dictionary defines “algorithm” as “A procedure or set of rules used in calculation and problem-solving”. There are various ways of classifying algorithms and their uses.

Proclaimed or Inferred (from data)

2. A classic example of a proclaimed algorithm is Income Tax: If (for a given set of circumstances) your net income is X, you pay f(X) in income tax, where f is laid down in the Finance Act. Another example comes from the practice of many universities, who have algorithms for determining the class of degree, given various marks at various stages.

3. A classic example of an inferred algorithm is the “wind chill”, which is an algorithm with a precise mathematical statement, but was experimentally determined by the multinational JAG/TI group of scientists.

4. Much of the current trend in “machine learning” is for a computer, sometimes using a well-understood algorithm (referred to as a meta-algorithm, since it is an algorithm to produce algorithms), to produce an algorithm without a precise mathematical statement, but based on a large amount of experimental data, generally referred to as “training data”. Most of these algorithms operate on data, often training or other background data (database) as well as foreground data (the question). They are at most as good as the background data they operate with, and bad data can lead to disastrously wrong results, as when the London Ambulance Service did not know where the Velodrome was. [Guardian2016]

Published or secretly understood, or not understood

5. Income tax, and wind chill, are published algorithms. Many companies operating in regulated industries, generally in finance, have secret algorithms for credit scoring, loan approval etc. These algorithms are part of their competitive edge. Yet, they need to be able to explain them to regulators, and to justify decisions if required. An example might
be a car insurance company, whose algorithm might include the step “To the base premium, we add a sum depending on the insured’s occupation, from this table”, where the table is computed based on past claims data. The table, the precise definition of occupations, and indeed whether it’s adding a sum, multiplying by a ratio, or both, are part of the insurer’s trade secrets. This example in insurance long predates the use of computers, never mind machine learning algorithms, but the advent of technology has permitted much more analysis and the use of many more factors. Equally, it has enabled much more precise calculations, rather than “gut feel” to populate these tables.

6. Conversely, many of the algorithms produced by machine learning, notably those based on “deep learning”, are not understood at all: no human being can say “why” the algorithm does what it does, nor can predict what it will do on data which are not the training data.

Advisory or Determinative?

7. The income tax algorithm determines the amount payable: it is not a suggestion to the tax inspector how much to charge. University degree classifications have become more determinative over the years. A recent adoption of a determinative algorithm is the Duckworth-Lewis algorithm in cricket. The algorithm is public, and understood by experts, but not by the general public, who just accept it.

8. Advisory algorithms produce a piece of advice to a human being who makes an ultimate decision. This may consist of evaluating several different scenarios, or may just be a simple “answer”. These answers may or may not (in practice, far too often do not) have some measure of confidence attached. Of course, an algorithm may in principle only be advisory but the human beings using it may in practice just rubber-stamp its “advice”, so in practice it’s determinative. In principle the GDPR (Article 22.1) talks about a “decision based solely on automated processing”, i.e. a determinative algorithm. One might expect a lot of litigation in this area.

9. For the purposes of this paper, we regard systems that call for further medical tests as advisory, so long as the final diagnosis is made by a human being.

Continuous or Discrete?

10. Many algorithms are used that yield numerical results varying continuously with their input. So if you earn £10 more, you pay £2 (or £4 or …) in income tax, and £20 more is twice that (unless you cross a threshold) and so on. If the wind speed increases by 1 km/h, then the wind chill changes by a certain amount. Other algorithms give one of a small number of discrete results, e.g. degree classification. In many cases the answer is
binary: bail or no bail; (referral for) melanoma or not. We use the shorthand “continuous algorithm” for “an algorithm whose results depend continuously on their input”.

11. For a continuous algorithm, a small error in the inputs should result in a small error in the answer, whereas for a discrete algorithm, even the smallest error in the inputs may result in a different answer. If we understand the algorithm, we can consider the question “how close are we to a boundary”, and possibly act on this information. For example, a university department might adopt (and some do) a rule that if a student fails by 1% or less, the scripts should be checked again. But if all we have is a black box that outputs decisions, we cannot ask this question.

Effect on people

12. It is a truism that every decision has consequences. The main concern is determinative (either de facto or de jure) algorithms making decisions that significantly affect people’s lives. Examples would include medical diagnosis, mortgage approvals and [in the USA] granting or otherwise of bail, and possibly sentencing.

