Enslaving the algorithm: from a 'right to an explanation' to a 'right to better decisions'?

Pre-print, January 2018. Accepted for publication in IEEE Security & Privacy.

Lilian Edwards
Strathclyde Law School, University of Strathclyde, Glasgow, UK. lilian.edwards@strath.ac.uk

Michael Veale
Department of Science, Technology, Engineering & Public Policy (STEaPP), University College London, UK. m.veale@ucl.ac.uk

As concerns about unfairness and discrimination in "black box" machine learning systems rise, a legal "right to an explanation" has emerged as a compellingly attractive approach for challenge and redress. We outline recent debates on the limited provisions in European data protection law, and introduce and analyse newer explanation rights in French administrative law and the draft modernised Council of Europe Convention 108. While individual rights can be useful, in privacy law they have historically unreasonably burdened the average data subject. "Meaningful information" about algosrithmic logics is more technically possible than commonly thought, but this exacerbates a new "transparency fallacy"—an illusion of remedy rather than anything substantively helpful. While rights-based approaches deserve a firm place in the toolbox, other forms of governance, such as impact assessments, "soft law", judicial review and model repositories deserve more attention, alongside catalysing agencies acting for users to control algorithmic system design.

Contents

Introduction

Businesses and governments are increasingly deploying ML systems to make and support decisions which have a crucial impact on everyday life: decisions about (inter alia) criminal sentencing and release on bail; medical treatment; eligibility for welfare benefits; what entertainment we see and can access; the price and availability of goods and services delivered online; and the political information to which we are exposed. These ML systems—colloquially entering public consciousness as just 'algorithms', or even just "AI"—have been extensively criticised in the last few years as a result of a number of well known "war stories" which have revealed patterns of discrimination embedded but invisible to casual users in such systems [5]. Because algorithms are trained on historical data, they risk replicating unwanted historical patterns of unfairness and/or discrimination. For example, in hiring systems, a lack of women being hired in the past may mean the systems fails to recognise the worth of female applicants, or even outright discriminate against them. Luxury goods may be advertised to people with certain profiles on social media and not to others, creating a consumer "under class". A severe obstacle to challenging such systems is that outputs, which translate with or without human intervention to decisions, are made not by humans or even human-legible rules, but by less scrutable mathematical techniques. A loan applicant denied credit by a credit scoring ML algorithm cannot easily understand if her data was wrongly entered, or what she can do to have a greater chance of acceptance in the future, let alone prove the system is illegally discriminating against her (perhaps on race, sex or age). This opacity has been described as creating a "black box" society [9].

Enter the 'right to an explanation'

Since the 1990s, the law in Europe has been concerned with this kind of opaque, and difficult to challenge decision making by automated systems. In consequence, the Data Protection Directive (DPD), a measure which harmonised relevant law across EU member states in 1995, provided that a "significant" decision could not be solely on based *solely* on automated data processing (art 15). Some EU members interpreted this as a strict prohibition, others as giving citizens a right to challenge such a decision and ask for a "human in the loop". A second right, embedded within art 12, which generally gives users rights to obtain information about whether and how their particular personal data was processed, gave users the specific right to obtain "*knowledge of the logic involved in any automatic processing*" of their data. Both these provisions, but especially the latter were not much noticed, even by lawyers, and scarcely ever litigated, but have revived in significance in the latest iteration of EU DP law within the General Data Protection Regulation (GDPR), which passed in 2016 and will come into operation across Europe in 2018.

In the GDPR, art 15 has been transformed into art 22 and arguably created what the media and some technical press have portrayed as a new "right to an explanation" of algorithms. Art 12 has also been revamped to a new art 15 and now includes a right to access to "*meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing*" (art 15(1)(h)). This provision, notably, only applies in the context of "*automated decision making in the context of*" art 22. This leaves it unclear if all the

constraints on art 22 (discussed below) are ported into art 15 (though our view is that it does not). Sadly, all this adds up to a reality considerably foggier than the media portrayal.

Several factors undermine the idea that art 22 contains a "*right to an explanation*".

