Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for

Lilian Edwards
Professor of Internet Law
University of Strathclyde
lilian.edwards@strath.ac.uk

Michael Veale
Doctoral researcher
University College London
m.veale@ucl.ac.uk

1 Introduction

Increasingly, algorithms regulate our lives. Decisions vital to our welfare and freedoms are made using and supported by algorithms that improve with data: machine learning (ML) systems. Some of these mediate channels of communication and advertising on social media platforms, search engines or news websites used by billions. Others are being used to arrive at decisions vital to individuals, in areas such as finance, housing, employment, education or justice.

The public has only relatively recently become aware of the ways in which their fortunes may be governed by systems they do not understand, and feel they cannot control; and they do not like it. Hopes of feeling in control of these systems are dashed by their hiddenness, their ubiquity, their opacity, and the lack of obvious means to challenge them when they produce unexpected, damaging, unfair or discriminatory results. Once, people talked in hushed tones about "the market" and how its invisible hand governed and judged their lives in impenetrable ways: now it is observable that there is similar talk about "the algorithm", as in: "I don't know why the algorithm sent me these adverts" or "I hate that algorithm".[1] Alternatively, algorithms may be seen as a magic elixir that can somehow mysteriously solve hitherto unassailable problems in society[2]. It seems that we are all now to some extent, "slaves to the algorithm". In his seminal book, Frank Pasquale describes this as "the black box society"[3], and the issue more broadly has become a subject of attention internationally by regulators, expert bodies, politicians and legislatures.[4]

---

[1] For qualitative interview work in this area, see Tania Bucher, "The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms" (2015) 20 *Information, Communication and Society* 1. (doi:10.1080/1369118X.2016.1154086)

[2] See with perfect concision, xkcd *Here to Help* (2017) Retrieved from https://xkcd.com/1831/

[3] Frank Pasquale *The Black Box Society: The Secret Algorithms that Control Money and Information* (2015, Harvard University Press) (hereafter "Pasquale").

[4] For regulators, see Information Commissioners Office (ICO) *Big data, artificial intelligence, machine learning and data protection* (2017, ICO) (hereafter "ICO Big Data"); European Data Protection Supervisor (EDPS) Opinion 7/2015 *Meeting the Challenges of Big data: a call for transparency, user control, data protection by design and accountability*, 19 November 2015.
For expert bodies see Royal Society *Machine learning: The power and promise of computers that learn by example*. (2017, Royal Society)*;* Wetenschapplijke Raad voor het Regeringsbeleid [Dutch Scientific Council for Government Policy] *Big Data voor een vrije en veilge samenleving* [Big Data for a free and safe society] (2016, WRR); Nesta *Machines that learn in the wild: Machine learning capabilities, limitations and implications*. (Nesta 2015)
For politicians see F Reinbold "Warum Merkel an die Algorithmen wil" (26 Oct 2016) *Spiegel Online*. Retrieved from http://www.spiegel.de/netzwelt/netzpolitik/angela-merkel-warum-die-kanzlerin-an-die-algorithmen-von-facebook-will-a-1118365.html ; See also pledges to regulate algorithms in the 2016 'Digital democracy' manifesto of Jeremy Corbyn, UK Labour Party leader (J Corbyn *Digital democracy* manifesto (2017)) and the 2017 presidential manifesto of Socialist candidate Benoît Hamon (B Hamon *Mon project pour faire batter le cœur de France* (2017)). For legislatures, see the 2017 inquiry into algorithmic decision-making of the UK Parliament's Commons Science and Technology Committee; in addition and more generally, see under the Obama administration, *Preparing For the Future of AI* , Executive Office of the President, National Science and Technology

There has been a flurry of interest in a so-called 'right to an explanation' that has been claimed to have been introduced in the EU General Data Protection Regulation (GDPR)[5]. This claim was fuelled in part by a short conference paper presented at a ML conference workshop[6], which has received considerable attention in the media[7]. However a similar remedy had existed[8] in the EU Data Protection Directive (DPD), which preceded the GDPR, since 1995[9]. This remedy held promise with its updated translation into the GDPR, yet in the highly restricted and unclear form it has taken, it may actually provide far less help for these seeking control over algorithmic decision making than the hype would promise.

Restrictions identified within the GDPR, arts 22 and 15(h) (the provisions most often identified as useful candidates for providing algorithmic remedies) include: carve-outs for intellectual property (IP) protection and trade secrets[10]; restriction of application to decisions that are "solely" made by automated systems(art 22 only); restriction to decisions that produce "legal" or similarly "significant" effects (ditto); the timing of such a remedy in relation to the decision being made; the authorisation of stronger aspects of these remedies by non-binding recitals rather than the GDPR's main text, leading to substantial legal uncertainty; and the practical difficulty in knowing when or how decisions are being made, particularly in relation to "smart" environments.[11] Given the volume of media and literature attention currently being paid to this possible "right to an explanation", our interest is threefold: what type of

---

Council, Committee on Technology, October 2016 (deleted from official White House page but now available in archive at https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf)

[5] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation, (hereafter "GDPR").

[6] B Goodman and S Flaxman "EU regulations on algorithmic decision making and "a right to an explanation", *2016 ICML Workshop on Human Interpretability in ML* (WHI 2016), New York, USA

[7] See as representative example from UK: "AI watchdog needed to regulate automated decision-making, say experts ", Guardian, Jan 27 2017 at https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions .

[8] Similarly as discussed in section 4 below, there is a long history of work into explanation facilities, previously referred to as "scrutability" in Web Science. See for example Judy Kay "Scrutable adaptation: Because we can and must" (2006) In: VP Wade and others (eds.) *Adaptive Hypermedia and Adaptive Web-Based Systems*, Lecture Notes in Computer Science, vol 4018. Springer, Berlin, Heidelberg.

[9] 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 (hereafter Data Protection Directive or "DPD"). For earlier discussions, see A Kobsa, "Tailoring privacy to users' needs". (2001) In: M Bauer and others (eds.) *User Modeling*, Springer (doi:10.1007/3-540-44566-8_52); Mireille Hildebrandt "Profiling and the rule of law." (2008) *Identity in the Information* Society, 55-70.

[10] Rosemary Jay "UK Data Protection Act 1998 - the Human Rights Context" (2000) 14:3 International Review of Law, Computers & Technology 385, DOI: 10.1080/713673366

[11] Sandra Wachter and others "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation" (2017) *International Data Privacy Law* (forthcoming); Mireille Hildebrandt "The Dawn of Critical Transparency Right for the Profiling Era" in J Bus and others (eds) *Digital Enlightenment Yearbook 2012* (IOS Press, 2012).

remedies currently exist in European law, how can they be meaningfully implemented, and are these the remedies one would really start from given a free hand.

This paper explores explanation as a remedy for the challenges of the ML era, from a European legal, and technical, perspective, and asks whether a right to an explanation is really the right we should seek. We open by limiting our scrutiny of "algorithms" in this paper to complex ML systems which identify and utilise patterns in data, and go on to explore perceived challenges and harms attributed to the growing use of these systems in practice. Harms such as discrimination, unfairness, privacy and opacity, are increasingly well explored in both the legal and ML literature, so here only highlighted to found subsequent arguments. We then continue on slightly less well travelled land to ask if transparency, in the form of explanation rights, is really as useful a remedy for taming the algorithm as it intuitively seems to be. Transparency has long been regarded as the logical first step to getting redress and vindication of rights, familiar from institutions like due process and freedom of information, and is now being ported as a prime solution to algorithmic concerns such as unfairness and discrimination. But given the difficulty in finding "meaningful" explanations (explored below), we ask if this may be a non-fruitful path to take.

We then consider what explanation rights the GDPR actually provides, and how they might work out in practice to help data subjects. To do this, we draw upon several salient algorithmic 'war stories' picked up by the media, that have heavily characterised academic and practitioner discussion at conferences and workshops. It turns out that because of the restrictions alluded to above, the GDPR rights would often likely have been of little assistance to data subjects "harmed" or "significantly" affected by algorithmic decision-making.

This exercise also identifies a further problem: DP remedies are fundamentally based around individual rights — since the system itself derives from a human rights paradigm — while algorithmic harms typically arise from how systems classify or stigmatise groups. While this problem is known as a longstanding issue in both privacy and equality law, it remains underexplored in the context of the "right to an explanation" in ML systems.

Next, we consider how a right to a "meaningful" explanation might practically be actioned given current technologies. First, we identify two types of algorithmic explanations: model-centric explanations (MCEs) and subject-centric explanations (SCEs). While the latter may be more promising for data subjects seeking individual remedies, the quality of explanations may be depreciated by factors such as the multi-

dimensional nature of the decision the system is concerned with; and the type of individual who is asking for an explanation.

On a more positive note though, we observe that explanations may usefully be developed for purposes other than to vindicate data subject rights. First they may help users to trust and make better use of ML systems by helping them to make better "mental maps" of how the model works; second, *pedagogical* explanations - constructed around the model, rather than made by decomposing it – may avoid the need to disclose protected intellectual property (IP) or trade secrets in the model, a problem often raised in the literature.

After thus taking legal and technological stock, we conclude that there is some danger of research and legislative efforts being devoted to creating rights to a form of transparency that may not be feasible, and may not match user needs. As the history of industries like finance and credit shows, rights to transparency do not necessarily secure substantive justice or effective remedies[12]. We are in danger of creating a "meaningless transparency" paradigm to match the already well known "meaningless consent" trope.

After this interim conclusion, we move on to discussing in outline what useful remedies relating to algorithmic governance may be derived from the GDPR other than a "right to an explanation". First, the connected rights-based remedies of erasure ("right to be forgotten") and data portability, in arts 17 and 20 respectively, may in certain cases be as useful, if not more so, than a "right to an explanation".

Second, we consider several novel provisions in the GDPR which do not, as is traditional, give individuals rights, but try to provide a societal framework for better privacy practices and design: requirements for Data Protection Impact Assessments (DPIAs) and privacy by design (PbD), as well as non-mandatory privacy seals and certification schemes. These provisions, unlike explanation strategies, may help produce both more useful *and* more explicable ML systems.

From these we suggest that we should perhaps be less concerned with providing individual rights on demand to data subjects and more concerned both with (a) building better ML systems *ab initio* and (b) empowering *agencies* such as NGOs, regulators or civil society scrutiny organisations not simply to challenge ML decisions on behalf of individuals, but to review the accuracy, lack of bias and integrity of a ML

[12] See for example Kate Crawford and Jason Schultz "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms" (2014) 55, 1 *Boston College Law Review*; Pasquale (n 3).

system in the round. US legal literature has begun to explore these options using its due process literature and public oversight experiences, with suggestions such as "an FDA for algorithms"[13] and variants on "big data due process".[14] However these solutions are currently largely aspirational, partly because the US lacks a clear omnibus legal regime around personal data to build on. European law, by contrast, provides a panoply of remedies in the GDPR that could be pressed into service immediately (or at least from May 2018 when it becomes mandatory law). Such approaches certainly come with their own challenges, but may take us closer to taming and using, rather than being 'enslaved' by, algorithms.

2 Algorithms, and how we might we be slave to them

Cast broadly, an algorithm is "any process that can be carried out automatically".[15] For our purposes, this definition is too wide to be helpful. Software has long been used for important decision-support, and this decision support has not existed within a governance vacuum. Algorithms have received plenty of unsung scrutiny in recent years across a range of domains. For example, in the public sector, a 2013 inventory of 'business critical models' in the UK government described and categorised over 500 algorithmic models used at the national level, and the quality assurance mechanisms that been carried out behind them.[16]

The algorithmic turn[17] that has been at the end of most recent publicity and concern relates to the use of technologies that do not model broad or abstract phenomena such as the climate, the economy or urban traffic, but model varied entities — usually people, groups or firms. They are primarily designed either to *anticipate* outcomes that are not yet knowable for sure, such as whether an individual or firm will repay a loan, or jump bail, or to *detect* and subjectively classify something unknown but somehow knowable using inference rather than direct measurement — such as whether a submitted tax return is fraudulent or not.

Lawyers involved with technology historically have experience in this area relating to rule-based "expert systems", although the substantive impact of these technologies on

---

13 Andrew Tutt "An FDA for Algorithms" (2016) Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=274799.
14 Crawford and Schultz (n 12), Danielle Keats Citron, "Technological Due Process" (2008) 85 *Wash. U. L. Rev.* 1249.
15 J-L Chabert and others, *A history of algorithms: From the pebble to the microchip* (Springer, 1999).
16 HM Treasury 2013. Review of quality assurance of government models: final report. HM Government: London
17 A variation on the more legally familiar "computational" turn: See M Hildebrandt and K de Vries (eds) *Privacy, Due Process and the Computational Turn* (Routledge, 2013)

lawyering has been relatively small compared to grand early expectations of wholesale replacement of imperfect human justice by computerised judges and arbitrators. Endeavours to create the "future of law" with expert systems in the 80s and 90s, whereby law would be formalised into reproducible rules, have largely been regarded as a failure except in some highly specific, syntactically complex but semantically un-troubling domains.[18] Not all scholars bought into this utopian vision uncritically — indeed, law was one of the earliest domains to be concerned about the application of ML systems without clear explanation facilities.[19]

2.1 The rise of learning algorithms

The lack of large-scale data and algorithmic architectures that could leverage them at the time initially scuppered progress in automated decision-making and decision support on anything but relatively simplistic problems. In recent years we have developed and deployed technologies capable of coping with more input data and highly non-linear correlations, and as a result we have been able to model social phenomena at a level of accuracy that is considerably more operationally useful. For a large part, this has been due to the move away from manually specified *rule-based algorithms*, where explicitly defined logics turn input variables, such as credit card transaction information, into output variables, such as a flag for fraud, toward *complex ML algorithms*, where output variables and input variables together are fed to an algorithm theoretically demonstrated to be able to learn from data. This process trains a model exhibiting implicit, rather than explicit, logics. The learning algorithms that make this possible are often not blazingly new, many dating from the 70s, 80s and 90s, but as we now have comparatively huge volumes of data that can be stored and processed cheaply, both the performance of and the speed of both deploying and further researching into ML systems has greatly increased.

Two main relevant forms of ML exist, which relate to the type of input data we have. *Supervised learning* takes a vector of variables[20], such as physical symptoms or

---

[18] See most notably Richard Susskind *Expert Systems in Law* (Clarendon Press, 1989); Zeleznikow J and Hunter D *Building Intelligent Legal Information Systems: representation and reasoning in law* (Kluwer, 1994); see also one of the authors' early rule based system efforts at L Edwards and JAK Huntley "Creating a civil jurisdiction adviser" (1992)1 *Information & Communications Technology Law* 1 5-40 (doi:10.1080/13600834.1992.9965640).

[19] For example John Zeleznikow and Andrew Stranieri "The split-up system: integrating neural networks and rule-based reasoning in the legal domain" (1995) *Proceedings of the 5th International Conference on Artificial Intelligence and Law, College Park, Maryland, USA, May 21 - 24, 1995*, doi:10.1145/222092.222235

[20] For those familiar with spreadsheet software, if you imagine a document where variables, such as age, income and marital status are columns, and observations, such as individuals in a particular year are rows, a vector of variables would be the same as selecting one row.

characteristics, and a 'correct' label for this vector, such as a medical diagnosis, known as a 'ground truth'. The aim of supervised learning is to accurately predict this ground truth from the input variables in cases where we only have the latter. *Unsupervised learning* is not 'supervised' by the ground truth: ML systems instead try to infer structure and groups based on other heuristics, such as proximity. Here, we might be interested in seeing which physical characteristics we could think of as 'clustered' together, without knowing immediately what such as cluster might mean.[21] Segmentation by market researchers, for example, would be a relevant field where unsupervised learning might be fruitfully applied, as here, we are interested in finding the most relevant groups for a given task.

Designers of ML systems formalise a supervised or unsupervised learning approach as a learning algorithm. This software is then run over historical training data. At various stages, designers usually use parts of this training data that the process has not yet "seen" to test its ability to predict, and refine the process on the basis of its performance[22]. At the end of this process, a *model* has been created, which can be queried with input data, usually for predictive purposes. Because these ML models are induced, they can be complex and incomprehensible to humans, as they were generated with predictive performance rather than interpretability as a priority. The meaning of "learning" in this context refers to whether the model improves at a specified task, as measured by a chosen measure of performance.[23] Evaluation, management and improvement of the resulting complex model is achieved not through the interrogation of its internal structure, but through examining how it 'behaves' externally — using performance metrics such as accuracy[24], for example.

ML is the focus of this piece, for several reasons. Partially, it is because in a more interconnected, more data-driven society, only they seem of performing functions we need. Because they can automate much more difficult or nuanced tasks than manually specified algorithms can, they form the technological basis for much of the critical infrastructure that we rely on to navigate, make sense and derive value from a society increasingly underwritten by data, such as search engines or voice recognition.

---

[21] Other types of learning exist, such as semi-supervised learning, where only some labels are present, and reinforcement learning, which learns from the feedback that results from action interaction, performs an action and receives feedback from this action. The former shares similarities with those discussed in the text above, and the latter has few current practical applications to draw on, so they are not discussed here.
[22] These separate collections of data are usually called 'test sets' or 'validation sets'.
[23] Tom Mitchell, *Machine learning* (1997 McGraw Hill).
[24] Accuracy alone is often not preferred, as a machine predicting rare events can get a high accuracy score simply by always predicting the event will not happen. One of the more common metrics in practice is AUC (area-under-curve), which signals a balance between avoiding false positives and false negatives. A variety of similar 'loss functions' exist, designed to measure performance in different ways and optimise on different grounds.

ML uptake is also driven by business models and political goals. Cheap computation has produced large datasets, often as by-products of digitised service delivery and so accruing to the Internet's online intermediaries and industrial giants as well as traditional nation-states. There has been a visible near-evangelical compulsion to 'mine' or infer insights from these datasets in the hope they might have social or economic value. New business models, particularly online, tend to offer services ostensibly for free, leaving monetisation to come from the relatively arbitrary data collected at scale along the way: a phenomenon some commentators refer to as 'surveillance capitalism'.[25] These logics of "datafication" have also lead to increasing uptake of ML in areas where the service offering does not necessarily require it, particularly in augmenting existing decisions with ML–based decision-support, in areas such as justice, policing, taxation or food safety.

