PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Opinion piece

Check for updates

Cite this article: Blacklaws C. 2018 Algorithms: transparency and accountability. *Phil. Trans. R. Soc. A* **376**: 20170351. http://dx.doi.org/10.1098/rsta.2017.0351

Accepted: 4 June 2018

One contribution of 13 to a discussion meeting issue 'The growing ubiquity of algorithms in society: implications, impacts and innovations'.

Subject Areas:
algorithmic information theory

Keywords:
algorithms, transparency, accountability, law

Author for correspondence:
Christina Blacklaws
e-mail: christina.blacklaws@lawsociety.org.uk

Algorithms: transparency and accountability

Christina Blacklaws

Law Society of England and Wales, The Law Society's Hall,
113 Chancery Lane, London WC2A 1PL, UK

This opinion piece explores the issues of accountability and transparency in relation to the growing use of machine learning algorithms. Citing the recent work of the Royal Society and the British Academy, it looks at the legal protections for individuals afforded by the EU General Data Protection Regulation and asks whether the legal system will be able to adapt to rapid technological change. It concludes by calling for continuing debate that is itself accountable, transparent and public.

This article is part of a discussion meeting issue 'The growing ubiquity of algorithms in society: implications, impacts and innovations'.

1. Algorithms, transparency and accountability

It is often argued that we are looking to a future in which decision-making based on automated processing of large datasets becomes increasingly common. This may be true, but it is also true that such decision-making is already with us. Indeed, according to computer scientist Pedro Domingos [1] 'People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world.' Perhaps there's some truth in that.

Big data, machine learning, algorithmic decision-making and similar technologies have the potential to bring considerable benefit to individuals, groups and society as a whole. They could also create new injustices and embed old ones in ways that allow them to be powerfully replicated across national and international networks.

This power should concern us. The law acts as a brake on power—subjecting power to its rules. That is what we mean by the rule of law. Powerful algorithms should be no exception to this. Tom Bingham, the former senior Law Lord, exploring the nature of the rule of law, argues

THE ROYAL SOCIETY
PUBLISHING

that the law must be accessible and, so far as possible, intelligible, clear and predictable [2]. Whether the rule of law does require such substantive content is a matter for debate but these are surely desirable objectives and they are characteristics to which the law aspires. We can contrast such accessibility and clarity (not always achieved) with the situation described by Cathy O'Neil in her book *Weapons of math destruction* [3] when she argues that 'verdicts from WMDs land like dictates from the algorithmic gods'. She points out that the models which these algorithms implement are often 'black boxes, their contents a fiercely guarded corporate secret'.

2. Royal Society and British Academy reports

Accountability and transparency are significant themes in the Royal Society's report on *Machine learning* [4] and the follow-up joint report by the British Academy and the Royal Society on *Data management and use* [5]. Two issues stand out. The first is that making computer code available—making it open source—is not enough. We might assume that, when faced with an algorithmic black box, the most revealing path to transparency would be to look inside. This may not tell us very much because the algorithm is learning from the data on which it was trained but the outcomes we want to understand are determined by how it actually weights that data. The second insight is that there is a trade-off between accuracy and interpretability—between the interpretability of hard-coded rules and the greater accuracy of neural networks.

Machine learning [4] is surely right to argue that, where the outputs from machine learning systems have a particularly significant social or individual impact, the trade-off between accuracy and interpretability should tilt in the direction of interpretability.

And is it true, as some people argue, that there are gaps and that new models of legal accountability—and liability—will be needed to bring decision-making algorithms within a proper legal framework? What does the law have to say?

3. The EU General Data Protection Regulation

The EU General Data Protection Regulation (GDPR), which comes into force in EU Member States in May 2018 [6], modernizes a European data protection regime that dates back a quarter century. It recognizes that rapid technological developments since the early 1990s present new data protection challenges. Among these challenges are the unprecedented scale of personal data use in the public and private sectors and the need for revised rules about automated decision-making and profiling.