Primarily, art 22 does not in its main thrust even contain a "*right to an explanation*": merely a right to stop processing unless a human is introduced to review the decision on challenge. However Art 22 does refer at points to a requirement of "*safeguards*", both where the right to prevent processing (paradoxically) does *not* operate, and where it does *but* sensitive personal data is processed. In relation to the first case, "safeguards" are partly listed in art 22(3) but in the second case, the only guidance is in recital 71. ("Sensitive" personal data in DP law refers to a restricted list of factors regarded as peculiarly important such as health, race, sex, sexuality and religious beliefs).

It is important to note that in European legislation, the *articles* in the main text are binding on member states but are accompanied by *recitals*, which are designed to help states interpret the articles and understand their purpose. Recitals are usually regarded as helpful rather than binding, but this is contested and differs among states. Unfortunately, in relation to art 22, recital 71 mentions some key matters not included in the main text. Art 22(3) mandates that "*safeguards*" include "*at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision*", but the safeguards listed in recital 71 "should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, *to obtain an explanation of the decision reached after such assessment* and to challenge the decision" [italics added].

This strange mishmash of texts thus cannot firmly be said to mandate a right to explanation in all or indeed any circumstances and may not be interpreted the same way from state to state.

This is a serious, but not the only, problem with art 22.

- Art 22 applies only to systems where decisions are made in a "*solely*" automated way—i.e. there is no "human in the loop"—and there are very few of these and fewer that are "significant" (see below). How "meaningful" this input has to be is subject to recent regulatory guidance (see [13]), but still unclear.
- What is a "*decision*"? The GDPR gives us no help with this at all other than that it includes a "measure" (recital 71). Is sending a targeted ad to a user using an algorithmic system a "decision"? It produces no binding effect; the advert may be ignored; it is hard to see what action causally flows from it in many cases. Yet as in the well-publicised Latanya Sweeney example [11], the sending of adverts promoting help with criminal arrests solely to "black-sounding" names, was worrying and offensive; and potentially dangerous, if these characterisations were then transferred to a system used to mandate and select for stop and search, or for airport security. Though a single advert delivery decision might not have a significant effect on an individual's life, the cumulative effect on an entire group or class may be worrying. Such "group privacy" impacts are not dealt with well by DP law—an area based on individualistic human rights—and exacerbated by a continuing lack of provision for class actions in EU states.

- Art 22 applies only to a decision that produces legal or other "*significant*" effects. This is vague in the extreme. Some would argue this could only apply to systems which make important, binding decisions on things like criminal justice, risk assessment, credit scoring, education applications or employment. Yet such systems are rarely if ever entirely automated, even if the human's involvement is often nominal. Furthermore some commercial decisions may seem trivial as a one-off, but significant in aggregate. Mendoza and Bygrave argue that advertising decisions can never be significant [6] which does not seem unreasonable. But is a system recommending buying choices, or targeting adverts not significant if over time it limits a user's worldview or choices? What about one which disseminates "fake news" via algorithmic filter bubbles? Arguably, such phenomena are becoming deeply destructive to our democracy. We have an obvious link here to the issue of to whom a decision needs to be "significant": the individual in question, or society as a whole?

Turning to new art 15 of the GDPR (right to information), this right, to "*meaningful information about the logic involved*" in *any* decision-making system may be more useful than art 22. It is (arguably) not directly as restricted as art 22 is to "solely" automated decisions having "significant" or legal effects. But there is an unresolved doubt about whether it only applies to information available *before* the system makes a decision about a particular data subject (see Wachter et al. [14]). In ML parlance, that means it is uncertain if the right is only to a general explanation of the model of the system as a whole ("model-based explanation"), rather than an explanation of how a decision was made based on that particular data subject's particular facts ("subject-based" explanation) [5].

But even if we agree that art 15 may give us some kind of functioning right to an explanation, we still have huge problems. The GDPR can only apply where decisions are made based on *personal data*. Personal data is defined in art 4(1), as "*any information relating to an identified or identifiable natural person*" and is certainly wider than data which has the name of a data subject attached to it. According to recital 26:

> *"To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly."*

Even allowing for this broad (and, still, controversial) approach to defining personal data, it is clear some data will be anonymous and so not fall within the scope of the GDPR. Yet algorithmic decisions which affect people may not involve personal data. The most obvious example is self-driving cars. They may kill people – passengers or pedestrians – as a result of algorithmic processing, yet the data involved in that decision may be entirely related to traffic, road conditions and other non-personal matters. Other circumstances may involve data that was once personal, but has been allegedly anonymised. This is very common, e.g., with profiles made from personal data collected by social networks and used to generate targeted marketing.