2.2 ML and society: issues of concern

Aspects of ML systems have raised significant recent concern in the media, from civil society, academia, government and politicians. Here, we give a high level, non-exhaustive overview of the main sources of concern as we see them, in order to frame the social, technical and legal discussions that follow.

2.2.1 Discrimination and unfairness

A great deal of the extensive recent literature on algorithmic governance has wrestled with the problems of discrimination and fairness in ML[26]. Once it was commonly thought that machines could not display the biases of people and so would be ideal neutral decision makers.[27] This had considerable influence on some early legal cases involving Google and other online intermediaries and their responsibility (or not) for algorithmic harms[28]. The drafting process of the 1995 European DPD explicitly recognised this — the European Commission noted in 1992 that

> "the result produced by the machine, using more and more sophisticated software, and even expert systems, has an apparently objective and

[25] S Zuboff, "Big other: surveillance capitalism and the prospects of an information civilization" (2015) 30 *Journal of Information Technology* 1 75-89.
[26] See the useful survey in Mittelstadt and others, "The Ethics of Algorithms: Mapping the Debate" (2016) 3(2) *Big Data & Society* especially at section 7.
[27] See Christian Sandvig, "Seeing the sort: The aesthetic and industrial defence of "the algorithm"" (2015) 11 *Media-N* 1.
[28] See this "neutrality" syndrome imported by analogy with common carrier status for online intermediaries and usefully traced by Uta Kohl, "Google: the rise and rise of online intermediaries in the governance of the Internet and beyond (Part 2)" (2013) 21 *Int J Law Info Tech* 2 187-234.

> incontrovertible character to which a human decision-maker may attach too much weight, thus abdicating his own responsibilities."[29]

As Mittelstadt et al put it, "this belief is unsustainable"[30] given the volume of evidence which has emerged in the last decade, mainly in the US in relation to racial discrimination. If ML systems cannot be assumed to be fair and unbiased, then some form of "opening up the black box" to justify their decisions becomes almost inevitable.

Some have argued that 'big data' will eventually give us a complete picture of society.[31] Even if this was true — we will come to our reservations about this — making decisions based on past data if often problematic, as the structures that existed in that data often contain correlations we do not wish to re-entrench. These correlations frequently relate to "protected characteristics", a varying list of attributes about an individual such as so-called race, gender, pregnancy status, religion, sexuality and disability, which in many jurisdictions are not allowed to directly (and sometimes indirectly[32]) play a part in decision-making processes[33]. Algorithmic systems trained on past biased data are thus inherently likely to recreate or even exacerbate discrimination. This may be based on explicit bias of the developers but will almost always rather be indirectly, unintentionally and unknowingly discriminatory.[34]

A troubling issue here is what Gandy calls "rational discrimination".[35] In many cases, protected characteristics like race might indeed statistically correlate with outcome variables of interest, such as propensity to be convicted of property theft, submit a fraudulent tax or welfare claim, follow an advert for a pay-day loan, or fail to achieve seniority in certain jobs. While these correlations may be "true" in the sense of statistical validity, we societally and politically often wish they weren't. ML systems are designed to discriminate — that is, to discern — but some forms of discrimination

---

[29] COM(92) 422 final — SYN 297 at 26
[30] Mittelstadt and others (n 26) at 25
[31] Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think* (Houghton Mifflin Harcourt 2013).
[32] See the UK Equality Act 2010, s. 19; for discussion of US law, cf. Barocas and Selbst (n 3333).
[33] See for a central US discussion of this topic in the context of ML discrimination, Solon Barocas and Andrew Selbst "Big Data's Disparate Impact" (2016) 104 *California Law Review* 671.
[34] See Calders, Toon, and Indrė Žliobaitė. "Why unbiased computational processes can lead to discriminative decision procedures." In: Bart Custers and others (eds.) *Discrimination and Privacy in the Information Society*. (2013 Springer))43-57; Hajian, S. (2013) *Simultaneous discrimination prevention and privacy protection in data publishing and mining.* (PhD Thesis, Universitat Rovira I Virgili 2013)
[35] See Oscar H Gandy Jr *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage* (Routledge 2009).

seem socially unacceptable. One way forward is to try to build fair or non-discriminatory ML systems where these characteristics are not explicitly fed into the system, even if they have some predictive value — eg, by omitting the data column containing race or gender. However, this may still not result in a fair system as these excluded variables are likely related to some of the variables that are included, e.g. transaction data, occupation data, or postcode. As a recent example, a *ProPublica* investigation uncovered the apparent use of "ethnic affinity", a category constructed from user behaviour rather than explicitly asked of the user, as a proxy for race (which had been deliberately excluded as illegal to ask) for advertisers seeking to target audiences on Facebook to use.[36] However cases around 'redlining' on the internet — "weblining", as it was known nearly 20 years ago[37] — are far from new. A spate of stories in 2000 during the heady years of the dot-com bubble surrounded racist profiling using personal data on the internet: in July consumer bank *Wells Fargo* had a lawsuit filed against it for using an online home-search system to steer individuals away from particular districts based on provided racial classifications[38], whilst in April the online 1-hour-media-delivery service *Kozmo* received a lawsuit for denying delivery to residents in black neighbourhoods in Washington, DC[39], which they defended in the media by saying that they were not targeting neighbourhoods based on race, but based on high Internet usage.[40]

It is worth noting that in the EU, there have been far fewer scare revelations of "racially biased" algorithms than in the US. While some of this may be attributed to a less investigative journalistic, civil society or security research community, or conceivably, a slower route towards automation of state functions, it may also simply

---

[36] See Julia Angwin and Terry Parris Jr, "Facebook Lets Advertisers Exclude Users by Race" (Oct 28 2016) *ProPublica*. Retrieved from https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race . ProPublica subsequently reported in an update to the above article on February 8 2017 that Facebook had amended their advert dashboard system so that it "will prevent advertisers from using racial categories in ads for housing, employment and credit". The system will also warn advertisers to comply with the law in other categories.

[37] Marcia Stepanek, "Weblining: Companies are using your personal data to limit your choices — and force you to pay more for products" (3 April 2000) *Bloomberg Business Week.* Retrieved from http://web.archive.org/web/20170516143710/https://www.bloomberg.com/news/articles/2000-04-02/weblining .

[38] Credit Union Times, "Wells Fargo yanks "Community Calculator" service after ACORN lawsuit" (July 19 2000) Retrieved from http://www.cutimes.com/2000/07/19/wells-fargo-yanks-community-calculator-service-after-acorn-lawsuit?slreturn=1494945213 .

[39] Elliot Zaret and Brock N Meeks "Kozmo's digital dividing lines", (April 11 2000) *MSNBC* Retrieved from http://web.archive.org/web/20001217050000/www.msnbc.com/news/373212.asp?cp1=1

[40] Kate Marquess. "Redline may be going online" (2000) 86 *ABA Journal* August at 81. Retrieved from EBSCOhost.com

reflect a less institutionally racist mass of training data[41]. While racism is problematic around the world, countries with deeper racial cleavages are naturally going to collect more racist datasets.

Not all problematic correlations that arise in an ML system relate to characteristics protected by law. This takes us to the issue of unfairness rather than simply discrimination. As an example, is it fair to judge an individual's suitability for a job based on the web browser they use when applying, for example, even if it has been shown to be predictively useful?[42] Potentially, there are grounds for claiming this is actually "true" discrimination: because the age of the browser may be a surrogate for other categories like poverty, since most such applications may be made in a public library. Indeed, is poverty itself a surrogate for a protected characteristic like race or disability? Unfair algorithms may upset individual subjects and reduce societal and commercial trust, but if legal remedies come into the picture then there is a worry of over extending regulatory control. Variables like web bowser might, even if predictively important, be considered to abuse short-lived, arbitrary correlations, and in doing so, tangibly restrict individuals' autonomy.

In the European DP regime, fairness is an overarching obligation when data is collected and processed[43] something which is sometimes overshadowed by the focus on legitimacy and particular user rights. The UK ICO's recent guidance on big data analytics seems to imply that ML systems are not unfair simply because they are "creepy" or produce unexpected results[44]. However, they may be where they discriminate against people because they are part of a social group which is not one of the traditional discrimination categories; e.g. where a woman was locked out of the female changing room at the gym because she used the title "Dr", which the system associated with men only.[45] The ICO report argues that unfairness may, on occasion, derive from *expectations,* where data is used for a reason apparently

---

[41] Note the recent report of the ICO (n 4) which pays serious attention to issue of fairness and bias but cites only US examples of such despite being a product of the UK regulator. (The German autocomplete cases – see n XX - are cited but referred to interestingly, as questions of error or accuracy, rather than discrimination or fairness.) Note also the Obama administration report on AI (n 40), specified (at 30) that "it is important anyone using AI in the criminal justice system is aware of the limitations of the current data".

[42] "How might your choice of browser affect your job prospects?" *The Economist*, 11 Apr 2013.

[43] GDPR, art 5(1)(a).

[44] ICO (n 4).

[45] ICO (n 4) at 20, referencing J Fleig "Doctor locked out of women's changing room because gym automatically registered everyone with Dr title as male" (18 March 2015) *The Mirror*, 18 March. Retrieved from http://www.mirror.co.uk/news/uk-news/doctor-locked-out-womens-changing-5358594 . This is one of the very few reports of algorithmic misbehaviour in the ICO report not emanating from the US, and from the original text it is not clear if the fault was algorithmic (the simplicity of the task means it is very unlikely to be ML based).

unconnected with the reason given for its collection;[46] and from lack of *transparency*, which we discuss in detail below in section 2.2.3. One interesting factor is whether it is fair to discriminate against people on the basis of *incorrect* information they volunteer to attempt to protect their own privacy, such as disposable email addresses — a well observed practice.[47]

'Rational discrimination' from ML mirroring society too perfectly is far from the only problem connecting algorithmic systems to fairness. A major problem arises from the subjective ways we collect training set data. We rarely, if ever, have the complete picture in the data we collect. Most data are not gathered at random from society, but collected in ways that can be problematically skewed. For example, we cannot measure who breaks the law — only who is convicted of it. Data collection in this area is far from impartial or neutral, instead linked heavily to broader issues such as access to justice, societal prejudices, and policing strategies.

The fault here clearly lies primarily with the training and input data, and the lack of awareness and care taken in the process of transforming it into algorithmic models. As one pundit put it: "Algorithms discriminate. It's not their fault, they're strings of math, but people program them".[48] Programming in this case consists of the myriad social processes touching upon data collection, curation, cleaning, to model training and deployment in user interfaces or decision systems.

A range of further issues compound this, which we will only cover briefly. The act of categorising people at all is subjective, political, and often distributive, but necessary for many of these systems to be implemented at all.[49] Where correlations are sought out manually, or variables combined, these processes could often also be done in many ways.

---

[46] See further Helen Nissenbaum's well known theory of contextual integrity (*Privacy in Context: Technology, Policy, and the Integrity of Social Life* (2009, Stanford Law Books)). In the EU DP regime, in a perfect world, all purposes for which data is to be used, including re-uses, should be notified in the privacy policy or otherwise (GDPR art 5(1)(b)). However as discussed below in section 6.1, this concept of "notice and choice" is increasingly broken.

[47] ICO (n 4) at 26 citing evidence that 60% of UK consumers intentionally provide false data online. See M Chahal "Consumers are 'dirtying' databases with false details" *Marketing Week*, 8 July 2015 at https://www.marketingweek.com/2015/07/08/consumers-are-dirtying-databases-with-false-details/ . This is in fact increasingly recommended as a privacy-protective practice: see eg Brunton F and Nissenbaum H *Obfuscation : a User's Guide for Privacy and Protest (*2015, MIT Press*)*

[48] Video by *Fusion* on "right to explanation" at http://fusion.kinja.com/eu-citizens-might-get-a-right-to-explanation-about-the-1793859992 .

[49] See Geoffrey C Bowker and Susan Leigh Starr *Sorting things out: Classification and its consequences*. (MIT Press 2000).

Reliance on past data additionally asks fairness questions that relate to the memory of algorithmic systems — how far back is it appropriate to judge people on? Are individuals entitled to a *tabula rasa* after a certain number of years, as is common in some areas of criminal justice?[50] There is a widely-held societal value to being able to 'make a fresh start', and technological change can create new challenges to this. Other common institutional frameworks for forgetfulness can be found in bankruptcy law and in credit scoring.[51] This observation is not new, and the role of computers in undermining the principle of 'forgive-and-forget' has been observed by a range of authors.[52]

2.2.2 Informational privacy

Privacy advocates and data subjects have long had concerns relating to profiling, which as a general notion, is a process whereby personal data about a class of data subjects is transformed into knowledge or "inferences" about that group, which can then in turn be used to hypothesise about a person's likely attributes or behaviour. These might include the goods and services likely to interest them, the social connections they might have or wish to develop, medical conditions or personality traits. As the GDPR, art 4(4) now defines it, profiling is:

> "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."

---

[50] Under the UK Rehabilitation of Offenders Act 1974, as in many other European countries, disclosure of convictions (with some exceptions) is not required as these convictions become "spent", in spheres such as employment, education, housing and other types of applications. Whether such convictions would be erased from training set data would not however necessarily follow, depending on who maintained the record, legal requirements and how training set data was cleaned. Notably official advice on spent convictions advises job applicants with spent convictions to check what is (still) known about them to employers via Google and also advises them of their "right to be forgotten) (see below); see https://www.nacro.org.uk/resettlement-advice-service/support-for-individuals/disclosing-criminal-records/criminal-record-checks/ , site visited 3 May 2017.

[51] Jean-François Blanchette and Deborah G Johnson "Data retention and the panoptic society: The social benefits of forgetfulness" *The Information Society* (2002) 18(1) 33–45; Tanne van Bree "Digital hyperthymesia — on the consequences of living with perfect memory" In L Janssens (ed) *The Art of Ethics in the Information Society* (Amsterdam University Press 2016).

[52] See AF Westin and MA Baker *Databanks in a free society: Computers, record-keeping, and privacy* (Quadrangle 1972); GT Marx "The iron fist and the velvet glove: Totalitarian potential within democratic structures." In JF Short (ed) *The social fabric: dimensions and issues* (Sage 1972) 135–161; in the legal community, this idea was popularised by V Mayer-Schönberger *Delete! The Virtue of Forgetting in the Digital Age* (2009, Princeton University Press). We discuss the connected emergence of the "right to be forgotten" below in section 6.2.1.

ML can build such profiles — which are now often implicit and relational, rather than clear-cut categories — but profiling is wider than ML, and many means of profiling common today remain grounded in manually defined classifications and distinctions which might even predate the digital era. As Hildebrandt notes, profiling is what all organisms do in relation to their environments, and is "as old as life itself".[53]

For data subjects, privacy concerns here embrace an enormous weight of issues about how data about them are collected to be bent into profiles; how they can control access to and processing of these data; and how they might control the dissemination and use of derived profiles. In particular, ML and big data analytics in general are fundamentally based around the idea of repurposing data, which is in principle contrary to the DP principle that data should be collected for named and specific purposes (GDPR, art 5(1)(b), "purpose limitation"). Data collected for selling books becomes repurposed as a system to sell adverts book buyers might like. Connected problems are that such "big data" systems encourage limitless retention of data; and the collection of "all the data" rather than merely a statistically significant sample (contra principles in art 5(1)(e) and (c)). These are huge problems at the heart of contemporary DP law[54], and we do not seek to review these fully here. We do however want to point out where these issues specifically affect ML.

Fist, an exceedingly trite point is that data subjects increasingly perceive themselves as having little control over the collection of their personal data to go into profiles. In the GDPR, collection falls under "processing" of data (art 4(2)) and is theoretically controlled by (inter alia) the need for a lawful ground of processing (art 6). Most lay people believe the only such ground is consent and thus consent defends their right to autonomous privacy management (though perhaps not in so many words). Yet consent is not the only lawful ground under art 6, and quite possibly, as much personal data is collected (at least in the private sector) on the grounds of the "legitimate interests" of the controller, or on the grounds that the data was necessary to fulfil a contract entered into by the data subject.[55] More importantly, consent has become debased currency given ever-longer standard term privacy policies, screen layout manipulation and network effects in markets, and is often described in terms

---

[53] Mireille Hildebrandt "Profiling and the Rule of Law", (2008) *Identity in the Information Society*, 55-70.
[54] See discussion in Art 29 WP *Opinion 03/2013 on purpose limitation*, 2 April 2013; ICO Big data, at 11-12; EDPS, supra, n 2.
[55] GDPR, art 6. The public sector in effect has its own lawful ground for processing in the public interest. Policing and national security are exempted from the GDPR , though see n XX below.

such as “meaningless” or “illusory”.[56] As a result polls of European data subjects tend to show around three-quarters no longer feel they are fully in control of their personal data.

The consent problem is aggravated by the rise of “bastard data”, a picturesque term coined by Joe McNamee.[57]

> "In a highly interconnected world of “big data” … [d]ata are merged and compared. New data are generated and these, in turn can be compared with new data sets, with further new data being collected. Data have become fertile and have bastard offspring that create new challenges that go far beyond what society previously (and, unfortunately, still) considered to be “privacy”."