The Article 29 Working Party [7] comprising data protection regulators across Europe in an opinion on profiling under the GDPR argued (p. 5) that 'advances in technology and the capabilities of big data analytics, artificial intelligence and machine learning have made it easier to create profiles and make automated decisions with the potential to significantly affect individuals' rights and freedoms.' It noted (p. 6) that the GDPR has provisions that seek to ensure that there is no unjustified impact on individuals' rights. These provisions include articles that impose:

- specific transparency and fairness requirements;
- greater accountability obligations;
- specified legal bases for processing;
- rights for individuals to oppose profiling—specifically for marketing; and
- in certain circumstances a requirement to carry out a data protection impact assessment.

4. Transparency and accountability in the General Data Protection Regulation

A number of provisions in the GDPR seek to promote a high degree of transparency in the processing of personal data [6]. In general these provisions require data controllers to provide data subjects with information about the processing of their personal data and to do so in a concise, transparent, intelligible and easily accessible form, using clear and plain language.

Where personal data are obtained from the data subject, Article 13(2)(f) requires data controllers to provide data subjects with information about 'the existence of automated decision-making, including profiling … and meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.' The purpose of such information provision is said to be 'to ensure fair and transparent processing'. This provision is mirrored in Article 14(2)(g) to cater for the situation in which personal data have not been obtained from the data subject. However, in the latter case, such information need not be provided if it 'proves impossible or would involve a disproportionate effort'.

The GDPR also introduces an explicit accountability principle that was arguably only implicit in the former Data Protection Directive [8]. This makes controllers responsible for, *and be able to demonstrate*, compliance with the GDPR (Article 5(2)).

5. Automated decision-making and the General Data Protection Regulation

The GDPR distinguishes general profiling, decision-making based on profiling and *solely* automated decision-making about individuals including profiling [6]. 'Profiling' is defined in Article 4(4) as 'any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular, to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'.

Article 22 of the GDPR provides certain safeguards in relation to profiling and automated decision-making to sit alongside the transparency and accountability obligations it imposes upon data controllers. The basic rule about solely automated decision-making or profiling with legal or other significant effects is that it is not allowed or, in the words of the Article 29 Working Party [7], 'a general prohibition on this type of processing exists to reflect the potentially adverse effect on individuals'.

But this is rightly not an absolute prohibition. Such processing can be undertaken if it is necessary for the performance of, or entering into, a contract; authorized by law which also lays down suitable protections for data subjects' rights and interests or based on a data subject's explicit consent.

Moreover, where the processing is based on contract or consent, data subjects have the right to obtain human intervention, to express a view and to contest the decision. Finally, there are further restrictions where decisions are based on special categories of personal data—the kind of personal data that could be particularly prone to being used to discriminate against individuals (discussed further below).

Consent might appear to offer a strong mechanism for legitimating automated decision-making and profiling. As a concept it is much used in the law. However, both *Machine learning* [4] and *Data management and use* [5] recognize its limits:

— Consent is rarely fully informed: how many people even read website terms and conditions?

— You need time, energy and understanding for meaningful consent and consent suffers from the so-called 'transparency paradox': detailed explanations may not be understood; simplified ones gloss over important details.

— An individualistic model of consent may fail to address such matters as the value of aggregated data, the relevance of individual genomic data to family members and the capacity to infer certain characteristics of the many from the data obtained by consent from the few.

To some extent these limitations in the traditional notion of consent are recognized in the GDPR [6]. 'Consent', which was less rigorously defined in the Data Protection Directive [8], is now defined as 'any freely given, specific, informed and unambiguous indication of a data subject's wishes by which he or she by a clear affirmative action signifies agreement'. This means there must be separate consents for different data uses and pre-ticked consent boxes are no longer acceptable. Where there is a clear power imbalance between a data controller and an individual, as where the controller is a public authority or an employer in relation to an employee, consent may not be freely given. There are also special provisions concerning the consent of children to information society services.

However, I think that Hannah Knox in her perspective 'Consent in a digital age' in *Data management and use* [5, pp. 36–37] is right to raise the question of whether, rather than approaching consent from just a legal or technical perspective, we may need a more human-centred understanding of what consent is and how it can be established.