Thirdly and lastly, a final restriction to these rights comes in the form of a carve-out in recital 63 (though not main text) for intellectual property (IP) and trade secrets. Explanations of how an algorithm works might reveal a firm's competitive advantage—its notorious "secret sauce".

Although this should not result in a "*refusal to provide all information to the data subject*", the lack of clear-cut provisions will likely continue to further muddy the waters. It is sometimes possible however to devise explanations of systems which do not need access to the proprietary innards of an ML model using "pedagogical" or "model-agnostic" methods [10, 12].

Improving the "right to an explanation" after the GDPR

Above, we have seen how the right to an explanation as we find it in the GDPR appears far from ideal. This is hardly surprising given that the provisions are still largely modelled on the 1995 Directive, which itself effectively predated the Internet and modern algorithm design.

More modern laws exist: one example is the French *loi pour une République numérique* (Digital Republic Act, Loi n 2016-1321) which gives a right to an explanation for administrative algorithmic decisions made about individuals. The new law provides that in the case of [author translation] *"a decision taken on the basis of an algorithmic treatment",* the rules that define that treatment and its *"principal characteristics"* must be communicated upon request. Further details were added by decree in March 2017 (R311-3-1-2) elaborating that the administration shall provide information about:

1. the degree and the mode of contribution of the algorithmic processing to the decision-making;
2. the data processed and its source;
3. the treatment parameters, and where appropriate, their weighting, applied to the situation of the person concerned;
4. the operations carried out by the treatment.

Some areas of government, such as national security and defence, are excluded.

The French approach has important advantages over the GDPR. First, looking at point 1, a decision is not, it seems, excluded from the right because it is only *partly* made by an algorithm.

Secondly, point 3 provides that where appropriate the *weightings* of factors in a system can be disclosed. This seems to imply the decision must be of a particular decision ("subject-based" explanation) rather than a vague overview of a complex model ("model-based" explanation), contradicting the interpretation by Wachter et al. [14]. Extracting estimates of the weightings within a complex algorithm is increasingly possible, particularly if only the area "local" to the query is being considered [10], which unlike the complex innards of the entire network, might display recognisable patterns [7].

Weights may help explain systems, but they are by no means a complete fast track to interpretability. There are at least two occasions when a court might say that weights are not useful for explaining a decision to a human user and therefore it is not "appropriate" to order disclosure. These are when the weighted inputs do not map to any real world features the user

will find intelligible, and, in older or restricted systems, where retrofitting an explanation system is unfeasible.

On the downside, the new French right applies only to administrative decisions. This makes its advances more comprehensible, for several reasons. Firstly, the number of discretionary decisions currently made by governmental algorithmic systems is relatively small compared to the increasing amount of ML profiling in the private/commercial sector. Secondly, there is a long established constitutional expectation that democratic governments will, to some extent, be transparent—e.g. via freedom of information (FOI) requests. By contrast the private sector is usually not required to disclose its secrets except on limited occasions such as financial disclosures.

Another new instrument in the planning is the modernisation of the Council of Europe *Convention 108 for the Protection of Individuals with regard to Automatic Processing of Personal Data* ("CoE 108"). CoE 108 is an international treaty relating to DP which, regardless of its name, can be signed by any state in the world. Its membership however remains relatively limited e.g. not currently including either the US or China.

In a recently circulated draft, it appeared that the CoE was considering as one alternative, a version of the "right to an explanation" for *all* automated decisions, without the restriction of art 22, GDPR. This would be an exciting development. While the outcomes of the CoE negotiations are far from settled, interestingly it seems this potentially expanded CoE right has already been ported to the draft UK transposition of the GDPR, the Data Protection Bill 2017 (see *https://perma.cc/X7X6-TXW9*). The Bill is oddly drafted but, in brief, the first two parts of the Bill reflect EU law since they transpose both the GDPR and the new EU Directive relating to DP in policing matters (Directive (EU) 2016/680). The third part of the Bill however applies DP law to UK intelligence services. These are outside of EU law, but the UK has chosen to make these provisions compliant, not with the GDPR but the draft modernised CoE. As a result the draft Bill's s96 currently contains one of the most advanced provisions on a right to explanation in the world! Before too much excitement is generated, however, it should be noted that any disclosures will still be controlled by overarching exemptions for national security in draft s108 and s109, and so are in fact never likely to be exercised.