This is what profiling in ML systems does. Typically, data about people, which are personal, are transformed into data which has often been seen instrumentally as non-personal and therefore outwith the scope of DP law, perhaps simply because the data subject name or other obvious identifier has been removed.[58] Many businesses, particularly those operating online in social networking, advertising and search, have regularly argued that their profiles, however lucrative, merely involve the processing of anonymised data and hence do not fall within the scope of DP control. In recent times, the anonymity argument has been parried on grounds of potential for re-identification[59]. This has become especially crucial in the emerging ambient environment deriving from the Internet of Things (IoT). Data we would once have regarded as obviously non-personal such as raw data from home energy meters or location data from GPS devices is now, often through ML techniques, able to re-

---

[56] See for a full discussion of the illusory nature of consent in the Internet world, Edwards L “Anti-Social networking: social networks, privacy, law and code” in Brown I and Mueller M eds *Research Handbook on Internet Governance* (Edward Elgar, 2013); Joergensen, R “The unbearable lightness of user consent”. (2014) *Internet Policy Review*, 3(4). DOI: 10.14763/2014.4.330 ; Brendan Van Alsenoy, Eleni Kosta & Jos Dumortier "Privacy notices versus informational self-determination: Minding the gap" (2014) 28(2)International Review of Law, Computers & Technology 185-203, DOI: 10.1080/13600869.2013.812594.

[57] J McNamee "Is Privacy Still Relevant in a World of Bastard data?", EDRi editorial, 9 March 2016 at https://edri.org/enditorial-is-privacy-still-relevant-in-a-world-of-bastard-data .

[58] “Personal data” is defined at art 4(1) of the GDPR as “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”. Note also the debate over “pseudonymous” data during the passage of the GDPR, which is defined as data processed “in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information..”(art 2(5)). After some debate, the final text recognises explicitly that such data *is* personal data, although it garners certain privileges designed to incentivise pseudonymisation, eg it is a form of “privacy by design” (art 25), and is excluded from mandatory security breach notification. See also refs at n 60.

[59] See the seminal text of P Ohm “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization” (2010) 57 UCLA Law Review 1701. In Europe, see Article 29 Working Party (hereafter”A29 WP”) *Opinion 05/2014 on Anonymisation techniques*, 2014.

connected to individuals, and identities established from it.[60] In practice, this has meant that the day-to-day actions that individuals undertake, especially in "smart" environments[61], leave trails of potentially sensitive latent personal data in the hands of controllers who may be difficult to identify. If controllers are not identifiable, data subjects may not be able to effectively exercise the DP rights we discuss in sections 3 and 6 below even if they overcome the personal data and consent hurdles.

Profiles assembled via ML or other techniques may be seen as "belonging" to a *group* rather than an individual data subject. A profile does not simply identify the characteristics of individual data subjects; rather they are constructed by contrast with the other data subjects in the dataset. In a system attempting to target people by their entertainment choices, I am not simply someone who likes music festivals, but someone who is modelled as 75% more likely (give or take a margin of statistical uncertainty) to attend a music festival than the rest of my cohort. "Persistent knowledge" over time links me into this class of interest to the platform that holds the data. Mittelstadt argues that big data analytics allow this new type of "algorithmically assembled" group to be formed whose information has no clear protection in DP law and possibly not in equality law.[62]

This idea of "group privacy" was an early, albeit marginalised, concern in DP, referred to as "categorical privacy" by some authors in the late 90s[63], and sometimes conflated with discussions of what is personal data. As Hildebrandt stated in an early 2008 paper, "data have a legal status. They are protected, at least personal data are… [p]rofiles have no clear legal status".[64] Hildebrandt argues that protection of profiles is very limited, as even if we argue that a profile *becomes* personal data when applied to an individual person to produce an effect, this fails to offer protection to (or,

---

[60] On reidentifiability of smart meter data, see M Jawurek and others "Smart metering de-pseudonymization." (2011) *ACSAC '11*, ACM, 227–36 and V Tudor and others "A study on data de-pseudonymization in the smart grid" (2015) *Proceedings of the Eighth European Workshop on System Security*. ACM, 2; on general reidentifiability from mundane data such as location or credit card transactions, see Yves-Alexandre de Montjoye and others "Unique in the crowd: The privacy bounds of human mobility" (2013) *Scientific Reports*, 3, (doi:10.1038/srep01376) and Yves-Alexandre de Montjoye and others "Unique in the shopping mall: On the reidentifiability of credit card metadata." (2015) *Science*, 347, 6221. (doi:10.1126/science.1256297)

[61] See further Mireille Hildebrandt *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar, 2015); Lilian Edwards "Privacy, Security and Data Protection in Smart Cities: A Critical EU Law Perspective" (2016) 1 European Data Protection Law Review 28-58.

[62] Brent Mittelstadt "From Individual to Group Privacy in Big Data Analytics" (2017*) Philos. Technol.*. (doi:10.1007/s13347-017-0253-7). See also Anton Vedder "KDD: the challenge to individualism" (1999) *Ethics and Information Technology* 275; Mantelero, A. "From group privacy to collective privacy: towards a new dimension of privacy and data protection in the big data era". In Taylor, L., van der Sloot, B., Floridi, L. (eds). *Group Privacy: New Challenges of Data Technologies*. (Springer, 2017);

[63] Vedder (n 62).

[64] Hildebrandt (n 53).

importantly, control over) the relevant *group* profile. A decade later, the GDPR possibly makes progress by asserting that if a profile can be used to target or "single me out"[65] — for example, to deny me access to luxury services or to discriminate about what price I can buy goods at — then the profile is my personal data as it relates to me and makes me identifiable.[66] This approach however remains emergent and will be applied with hesitation even in some parts of Europe, given it is founded on a recital not a main text article.[67]

A final key issue is the ability of such systems to transform data categorised as "*ordinary*" personal data at the time of collection into data perceived as especially *sensitive*. In European DP law, "special" categories of data (known as "sensitive data" in the UK) receive special protection. These are defined as restricted to personal data relating to race, political opinions, health and sex life, religious and other beliefs, trade union membership and (added by the GDPR for some purposes) biometric and genetic data.[68] As already noted, in other areas of law, such as equality or employment law, subtly different "protected" characteristics may appear. In US privacy law, no general concept of sensitive data as such applies but it does have specific and highly regulated statutory privacy regimes for health, financial and children's data.[69] Sensitivity itself is thus a legally and culturally constructed concept.

A relevant and well-publicised "war story" is the 2012 story of how the American supermarket *Target* profiled its customers to find out which were likely to be pregnant, so that relevant offers could then be targeted at them. As a result, according to urban myth, a teenage daughter was targeted with pregnancy related offers before her father with whom she lived knew about her condition.[70] In DP law, if consent is used as the lawful ground processing of special categories of data, that

---

[65] See GDPR, recital 26.
[66] This discussion is important as whether a profile is seen as the personal data of a person also determines if they have rights to erase it or to port it to a different system or data controller. See discussion at sections 6.2.1 and 6.2.2.
[67] GDPR recital 26. See discussion of the status of recitals, below p xx. This approach to personal data has however been championed by the A29 WP for many years: see eg *Opinion 4/2007 on the concept of personal data* 01248/07/EN WP 136 at 13.
[68] GDPR, art 9 and see *Lindqvist v Kammaraklagaren*, European Court of Justice, Case C-101/01, 6 November 2003.
[69] See HIPAA (Health Insurance Portability and Accountability Act of 1996); Sarbanes–Oxley Act of 2002 (also known as the "Public Company Accounting Reform and Investor Protection Act"); and COPPA (Children's Online Privacy Protection Act of 1998) .
[70] Hill K, "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did", *Forbes,* February 16, 2012 at https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#6c8eb8fe6668 .

consent must be "explicit".[71] If ordinary data about purchases are collected and algorithmically transformed into insights that are sensitive, such as those related to health, or "protected", such as those relating to pregnancy, what is the correct standard of safeguard? For additional complication, the GDPR lays down a basic rule that profiling "shall not be based" on the special categories of personal data, although with an exception where there is explicit consent.[72] Does this apply to ML systems where the inputs are non-sensitive but the output inferences may be, as was the case in the *Target* profiling? These concerns are not just about race. How could such explicit consent to profiling be given where, say, data is legitimately gathered from public social media posts using the "legitimate grounds" ground of processing[73] and transformed into data about political preferences which is "sensitive" data in the GDPR (art 9(1))?[74] What about when ordinary data (e.g. number of miles walked per day) collected via a wearable like a Fitbit is transformed into health data used to reassess insurance premiums?[75]

---

[71]In an analogue world, this was a significant safeguard, as it meant that consent would often be taken with especial care, such as in written form. In an online world, where consent is most typically taken by reference to largely unread and ignored privacy policies, this is of limited protection, though it may push controllers towards more consumer friendly methods of consent collection such as opt-in rather than opt-out.

[71] GDPR, art 9(2)(a). Other grounds are available but what is noticeable for commercial data controllers is that processing cannot be justified on the basis that it was necessary for the "legitimate interests" of the controller, nor because it was necessary for the performance of a contract between data subject and controller – these are probably the two prevalent grounds for processing used in the commercial world. For executive and judicial processing of special categories of data (excluding policing which is excluded from the GDPR), the main grounds are art 9(2)(c) (emergency health situations),(f) (re legal claims or defences or judicial action)and (g) (substantial public interest).

[72] GDPR, art 22(4). It is also not clear if a controller can simply request a blanket consent to profiling of sensitive personal data in a privacy policy – which would tend to make this provision nugatory - or if something more tailored is needed. It is interesting that in the Common Statement of a number of EU DPAs on 16 May 2017 (https://www.cnil.fr/en/common-statement-contact-group-data-protection-authorities-netherlands-france-spain-hamburg-and , retrieved 20 May 2017) announcing a number of privacy breaches by Facebook, one issue is that FB "uses sensitive personal data fom uses without their explicit consent. For example, data relating to sexual preferences were used to how targeted advetisements" (Dutch DPA. It is not said if that data was created algorithmically or existed as a user input. See further on this issue, (n 125).

[73] GDPR, art 6 (1)(f)). Note that these interests may be over-ridden by overridden by the "interests or fundamental rights and freedoms of the data subject" and that this ground is *not* available for special categories of data under art 9 (see n 71 supra).

[74] This is quite likely part of the data collection and processing to produce political targeted ads pushed out via Facebook, which is being undertaken in the UK and elsewhere by companies like Cambridge Analytica. This whole area is currently highly sensitive, given concerns about recent elections and referenda, and is under investigation by the UK's ICO: see "Watchdog to launch inquiry into misuse of data in politics ", *Guardian*, 4 March 2017 at https://www.theguardian.com/technology/2017/mar/04/cambridge-analytics-data-brexit-trump . Interestingly the article itself is currently the subject of legal dispute! (retrieved at 19 May 2017.)

[75] This of course raises the issue of what we define as "health data" which the CJEU has not yet decided on. Similar issues have risen in US in relation to the scope of HIPPA. In an interesting example of "counter-profiling" obfuscation, see "Unfit Bits" in Khazan O "How to Fake Your Workout" *The Atlantic*, 28 September 2015 at https://www.theatlantic.com/health/archive/2015/09/unfit-bits/407644/

2.2.3 Opacity and transparency

Users have long been disturbed at the idea that machines might make decisions for them, which they could not understand or countermand, on the basis of flawed or incomplete data; a vision of out of control authority which derives from earlier notions of unfathomable bureaucracy found everywhere from Kafka to Terry Gilliam's *Brazil*. Such worries have emerged from the quotidian world (credit scoring, job applications, speeding camera tickets) as well as the emergent, fictional worlds of technology (wrongful arrest by *Robocop*, *2001*'s HAL, automated nuclear weapons launched by accident in *Wargames*).

In Europe, one of the earliest routes to taming pre-ML automated processing was the creation of "subject access rights" (SARs) empowering a user to find out what data was held about them by a company or government department, together with a right to *rectify* one's personal data — to set the record straight. These rights, harmonised across Europe in the DPD, art 12, included the right to rectify, erase or block data the processing of which did not comply with the Directive — in particular where they were incomplete or inaccurate. These rights were, as we shall discuss below, fused and extended into the so-called "right to be forgotten" in the GDPR, which succeeded the DPD in 2016. As Citron and Pasquale have mapped, although the US lacked an omnibus notion of DP laws, similar rights emerged in relation to credit scoring in the Fair Credit Reporting Act 1970.[76]

Domains such as credit scoring, public or rented housing applications and employment applications have entrenched in the public mind the intuition that challenging a decision, and possibly seeking redress, involves a preceding right to an explanation of how the decision was reached. This lead in Europe to a specific though rather under-used right in the DPD (art 15) to stop a decision being made solely on the basis of automated processing.[77] Data subjects had a right to obtain human intervention (a "human in the loop"), in order to express their point of view but this right did not, notably, contain an express right to an explanation.[78] This right was updated in the GDPR to extend to a more general concept of profiling[79] as discussed

---

[76] See Citron DK and Pasquale F "The Scored Society: Due Process for Automated Predictions" (2014) 89 Washington Law Review 1 .

[77] This is interestingly interpreted by Jones to imply that European systems are more interested in the human dignity of data subjects than the US system: see Jones, ML "Right to a Human in the Loop: Political Constructions of Computer Automation & Personhood from Data Banks to Algorithms" (2017) 47 Social Studies of Science 216.

[78] But see the information rights in art 12(a) DPD which became art 15(h) in the GDPR: discussed supra at XX.

[79] GDPR, art 4 (4). "Profiling" includes "any form of automated processing of PD consisting of the use of PD to evaluate certain personal aspects relating to a natural person in particular to analyse or predict [...] Performance

above. As Citron and Pasquale[80] map in detail, credit scoring has been a canonical domain for these issues in the US as well, as it has evolved from 'complicated' but comprehensible rule based approaches embodying human expertise, to 'complex' and opaque systems often accused of arbitrary or unfair decisions. As such this domain foreshadows the difficulties routinely encountered now in trying to interpret many modern ML systems.

Explanation rights of a sort are common in the *public* sphere in the form of freedom of information (FOI) rights against public and governmental institutions. Transparency is seen as one of the bastions of democracy, liberal government, accountability and restraint on arbitrary or self-interested exercise of power. As Brandeis famously said, "[s]unlight is said to be the best of disinfectants; electric light the most efficient policeman."[81] Transparency rights against public bodies enable an informed public debate, generate trust in and legitimacy for the government, as well as allow individual voters to vote with more information. These are perhaps primarily societal benefits, but citizens can clearly also benefit individually from getting explanations from public bodies via FOI: opposing bad planning or tender decisions, seeking information on why hospitals or schools were badly run leading to harm to one self or one's child, and requiring details about public funding priorities are all obvious examples.

By comparison with FOI, transparency rights are less clearly part of the apparatus of accountability of *private* decision-making. As Zarsky says, "the "default" of governmental action should be transparency".[82] The opposite is more or less true of private action, where secrecy, including commercial or trade secrecy (and autonomy of business practices[83]) and protection of intellectual property (IP) rights, are *de facto* the norm. Disclosure of personal data to its subject, from both public and private data controllers, is justified at root in Europe by the fundamental nature of privacy as a human right, sometimes extended to a separate right to DP.[84] DP law might seem quite odd when looked at from outside the informational privacy ghetto, as it is one

---

at work, economic situation, health, personal preferences, interests, reliability or behaviour, location or movements". Note such profiling may be achieved other than by ML; see discussion in section 2.2.2.

[80] Citron and Pasquale, supra n 76.

[81] Louis Brandeis *Other People's Money, and How Bankers Use it*. (National Home Library Foundation 1933) at 62.

[82] Zarsky T "Transparency in Data Mining : From Theory to Practice" in Bart Custers and others (eds) (supra n 34) at 301-324.

[83] Note that in the EU freedom to conduct a business is a fundamental right (art 14, Charter of Fundamental Rights of the European Union (CFEU) 2012/C 326/02).

[84] ECHR, art 8; CFEU arts 7 and 8.

of the few bodies of law that applies a general principle of transparency[85] even-handedly to private and public sector controllers, with more exceptions for the latter than the former in terms of policing[86]. and national security.[87]

Yet an explanation, or some kind of lesser transparency, is of course often essential to mount a challenge against a private person or commercial business whether in court or to a regulatory body like a privacy commissioner, ombudsman, trading standards body or complaints association. On a societal level, harmful or anti-competitive market practices cannot be influenced or shut down without powers of disclosure. The most obvious example of transparency rights in the private[88] sphere outside DP, and across globally disparate legal systems, lies in financial disclosure laws in the equity markets; however arguably these are designed to protect institutional capitalism by retaining trust in a functioning market rather than protecting individual investors, or less still, those globally affected by the movements of markets. Disclosure is also reasonably common in the private sector as a "naming and shaming" mechanism[89] — e.g. the introduction in the GDPR of mandatory security breach notification[90], or the US EPA Toxics Release Inventory.[91] Disclosures may also be made voluntarily to engage public trust as in programmes for visible corporate social responsibility (CSR), and standards for this exist with bodies such as the Global Reporting Initiative (GRI).

---

[85] GDPR, art 5(1)(a).

[86] See now GDPR art 2(2)(d) but note the new DP Policing Directive, Directive EU/2016/680 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data.

[87]Arguably despite these exceptions, European countries have traditionally been more transparent than the US in the development of ML systems used for judicial or penal decision support. It is not uncommon to find ML systems in Europe developed in-house, rather than privately procured and subject to proprietary secrecy. Indeed, risk scoring based on ML has been used in the UK and the Dutch justice system for several years (the Offender Assessment System (OASys) in the former and Statisch Recidiverisico (StatRec) in the latter), and both governments have historically published both model weights and detailed analysis of different aspects of their performance — for example in the way that predictive validity differs by race, age or gender. See Robin Moore (ed), *A compendium of research and analysis on the Offender Assessment System (OASys)*. (Ministry of Justice Analytical Series, 2015); Tollenaar N and others, "StatRec — Performance, Validation and Preservability of a Static Risk Prediction Instrument" (2016) 129 *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 25

[88] An ancillary question which cannot be pursued here is the question of how many of the functions of the state are now carried out by private bodies or public-private partnerships and what the resulting susceptibility to FOI requests (or other public law remedies, such as judicial review) should be.

[89] Zarsky (n 82) at 311.

[90] GDPR, arts 33-34.

[91] Khanna M and others "Toxics release information: A policy tool for environmental protection" (1998) *Journal of Environmental Economics and Management* 36(3), 243-66.