6. Algorithmic discrimination and the General Data Protection Regulation

We should also explore what the law will do to address algorithmic discrimination. The GDPR's [6] starting point is to prohibit processing of special categories of personal data—for example, personal data *revealing* racial or ethnic origin, political opinions, health, religious or philosophical beliefs, trade union membership or sexual orientation. It also prohibits the processing of genetic data, and biometric data for the purpose of uniquely identifying a natural person. The GDPR recognizes that processing these special categories of data could create significant risks to fundamental rights and freedoms. It lifts the prohibition on processing these data in only a limited number of circumstances—for example, where processing is necessary for the establishment, exercise or defence of legal claims or whenever courts are acting in their judicial capacity (Article 9).

The exceptions to the general prohibition in the GDPR on solely automated decision-making or profiling with legal or other significant effects, which I have already outlined, are further tightened in relation to the special categories. The exceptions do not apply unless the data subject has given explicit consent for them to be processed for specified purposes or processing is necessary for reasons of substantial public interest.

What does this mean in practice? Article 9(1) of the GDPR applies to the '[p]rocessing of personal data *revealing* [special categories of personal data]' (my emphasis). Bryce Goodman's paper [9] on accountable algorithms, algorithmic discrimination and the GDPR points out that the effect of these provisions concerning special categories of data depends on the interpretation of *revealing*. As a minimum, the explicit use of special category data—for example, political opinions—would be excluded from the datasets used for machine learning. But what about the maximal limit? Would the use of proxy variables (age, income, choice of newspaper, etc.) that might correlate highly with political opinions also be excluded? As *Machine learning* [4] explains 'machine learning ... destabilises the current distinction between "sensitive" or "personal" and "non-sensitive" data' because 'it allows datasets which at first seem innocuous to be employed in ways that allow the sensitive to be inferred from the mundane.' The example cited is of how Facebook 'Likes' could be used to predict sensitive user characteristics, including sexual orientation, ethnicity and religious and political views.

7. The limits of the General Data Protection Regulation

The GDPR [6] wrestles with automated processing and profiling and it clearly imposes some important transparency and accountability requirements. But the GDPR has its limits. Its realm is the realm of personal data. 'Personal data' is broadly defined under both the former EU Data Protection Directive [8] and the GDPR. The GDPR definition in Article 4(1) explicitly adds name, location data, online identifiers and genetic factors as examples of personal data to a similar definition in the Directive to give the following definition:

> ‘personal data’ means any information relating to an identified or identifiable natural person (data subject); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

The breadth of ‘personal data’ has been emphasized in recent European jurisprudence. In Nowak v Data Protection Commissioner [10], in a matter referred to them for a preliminary ruling by the Supreme Court of Ireland, the question for the Court of Justice of the European Union (CJEU) was whether a candidate’s corrected examination script was personal data (it was agreed that the result of the examination was such data).

In its judgment [10] the CJEU noted that ‘[t]he use of the expression “any information” in the definition of the concept of “personal data”, within Article 2(a) of Directive 95/46, reflects the aim of the EU legislature to assign a wide scope to that concept, which is not restricted to information that is sensitive or private, but potentially encompasses all kinds of information, not only objective but also subjective, in the form of opinions and assessments, provided that it “relates” to the data subject’ (para. 34). Their conclusion was that ‘the written answers submitted by a candidate at a professional examination and any comments made by an examiner with respect to those answers constitute personal data . . . ’.

Clearly, vast swathes of data fall within the definition of ‘personal data’, but not all. The CJEU also noted that ‘the aim of any examination is to determine and establish the individual performance of a specific person’ and it distinguished ‘a representative survey, to obtain information that is independent of that person’ (para. 41). A great deal of evidence-based social policy draws on research using essentially anonymized and aggregated data about individuals that cannot be classified as personal data. Decisions affecting individuals are also based on data about the natural environment or the performance of technological systems (for example diesel emissions). And so the most important of the limits to the GDPR, of course, is that it does not apply to non-personal data.