Is a 'right to an explanation' the best remedy we can have?

The 'right to an explanation' is only one tool for scrutinising, challenging and restraining algorithmic decision making. While it has rhetorical strength in demanding transparency to enable user challenge, it has serious practical and conceptual flaws.

First, reliance on an explanation to bolster individual rights places a primary and heavy onus on users to challenge bad decisions. Even ordinary DP subject access requests (SARs) demand an enormous amount of time and persistence and in reality are mainly used effectively only by journalists and insiders who know how the company in question organises its data processing systems. Very few ordinary users historically make use of SARs and still less will probably use the right to an explanation. These issues are abetted by the common problems of consumer

access to justice, including a general lack of access to legal aid, and, in the EU, to class actions or collective redress. As noted below, users are supposed to be represented by their state Data Protection Authorities (DPAs) in Europe—but in reality this support may be lacking due to a lack of manpower and expertise in these bodies.

Second, an explanation even if obtained may not be helpful in mounting a challenge. This may not be just because of the well-known difficulties about expressing machine logic in human comprehensible form [5]; despite the progress that is being made in overcoming the hurdles to producing "meaningful information" about algorithmic logic. But algorithmic models, inputs and weightings, however disclosed, may still not show that a system has been designed to be biased, unfair, or deceptive. Most algorithms will display inadvertent bias rather than explicitly coded-in bias: designers will not want to be sued or prosecuted for illegal action even if their own ethics do not forbid it. Researchers are discovering now how difficult many of these problematic but non-obvious issues can be to spot even when they have the whole dataset to hand. Biased or discriminatory behaviour may only become apparent looking at the corpus of users as a whole—something which will not happen through individual user challenges.

It might be possible to better understand these aspects of a model as a whole if many individuals could utilise their individual rights to explanation at once. This would require ideally far better legal *and* technical mechanisms for collective action and challenge than we have now. At the moment even gathering information about the collective impact of algorithmic systems on users is difficult and unusual. A good example is the sending of "dark ads" to voters during recent political campaigns via political profiling of voters on social media platforms. Because these adverts were personally targeted to users, outside agencies could not even know what ads were being deployed less still count them or their influence. "Dark ads" during the 2016 British general election were tracked to some extent by volunteers who installed and browsed the internet with the WhoTargetsMe tool, but naturally this counting was very partial and non-representative. Given the already weighty burden on the individual, also requiring them to coordinate collective action to expose unfair algorithms seems unlikely to succeed.

In short, a legal right to an explanation may be a good place to start but it is by no means the end of the story. Rights become dangerous things if they are unreasonably hard to exercise, or ineffective in results, because they give the illusion that "something has been done" while in fact things are no better. It is instructive here to compare the history of consent to sharing of data, which has moved in the online world from a real bulwark of privacy to something most often described as "meaningless" or "illusory." Consent is usually given via privacy policies which are largely never read, and if read not understood; cannot be negotiated and change from time to time. Thus consent has become a formality validating the actions of the data controller rather than something empowering the user. This is sometimes known as the "notice and choice fallacy". It would be worrying and dangerous to see the "right to an explanation" become a similar empty formality. This "transparency fallacy" is something we should both guard against and which should spur us on, as in the next section, to look at alternative and supplementary ways to build better systems.

In particular any remedy given by a right to an explanation will often come too late for that particular user. The current investigations into automated recidivism decisions in the US biased against black data subjects show very clearly a history of discrimination for years which is only now becoming apparent. In many cases, we would *rather the system was never broken in first place*, or at very least that the individual decision concerned swung in a more just direction. This takes us to a range of *ex ante* as well as *ex post* governance tools.