Despite the sometimes almost unthinking association of transparency and accountability, the two are not synonymous.[92] Accountability is a contested concept, but in essence involves a party being held to account having to justify their actions, field questions from others, and face appropriate consequences.[93] Transparency is only the beginning of this process. It is interesting that in the context of open datasets as a successor to FOI, there is considerable evidence that disclosure (voluntary or mandated) of apparently greater quantities of government data does not necessarily equal more effective scrutiny or better governance.[94] O'Neill calls this a "heavily one-sided conversation" with governments able to minimise the impact of disclosures by timing of release, difficulty of citizens in understanding or utilising the data, failures to update repositories and resource agencies who use and scrutinise open data, and general political obfuscation.[95] Heald terms this a "transparency illusion" which may generate no positive results while possibly creating negative impacts, such as privacy breaches and loss of trust if disclosures of maladministration are not met with punishment.[96]

Notwithstanding these doubts, and turning to ML systems, transparency rights still seem intimately linked to the ideal of effective control of algorithmic decision-making. Zarsky argues that the individual adversely affected by a predictive process has the right to "understand why" and frames this in familiar terms of autonomy and respect as a human being; Hildebrandt has long called for Transparency Enhancing Tools to control the impacts of profiling.[97] Similar ideas pervade the many calls for reinstating due process in algorithmic decision making[98], for respecting the right to a "human in the loop" as an aspect of human dignity[99] and for introducing "information accountability" in the form of "policy awareness" which will "make bad acts visible to all concerned"; or varied similar ideas of "algorithmic accountability."[100]

---

[92] See for example Pasquale's rejection of the idea that transparency has created any real effects on or accountability of the financial sector, especially after the financial crash and recession of the last decade. See "Finance's Algorithms: the Emperor's New Codes" in Pasquale (n 3).
[93] Mark Bovens "Analysing and Assessing Accountability: A Conceptual Framework" (2007) *European Law Journal* 13(4) 447–468 doi: 10.1111/j.1468-0386.2007.00378.x
[94] See Helen Margetts "Transparency and Digital Government" in C Hood and D Heald (eds.). *Transparency: the Key to Better Governance?* (Oxford University Press 2006) 3-23;
[95] O O'Neill "Transparency and the Ethics of Communication" in C Hood and D Heald (n 94) at 75-90.
[96] D Heald. "Transparency as an Instrumental Value" in C Hood and D Heald (n 94) 59-73.
[97] Zarsky (n 82) at 317; Hildebrandt, supra n 53 and in her subsequent work.
[98] For example, see Crawford and Schultz (n 12); Citron (n 14).
[99] See Jones (n 77)
[100] DJ Weitzner et al "Information Accountability" (2007) *Computer Science and AI Lab, Technical Report, MIT-CSAIL-TR-2007-034*. Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/37600/MIT-CSAILTR2007-034.pdf; M Perel and others "Accountability in Algorithmic Copyright Enforcement" (2016) 19 *Stanford Technology Law Review*; Diakopoulos N "Algorithmic Accountability Reporting: On The Investigation of Black

Yet this connection has never really been justified in terms of practical efficacy in relation to the broad range of algorithmic decisions. If we return to the notion of algorithmic "war stories" that strike a public nerve, in many cases what the data subject wants is *not* an explanation — but rather for the disclosure, decision or action simply not to have occurred. Consider, in relation to an individual, the *Target* pregnancy case mentioned in section 2.2.2 above, or another recent case of outrage affecting a group, when Google wrongly categorised some black people in its Photos app as gorillas.[101]

In the few modern EU legal cases we have on controlling algorithmic governance, an explanation has not usually been the remedy sought. An interesting example is the seminal CJEU *Google Spain*[102] case which introduced the "right to be forgotten" and is one of the few cases of algorithmic harm to have come to the highest EU court. In this case, the claimant, Mr Costeja, asking Google to remove as top link in searches on his name, a link to an old and outdated page in a newspaper archive recording his long-repaid public debt. Mr Costeja's successful ambition when he went to court was to remove the "inaccurate" data; he had, apparently, no interest in *why* Google's search algorithm continued to put long outdated results at the top of its rankings (even though arguably this was inexplicable in terms of how we think we know Page Rank works). A similar desire for an action, not for an explanation, can be seen in the various European "autocomplete defamation" cases.[103]

In all these cases, an explanation will not really relieve or redress the emotional or economic damage suffered; but it will allow developers not to make the same mistake again. Clearly these cases may not be typical. An explanation may surely help overturn the credit refusal issued by a machine, or an automated decision to wrongfully refuse bail to a black person or welfare to someone with medical symptoms — and these are obviously important social redresses — but it will not help in all cases. And even in these more mainstream cases, as Pasquale correctly identifies, transparency alone does not always produce either redress or public trust

---

Boxes" (2013) *Tow Centre for Digital Journalism*. Retrieved from http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf. For a rejection of rights of transparency as the answer to algorithmic accountability, see Kroll and others "Accountable Algorithms" (2017) *University of Pennsylvania Law Review*, 165, 633.

[101] See "Google Photos labels black people as 'gorillas' ", *The Telegraph*, 4 May 2017 at http://www.telegraph.co.uk/technology/google/11710136/Google-Photos-assigns-gorilla-tag-to-photos-of-black-people.html .

[102] *Google Spain v Agencia Española de Protección de Datos (AEPD) and González*, Case C‑131/12, 13 May 2014.

[103] Discussed in detail in Kohl (n 28) and Jones (n 77)

in the face of institutionalised power or money[104], just as David Brin's *Transparent Society* does not in fact produce effective control of state surveillance when the power disparity between the state and the *sousveillant* is manifest.[105]

Thus it is possible that in some cases transparency or explanation rights may be overrated or even irrelevant. This takes us to the question of what transparency in the context of algorithmic accountability actually means. Does it simply mean disclosure of source code including the model, and inputs and outputs of training set data? Kroll et al argue that this is an obvious but naïve solution since transparency in source code is neither necessary to, not sufficient for algorithmic accountability, and it moreover may create harms of its own in terms of privacy disclosures and the creation of "gaming" strategies which can subvert the algorithm's efficiency and fairness.[106] Instead they point out that auditing, both in the real and the digital world can achieve accountability by looking at the external inputs and outputs of a decision process, rather than at the inner workings. Even in the justice system, it is common for courts to adjudicate based only on partial evidence, since even with discovery, evidence may be unavailable or excluded on grounds like age of witness, hearsay status or scientific dubiety. We often do not understand how things in the real world work: my car, the stock market, the process of domestic conveyancing. Instead of (or as well as) transparency, we often rely on expertise, or the certification of expertise (e.g., that my solicitor who does my house conveyancing, is vouched for both by her law degree and her Law Society affiliation, as well as her professional indemnity insurance if things go wrong (see further, section 6.2.3).) Transparency may at best be neither a necessary nor sufficient condition for accountability and at worst something that fobs off data subjects with a remedy of little practical use.

We return to this question of "transparency fallacy" below at section 6.1, and to the question of what types of explanation in what circumstances may actually be useful, and to whom (sections 4 and 5). First however we consider the recent legal debate on whether a "right to an explanation" of algorithmic decisions does indeed exist in EU DP law.

---

[104] Pasquale (n 3) at 212.
[105] See B Schneier "The Myth of the "Transparent Society" *Wired* , March 6, 2008.
[106] See Kroll and others (n 100). Further discussion of "gaming" is found at section 5.1.

3 Seeking a right to an explanation in European data protection law

In 2016, to the surprise of some EU DP lawyers, and to considerable global attention, Goodman and Flaxman[107] asserted in a short paper that the GDPR contained a "right to an explanation" of algorithmic decision making. As Wachter and others have comprehensively pointed out[108], the truth is not quite that simple. In this section we consider the problems involved in extracting this right from the GDPR, an instrument still heavily built around a basic skeleton inherited from the 1995 DPD [109] and created by legislators who, while concerned about profiling in its obvious manifestations such as targeted advertising, had little information on the detailed issues of ML. Even if a right to an explanation can viably be teased out from the GDPR, we will show that the number of constraints placed on it by the text (which is itself often unclear) make this a far from ideal approach.

3.1 GDPR, article 22: automated individual decision-making

Our starting point is art 15 of the now-replaced DPD which was explicitly aimed at protecting users from unsupervised automated decision making. This rather odd little provision[110] was mainly overlooked by lawyers and commentators to the point of non-significance and few saw the potential it had as applied to algorithmic opacity. It is clear that art 15 of the DPD did not contemplate dealing with the special opacity found in complex, ML systems, and very little was changed to manage this in the new GDPR, art 22 which provides:

> "the right not to be subject to a *decision* based *solely* on automated processing, including profiling, which produces *legal effects, concerning him or her*, or *significantly affects him or her*". [italics added]

---

[107] Goodman and Flaxman (n 6)
[108] Wachter and others (n 11).
[109] Itself based to some extent on preceding national laws as well as the Council of Europe's work. For a history of DP law, see G Fuster *The emergence of personal data protection as a fundamental right of the EU* (Springer 2014).
[110] Mendoza and Bygrave describe art 15 as "a second class data protection right: rarely enforced, poorly understood and easily circumvented", not included in other fair information privacy schemes such as the OECD guidelines nor demanded by safe harbor (Mendoza I and Bygrave L "The Right Not to Be Subject to Automated Decisions Based on Profiling" in Tatiani Synodinou, Philippe Jougleux, Christiana Markou, Thalia Prastitou (eds.), *EU Internet Law: Regulation and Enforcement* (Springer, 2017, forthcoming); University of Oslo Faculty of Law Research Paper No. 2017-20 . Their article draws on Bygrave's own early work in Bygrave L "Mining the machine: Article 15 of the EC data protection directive and automated profiling" (2001) 17 Computer Law & Security Rev 17–24 . See also Kobsa A "Tailoring Privacy to Users' Needs" in M Bauer and others (eds) *User Modeling* (Lecture Notes in Computer Science vol 2109, Springer 2001); Hildebrandt supra n 53.

Importantly, art 22 like art 15 before it is a very delimited right. Crucially, the remedy it provides is primarily to prevent processing of a particular kind and secondly, to require that a "human in the loop" be inserted on challenge. The remedy is not, *prima facie*, to any kind of explanation of how processing was carried out or result achieved, that being the province of the information rights of the data subject (see below).[111]

Even after this there are a number of hurdles to get over. First, at 22 applies only when the processing has been *solely* by automated means. ML systems that affect people's lives significantly are usually not fully automated — instead used as decision support[112] — and indeed in a great deal of these cases – for example involving victims of crimes or accidents - full automation seems inappropriate or far off. Art 22 would be excluded from many of the well-known algorithmic "war stories" on this basis : for example the algorithmic decisions on criminal justice risk assessment reported by *Pro Publica* in 2016.[113]. While the racial bias in these systems is clearly objectionable, the important point here is that these systems were always at least nominally advisory.

When does "nominal" human involvement become no involvement? A number of European DP authorities are currently worrying at this point.[114] Human involvement can also be rendered nominal by "automation bias", a psychological phenomenon where humans either over or under-rely on decision support systems.[115] The Dutch Scientific Council for Government Policy in early 2016 specifically recommended that

---

[111] Mendoza and Bygrave argue it is implicit in art 22 and from articles 13-15 that there is a right to be informed that automated decision is being made (supra n 110 at 13).

[112] Cabinet Office *Data Science Ethical Framework* (May 2016) Retrieved from https://www.gov.uk/government/publications/data-science-ethical-framework . Even where decisions can be taken autonomously by systems, the framework, specifically advises human oversight in non-trivial problems.

[113] See Julia Angwin et al: Pro Publica "Machine Bias", 23 May 2016 at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing . Exactly because these systems might be used without enough human (judicial) supervision (or ability to supervise) they have been banned in the courts of some states such as Wisconsin : see https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html?_r=0 .

[114] The UK ICO is consulting on this point at time of writing: see *Feedback Request – profiling and automated decisonmaking (*v 1.0 ,2017/04/06*)* at https://ico.org.uk/media/2013894/ico-feedback-request-profiling-and-automated-decision-making.pdf at 20. "Do you consider that "solely" in Article 22(1) excludes any human involvement whatsoever, or only actions by a human that influence or affect the outcome? What mechanisms do you have for human involvement and at what stage of the process?"

[115] Linda J Skitka and others "Accountability and automation bias" *International Journal of Human-Computer Studies* (2000) 52 701-717; Kate Goddard and others "Automation bias: a systematic review of frequency, effect mediators, and mitigators" *Journal of the American Medical Informatics Association* (2012) 19(1) 121–127,. Literature that indicates we tend to overrely on algorithms includes Dijkstra, J. J. (1999). "User agreement with incorrect expert system advice". Behaviour & Information Technology, 18(6), 399–411. doi:10.1080/014492999118832. Those that think we do not include Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International Journal of Human-Computer Studies, 58(6), 697–718. doi:10.1016/S1071-5819(03)00038-7; Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144 (1), 114. doi:10.1037/xge0000033

more attention be paid to "semi-automated decision-making" in the GDPR, in relation to profiling.[116]

Second, art 22 requires there to have been a "*decision*" which "produces legal effects, concerning him or her, or significantly affects him or her". There is little clue what a "decision" is in art 22 beyond the brief statement of the GDPR that it "may include a measure" (recital 71). This takes us to two sub-issues. First, is a "decision" what a ML system actually produces? ML technologists would argue that the output of an algorithmic system is merely something which is then *used* to make a decision, either by another system, or by a human (such as a judge). When queried, ML models mostly output a classification or an estimation, generally with uncertainty estimates. On their own they are incapable of synthesising the estimation and relevant uncertainties into a decision for action.[117]

Second, even if we posit that algorithmic "output" and human "decision" may be conflated in art 22 for purposive effect, when does an ML "decision" affect a specific individual? What if what the system does is classify subject X as 75% more likely than the mean to be part of group A, and group A is correlated to an unwelcome characteristic B (poor creditworthiness, for example)? Is this a decision "about" X? It is interesting that in relation to a "legal" effect the decision must be "concerning him or her" but not in relation to a "significant" effect. In the paradigmatic domain of credit scoring, there seems no doubt to the ordinary person (or lawyer) that there is a decision (by the credit offering company) and that it affects an individual data subject (the person seeking credit). But in many cases using ML systems, as we see below, this is not so clear.

3.1.1 Article 22 in the context of 'algorithmic war stories'

Consider two well-known and influential early examples of 'algorithms gone bad'. In 2013[118], Latanya Sweeney, a security researcher at Harvard University, investigated the delivery of targeted adverts by Google AdSense using a sample of racially associated names. She found statistically significant discrimination in ad delivery based on searches of 2,184 racially associated personal names across two websites. First names associated predictively with non-white racial origin (such as DeShawn, Darnell and Jermaine) generated a far higher percentage of adverts associated with

---

[116] Wetenschappelijke Raad voor het Regeringsbeleid *WRR-rapport nr. 95: Big data in een vrije en veilige samenleving* [WRR report no. 95: Big data in a free and safe society] (2016), 142. Retrieved from https://www.wrr.nl/onderwerpen/big-data-privacy-en-veiligheid/documenten/rapporten/2016/04/28/big-data-in-een-vrije-en-veilige-samenleving

[117] Heather Douglas *Science, Policy and the Value-Free Ideal* (University of Pittsburgh Press, 2009), We return to the issue of "decisions" and ML below.

[118] Latanya Sweeney "Discrimination in Online Ad Delivery" (2013) *Comms of the ACM*, 56, 5, 44-54

or using the word "arrest" when compared to ads delivered to "white" first names. On one of the two websites examined, a black-identifying name was 25% more likely to get an ad suggestive of an arrest record. Sweeney also ruled out knowledge of any criminal record of the person to whom the ad was delivered. Acknowledging that it was beyond the scope of her research to know what was happening in the "inner workings of Google AdSense"[119], and whether the apparent bias displayed was the fault of society, Google or the advertiser, Sweeney still asserted her research raised questions about society's relationship to racism and the role of online advertising services in this context.

In an even earlier incident of notoriety in 2004, the Google search algorithm(s) placed a site "Jew Watch" at the top of the rankings for many people who searched on the word "Jew". Google (in stark contrast to its more recent attitudes[120]) refused to manually alter their ratings and claimed instead that the preferences of a particular group of searchers had put Jew Watch to the top rather than any normative ranking by Google. "[B]ecause the word "Jew" is often used in an anti-Semitic context, this had caused Google's automated ranking system to rank Jew Watch — apparently an anti-Semitic web site — number one for the query" [121]. In the end Google refused to remove the site from the rankings but collective effort was encouraged among users to push up the rankings of other non-offensive sites, and eventually the site itself disappeared from the Internet.

In each of these cases, did a relevant, "legal", or "significant", decision take place affecting a *person* — or only a group? Here we have one of the rare examples of a system apparently making a "decision" solely by automated processing so the first hurdle is surmounted, but is the second? In the *Google AdSense* example, was a "decision" taken with particular reference to Sweeney? Clearly there was no effect on

---

[119] For example, Sweeney raises the following possibilities: did the advertiser provide multiple templates which themselves "targeted" a list of black sounding names? Did Google's algorithm adjust as it received hits to serve the ads to people with black-associated names more frequently? Did people with black-identifying names click on the "arrest" related ads more often? In each scenario, the combinatorial "approach aligns the financial Interests of Google, as the ad deliverer, with the advertiser".

[120] Google has rethought its approach to such cases, especially after unfavourable press reports, especially a 2016 Observer investigation: see "Google alters search autocomplete to remove 'are Jews evil' suggestion", *Guardian*, 5 December 2016 at https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-suggestion . Interestingly it seems Google's preferred approach is to add "quality rating" to pages to downgrade them in their search algorithms rather than removing links *per se* : this is interesting considering the issues raised later over how it might be possible to alter ML models using the art 17 "right to be forgotten". See "Google launches new effort to flag upsetting or offensive content in search", March 14 2017, *SearchEngine Watch* at http://searchengineland.com/google-flag-upsetting-offensive-content-271119 .