13. However, there are also effects on people’s prospects (being shortlisted for jobs), finances, e.g. gender discrimination in insurance, which in theory is illegal, but can be perpetuated through use of gender-correlated data, and other important aspects.

Data

14. An inferred algorithm is based on data. In the case of a scientific experiment, where the algorithm is often called a Law, as in Boyle’s Law, one expects the experiment to the reproduced, and these repeat experiments to confirm the algorithms. In the case of other fields, the nature of the data is more variable. The term “ground truth” is used to refer to absolute facts, such as “this skin lesion is a melanoma”, or “this person did not offend while on bail”, whereas we use “proxy data” to refer to what is actually used as the training data, “did a dermatologist diagnose skin cancer” or “did a bail judge refuse bail [on the grounds of likelihood of re-offending]”.

15. Clearly ground truth is better, but it may be too difficult, or even impossible, to obtain. The study in [Grin et al. 1990] (which deduces that dermatologists have a 64% accuracy rate, but a sensitivity of over 80%) was very rare in actually conducting biopsies on patients with a negative diagnosis by the dermatologist, so that we have ground truth. In many circumstances it may be impossible, e.g. we can’t obtain “did this person [who was bailed] offend while on bail”, merely “was this person [who was bailed] caught offending while on bail”. This may seem pedantic, but note that we can’t obtain the opposite datum
“would this person [who was not bailed] have been caught offending while on bail had he been released”.

16. This use of background data to derive an algorithm has two important consequences. The first is that any biases in the background data will be perpetuated in the algorithm. The second is that we have no guarantee at all that the algorithm produced will remain valid outside the range of those data. This means that every such algorithm needs to be traceable back to the background data that produced it.

Extent

17. It is hard to know the extent of algorithmic decision-making, either advisory or determinative, anywhere. However, we can quote some examples from Government, and the BCS submission talks about industrial applications.

UK Government, QE Carrier.

18. The business case for this drew heavily on the Cost Capability Trade Off Model, which uses a variety of modern algorithmic techniques such as Hybrid Model of Non-Linear Regression, Optimisation, Monte-Carlo Simulation and Design of Experiment in order to forecast the optimum performance within budgetary constraints. It is a mixture of proclaimed and experimental, is the company’s intellectual property, but was explained to the customer, and was advisory, exploring a variety of scenarios. See [ChamberlainPinker2016] and http://www.telegraph.co.uk/education/stem-awards/stem-hq/the-beauty-of-maths/.

Other Government uses

19. Most US states use some kind of algorithmic risk-based assessment in the criminal justice system, sometimes for bail, parole or sentencing. These have been the subject of substantial controversy [Angwin2016; Floresetal2016]. In particular, there is controversy over whether, and at what level, they should be advisory or determinative: the clearest decision currently [Wisconsin2016] is that at sentencing they should be advisory only.

20. There are many other potential uses in Government, particularly in trying to produce “joined-up thinking” between different agencies, in areas as diverse as nuclear disaster response and food insecurity mitigation. See [Smith et al., 2015].
21. Firstly, it must be noted that ‘biases or discrimination’ are human perceptions of a process, and what is, or is not, discrimination depends on society, and changes over time. Consider, for example, the recent ruling about gender bias in insurance. It’s also a very nebulous question when pushed to precise specification, as doing something algorithmically requires.

22. A. “Eliminate”. If the algorithm is known and fully understood, it is possible to assert that it is no more biased than the data fed into it. New algorithms for combining panelists’ scores [MacKayeetal2017] could be used for many decision-making processes, e.g. the Research Excellence Framework, and would improve the openness of the decision-making. If the algorithm is not known and fully understood, it is only possible to eliminate those biases which are known about and checked for, and then only if the algorithm has the appropriate mathematical properties, e.g. linearity.

23. B. “Introduce”. It is certainly possible to introduce biases through an algorithmic process, just as it is possible to do so through a manual process.
   a. As pointed out [Angwinetal2016], it is possible for such processes to be biased by considering things that human beings would refuse to consider publically: the system used in Broward County Florida asks “Was one of your parents ever sent to jail or prison?”, while it’s hard to imagine a judge accepting the prosecution argument “the defendant deserves a harsher sentence because his father went to prison”.
   b. Consider, again, insurance, more specifically car insurance. It is no longer legal to discriminate on the basis of gender. It is, currently, legal to discriminate on the basis of occupation, even though some occupations are predominantly occupied by one gender, giving rise to indirect discrimination [McDonald2015]. We refer to the BCS submission for practical measures. See also recommendation 4.