In the next sections, we begin to consider what other regulatory tools have been created in the GDPR and elsewhere which might be pressed into use to try to assure, audit or instigate the creation of algorithms which are fairer, less discriminatory and, ideally, also less opaque.

Investigating *before* the system is built

Privacy by design, data protection by design and impact assessments

The GDPR introduces a number of new provisions which, rather radically, do not confer individual rights, but rather attempt to create an environment in which less "toxic" automated systems will be built in future. These ideas come out of the long evolution of "privacy by design" (PbD) engineering as a way to build privacy-aware or privacy-friendly systems, generally in a voluntary rather than mandated way. They recognise that a regulator cannot do everything by top down control, but that controllers must themselves be involved in the design of less privacy-invasive systems. These provisions include requirements that:

- controllers must, at the time systems are developed as well as at the time of actual processing, implement "appropriate technical and organisational measures" to protect the rights of data subjects (GDPR, art 25). In particular, "data protection by default" is required so that only personal data necessary for processing are gathered. Suggestions for PbD include making use of pseudonymisation and data minimisation;
- when a type of processing using "new" technologies is "likely to result in a high risk" to the rights of data subjects, then there must be a prior Data Protection Impact Assessment (DPIA) (art 35);
- every public authority and every "large scale" private sector controller and any controller who processes the "special" categories of data under art 9 (sensitive personal data) must appoint a Data Protection Officer (DPO) (art 37).

DPIAs especially have tremendous implications for ML design. Impact assessments are tools used in many domains to assess or estimate impacts of particular interventions or courses of action. GDPR, art 35 notes that:

> *"Where a type of processing in particular using new technologies [...] is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data".*

Where a DPIA "indicates that the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk", the controller is obliged to "consult [the data protection authority] prior to processing" (art. 36(1)) with a view to putting in measurers to mitigate the risk. Realistically, this seems only likely in cases of highly novel technologies or the use of existing technology in a new context; DPIAs are not intended as a tool to stop processing, but rather as a way to refine, or provide points of accountability for the future operation of, complex systems.

The Article 29 Working Party (A29 WP) (a body made up of national DP supervisory authorities which gives authoritative but not binding recommendations on how to interpret DP laws—soon to be revamped as the European Data Protection Board) has issued draft guidance (17/EN WP 248) further elaborating the conditions under which a data controller must carry out a DPIA. These include if two or more of the following conditions are met as part of processing:

- Evaluation or scoring of individuals;
- Automated decision-making with legal or similar significant effect;
- Systematic monitoring, including of a public area;
- Sensitive data processing, as defined in the GDPR;
- Data processed on a large scale;
- Matched or combined datasets, particularly if data subjects might have had different expectations about their use;
- Data concerning vulnerable data subjects;
- Innovative use of technological or organisational solutions;
- Data transfer outside the EU;
- Processing that prevents rights being exercised, such as in a public area people cannot avoid, or as a necessary prerequisite to service provision.

Taken together with the GDPR, this guidance indicates that a DPIA will be an obligatory precursor for many ML systems with sizeable anticipated risks or consequences for individuals or groups.

DPIAs are not a replacement for explanations of algorithmic systems, not least because they are aimed at helping builders and regulators, not, directly, users (although art 35(9) does provide that "*where appropriate, the controller shall seek the views of data subjects*".) DPIAs are also not required to be public documents, although it is considered good practice by the A29 WP to do so at least in part. However they may be of considerable value in leading to the building of better systems overall.

Arts 35 and 36 do not specifically require a DPIA to combat potential discrimination: early drafts made a DPIA mandatory where there was a "*risk of discrimination being embedded in or reinforced by the operation*" [1]. This amendment in the final text was relegated to recitals (see Recitals 71 and 75), which as discussed, have a murky status in EU law. However in its draft guidance, the A29 WP explicitly clarified that "rights and freedoms" in art. 35(1) "may also

involve other fundamental rights such as [...] prohibition of discrimination". The UK's Information Commissioner's Office (the ICO), in their influential guidance report on *Big data, artificial intelligence, machine learning and data protection* [8] note firmly that "potential privacy risks" have already been identified with "the use of inferred data and predictive analytics" and go on to provide a draft DPIA for big data analytics (Annex 1). It seems clear that, despite the uncertainty of the "high risk" threshold, DPIAs are quite likely to become the required norm for algorithmic systems, especially where sensitive personal data, such as race or political opinion, is processed on a "large scale" (GDPR, art 35(3)(b)).