[121] See "Google In Controversy Over Top-Ranking For Anti-Jewish Site", 24 April 2004 , *SearchEngine Watch* at https://searchenginewatch.com/sew/news/2065217/google-in-controversy-over-top-ranking-for-anti-jewish-site .

her legal status (which implies changes to public law status such as being classified as a US citizen, or private law effects such as having capacity to make a will[122]) but did the delivery of the advert significantly affect her as an individual? The most obvious takeaway is that a racial *group* was affected by an assumption of above average criminality, and she was part of that group, which although a familiar formulation in discrimination laws, takes us to somewhere very different from the individual subject-focused rights usually granted by DP and the GDPR.

Even if we accept an impact on Sweeney as an individual constructed through group membership, was it "significant"? She did after all merely have sight of an advert which she was not compelled to click on, and which could even have been hidden using an ad blocker. Mendoza and Bygrave (n 110 at 12) express doubts that targeted advertising will "ordinarily" generate significant consequences (though it might if aimed at a child) and point to the two examples given by recital 71 of automated credit scoring and e-recruitment. Was she significantly affected by pervasive racism as exemplified by the advert delivery? This sounds more important to be sure but surely responsibility should lie with the society that created the racist implications rather than the "decision" taken by Google AdSense, or Google alongside the advertiser? Is it relevant that almost certainly no human at Google could have known Sweeney would be sent this advert, or is that merely another example, as Kohl (n28) discusses, of confusing automation with lack of responsibility? Does it matter that Sweeney could have conceivably asked not to be shown this kind of advert (though perhaps not in 2013) using Google's own tools?

In the *Jew Watch* example, it is even harder to say a "decision" was made affecting any one individual significantly. Given the complexity of the search algorithms involved, dependent not only on variables derived from the searcher but also the general search environment, it is very hard to predict a particular ranking of sites being shown to a particular user in advance. Furthermore quite likely given the evidence quoted above, the searcher might not themselves be of the class affected (Jews).[123]

---

[122] See discussion in Mendoza and Bygrave, (n 110 at 10) who suggest a decision must have a "binding effect". It is hard to see how an advert could have that. On the other hand art 22 clearly applies to "profiling" which as we have seen (p XX) includes in its definition in art 4(4) the evaluation of "personal aspects" of a person including their "personal preferences". This sounds a lot like targeted advertising, though see below on whether that decision would be "significant"..

[123] We might compare this example to cases in some European courts concerning algorithmic defamation, where Google autocomplete appeared to generate a suggestion that a particular name was falsely associated with unsavoury, and hence reputation-harming. In such cases, however, the causal connection between the autocomplete text produced by the algorithm, and the reputational harm suffered by the data subject whose name

3.1.1.1 Re-enter the "right to an explanation"?

Art 22 operates only under certain conditions. It does not apply when the data founding the decision was lawfully processed on the basis that it was necessary for entering a contract, authorised by law or, most crucially, based on explicit consent (art 22(2)).[124] In these cases art 22 is excluded *but*, instead, "suitable measures to safeguard the data subject's rights" *must* be put in place, which *should* include "at least the right to obtain human intervention […] to express [the data subject's] point of view, and to contest the decision."(art 22(3)) [italics added]

Recital 71, explaining further art 22, then mentions all of the above safeguards but also *adds* an explicit "right to an explanation". Is this therefore another route to a "right to an explanation" in art 22? This seems paradoxical. Art 22 gives a primary right, i.e. to stop wholly automated decision making. Would it give what seems an equally powerful right — to an explanation — in circumstances where the primary right is excluded because the data subject has already consented to the processing?

To complicate matters further, under art 22(4), solely automated decisions based on *sensitive* personal data are illegal *unless* based on explicit consent or "substantial public interest". In both cases again, the main text requires the implementation of "suitable measures" to safeguard the data subject's rights, but does not list what these include, referring the reader back again to recital 71 for assistance.[125] So it may be possible to read a "right to an explanation" into these cases as well.

Does it matter that the "right to an explanation" is only mentioned in the recital text not the main article text? Here we encounter a pervasive problem in the GDPR in particular, and European legislation in general, which is the status of recitals. Recitals, while a part of the text, are assumed to be interpretative of the main text

---

was searched on, seems rather more obvious and is probably both more predictable and not dependent as much on the characteristics of the searcher. See discussion in Jones (n 77) at X.

[124] GDPR, art 9(2). Every act of processing personal data in the GDPR requires a lawful ground of processing: see above discussion of consent as such a ground, p XX.

[125] Note also that para 2 of recital 71 details a long list of further suggestions to the data controller to "ensure fair and transparent processing". These involve "appropriate mathematical or statistical procedures for the profiling, [and]technical and organisational measures". These seem only to be required (if they indeed are) in relation to processing of special categories of data (see art 22(4). Interestingly these move in functionality from merely fixing errors in functionality, to ensuring security, to "prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation" (i.e. the special categories of data). This appears to point to the field of discrimination-aware data mining, still nascent in the research community at the time of the drafting of the GDPR, and can be seen as a transition from the traditional function of individual subject access rights (to ensure accurate and secure processing) to a more aspirational function.

rather than creating free standing extra obligations.[126] In the GDPR however, as a matter of political expediency, many issues too controversial for agreement in the main text have been kicked into the long grass of the recitals, throwing up problems of just how binding they are. Wachter and others argue that the history of art 22 in the preliminary drafts indicates a deliberate omission of a "right to an explanation" from the main text of art 22, not an accidental or ambiguous omission[127] which implies the main text omission should rule out the "right to an explanation" in the recital. However the use of the mandatory "should" in recital 71 muddies the waters further.[128]

Our view is that these certainly seem shaky foundations on which to build a harmonised cross EU right to algorithmic explanation.

Thus, returning to the Sweeney *Google AdSense* case study, we find several further issues. Firstly, if we accept for argument's sake that a "decision" was made regarding her which had "significant effects" *, then* was it "based on" (art 22(4)) an art 9 "special" category of data (in this case, race)? If so, it worth noting that art 9(2) of the GDPR probably required that she had given that data to Google by explicit consent. If that was so, she could potentially claim under art 22(4) the "right to an explanation" of how the advertising delivery algorithm had worked.

But was the decision based on race? Was it not more likely instead based on a multiplicity of "ordinary" information that Sweeney provided as signals to the ranking algorithm, plus signals from the rest of the "algorithmic group"[129], which together might statistically proxy race? Perhaps it was based on information the advertiser provided to Google — trigger names or keywords, for example? Ironically it seems like we are stuck in a Catch 22–like situation: to operationalise this 'right to explanation', you need to know what its relevant input variables were, which itself may require access to something resembling an algorithmic explanation.

Finally looking at the primary remedy art 22 provides, a "human in the loop", how valuable actually is it? Certainly, for issues of abusive or upsetting content thrown up by search or advertising algorithms, as in the Sweeney case, pretty useful : this is why Google and Facebook are both currently hiring many workers to manually trawl

---

[126] See Tadas Klimas and Jurate Vaiciukaite "The Law of Recitals in European Community Legislation" (2008) 15 (1) ILSA Journal of International and Comparative Law 63 at 92. They admit the usage of recitals in EU law can be perplexing and is at core politicised. "Recitals in EC law are not considered to have independent legal value, but they can expand an ambiguous provision's scope. They cannot, however, restrict an unambiguous provision's scope, but they can be used to determine the nature of a provision, and this can have a restrictive effect."
[127] Wachter and others, supra, at 9-11.
[128] Interestingly the French text of recital 71 appears to replicate the use of "should" (*devrait*) while the German text is differently constructed so that it does not.
[129] See Mittelstadt and others (n 26).

through their outputs using both real and hypothetical queries (see n 120 below). In such circumstances an intuitive response is likely to be correct and this is something machines do badly. But typically, (see section 4.2) the types of ML algorithms that are highly multidimensional make "decisions" with which humans will struggle as much as, it not more than, machines : simply because of human inability to handle such an array of operational factors. In some kinds of cases – for example, the much discussed "trolley problem" – humans are as likely to make spur of the moment decisions as reasoned ones. For these reasons, Kamarinou and others have suggested that machines may in fact soon be able to overcome certain "key limitations of human decision-makers and provide us with decisions that are demonstrably fair".[130] In such an event they recommend it might be better, not to have the "appeal" from machine to human which art 22 implies, but the reverse.[131]

3.2 GDPR, article 15: a way forward?

A right which might be more usefully employed to get a transparent explanation of a ML system is part *not* of art 22 but rather a provision not specially related to automated decision making, ie, art 15. Article 15 provides that the data subject shall have the right to confirm whether or not personal data relating to him or her are being processed by a controller and if that is the case, access to that personal data and the "following information." This includes in the context of "automated decision making […] referred to in art 22(1) and (4)" access to "*meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing*" (art 15(1)(h)).

As noted above, this version of the 'right to an explanation' is not new, but has existed in the DPD since 1995[132]. While this may seem a more straightforward way to get to such a right than via art 22, it has its own problems.

A first issue is timing. Wachter and others suggest that art 15 "subject access rights" should be contrasted with the "information rights" of the GDPR, arts 13 and 14. Articles 13 and 14 require that information of various kinds should be made available to the data subject when data are collected from either her (art 13), or from another party (art 14). This information is reminiscent of that required to inform consumers before entering, say, distance selling contracts. In contrast, art 15 refers to rights of "access" to data held by a data controller. This seems to imply data has been collected

[130] See Custers and others (n 34); see also the resources and papers presented at the Fairness, Accountability and Transparency in Machine Learning workshop series (www.fatml.org).
[131] Kamarinou et al, supra n xx. This is endorsed by the editorial by Kuner et al "Machine learning with personal data: is data protection law smart enough to meet the challenge?" (2017) 7 International Data Privacy Law 1-2.
[132] DPD, art 12(a).

and processing has begun or taken place. From this Wachter et al argue that the information rights under arts 13 or 14 can only refer to the time before (*ex ante)* the subject's data is input to the model of the system. As such the only information that can be provided then is information about the general "system functionality" of the algorithm, i.e. "the logic, significance, envisaged consequences and general functionality of an automated decision-making system".

In the case of art 15 access rights, however, it seems access comes after processing. Therefore *ex post* tailored knowledge about *specific decisions* made in relation to a particular data subject can be provided, ie, "the logic or rationale, reasons, and individual circumstances of a specific automated decision".

This division seems moderately sensible and seems to promise a right to an explanation *ex post*, despite some textual quibbles.[133] However, whether such an explanation can be "meaningful" in substance is another story as we discuss below in section 4.

Secondly art 15(h) has a carve out, albeit only in recitals, for the protection of trade secrets and IP. "That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software." (recital 63). This probably explains the lack of use of this right throughout the EU, as a similar defence was included in the DPD. Recital 63 of the GDPR does progress things a little given it now states that this should not justify "a refusal to provide *all* information to the data subject" (emphasis added). However as we discuss below (section 5.2), some explanation systems need not in any case necessarily infringe IP rights.

Next, we turn to some of the practical challenges around the right to "meaningful information about the logic involved" art 15 potentially provides.

4 Implementing the right to an explanation

Explanations and the demand for them in machine learning systems are not new. Computer scientists have been long concerned that neural networks "afford an end

[133] Wachter and others (n 11) argue that the art 15(h) *ex post* right still seems dubious given that it includes the right to the "*envisaged* consequences of such processing"[italics added], which, particularly when considered alongside the German version of the text, seems "future oriented". However recital 63, which annotates art 15, refers merely to the "consequences of processing" *not* the "envisaged" consequences. Is this an accidental or inconsequential small textual difference, or is it enough to restrict the apparent scope of art 15(1)(h) to "system logic"? As we have already noted, the text of main article normally takes precedence over that of recitals. However it could be argued that EC laws should be interpreted teleologically and restricting art 15(h) to *ex ante* explanations seems against the purpose indicated by the recital.

user little or no insight into either the process by which they have arrived at a given result"[134], and that "people should be able to scrutinise their user model and to determine what is being personalised and how."[135]

ML explanations are not just good for users but for system designers too. Such systems often do not work perfectly at the time of deployment. Given their probabilistic nature, we must *expect* them to fail in some cases. A system which has predictive accuracy of 90% on unseen data used to test it, would, in a simple case, be expected to fail 10% of the time on new unseen data. In the real world, this is usually worsened by the changing nature of the world and the phenomena ML systems are often expected to accurately model.[136] Explanations can be used to help assess the reliability of systems: for example assessing if the correlations that are being used are spurious, non-generalisable, or simply out-of-date. These systems of user and peer feedback can help to both ensure quality, and a better system[137].

Here, our focus is however mainly on users. Below we discuss (a) what types of explanation are possible (and which might be desirable) and (b) in what situations and for what users, an explanation of an ML system may be difficult, limited or impossible. Finally in section 5 we suggest some positive avenues for explanation facilities including (a) explanations aimed at helping users to form better mental maps of how algorithms work, and thus to develop better trusted relationships with them; and (b) pedagogical rather than decompositional explanations as a way to avoid the "IP" restraint on ML algorithms.

4.1 Types of explanation: Model-centric v subject-centric explanations

4.1.1 Model-centric explanations (MCEs)

Model-centric explanations (MCEs) provide broad information about a ML model which is not decision or input-data specific. Computer scientists would refer to this explanation as 'global', as it seeks to encapsulate the whole model — although we deliberately avoid this terminology here, as it is likely to cause more confusion across

---

[134] AB Tickle and others, "The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks" (1998) 9 *IEEE Transactions on Neural Networks* 6 1057–1068 (doi:10.1109/72.728352) at 1057. See also Zelznikow and Stranieri's work in 1995 (n 19).
[135] Kay (n 8) at 18.
[136] J Gama and others "A survey on concept drift adaptation" (2013) *ACM Computing Surveys*, 1(1). (doi:10.1145/2523813).
[137] In the ML field of recommender systems, this reason for explanation has been discussed under the term 'scrutiny', and is considered a hallmark of good user design. See N Tintarev and J Masthoff "Explaining Recommendations: Design and Evaluation" in *Recommender Systems Handbook* (Springer, 2015).

disciplines than clarity. This provides strength in that one set of information can be provided to everyone, but there are limitations on how detailed and practical such an explanation can be.

This information could include:

- *setup information*: the intentions behind the modelling process, the family of model (neural network, random forest, ensemble combination), the parameters used to further specify it before training;
- *training metadata*: summary statistics and qualitative descriptions of the input data used to train the model, and the output data or classifications being predicted in this model;
- *performance metrics:* information on the model's predictive skill on unseen data, including breakdowns such as success on specific salient subcategories of data;
- *estimated global logics*: these are simplified, averaged, human-understandable forms of how inputs are turned into outputs, which by definition are not complete, else you could use them instead of the complex model to achieve the same results. These might include variable importance scores, rule extraction results, or sensitivity analysis;
- *process information*: how the model was tested, trained, or screened for undesirable properties.

Some work around algorithmic decision-making concerned with the consistency, or procedural regularity of the decisions being undertaken falls into this category[138]. Information about the logics, which might be provided in the form of cryptographic assurances[139], might help ensure consistency against an adversary intent on switching algorithmic systems behind-the-scenes, or making arbitrary decisions under the guise of a regular automated system. However for "meaningful information" for individual data subjects we are probably going to look towards subject-centric explanations.

4.1.2 Subject-centric explanations (SCAEs)

Subject-centric explanations (SCEs) are built on and around the basis of an input record. They can only be provided in reference to a given query — which could be

---

[138] Kroll and others (n 100).

[139] Cryptographic assurances often take the form of 'zero-knowledge proofs', where one party can provide information that mathematically verifies that a given statement is true, without conveying any further information (such as the structure of an algorithmic model). Some, limited prior work has demonstrated the feasibility of verifying certain types of ML systems with cryptographic methods, although this has largely been with a view to creating systems where analytics can be decentralised in a trusted manner, rather than centralised and verified. See Danezis G., Kohlweiss M., Livshits B., Rial A "Private Client-Side Profiling with Random Forests and Hidden Markov Models" in Fischer-Hübner S. and Wright M. (eds) *Privacy Enhancing Technologies*. PETS 2012. Lecture Notes in Computer Science, vol 7384. (Springer, 2012).

real or could be fictitious or exploratory. Computer scientists would refer to this type of explanation as 'local', as the explanation is restricted to the region surrounding a set of data. Complex models cannot be explained effectively in their entirety — which is why they have obtained a reputation as 'black boxes'. They might, however, be able to me more usefully explained by only considering certain relevant parts of them at any one time.

One reason for this is known as the 'curse of dimensionality' in computer science. Data can be thought of geometrically: with two numeric variables, you can display all data on a two-dimensional scatter plot. With three variables, a three-dimensional one. Conceptually, you can scale this up to however many variable you have in your data. However, the volume does not scale in the same way as the increase in variables. The volume increases exponentially, as the variables increase linearly. As a result, the 'space' the data can occupy grows massively very quickly, and explaining the patterns inside it all in one go, as MCEs try to, quickly becomes unwieldy.

Despite this, explanations are possible if we zoom in to the part of the space in and around a vector of variables that interest us. By doing this, the system can become considerably more interpretable. This is an active field of research which we believe needs more consideration from a legal perspective. Here, we distinguish between four main types of SCEs:

- *Sensitivity-based* subject-centric explanations: what changes in my input data would have made my decision turn out otherwise?[140]
- *Case-based* subject-centric explanations: which data records used to train this model are most similar to mine?[141]
- *Demographic-based* subject-centric explanations: what are the characteristics of individuals who received similar treatment to me?[142]
- *Performance-based* subject-centric explanations: how confident are you of my outcome? Are individuals similar to me classified erroneously more or less often than average?

Unlike MCEs, SCEs are less suited for discussing aspects such as procedural regularity. Instead, they are more about building a relationship between these tools and their users or decision subjects that can provide "meaningful" explanation. In this sense,

---

[140] W Samek and others "Evaluating the visualization of what a deep neural network has learned" (2016) *IEEE Transactions on Neural Networks and Learning Systems*; MT Ribeiro and others " 'Why should I trust you?': Explaining the predictions of any classifier" (2016), eprint: arXiv:1602.04938.
[141] D Doyle and others "A review of explanation and explanation in case-based reasoning". (2003) *Department of Computer Science, Trinity College, Dublin*.
[142] L Adrissono and others "Intrigue: Personalized recommendation of tourist attractions for desktop and handheld devices". (2003) *Applied Artificial Intelligence, 17*, 687–714; Tintarev and Masthoff (n 137).