When we survey the wider landscape, questions of transparency and accountability take on a different hue. In thinking about machine learning and cyber-physical systems—systems that have a physical effect in the material world—we move away from concerns about individuals’ data protection rights and into the sphere of liability, including contractual and tortious liability, compensation for harm and insurance. These considerations are familiar enough in relation to today’s automobiles and they are likely to be equally relevant to autonomous vehicles—indeed, to the whole swathes of infrastructure, systems and artefacts. How the existing law copes and develops through case law or legislation remains to be seen. The instinct of many lawyers trained in the common law tradition is that the law of England and Wales offers a robust conceptual structure which can be adapted over time. What they may be less sure about is how much time we have. If, as so many experts appear to be warning us, the pace of technological innovation is accelerating in ways that will have a tangible impact, then it may be that, just as the first industrial revolution of the eighteenth and early nineteenth centuries led to significant developments in the law and in legal systems—including the rise of the profession of solicitor—so this twenty-first century revolution will demand similarly major changes.

8. Dialogue and democracy

The individual and social benefits that machine learning systems will bring must not come at the price of either justice or fairness. Powerful machine learning systems are, and must continue to be, subject to the rule of law. How far the legal framework will need to adapt in the future remains an open question and one that we need to continue to explore through ongoing dialogue.

Rigorous research, cross-disciplinary perspectives, public debate and democratic accountability are important components of this dialogue. The engagement of the British Academy and the Royal Society, the involvement of Parliament through the House of Commons Science and

Technology Committee's inquiry into algorithms and decision-making along with the recent report of the House of Lords Select Committee on artificial intelligence *AI in the UK: ready, willing and able?* [11] are all contributing to this debate.

In exploring the intelligibility of artificial intelligence (AI), the latter report makes a distinction between technical transparency (which, notwithstanding that experts might be given access to source code, would often be difficult or impossible to achieve) and explainability of a kind which relates to Article 22 of the GDPR and which the Select Committee [11] regarded as 'a more useful approach for the citizen and the consumer' (para. 105).

In moving forward, however, it is interesting that the Select Committee recommended an institutional approach in which the Centre for Data Ethics and Innovation in consultation with a range of other expert bodies, including the Alan Turing Institute, the IEEE and the BSI, should produce guidance on the requirement for AI systems to be intelligible. Alongside this work, and as part of the wider debate on transparency, accountability and algorithms, the Law Society will seek to play its part too.

Data accessibility. This article has no additional data.
Competing interests. I declare that I have no competing interests.
Funding. I received no funding for this study.

References

1. Domingos P. 2015 *The master algorithm: how the quest for the ultimate learning machine will remake our world*. New York, NY: Basic Books.
2. Bingham T. 2010 *The rule of law*. London, UK: Allen Lane.
3. O'Neil C. 2016 *Weapons of math destruction*. London, UK: Allen Lane.
4. Royal Society. 2017 *Machine learning: the power and promise of computers that learn by example*. London, UK: Royal Society. (https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf)
5. British Academy and Royal Society. 2017 *Data management and use: Governance in the 21st century*. London, UK: British Academy and Royal Society. (https://royalsociety.org/~/media/policy/projects/data-governance/data-management-governance.pdf)
6. European Parliament and Council of the European Union. 2016 *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Brussels, Belgium: European Commission. (https://publications.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en)
7. Article 29 Data Protection Working Party. 2017 *Guidelines on automated individual decision-making and profiling for the purpose of regulation 2016/679*, Report 17/EN, WP 251. Brussels, Belgium: European Commission. (https://ec.europa.eu/newsroom/document.cfm?doc_id=47742)
8. European Parliament and Council of the European Union. 1995 *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*. Brussels, Belgium: European Commision. (https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046)
9. Goodman B. 2016 A step towards accountable algorithms: algorithmic discrimination and the European Union General Data Protection. In *Machine Learning and the Law. 29th Conf. on Neural Information Processing Systems, Barcelona, Spain, 8 December*. (https://www.mlandthelaw.org/papers/goodman1.pdf)
10. CJEU. 2017 Nowak v Data Protection Commissioner. Case C-434/16. Court of Justice of the European Union. (http://curia.europa.eu/juris/liste.jsf?language=en&num=C-434/16)
11. Artificial Intelligence Select Committee. 2018 *AI in the UK: ready, willing and able?* 16 April 2018, HL 100 2017-19. London, UK: House of Lords. (https://publications.parliament.uk/pa/Id201719/Idselect/Idai/100/100.pdf)