Impact assessments that deal with risks of discrimination do already exist. The UK has extensive experience with Equality Impact Assessments (EqIAs), which used to be required for every new governmental policy, and were considered one of the main ways of documenting fulfilment of the Public Sector Equality Duty, itself brought into law by the Equality Act 2010. The Public Sector Equality Duty requires due regard to be given to impacts on protected classes before a policy is finalised or implemented, and it primarily accessed through judicial review. This requirement does not always need to be met via an EqIA (see *R (Brown) v Sec of State for Work and Pensions* 2008 EWHC 3158 (Admin)), but must be carried out with rigour before a policy is implemented, with documentation to show the process if it is not otherwise clear. Arguably, where new public sector decision support systems are built, an EqIA would be highly appropriate and could be combined with the DPIA process.

Certification systems

The GDPR also introduces the idea of voluntary certification for ML systems. Article 42 proposes voluntary certification of controllers and processors to demonstrate compliance with the Regulation, with "certification mechanisms" and the development of "seals and marks" to be encouraged by EU member states. In the UK, a tender has already been advertised by the ICO for a certification authority to run a UK privacy seal, although progress has been interrupted by the vote to exit the European Union, and the subsequent political turmoil.

Taken together, these provisions offer exciting opportunities to operationalise what in the US have been called "big data due process" rights [3, 4]. Certification could be applied to two main aspects of algorithmic systems:

1. certification of the algorithm as a *software object* by
    1. directly specifying either its design specifications or the process of its design, such as the expertise involved ("technology-based standards") and/or
    2. specifying output-related requirements that can be monitored and evaluated ("performance-based standards");
2. certification of the whole person or process using the system ("*system controller")* to make decisions, which would consider algorithms as situated in the context of their use.

One notable advantage is that certification standards could be set on a per-sector basis. This is already very common in other socio-technical areas, such as environmental sustainability standards.

The downside of what seems an exciting approach is that the history in the privacy domain of self-regulation of the private sector by seal and certificates is dispiriting. Essentially this involves privatisation of regulation and scrutiny. Certification scheme and trust seals have to make money to survive, which can only be obtained by asking fees from members. Given this self-interest, it is hard to punish members too hard when they breach the rules of the seal or certificate, for fear they will leave, either altogether or for a less demanding trust seal (in a plural market, which is generally what is envisaged). This in turn tends to diminish the value of the seal or certificate as a guarantee of trustworthiness. There is also little proof users regard seals and certificates as indicators of trust which may mean organisations are unwilling to pay or make an effort to belong to them, unless by doing so they can avoid more stringent "top down" regulation.

The clear worked example here is the US–EU "Safe Harbor" agreement for the export of personal data to the US which came ignominiously to an end in *Schrems v DPC of Ireland* in the CJEU in 2015 (Case C-362/14). The US as a country had been deemed not to have "adequate" protection in its law for personal data, and so in principal, EU data could not be exported to the US as far back as the DPD in 1995. A solution was found however in the "Safe Harbor" agreement whereby US companies could receive EU data if they joined a trust seal, membership of which guaranteed they were meeting adequate privacy standards. One of the largest and most prominent seals used in this way was TrustE. Yet as Charlesworth showed back in 2000 [2], TrustE had a long history of overlooking major data breaches by members or imposing only desultory sanctions. This, as well as the US's own cavalier attitude to covert state surveillance exposed in the Snowden revelations, lead to the agreement's judicial annulling.