SCEs are considerably more linked to communities of interface design than communities concerned with engineering issues, such as the cryptographic assurances discussed above.

4.2 Domain: some tasks are easier to 'explain' than others

Meaningful explanations of ML do not work well for every task. As we began to discuss above, the tasks they work well on ideally have only a few input variables: the "curse of dimensionality". Systems with more variables will typically perform better than simpler systems so we may end up with a trade off between performance and explicability.

One way to deal with this is if different input variables can be combined in a clear and visual way. Images are a good example of the latter: for a ML system, and especially since the rise in popularity of deep learning, pixels are treated as individual inputs. While we struggle to read a table full of numbers at a glance, the brain can process thousands of pixels at once, meaningfully and in relation to one another. Similarly, words hold a lot of information, and a visual displaying 'which words in a cover letter would have got me the job, were they different' is also meaningful. Design might help us with some of these challenges. As an example, smartphones produce a great number of data points about movements. In a ML system we might try to predict whether individuals are sitting, standing etc. from the accelerometer and gyroscope in their smartphone: such a system might have 561 variables, after processing for time and frequency.[143] This is not on the face of it a system whose inferences are easy to explain to humans. Yet were we to connect this dataset to a visualisation about the phone's position in space, we might be able to collapse these 561 variables into something visually compelling.

Even visualisation cannot deal with the basic problem that in some systems there is no theory correlating input variables to things humans understand as causal or even as "things". In ML systems, unlike simulation models, the features that are being fed in might lack any convenient or clear human interpretation in the first place, even if we are creative about it. LinkedIn, for example, claim to have over 100,000 variables held on every user that feed into ML modelling.[144] Many of these will not be clear variables like "age", but more abstract ways you interact with the webpage, such as

---

[143] See Jorge-L Reyes-Ortiz and others. "Transition-Aware Human Activity Recognition Using Smartphones" (2015) 171 *Neurocomputing* 1 (doi:10.1016/j.neucssom.2015.07.085).
[144] K Liu, "Developing Web-scale ML at LinkedIn—from Soup to Nuts." (2014) *Presented at the NIPS Software Engineering for ML.*

how long you take to click, the time you spend reading, or even text you write but later delete without posting.[145] These variables may well hold predictive signals about individual characteristics or behaviours, but we lack compelling ways to clearly display these explanations for meaningful human interpretation. In these cases, we must ask — what could a satisfactory explanation even look like for decisions based on this data?

4.3 Users: explanations might fail those seeking them most

It is worth considering the typical data subject that might seek an explanation of a ML-assisted decision. We might expect them to have received outputs they felt were anomalous. They might feel misclassified or poorly represented by classification systems — hardly uncommon, as literatures on the problematic and value-laden nature of statistical classification note.[146] While some might wholesale reject the schema of classifications used, others might want to know if such a decision was made soundly. For these decision subjects, an explanation might help.

Comment [A1]: Ive tidied this a little as it was a little rambly. I still feel a bit confused. Is the problem that outlies are cleaned from the training set so aren't easily classified, or that ML does better classifying things which have many occurences rather than few?

However, it also seems reasonable to assume that individuals with outputs they felt were anomalous are more likely than average to have provided inputs that can genuinely be considered statistically anomalous compared to the data an algorithmic system was trained on. To a ML system, they are "weirdos".

Researchers have long recognised that some queries of ML systems are more difficult to predict than others, given their relative individual complexity.[147] Given the many variables being used for each record, spotting these individuals cannot be done with methods such as visualisation, which we often use to detect outliers. Most of the phenomena we are interested in modelling, such as burglary, child abuse, terrorism or loan defaults, are rare, at least in comparison to their non-occurrence, and this also makes prediction harder.[148]

ML practitioners expect this kind of dynamic within the data they use. The common technique of *boosting* relies on this type of distribution of patterns. Boosting involves training a ML algorithm, then looking at which cases it gets wrong. These cases are

---

[145] S Das and AD Kramer, "Self-Censorship on Facebook" (2013) *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* 120–127.
[146] J Scott, *Seeing like a State* (Yale University Press, 1998); Bowker and Starr (n 49); SE Merry, *The seductions of quantification* (University of Chicago Press, 2016).
[147] Gary M Weiss, "Mining with rarity: a unifying framework" (2004) 6 *ACM SIGKDD Explorations Newsletter* 7–19; Gary M Weiss "Mining with Rare Cases" (2009) In Oded Maimon and Lior Rokach (eds.) *Data Mining and Knowledge Discovery Handbook* 747–757.
[148] T Jo and N Japkowicz, "Class imbalances versus small disjuncts." (2004) 6 *ACM SIGKDD Explorations Newsletter* 1 40-49.

assumed to hold different patterns that have not been adequately picked up by the current system, so are then weighted to seem more important for a subsequent round of training to be combined with the first later. In this round, the higher weighting on those previously misclassified emphasises to the system being trained that these contain patters that are important, but perhaps less straightforward, to learn.[149]

Why might this challenge meaningful explanations? SCEs practically focus on taking the model you have, selecting a certain part of it, and modelling it in a simpler and more interpretable way. This simplification necessarily discards the complex outlier cases, just as you might do when you simplify a scatterplot into a smooth trend-line or a 'line of best fit'. Optimising an explanation system for human interpretability necessarily means diluting predictive performance to capture only the main logics of a system: if a more interpretable system with exactly the same predictive performance existed, why use the more opaque one? Traditionally, this has been described as the "fidelity" of an explanation facility for a machine learning system: how well does it mimic the behaviour of the system it is trying to explain?[150] The more pressing, related question is, are the cases that an interpretable model can no longer predict simply distributed at random, or are they correlated with those we might believe to have a higher propensity to request a right to explanation? We lack empirical research in this area. If the users of complex ML systems who seek explanations are likely to be these "rare birds", then it is worrying that they are the most likely to be failed.

5 Setting a course for better explanations

Better explanations are possible, although it may involve rethinking *how* we use explanations. As emphasised above, this is far from a new field of research. In this section, we highlight two promising avenues; the first being long understood in the design field, that explanations that allow users to interactively explore algorithmic systems can strongly enable individuals to develop good and trustworthy (although not perfect or complete) mental models of the systems they use and are subject to. The second rests on another insight, harking back to the days of expert systems — you do not have to have access to the innards of a model to attempt to explain it. Many of the most promising methods to explain algorithmic systems do not try to "decompose" or open the model, but try to "pedagogically" fit a simpler model around

[149] Robert E Schapire, "The Boosting Approach to ML: An Overview" (2003) *MSRI (Mathematical Sciences Research Institute) Workshop on Nonlinear Estimation and Classification*.
[150] Tickle and others (n 135) at 1058.

it, querying the black box like an oracle to estimate its core logics in an intelligible form.

5.1 Exploring with explanations

Above we introduced the idea of model-centric (MCEs) vs subject-centric (SCEs) explanations. Which are best for helping users understand complex ML systems? Unsurprisingly, the best explanations of complex systems are what are known as "exploratory" and which use subject-centric inputs. Experimental tests have found that interfaces that provided SCEs repeatedly were far more effective at helping users complete tasks, even where the same amount of information was provided in total using both focuses.[151]

Drawing on the literature on human–computer interaction (HCI), SCEs can be thought of as "seams" in the design of a ML system.[152] Seamless design hides algorithmic structures, providing certain kinds of effortlessness and invisibility. This promotes an acceptance of technology based on its effect : the idea that when a machine runs efficiently and appears to settle matters of fact, attention is often drawn away from its internal complexity to focus only on the inputs and outputs.[153] Yet "seamful" algorithms, where individuals have points in the designed systems to question, explore and get to know them, help build important, albeit partial, mental models that allow individuals to better adapt their behaviour and negotiate with their environments.[154] By introducing these "seams" of explanation, it has been demonstrated that even new users can quickly build mental models of ML systems to the level of those with seasoned experience.[155]

The GDPR, as discussed in detail below at 6.2.3, mandates Privacy by Design (PbD). This is unlikely to be implemented anytime soon as a detailed technology mandate for designers backed by sanctions like fines, but it might encourage us to think of how we might build ML systems so as to best allow users to understand them and how they make decisions. Exploratory systems allow individuals to explore the logics of algorithms for themselves, not "ex ante" or "ex post" decisions but as they use and interact with the systems. As discussed above, one subject-centric approach is to allow

Comment [A2]:

Comment [A3]: for me this is still the hardest section. I think i get the main point but could gaming (and whether it really matters that much – paras highlighted) be generalised to a general point about ALL explanations , perhaps in section 4 at the top? (where I have now put expln to help system designers)
I also stil don't understand how a MCE CAN provided the same amount of info as SCEs (your n 152)

[151] Dianne C Berry and Donald E Broadbent "Explanation and verbalization in a computer-assisted search task" (1987) 39 *The Quarterly Journal of Experimental Psychology Section A* 4, 585-609 (doi: 10.1080/14640748708401804)
[152] Matthew Chalmers and Ian McColl, "Seamful and Seamless Design in Ubiquitous Computing" (2003) *Workshop At the Crossroads: The Interaction of HCI and Systems Issues in UbiComp.*
[153] See Bruno Latour, *Pandora's hope: Essays on the reality of science studies*. (Harvard University Press 1999)
[154] Kevin Hamilton and others, "A path to understanding the effects of algorithm awareness" (2014) *CHI '14*. (doi: 10.1145/2559206.2578883)
[155] Motahhare Eslami and others. "First I "like" it, then I hide it: Folk Theories of Social Feeds" (2016) *CHI' 16.* (doi: 10.1145/2858036.2858494)

a user to query a model with 'what it would have done' with a certain set of data points, and what would have made it different; or which types of 'nearby' individuals or data-points would have received similar or different treatment.

Some SCEs might just let individuals see the logics happening around their own data points, but this would risk keeping individuals inside their own 'filter bubbles', particularly in complex systems. Unfortunately, it will be easier to build SCEs that let you explore the logics around yourself rather than around others. For example, tools already exist to let you "try out" what your credit score might be online, through filling in a questionnaire, for example, or signing into these using your data profile (for example, by authorising a 'soft' check on your credit file, or potentially one day, by giving access to your social media API).

Comment [A4]: Is this right> What was there a bit garbled I think.

Exploring systems by simulating the inputs of others is harder. In relatively simple experiments, researchers have attempted to 'reverse engineer' algorithmic systems online in order to study phenomena such as price discrimination, by simulating the profiles of diverse individuals while browsing.[156] However, presenting valid hypothetical subjects other than yourself to many of these systems is becoming increasingly difficult in an era of personalisation. British intelligence services have noted the challenge in providing data such as "a long, false trail of location services on a mobile phone that adds up with an individual's fake back-story", with the former director of operations for MI6 noting that "the days in which intelligence officers could plausibly adopt different identities and personas are pretty much coming to an end."[157] . Individuals everywhere, not just MI6, will find it harder to "fake" a new persona without changing their lifestyle, haunts, friends etc , in these days of the "digital exhaust".

A problem frequently raised with this kind of repeated querying of ML systems to establish a "mental model" is that it might be used by users to "game the system". In fact this is unlikely. In private sector systems such as targeted advertising, as we have already seen above, uses do often try to "game" or self-optimise systems with false data such as birthdates or locations. Yet in public sector cases, such as ML sentencing and parole systems, it seems unlikely that gaming will be a large problem. As the criminological literature has noted, any evidence that the severity of sentencing deters crime, as opposed to the probability of apprehension, is patchy at best.[158] If

---

[156] A Hannak and others, "Measuring Price Discrimination and Steering on E-Commerce Web Sites" (2014) *Proceedings of the 2014 Conference on Internet Measurement Conference*. (doi: 10.1145/2663716.2663744)
[157] Sam Jones "The spy who liked me: Britain's changing secret service" (29 September 2016) *Financial Times*. Retrieved from https://www.ft.com/content/b239dc22-855c-11e6-a29c-6e7d9515ad15
[158] See amongst the broad literature, for example, Daniel S Nagin. "Deterrence in the twenty-first century" (2013) 42 *Crime and Justice* 1 (doi:10.1086/670398).

sentencing itself does not deter crime, then it seems unlikely that prisoners will change their characteristics to attempt to game a recidivism algorithm that will not even be used until after they have been apprehended. Perhaps within prison, individuals might seek to 'game' an algorithm used during parole, by behaving well, or taking specified courses, for example. Yet for this to be gaming, we would need to assume that the act of taking these courses, or behaving well, would not be useful or transformative in and of itself.

For important decisions, the danger of gaming can say more about the need to deal with the fragility of the prediction system and the shallowness of the policy solution, rather that the need to keep the mechanisms at play under wraps. Where systems work, but can be gamed, they rely on information asymmetries to keep them predictively useful. This should already make us wary: if all that is preventing misuse is 'keeping the lid' on the logic, then it open up large potential for misuse from those individuals that have managed to, likely with money and power, pry the lid open more than others.

Explanation facilities might help here to allow decision subjects to build more effective and relevant mental models, and work better with algorithmic systems[159]. Some evidence has shown that the availability of explanations of this sort can build trust both in users and in designers.[160]

5.2 Explaining black boxes without opening them

Since early research into "expert systems" in the late 80s onwards, there has been awareness that a mere *trace* of the "logic" of how an automated system transformed an input into an output was not "meaningful" to a human, let alone a non-expert. As we have seen, the way that ML systems optimise for performance usually comes at the expense of internal interpretability. Researchers since have generally seen

[159] Perel and Elkin-Cohen describe this as "black box tinkering" and are positive about it for empowering users in the field of algorithmic copyright enforcement: see Perel (Filmar), Maayan and Elkin-Koren, Niva, BLACK BOX TINKERING: Beyond Transparency in Algorithmic Enforcement (March 3, 2016). Florida Law Review, Forthcoming. Available at SSRN: https://ssrn.com/abstract=2741513 or http://dx.doi.org/10.2139/ssrn.2741513

[160] Evidence in this field is mixed, with some suggestions that explanation facilities help acceptance of decisions without building systemic trust (H Cramer and others "The effects of transparency on trust in and acceptance of a content-based art recommender" (2008) *User Modeling and User-Adapted Interaction*, 18(5), 455-496); while others find links to trust building overall (A Busone and others "The role of explanations on trust and reliance in clinical decision support systems" (2015) *ICHI '15,* 160-169.)

explanation as an entirely separate optimisation challenge — *decoupling* algorithmic reasoning from algorithmic explanation.[161]

There are two main styles of decoupled algorithmic explanations[162]. The first type is the *decompositional* explanation. Decompositional approaches attempt to open the black box, and understand how the structures within, such as the weights, neurons, decision trees and architecture, can be used to shed light on the patterns that they encode. Decompositional approaches require access to the source code of the model, or at least certain aspects of its weights, in order to build explanation systems. Some ML systems are decompositionally more explainable than others. Many regression methods — which are, in fact, ML — are so explainable that they are commonly used within social sciences to build models of the world to better *understand* phenomena from obesity to voting, rather than to predict them. Other systems are more difficult to pull apart, although there is considerable research progress being made even in decompositional approaches to complex deep learning systems.[163]

There are also decompositional-style methods to obtain further information on explanations *during* training of a model, rather than afterwards. These are more restricted, as they also require the original data, rather than just the trained model. Random forests, which are ensembles of individually interpretable decision-trees trained on a dataset, but that 'vote' on the correct classification (a computational approach to the "wisdom of the crowd"), have a variety of commonly used "variable importance" measures. These measures take each of the variables that feed into a model in turn, add a random amount of noise to them, and see what effect it has on the accuracy of the model. For example, a credit scoring model might take age, income and education as input data. We would first train a model correctly, then add some random noise to 'age', train the model again, and see if it performs significantly worse. If it does, we might say that age is an important variable.[164] These approaches are often used operationally to publicly explain ML systems, such as in a Durham Police recidivism risk system, currently being rolled out to support custody officers as

---

[161] MR Wick and WB Thompson, "Reconstructive expert system explanation" (1992) 54 *Artificial Intelligence* 1–2 33–70 (doi:10.1016/0004-3702(92)90087-E). This corresponds to the "naïve" approach Kroll et al talk about of merely dumping source code, inputs and outputs (see n 106).

[162] Combinations between these two styles are also possible. See Tickle and others (n 134).

[163] For example, see G Montavon and others, "Explaining nonlinear classification decisions with deep Taylor decomposition" (2017) 65 *Pattern Recognition* 211-222.

[164] See Carolin Strobl and others, "Bias in random forest variable importance measures: Illustrations, sources and a solution." (2007) 8 *BMC Bioinformatic*s 1 25.

to whether they should recommend bail upon arrest.[165] Explanations are provided by the weights of different variables used in the random-forest—driven recidivism scores.

Comment [A5]: We've lost fn 170?

On the other end of the spectrum, *pedagogical* systems — more recently also referred to as *model agnostic* systems — do not even need to open the black box. They can get the information they need by simply querying it, like an oracle.[166] Pedagogical systems have the great advantage that since they demand a much lower level of model access and are thus less likely to run into the IP or trade secrecy barriers embedded in art 15(h) (see section 3.2 above). Indeed, for firms that provide remote access to querying their models — for example, through an API — it might be technically possible to build pedagogical explanations even if the firm does not directly condone it. Furthermore pedagogical systems cannot easily be reverse engineered to construct a model of equal performance, as some might fear. In particular, the subject-specific nature of the vast majority of pedagogical explanation systems means that even if an algorithm could be siphoned and rebuilt elsewhere, that reconstruction would be limited to individuals similar to those to which the explanations related. More critically, if a more explainable system was similarly accurate, why use a pedagogical system in the first place? Statistical controls also exist that might be fruitfully repurposed to prevent 'over-explaining' to any one person or organisation, notably in the area of "differential privacy" guarantees.[167]

6 Looking for better remedies than explanations in the GDPR

6.1 Avoiding a "transparency fallacy"

We suggest that it may be better to think about alternate methods for providing accountability, redress or reassurance to data subjects who are affected by algorithmic decision-making rather than concentrating on the search to extract a "right to an explanation" either from art 22 or, better, art 15 of the GDPR. A useful warning can be taken from the history of consent in information privacy.