24. C. “Amplify”. If one can introduce a bias, one can certainly amplify it. But there is an effect, known as “uncertainty bias”, by which an unbiased algorithm can become biased by considering factors that are not uniformly distributed, and this bias can grow over time as an “active learning” algorithm learns and reinforces its bias [GoodmanFlaxman2016].

25. D. “Detected”. Currently, detecting bias relies on human effort, sometimes aided itself by machine learning. We have quoted various instances of bias, but these are merely those that researchers have chosen to investigate. Direct bias, e.g. highly-paid jobs being offered to explicitly male searchers [Dattaetal2016] can be obviated by search
engines never taking gender into account (but this is highly improbable), but even so this would not eliminate indirect bias [McDonald2015].


27. F. “Transparent and Accountable Way”. Many of the most efficient algorithms for certain tasks are not transparent or accountable. It is, with the current state of technology, impossible to understand how a typical neural net reaches its decisions. Though the state of technology may change, it is currently the case that only algorithms which can be reduced to a (possibly complex) formula can be understood, and even then possibly only by experts. There is growing evidence that “deep neural networks”, often the method of choice for building machine vision systems and other sorts of pattern recognition, simply do not operate the way humans do, despite performing well on many experiments: as the example below shows and [Nguyenetal2015] put it “Deep Neural Networks are Easily Fooled”. This is particularly relevant in the area of automation of vehicle control (“driverless cars”), where [Evtimovetal2017] show that road signs can be subtly altered to be recognised as completely different.

(Left) is a correctly predicted sample, (centre) difference between correct image, and image predicted incorrectly magnified by 10x (values shifted by 128 and clamped), (right) adversarial example. All images in the right column are predicted to be an “ostrich, Struthio camelus”. Taken from [Szegedyetal2013].

28. G. “The implications of increased transparency”. The background data should have been thoroughly anonymised (easier said than done). The foreground data should be subject to the usual data protection rules. There is a real challenge with “active learning” algorithms, where today’s foreground data becomes tomorrow’s background
data. To the best of our knowledge, no satisfactory research has been carried out here. See BCS submission.

29. H. Good practice is the opposite of bad practice, and there is much bad practice. Two specific elements of bad practice that should be flagged and users of machine learning warned about are these.

a. Confusing ‘correlation’, which is what algorithms, machine learning or otherwise, can detect, with ‘causality’, which is what people are generally interested in. Consider [Grindrod2016, page 16] the example of “Screen time [at 14 years] was associated with lower academic performance [at 16 years]”. Media reporting generally converted correlation into causality, as in “programmes aimed at reducing screen time could have important benefits for teenagers' exam grades,”, whereas increased screen time might be an early indicator of poor performance, and the investment should be in coaching/catching up.

b. Splitting the background data into training and testing data (which many machine learning protocols require) and then believing that this split is significant. At the very least, this process should be repeated several times, known as cross-validation. The state of the art is not yet able to give good guidelines for “several”.

Methods for providing regulatory oversight of Algorithmic decision-making

30. GDPR Article 15.1(h) provides that, even if “automated processing including profiling” is allowed, then the data subject, “shall have the right to obtain from the controller [...] meaningful information about the logic involved, as well as the significance and the envisaged consequences”. As it stands, this is between the data subject and the controller. The IMA leaves it to lawyers to say what general regulation, if any, is or needs to be imposed beyond GDPR itself.

31. The Government, local authorities etc., may well need to impose internal purchasing and/or commissioning regulations to give effect to the recommendations below.

Recommendations

32. 1. The Government, and all its agencies and subcontractors, need to review as a matter of urgency the use of all automatic processing to ensure it

a. is only based on legitimate personal data of the individual;

b. carries a precise description of any training data it was built from;
Written evidence submitted by the
Institute of Mathematics and its Applications (ADM0008)

c. has been tested for indirect discrimination;
d. is capable of being explained in line with GDPR 15.1(h). In particular, buzz-phrases like “deep learning algorithm”, “data-based algorithm” should act as warning signs that the “algorithm” is in fact no more than unscientific reasoning by generalisation.