Enabling review and challenge after a system is built

Representation bodies for data subjects

The aim of a right to an explanation is fundamentally to enable challenge to poor or wrongly made decisions by "black box" systems. Yet, as we have seen, individual users typically struggle to assert their rights, especially in complex and opaque areas such as ML systems. Moreover, if we are talking about a harm such as embedded systemic discrimination which typically affects a whole class of people, then it may seem far more appropriate for a representative body to accept and mount challenges than each individual user. This is a common problem in consumer law, and was long anticipated in DP law by the creation of state DPAs, which already have a role to investigate user complaints and enforce breaches against data controllers. One historic problem here has been the low level of sanctions a DPA could dish out: this has famously been met in the GDPR by the creation of a maximum fine of up to EUR 20 million or 4% of global annual turnover. This welcome change does not alter the fact however that the volume of DP breaches both deliberate and inadvertent is now so huge that

one agency alone cannot combat it. Furthermore, state DPAs throughout the EU are wildly underfunded, since in their nature (and by law), they have to be seen to be independent of both state and the commercial sector. A final major problem germane to ML-related complaints is that DPAs are typically staffed by civil servants and/or lawyers and rarely have much technical understanding or capacity.

The GDPR tries to help here in two main ways. Article 80(1) provides for all member states that a data subject can mandate a third body to lodge a complaint, exercise the right to judicial remedy and to receive compensation on his or her behalf. This is a useful step, particularly if civil society bodies can find an exemplary case to support: but it still requires a data subject to notice a breach, and have the time and effort to reach out to a third body and enlist their help. Given a third body can receive compensation on behalf of a data subject, this might also result in a dubious "ambulance chasing" industry (similar to the UK furore over wrongly paid payment protection insurance, or PPI).

Article 80(2), GDPR, by contrast, permits member states to allow third party bodies to take up complaints, for example against a data controller, *without being mandated by a data subject*. Through this provision, civil society bodies could monitor sectors and controllers for breaches and pursue suspected infringements of their own accord through judicial remedies.

Article 80(1) is mandatory for GDPR-implementing states to put into law, while Article 80(2) is not. While Germany looks to have implemented this, the UK's daft Data Protection Bill 2017 (thus far) does not. This seems odd, given the UK has already nominated 'super-complaint' NGOs that can, for example raise consumer or financial rights issues to regulators on behalf of groups they perceive as affected.

Bodies like this might become effective watchdogs in particular areas or sectors, but they will find it difficult to do this without having the capability for some access in the round to training data, input data, outputs and models of algorithmic systems in order to establish whether breaches are occurring. A large problem here is that courts have typically been reluctant to order access to source code for decision making systems even in relation to traditional non algorithmic systems. This is of course partly because of the issue of proprietary IP rights in code already noted above (see e.g. *Viacom v YouTube* Civ. 2103 (LL) in the US) but other issues may also be implicated.

In the UK there appears to be no reported case where a court has ordered disclosure of the source code of a decision support system to litigants, even in the surprisingly high number of disputes involving public sector systems, where issues of copyright might have been thought to be less prominent. It is interesting though that in at least one case, *Northern Metco Estates Ltd v Perth & Kinross District Council* 1993 S.L.T. (Lands Tr) 28, the output of a conventional though complex automated decision support system was doubted in respect to its value without more information as to how it was generated. The system in question calculated one factor ("economic rent") to feed into a compensation valuation in cases of compulsory acquisition of land. The court was disturbed at the lack of evidence they received as to exactly how this calculation had been done, but, interestingly, did not seem interested to find out more but rather to exclude it from influence. It seems quite likely that courts will be reluctant to become activist about disclosures of source code let alone algorithmic training sets and

models until they feel more confident of their ability to comprehend and use such evidence—which may take some time.

This likely judicial reluctance to become involved in unpicking algorithmic systems is unfortunate because at least in the public sector, a very valuable tool towards transparency might be found in the institution of judicial review. In the UK courts have the power on petition to review if administrative acts and discretion were legal and reasonable. In principle there seems no reason why the court's powers in such review might not extend to demanding access to data and models. However as noted above there is as yet no record of this having been done. This may be down to a combination of proprietary code issues and technophobia but there may also be worries about exposing sensitive personal data to the public eye (court documents are usually public).

One way forward here may be drawn from developing practice around sensitive public sector data and access to it by non-governmental data analysts. In recent years, much work has been done into building methods by which sensitive data can be accessed and analysed securely, proportionally and safely. National statistics offices have been working on ways to enable access to these rich troves of information to external data analysts without compromising security and privacy. Although several methods have been used here, the most interesting for our purposes is the development of secure environments for access to sensitive data. These locations are usually disconnected from a network and may be physically located in a public agency, private company or increasingly in academia.