---

[165] Sheena Urwin, "Algorithms in Durham Constabulary custody suites - How accurate is accurate?" (2017) *Presentation at TRILCon '17*, University of Winchester.

[166] For an example of a pedagogical system, see MT Ribeiro and others, "'Why should I trust you?': Explaining the predictions of any classifier" (2016) *arXiv:1602.04938*.

[167] Cynthia Dwork, "Differential privacy: A survey of results." (2008) *International Conference on Theory and Applications of Models of Computation*, Springer (doi: 10.1007/11787006_1).

Privacy scholars are already over-familiar with the notion that consent, often regarded by lay audiences as the primary safeguard for control of personal data, has in the online world become a mere husk of its former self, often described as "meaningless" or "illusory."[168] Online consent is most often obtained by displaying a link to a privacy policy at the time of entry to or registration with a site, app or network, and asking the user to accede to these terms and conditions by ticking a box. As there is no chance to negotiate and little evidence that the majority of users either read or understand these conditions, it is hard to see how this consent is either "freely given, specific, informed and unambiguous" despite these being conditions for valid consent under the GDPR.[169] Consent as an institution in fact only encourages data subjects to give up their data when perhaps they should not, since many users have a faulty understanding of the privacy risks involved, due to asymmetric access to information and hard-wired human failure to properly assess future, intangible and contingent risks. Even in the real rather than online world, it is manipulated by those such as employers or insurers who can exert pressures that render "free" consent imaginary. Even if consent is given in a free and informed way, constant vigilance is needed as privacy policies and practices change frequently. It is unreasonable and increasingly unsustainable to abide by the liberal paradigm and expect ordinary users to manage their own privacy in the world of online dependence and "bastard data."[170] As a result, it is now beyond trite to talk about a "notice and choice fallacy."[171].

Relying on individual rights to explanation as the means for users to take control of ML systems risks creating a similar "transparency fallacy" (adapting Heald's notion of a "transparency illusion"[172]). Individual data subjects are not empowered to make use of the kind of algorithmic explanations they are likely to be offered even if (unlikely as it seems) the problems identified in section 4 are overcome. Individuals are mostly too time-poor, resource-poor, and lacking in the necessary expertise to meaningfully make use of these individual rights. In some ways the transparency fallacy is *even*

---

[168] See discussion and references supra section 2.2.2 and n 56.
[169] GDPR, art 4(11). The GDPR does attempt to improve the quality of consent with some new measures such as the requirement that the data controller must be able to prove consent was given (art 7(1)), that terms relating to consent in user contracts must be distinguishable from other matters, and written in "clear and plain language" (art 7(2)); and that in determining if consent was given "freely" ,account should be taken of whether the provision of the service was conditional on the provision of data not necessary to provide that service(art 7(4)). It is submitted however that these changes are not major, and that much will depend on the willingness of EU member state DP regulators to take complex, expensive and possibly unenforceable actions against major data organisations (Google, Facebook, Amazon and others) emanating from non-EU origins with non EU law norms. The Common Statement of 5 DPAs (n72) is certainly an interesting first shot over the bows.
[170] McNamee (n 57).
[171] See full discussion in Edwards, supra n 56.
[172] Heald (n 96).

worse than its consent cousin, since the explanation itself may not be meaningful enough to confer much autonomy even on the most empowered data subject.

Next, we consider if in the stampede to find a right to an explanation, other new user rights and tools in the GDPR have been given undeservedly little attention.

6.2 Better machine learning with the tools of the GDPR

Might some GDPR rights that relate to personal data help us negotiate with and govern algorithmic systems that we help to train? In this subsection, we explore this direction, with reference to two main rights: the right to erasure (colloquially often called "right to be forgotten") in art 17, and the right to data portability in art 20, as well as a proposed supporting environment for enforcement the GDPR establishes using a varied range of instruments, such as Data Protection Impact Assessments and privacy seals.

6.2.1 GDPR, article 17: the right to erasure ("right to be forgotten")

Comment [A6]: Going forward need to restructure this section: the reasons I think are about PD and about inferences but the structure isn't tight around this

Article 17 of the GDPR states that the "data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay".[173] In the context of ML, we believe a data subject might usefully seek erasure as a remedy for three main reasons.

Firstly, a data subject might seek erasure of her personal data simply because she does not wish the data controller to have a copy of it any longer. This is not an unrestricted right. Erasure can be obtained on one of various grounds[174], including that the data are no longer necessary in relation to the purposes for which they were collected; that the data subject has withdrawn her consent to processing; that the personal data have been unlawfully processed; that the data must be erased under local state or EU law (e.g. because of rehabilitation of offenders or bankruptcy rules); or that the data was provided while a child under 16. Most usefully, erasure can be sought if the data was being used to profile the data subject and had been collected

[173] The right to erasure ("right to be forgotten") in arts 17 and 18 (restriction of processing) emerged after the landmark CJEU case of *Google Spain v González* , Case C‑131/12, 13 May 2014 and is both wider in effect and more specified than the rule elaborated in that case out of the DPD.
[174] Art 17(1).

lawfully but without her consent.[175] The right can conceivably be repelled by the controller on "compelling legitimate grounds".[176]

An important issue here is what personal data in the ML system an individual data subject has rights over. Clearly she has the right to erase her explicit data used as inputs to an ML system (e.g. name, age, medical history) but does she have the right to erase *metadata* about her behaviour and movements both in real and virtual world? We have already seen that ML systems such as those run by Facebook or LinkedIn make heavy use of this e.g. type of links clicked on on-site, photos viewed, pages "Liked"; or, in the real world, location and movement as perhaps tracked by a GPS in smartphone or wearable is a commonly used variable. While metadata should qualify as personal data if it clearly allows a data subject to be identifiable (e.g. by "singling out") it does not appear the history of art 17 ever contemplated its use for such purposes. Finally, what about the *inferences* that are made by the system when the data subject's inputs are used as query? These seem what a use would perhaps most like to delete – especially in a world of "bastard data" where one system's output becomes another's input. Somewhat surprisingly, the A29 WP, in the context of the right to portability (see below) have already issued guidance that the inference of a system is *not* the data of the subject but "belongs" to the system that generated it (n 197 below). It is not yet clear if this approach would be advised re the right to erase though it logically would as the two rights (17 and 20) are seen as complementary. In that case we seem to have a clear conflict with the already acknowledged right of a data subject to erase an inference from Google's search algorithm (i.e. the "right to be forgotten as vindicated in *Google Spain*).

6.2.1.1 Machine "unlearning"

Secondly, a data subject might seek erasure of her data from the model of a trained ML system because she was unhappy with the inferences about her that the model produced. In other words she wants to alter the model. This is unlikely to be helpful because it is unlikely that one data subject withdrawing their personal data would make much difference to a trained model — ML systems often require multiple examples of a phenomenon of interest to recognise the pattern. They are calibrated ('regularised') this way to avoid modelling the "noise" or random elements in the data ('overfitting'), rather than just capturing the main "signal" hoped to be fruitful in

---

[175] i.e. on the ground of the legitimate interests of the data controller under art 6(1)(f) or, for a public data controller, the public interest under art 6(1).

[176] There is no guidance in recital 69 on what this might mean. Note that art 17 rights can also be excluded by EU states where exercising them affects important public interests (art 17(3)): these include freedom of expression, 'public interest' in the area of health, public archives and scientific, historical, and statistical research, and legal claims.

analysing future cases after the model is built. To make effective use of this right to alter models, whole groups would need to collaborate explicitly or implicitly to request erasure. We might imagine a data subject whose data generated by a wearable fitness tracker phenomena have been correlated with a rare medical condition. She might persuade the rest of her "algorithmic group" to withdraw their personal data from the system so that the model could no longer make this correlation. This seems extremely difficult to organise in practice, as well as probably also involving unwanted privacy disclosures.

Thirdly, a data subject might seek erasure of an entire model (or aspects of it) on the grounds that *it* is her personal data. This might be based on the assertion that the model itself is the personal data of each and every data subject whose input data helped train and refine it. On the face of it this seems implausible. To a lawyer, a ML model resembles more a structure of commercial use which will probably be protected by trade secrets or possibly, by an IP right such as a patent or, in Europe a database right[177], which is a right essentially over the arrangement of data in a certain system, personal or otherwise, rather than the data itself. For ML specialists, an argument might be made that personal data used to create a trained model might be fully or partially reconstructed by querying the model.[178] Attempts have already been made by researchers to extract personal data in this way as a form of "adversarial" ML. An attacker might attempt to query, observe or externally influence a ML system to obtain private information about some or all individuals within its training set.[179] In this type of attack, individual records can be recovered from a model with high

[177] See Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases. Case law on the EU database right both in the CJEU and national courts has been generally restrictive and it is by no means sure it would operate here, at least in the UK.

[178] See relevant literature on small disjuncts in ML. Weiss (n 147); Andrea Pohoreckyj Danyluk & Foster Provost, "Small disjuncts in action: learning to diagnose errors in the local loop of the telephone network" (2014) *Proc. of Tenth International Conference on ML* 81–88.

[179] See Ling Huang and others, *Adversarial ML* (2011) (doi:10.1145/2046684.2046692). The main approach to counter this characteristic has been differential privacy, a statistical technique that seeks to ensure that the properties of a model do not significantly change if a data point is added or omitted. Differential privacy can be applied to ML provide some assurance over the leakage of individually private statistics. Applying it is not easy, costless or straightforward, often requiring significant modifications to modelling processes and choice of algorithms. Differential privacy approaches in ML are currently not widely taught to practitioners, not broadly implemented in common software, and usually come with costs to model performance. See further D Ananad and others, "Signal Processing and ML with Differential Privacy: Algorithms and challenges for continuous data" (2013), *IEEE Signal Process. Mag.* 86–94; Ji Zhanglong and others, "Differential Privacy and ML: a Survey and Review"(2014) *arXiv [cs.LG]*, http://arxiv.org/abs/1412.7584.

probability. Indeed, some applications of ML specifically utilise this characteristic to try and improve or better understand data compression techniques.[180]

Assuming that some grounds for erasure *were* established, for a data controller, requests for erasure of personal data from an ML model would not always be straightforward as it might involve retraining the model and, especially, revising the features of that model.[181] This would be problematic as the high computational and labour costs of ML systems restrict many organisations' practical capacities for constant retraining of the model when either new data, or indeed, requests for erasure come in. In these situations, swift and easy erasure is likely difficult to achieve. Computationally faster approaches to 'machine unlearning' have been proposed, but still require retraining and would require foundational changes to model architectures and processes to use.[182]

6.2.1.2 Model trading and the right to erasure

A rising business model involves the trading or publishing[183] of trained models without the data which was used to train them. For example, Google's ML models *syntaxnet* for parsing sentences (into the relations between verbs, propositions and nouns, for example) is based on proprietary treebank data[184], while the word embedding model *word2vec* (to map which words have similar meanings to each other, in which ways) uses closed access text from Google News[185] is also available. Can a data subject withdraw their personal data in some useful way from a model which has been traded? This presents interesting and extremely difficult *legal* challenges to the right to erasure.

Comment [A7]: Do people pay for these or what? How common is this?

Art 17(2) of the GDPR is an obvious starting point. It provides that where a controller has made personal data "public" but is asked to erase, then they are to take "reasonable steps, including technical measures" to inform other controllers processing the same personal data that the data subject has requested the erasure by them of "any links to, or copy of, or replication of, those personal data".

---

[180] D Sculley and CE Brodley, "Compression and ML: a new perspective on feature space vectors" (2006) *Data Compression Conference (DCC'06)* 332–341; George Toderici and others, "Full Resolution Image Compression with Recurrent Neural Networks" (2016) *arXiv [cs.CV]* Retrieved from http://arxiv.org/abs/1608.05148.

[181] Pedro Domingos, "A few useful things to know about ML" (2012) 55 *Commun. ACM* 78–87. Retraining might only involve a single piece of data, such as transforming a postcode into geospatial coordinates. In this kind of case, an erasure request is simple. However if a variable is constructed by reference to other inputs – e.g. the distance of an input from the mean, which involves all data points — then complete erasure might require recalculation of the whole dataset.

[182] Yinzhi Cao and Junfeng Yang, "Towards Making Systems Forget with Machine Unlearning" (2015) *IEEE Symposium on Security and Privacy* 463–480.

[183] Alternately, access to them may be provided through APIs.

[184] See https://github.com/tensorflow/models/tree/master/syntaxnet [Retrieved 15 February 2017].

[185] See https://code.google.com/archive/p/word2vec/ [Retrieved 15 February 2017].

This is a difficult provision to map to ML model trading. It clearly had in contemplation the more familiar scenarios of, say, reposted social media posts, or reposted links to webpages. First, are models sold under conditions of commercial confidentiality, or within proprietary access-restricted systems, made “public”? If not, the right does not operate. Was a "copy" or “replication” of the personal data made? Again, if we regard the model as a structure derived from personal data rather than personal data itself, neither of these applies. Was there a “link to” that original personal data? This seems more possible, but it is still rather a linguistic stretch.

Finally art 17(2) makes it plain that a controller is only obliged to do this as far as is reasonable, “taking account of available technology and the cost of implementation”. Even if all these problems are met, the obligation is only on the model-seller to “inform”. There is no obligation on the controller to whom the model was traded to do anything with this information. The data subject would, it seems, have to make another erasure request to that controller, unless they chose to redact the model voluntarily.

6.2.2 GDPR, article 20: the right to data portability

Article 20 provides that data subjects have the right to receive their personal data, “provided” to a controller, in a “structured, commonly used and machine readable format”, and that they then have the right to transmit that data to another controller “without hindrance”. Data portability is conceptually a sibling right to art 17. In theory, a data subject can ask for their data to be erased from one site (e.g. Google) and at the same time ported into their own hands.[186] Data subjects can also ask for data to be ported directly from controller A who currently is processing it to a controller B of their own choice.[187] Data portability is aimed explicitly at allowing data subjects to gain greater control over their personal data for consumer protection more than privacy purposes — e.g by allowing them to retrieve billing or transaction data from energy companies or banks — and re-use it in their own preferred ways to save money or gain advantages.[188]

In the context of ML, it is possible to imagine art 20 rights being used to facilitate user control over their personal data and possibly, the inferences drawn from it. It has often been suggested that data subjects might safeguard their privacy by adopting use

---

[186] Art 20(3).
[187] Art 20(2).
[188] See as an example of this kind of thinking, the UK’s *midata* scheme which on a voluntary basis preceded art 20. There are provisions in the UK’s Enterprise and Regulatory Reform Act 2013 for requiring companies to release data, but a 2014 government review concluded there was no case to use them. See Department for Business, Innovation and Skills (2014*) Review of the midata voluntary programme*. HM Government. Retrieved from https://www.gov.uk/government/publications/midata-voluntary-programme-review

of what are sometimes known as Personal Data Containers (PDCs). Using these technologies, the idea is that personal data need not be shared to secure desired services from giants such as Google or Facebook, who then use that data for their own profiling purposes, but rather the subject only provides an index of the data, keeping their own data either on their own server or perhaps in a trusted cloud storage. The philosophy behind this goes back several decades, to the idea that an "end-to-end" principle on the internet would empower the edges of a network, and avoid centralisation.[189] Proponents of data containers, which encompass research projects such as DataBox and Hub of all Things (HaT)[190], argue that these devices in your own homes or pockets might help you to archive data about yourself, coordinate processing with your data, and guard against threats[191]. Art 20 rights might enable data subjects to withdraw their personal data into PDCs in order to establish more informational self-determination in comparison to suffering the vagaries of profiling. However, as Hildebrandt points out, what we increasingly want is *not* a right not to be profiled — which means effectively secluding ourselves from society and its benefits — but to determine *how* we are profiled and on the basis of what data — a "right how to be read"[192]. Using art 20 portability rights, a data subject might choose to take their data to a controller whose model appealed to them from a market of choices: perhaps on the basis of a certification against particular values (see below) — rather than simply accept the model used by Google or its ilk.

This is no panacea, and there are a number of clear problems with using art 20 this way. First, is it likely the ordinary consumer would have either the information or the motivation to "shop around" for models in this way? Given the well-known inertia of consumers even about quite straightforward choices (e.g. switching energy suppliers, ISPs or banks to save money or get better service), it seems difficult to believe they could make this fairly esoteric choice without considerable improvements such as labelling or certification of algorithms (see section 6.2.3 below). It will take a long time for a competing marketplace of algorithmic model choices to emerge and indeed it is hard to see the current marketplace taking to such voluntarily.[193] Sometimes, as

---

[189] See Larry Lessig (2006) *Code 2.0*. Basic Books at 111; see also visions of this in the marketing literature, such as Alan Mitchell (2002) *Right Side Up*, HarperCollins.
[190] See discussion in Lachlan Urquhart and others, *Realising the Right to Data Portability for the Internet of Things* (March 15, 2017). Available on SSRN (doi:10.2139/ssrn.2933448).
[191] Richard Mortier and others "The Personal Container or Your Life in Bits" (2010) *Digital Futures '10, October 11–12, 2010, Nottingham, UK*. Retrieved from http://mor1.github.io/publications/pdf/de10-perscon.pdf
[192] Mireille Hildebrandt, *Smart technologies and the end(s) of law* (2005 Edward Elgar).
[193] It is beyond the scope of this paper to get into the economic and competition arguments here but it is already clear that information intermediaries occupying monopolistic or oligopolistic positions in the marketplace will not be keen on relinquishing them without regulatory command: this has already been seen in the UK in its attempts to introduce the *midata* scheme to the energy markets (n 188).

in criminal justice systems, it is hard to see how competing suppliers of models could emerge at all. On a practical point, it is quite possible that although the data subject may in theory gain greater control over their personal data, in reality they may not have the knowledge or time to safeguard their data against emerging threats.