33. 2. The Government should make more use of advanced decision support tools, such as were used (as a novelty) for the QE carrier decision, but which have been used routinely in Intel (and many other corporations) for many years. These tools are not limited to capital purchases: [Li et al. 2015] show a 30% reduction in ambulance response times through use of optimisation algorithms based on actual traffic data. They have potential to improve “joined up thinking” [Smithetal2015].

34. 3. Any algorithmic process, whether advisory or determinative should only have those data that a manual process would be allowed to consider (see Ba para23).

35. 4. An advisory or deterministic system having significant effects (e.g. bail or criminal sentencing) should, in the current state of technology, be reducible to a formula (F para27). In all circumstances, it should be explicable, at least to an independent expert (GDPR 15.1(h)). It should undergo formal testing for indirect discrimination (as in Bb para23). This should be required in the purchasing process, but also verified by an independent competent authority.

36. 5. Since the process should be explicable, it should be possible to produce some a posteriori estimate of accuracy for a discrete algorithm, or of the extent to which this is a marginal decision.

37. 6. In any such process, the person concerned should automatically be shown the data on them (the foreground data) being fed to any algorithm. This is the algorithmic equivalent of being present in court, and should be automatic, not just “on request”. There should be the same rights of objection/appeal here as with anything else said in court.

38. 7. Personal data should not be fed into algorithms doing online learning from new data (see G para28). This restriction may change when we have sufficient understanding of the issues. In addition, there is no guarantee that an online algorithm will remain unbiased, or relevant. Hence we recommend that truly online algorithms not be used, and replaced by ones where the data are updated and full revalidation carried out.
39. In machine learning algorithms, the background data contribute to the decision, so every such algorithm should be prominently labelled with the data that created it, both as a statement the lay person can understand (e.g. “based on London traffic data 1982-2002”) and question (“but that was before the congestion charge”), and ultimately such that an expert can analyse it. This is an obvious (to mathematicians) consequence of the 15.1(h) right.

40. Commissioners of machine learning algorithms should be aware of the pitfalls in H (para 29) above.

October 2017

References

   http://mobile.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html

42. [Angwinetal2016] Angwin, J., Larson, J., Mattu, S. & Kirchner, L.,
    Machine Bias. There is software that is used across the county to predict future criminals. And it is biased against blacks.

43. [ChamberlainPinker2016] Chamberlainb, N. & Pinker, E.,


45. [Evtimovetal2017] Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A. & Song, D.,

46. [Floresetal2016] Flores, A.W., Bechtel, K. & Lowenkamp, C.T.,
False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.


47. [GoodmanFlaxman2016] Goodman,B. & Flaxman,S.,
   European Union regulations on algorithmic decision-making and a "right to explanation".

   Accuracy in the Clinical Diagnosis of Malignant Melanoma.

49. [Grindrod2016] Grindrod,P.,
   Beyond privacy and exposure: ethical issues within citizen-facing analytics.
   DOI: 10.1098/rsta.2016.0132.


51. [Lietal2015] Li,Y., Zheng,Y., Ji,S., Wang,W., Hou U,L., Gong,Z.,
   Location selection for ambulance stations: a data-driven approach.
   Proc. 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, New York, NY, USA, Article 85.
   DOI: 10.1145/2820783.2820876.

52. [MacKayetal2017] MacKay,R.S., Kenna,R., Low,R.J. & Parker,S.,
   Calibration with confidence: A principled method for panel assessment.
   DOI [http://dx.doi.org/10.1098/rsos.160760](http://dx.doi.org/10.1098/rsos.160760)

53. [McDonald2015] McDonald,S.,
   [https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=RES2015&paper_id=791](https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=RES2015&paper_id=791)

54. [Nguyenetal2015] Nguyen,A., Yosinski,J. & Clune,J.,
Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images.

55. [Smithetal2015] Smith,J.Q., Barons,M.J. & Leonelli,M.,

56. [Szegedyetal2013] Szegedy,C., Zaremba,W., Sutskever,I., Bruna,J., Erhan,D.,
Goodfellow,I. & Fergus,R.,
Intriguing properties of neural networks.
https://arxiv.org/abs/1312.6199