We imagine that such systems might scale to become depositories for the scrutiny of the models and data associated with public sector algorithmic systems. Challenges to private sector systems might also benefit from such systems, albeit with IP issues to overcome. Archival or specialist libraries may yet become homes for infrastructure to publicly scrutinise models and code, just as they already often act as code escrow agencies.

Conclusions

In this article, we addressed the challenges of algorithmic systems which appear to deliver biased, discriminatory, unfair or opaque decisions and the rise of the "right to an explanation" as a possible legal remedy. We explored the limits of the right to an explanation in the GDPR, as well as in legislation drafted subsequently.

As things stands, the "right to an explanation" found in the GDPR, though beguiling, is uncertain, convoluted, rife with technical difficulties, and likely to be interpreted differently in different member states. There are key restrictions in the remedies it gives which are likely to exclude the decision support systems where transparency would be most welcomed: in crucial sectors such as criminal justice, policing, and child protection where decisions are not "solely" based on automated processing. As a rights based remedy, it places a large burden on individual users to not only seek their own explanations but follow up with challenges. It also does not work well as a remedy for harms experienced in aggregate by groups or protected classes. There is a danger the hype around the "right to an explanation" may create the belief

there is no need for further legal and non-legal solutions. Some newer legal instruments such as the new French law, offer refinements to the GDPR, but may also add their own restrictions.

Accordingly we proposed a wider sweep of attention towards other legal and para-legal remedies which may also have scope to impel the creation of better and more scrutable algorithmic systems. These include privacy by design, Data Protection Impact Assessments, and certification or seal schemes, all of which might offer scrutiny during the process of building systems rather than merely after they cause harm. We also suggested that redress for users after a system has caused harm could be improved by taking the opportunities the GDPR gives for representation and collective action by non-governmental bodies. Finally we noted that access to source code and data of algorithms will remain both a practical issue for NGOs as well as in contested legal cases for as long as courts are reluctant to order its disclosure, and suggested secure model depositories might help in this area.

Acknowledgements

Lilian Edwards was supported in part by the Arts and Humanities Research Council (AHRC) centre CREATe [grant number AH/K000179/1], and the EPSRC Digital Economy Hub Horizon at the University of Nottingham [grant number EP/ G065802/1]. Michael Veale acknowledges funding from the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/ M507970/1].

References

[1] Binns, R. 2017. Data protection impact assessments: A meta-regulatory approach. *International Data Privacy Law*. 7, 1 (2017), 22–35. DOI: 10.1093/idpl/ipw027.

[2] Charlesworth, A. 2000. Clash of the data titans—US and EU data privacy regulation. *Eur. Pub. L.* 6, (2000), 253.

[3] Citron, D.K. 2008. Technological due process. *Washington University Law Review*. 85, (2008), 1249–1313.

[4] Crawford, K. and Schultz, J. 2014. Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review*. 55, (2014), 93–128.

[5] Edwards, L. and Veale, M. 2017. Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law & Technology Review*. 16, (2017), 18–84. DOI: 10.2139/ssrn.2972855

[6] Mendoza, I and Bygrave, L.A. 2017. The right not to be subject to automated decisions based on profiling. *EU internet law: Regulation and enforcement*. T.-E. Synodinou et al., eds. Springer.

[7] Montavon, G. et al. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognit.* 65, (2017), 211–222.

[8] Information Commissioner's Office 2017. *Big data, artificial intelligence, machine learning and data protection*. ICO.

[9] Pasquale, F. 2015. *The black box society: The secret algorithms that control money and information*. Harvard University Press.

[10] Ribeiro, M.T. et al. 2016. Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1135–1144.

[11] Sweeney, L. 2013. Discrimination in online ad delivery. *Queue*. 11, 3 (2013), 10.

[12] Tickle, A.B. et al. 1998. The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Trans. Neural Netw.* 9, 6 (1998), 1057–1068.

[13] Veale, M. and Edwards, L. 2018. Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law & Security Review*, (2018). DOI: 10.1016/j.clsr.2017.12.002

[14] Wachter, S. et al. 2017. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*. 7, 2 (2017), 76–99. DOI: 10.1093/idpl/ipx005.