Secondly, from a legal perspective, art 20 is (much like art 22) hedged around with what often seem capricious restrictions. It only applies to data the subject "provides". There seems no clear consensus on whether this covers just the explicit data a person provides (e.g. their name, hobbies, photos etc. on Facebook); the meta data the user supplies unknowingly (e.g. which pictures they look at, what links they click on, who is in their friends graph); or most damningly, the inferences that are then drawn from that data by the ML or profiling system itself. The Article 29 Working Party suggests that both the data a data subject provides directly, and data provided by "observing" a data subject, is subject to portability; but data *inferred* from these are not.[194] Furthermore art 20 only applies to data provided by "consent" (art 20(a)) — accordingly if data has been collected and profiled under another lawful ground such as the legitimate interests of the data controller, no right to portability exists.[195] Lastly it is worth emphasising this right only covers data which was being processed by "automated means" (art 20(1)(b)) — though not, as in art 22, "*solely*" automated means!

6.2.3 Privacy by design, supported by co-regulatory provisions

The GDPR discussion so far has revolved around rights given to individual data subjects. Although section 2.2 above demonstrates that algorithms create *societal* harms, such as discrimination against racial or minority groups, a focus on DP remedies makes an individualised approach inevitable. DP is a paradigm based on human rights which means it does not contemplate, as discussed above, remedies for groups (or indeed, for non-living persons such as corporations, or the deceased[196])

This means that even if the rights we have discussed above – arts 22, 15(h), 17 and 20 especially – do become valuable tools for individuals to try to "enslave" the algorithm, it is still up to individual data subjects to exercise them. This is not easy in

---

[194] See early guidance from the A29 WP: *Guidelines on the right to data portability*, 16/EN. WP 242, 13 December 2016. See possible conseqences of this for the right to erasure in section 6.2.2 above.
[195] This bizarre choice can only be explained by thinking of at 20 as a solution to promote competition by allowing data subjects to make active choices to retrieve their voluntarily posted data from social networks.
[196] Lilian Edwards and Edina Harbinja "Protecting Post-Mortem Privacy: Reconsidering the Privacy Interests of the Deceased in a Digital World" (2013) *Cardozo Arts & Entertainment Law Journal*, 32, 1. (doi:10.2139/ssrn.2267388)

the EU where consumers are on the whole far less prepared and empowered to litigate than in the US. The UK and many other EU nations have no generic system of class actions; this has been seen as a problem for many years, but attempts to solve it on an EU wide basis have repeatedly stalled.[197] Individuals are further hampered in meaningfully attaining civil justice by a general prejudice against contingency lawyering combined with dwindling levels of civil legal aid.

The DP regime contemplates that data subjects may find it hard to enforce their rights by placing general oversight in the hands of the independent regulator each state must have[198] (its DP Authority or DPA). However, DPAs are often critically underfunded since they must be independent of both state and commerce. They are often also significantly understaffed in terms of the kind of technical expertise necessary to understand and police algorithmic harms. In fact, financial constraints have in fact pushed DPAs such as the UK's ICO towards a much more "public administrative" role than one would expect, where problems (e.g. spamming, cold calling, cookies) are looked at more in the round as societal ills, than via championing individual data subject complaints.

Is it possible to derive any ways forward from the GDPR that are more likely to secure a better algorithmic society as a whole, rather than merely providing individual users with rights as tools which they may find impossible to wield to any great effect?

6.2.3.1 "Big data due process" and neutral data arbiters

It is interesting, looking from Europe, to observe how the predominantly North American legal literature has tried to solve the problems of algorithmic governance without the low-hanging fruit of a DP–based "right to an explanation". One notable bank of literature explores the idea of "big data due process". Crawford and Schultz[199] (drawing on early work by Citron[200]) interestingly attempt to model how due process rights already familiar to US citizens could be adapted to provide fairness, agency and transparency in cases around algorithmic automated systems in the governmental sector. Citron's work argues[201] for a number of radical adaptations to conventional due process which might include:

---

[197] It is of course possible for individual EU states in domestic law to provide class action rights but it is simply not a traditional or prevalent feature of these legal systems. The UK has made some legislative exceptions lately: e.g. the Consumer Rights Act 2015 allows collective proceedings and collective settlement orders in the Competition Appeal Tribunal – but this is only in relation to breach of competition rules, and does not apply to all consumer rights in the 2015 Act let alone all legal rights in the UK system, including the DP regime.
[198] See GDPR, art 51.
[199] Crawford and Schultz (n 12) at 123.
[200] Citron (n 14).
[201] Interestingly, she rejects as part of the "opportunity to be heard" a simple right to access to the algorithm's source code and/or a hearing on the logic of its decision as too expensive under the balancing test in *Matthews v Eldridge* (Citron (n 14) at 1284.*)*

- extra education about the "biases and fallacies" of automation for government agencies using automated systems;
- agencies to hire "hearing officers" to explain in detail their reliance on the outputs of such systems to make administrative decisions, including any "computer generated facts or legal findings";
- agencies to be required to regularly test systems for bias and other errors;
- audit trails to be issued by systems and notice to subjects that they have been used to make decisions, such that judicial review is possible.

Crawford and Schultz take these ideas of re-modelled due process and note they fit better into a model of structural rather than individualised due process.[202] For opaque predictive systems where data subjects never become aware of opportunities they might have had, reliance on individual rights and awareness is deeply problematic. In a structural approach, oversight and auditing can primarily be driven by public agencies. They suggest a "neutral data arbiter" with rights to investigate complaints from those whose data is used in predictive automatic systems, and provide a kind of "judicial review" by reviewing audit trails to find bias and unfairness that might render automated decisions invalid. This idea of an external regulator or audit body which might investigate complaints and provide mediation or adjudication is one with clear appeal in the literature: Crawford and Schultz suggest the FTC might act as a model but Tutt, for example, suggests an "FDA for algorithms".[203]

Seen through European eyes, two problems quickly emerge. One, the EU DP regime applies to private and public sector alike and in the private sector, it is harder to see these "due process" measures being taken on-board without compulsion or external funding. As we noted above, whereas transparency is a default in the public sector, the opposite is true in the private sector. Two, we essentially already have "neutral data arbiters" in the form of the state DPAs, and as just discussed, they are already struggling to regulate general privacy issues now let alone these more complex and opaque societal algorithmic harms.

6.2.3.2 Privacy by Design, DPIAs and certification schemes

However the GDPR introduces a number of new provisions which 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

---

[202] Crawford and Schultz (n 12)at 124.
[203] See Tutt (n 13). Other suggestions for algorithmic audit are usefully compiled by Mittelstadt and others (n 26) at 49.

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 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. PIAs (as they were formerly known) have traditionally been voluntary measures, in practice largely applied by public bodies bound to compliance and audit, such as health trusts. Attempts to expand their take up in Europe into areas like RFID[204] and the Internet of Things[205] by the private sector have in the main been unsuccessful. However the new art 35 is compulsory, not voluntary, and its definitions of "high risk" technologies are almost certain to capture many if not most ML systems. Art 35(3) (a) requires a DPIA where in particular there is a

> "systematic and extensive evaluation of personal aspects relating to natural persons […] based on automated processing, including profiling […] and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person."

This is almost identical to the formulation used in art 22. The ICO in their guidance report on *Big Data, AI and ML*[206] note firmly that "potential privacy risks" have already been identified with "the use of inferred data and predictive analytics". Accordingly, they provide a draft privacy impact assessment for big data analytics (Annex 1). It seems clear that, despite the uncertainty of the "high risk" threshold, DPIAs are quite

---

[204] See Privacy and Data Protection Impact Assessment Framework for RFID Applications 12 January 2011 at http://cordis.europa.eu/fp7/ict/enet/documents/rfid-pia-framework-final.pdf .
[205] See Data Protection Impact Assessment Template supported by Commission Recommendation 2014/724/EU at https://ec.europa.eu/energy/sites/ener/files/documents/DPIA%20template_incl%20line%20numbers.pdf .
[206] ICO (n 4)

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" (art 35(3)(b)).[207]

Where a DPIA is carried out and indicates a "high risk", then the local member state DPA must be consulted before the system can be launched. The impact assessment must be shared and the DPA must provide written advice to the controller and can use their powers to temporarily or permanently ban use of the system. Given the fines that can also be levied against non-compliant controllers under the GDPR (in the worst cases, up to 4% of global turnover[208]) this is potentially a very effective method to tame unfair ML systems[209]. Reuben Binns describes this as a kind of regulatory "triage".[210]

The voluntary measures of the GDPR may be equally influential 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.[211] In the UK, a tender has already been advertised by the ICO for a certification authority to run a UK privacy seal[212], 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 Citron's "big data due process" rights and Crawford and Schultz's "procedural due process". Certification could be applied to two main aspects of algorithmic systems:

a) certification of the algorithm as a software object by

---

[207] See also A29 WP A29 2016/679. Judging by this guidance, almost every ML system seems likely to require a DPIA.
[208] Art 83, GDPR. For potentially less severe transgressions, the maximum fine is the higher of €10m or 2% of global turnover, while for potentially more severe transgressions, the maximum fine is the higher of €20m or 4% of global turnover.
[209] In other work, one author has suggested that PIAs could be developed into more holistic Social Impact Assessments (SIAs) and although this was developed to deal with the IoT it might also have considerable application to ML systems: see L. Edwards, D. McAuley and L. Diver, "From Privacy Impact Assessment to Social Impact Assessment," 2016 *IEEE Security and Privacy Workshops* (SPW), San Jose, CA, 2016, pp. 53-57. doi: 10.1109/SPW.2016.19 .
[210] Reuben Binns "Data protection impact assessments: a meta-regulatory approach" (2017) 7 (1) *International Data Privacy Law* 22. (doi:10.1093/idpl/ipw027)
[211] For an early analysis of these provisions, see Rowena Rodrigues and others "Developing a privacy seal scheme (that works)" (2013) 3 *International Data Privacy Law* 2 100–116 (doi: 10.1093/idpl/ips037)
[212] Gemma Farmer "What's the latest on the ICO privacy seals?" (2015) *Information Commissioner's Office Blog*. Retrieved from https://iconewsblog.wordpress.com/2015/08/28/whats-the-latest-on-the-ico-privacy-seals/

   a. directly specifying either its design specifications or the process of its design, such as the expertise involved (technology-based standards, assuming good practices lead to good outcomes)
   b. and/or specifying output-related requirements that can be monitored and evaluated (performance-based standards);

b) certification of the whole person or process using the system to make decisions, which would consider algorithms as situated in the context of their use. Citron's "hearing officers", for example, might be provided by such provisions, perhaps as a form of alternate dispute resolution.

In these cases, not only could fairness and discrimination issues be considered in the standards to certify against[213], but it could be an opportunity to proactively encourage the creation of more scrutable algorithms.

One notable advantage is that certification standards could be set on a per-sector basis. This is already very common in other sociotechnical areas, such as environmental sustainability standards, where the standards for different environmental and labour harms in different certification systems such as SAN/Rainforest Alliance and Fair Trade also differ by crop. As we note, explanations and their effectiveness differ strongly by type (sections 4.1 and 5) domain (section 4.2) and user seeking explanation (4.3), and it is likely that the exact form of any truly useful explanation-based remedy would vary strongly across both these and other factors. Certification could be augmented by the development of codes of conduct (arts 40 and 41, GDPR) for any specified sector, such as for algorithms considering housing allocation systems, targeted advertising, tax fraud detection or recidivism.

Promising as this may sound, voluntary self-or co-regulation by privacy seal has had a bad track record in privacy, with recurring issues around regulatory and stakeholder capture. The demise of *Safe Harbor* alone[214], which was externally validated for years by trust seals like *TrustE*, means that many Europeans will be rightly sceptical about the delivery of real corporate change and substantive compliance with privacy rights by certification[215].

Another issue is that DPIAs, PbD, certification and the general principle of "accountability"[216] in the GDPR bring with them a real danger of formalistic

---

[213] Issues of algorithmic fairness are specifically discussed in GDPR, recital 71.
[214] See CJEU case of *Schrems v Data Protection Commissioner of Ireland*, Case C-362/14, 6 October 2015.
[215] See on the failure of TrustE and similar privacy seals to meet European privacy standards, Charlesworth (cite from 2nd edn *Law and Internet*).
[216] GDPR art 5(2). There is not time to discuss this fully in this article, but it is likely to support the creation of a new world of form-filling for data controllers.

bureaucratic overkill alongside a lack of substantive change: a happy vision for more form-filling jobs and ticked boxes, but a sad one for a world where automated algorithms do their jobs quietly without imperilling human rights and freedoms, especially privacy and autonomy.

6.3 Conclusions

Algorithms, particularly of the ML variety, are increasingly consequential to individuals' lives but have caused a range of concerns. Transparency in the form of a "right to an explanation" has emerged as a compellingly attractive remedy since it intuitively presents as a means to "open the black box", hence allowing individual challenge and redress, as well as possibilities to foster accountability of ML systems. In the general furore over algorithmic bias, opacity and unfairness laid out in section 2, any remedy in a storm has looked attractive.

In this article, we traced how, despite these hopes, a right to an explanation in the GDPR seems unlikely to help us find complete remedies, particularly in some of the core "algorithmic war stories" that have shaped recent attitudes in this domain. A few reasons underpin this conclusions. Firstly (section 3), the law is restrictive on when any explanation-related right can be triggered, and in many places is unclear, or even seems paradoxical. Secondly (section 4), even were some of these restrictions to be navigated (such as with decisive case law), the way that explanations are conceived of legally - as "meaningful information about the logic of processing" - is unlikely to be provided by the kind of ML "explanations" computer scientists have been developing.

ML explanations are restricted both by the type of explanation sought, the multi-dimensionality of the domain and the type of user seeking an explanation. However (section 5) "subject-centric' explanations (SCEs), which restrict explanations to particular regions of a model around a query, show promise. In particular we suggest these are not just usable, as Wachter and others[217] argue, "*after* an automated decision has taken place", but might be put into interactive systems that allow individuals to explore and build their own mental models of complex algorithms. Similarly "pedagogical" systems which create explanations around a model rather than from decomposing it may also be useful and benefit from not relying on disclosure of proprietary secrets or IP.

As an interim conclusion then, while convinced that recent research in ML explanations shows promise, we fear that, given the preconceptions in the legal

[217] Wachter and others (n 11)

wording of provisions like the GDPR art 15(h), the search for a right to an explanation may be at best distracting and at worst nurture a new kind of "transparency fallacy" to match the existing phenomenon of "meaningless consent" (section 6.1). So as our last exercise, we turn our focus to the other legal rights of the GDPR which might aid those impacted adversely by ML systems. We note with caution some possible uses of the GDPR's "right to erasure" and the "right to data portability" to "slave" the algorithm, but find that, like the "right to an explanation", these rely too much on individual rights for what are too often group harms.

However, radically, in section 6 we find that some of the new tools in the GDPR, in particular the requirements for Privacy by Design and DPIAs, and opportunities for certification systems, might go beyond the individual to focus *a priori* on the creation of better algorithms, as well as creative ways for individuals to be assured about algorithmic governance e.g. by certification of performance, or of the professionals building or using algorithms. Starting from a notion of creating better systems, with less opacity, clearer audit trails, well and holistically trained designers, and input from concerned publics[218] seems eminently more appealing than grimly pursuing against the odds a "meaningful" version of the interior of a black box.

6.3.1 Further work

There are other matters which have only been hinted at in this already long article and which we hope to explore in further work. One is oversight and audit. Any system based on GDPR rights ultimately puts the supervisory burden on the state DPA. Is this correct? We have already seen that DPAs are overwhelmed by the task of managing privacy enforcement in the digital era. Is every algorithmic harm also their bailiwick? Does this extend to datasets steeped in societal racial bias, driverless trolley-cars that cannot understand whether to mow down one person or five[219], identification systems that think only light skinned people are beautiful[220] and social media algorithms that distribute fake news? All of these involve the processing of personal data at some level, but they do not relate to privacy except in the loosest sense. There is an issue here about whether simply because "data protection" has the word data in it, should it acquire hegemony over all the ills of data-driven society?

---

[218] See GDPR art 35(7)(9) which suggests when conducting a DPIA that the views of data subjects shall be sought when appropriate but (always a catch) "without prejudice to" commercial secrecy or security.
[219] See *passim* the glorious *Trolley problem memes* page at https://www.facebook.com/TrolleyProblemMemes/ .
[220] See "FaceApp sorry for suggesting that light skin is 'hotter' than dark skin" , *The Inquirer*, 25 April 2017, at https://www.theinquirer.net/inquirer/news/3008961/faceapp-sorry-for-suggesting-that-light-skin-is-hotter-than-dark-skin .

Furthermore, what about ML systems that mainly deal with non-personal data? Should they be excluded from any DP based governance system? The EU already thinks, from an economic perspective, that the lack of rights over non-personal data is a problem waiting to happen[221], particularly as we likely want critical systems to be reliable, rather than simply non-discriminatory. On the other hand, that could be seen as an advantage: in a recent UK Parliamentary consultation on how to regulate algorithms, the Royal Society complained that:

> "Machine learning algorithms are just computer programs, and the range and extent of their use is extremely broad and extremely diverse. It would be odd, unwieldy, and intrusive to suggest governance for all uses of computer programming, and the same general argument would apply to all uses of machine learning.
>
> […] In many or most contexts machine learning is generally uncontroversial, and does not need a new governance framework. How a company uses machine learning to improve its energy usage or warehouse facilities, how an individual uses machine learning to plan their travel, or how a retailer uses machine learning to recommend additional products to consumers would not seem to require changes to governance. It should of course be subject to the law, and also involve appropriate data use.
>
> Many of the issues around machine learning algorithms are very context specific, so it would be unhelpful to create a general governance framework or governance body for all machine learning applications. Issues around safety and proper testing in transport applications are likely to be better handled by existing bodies in that sector; questions about validation of medical applications of machine learning by existing medical regulatory bodies; those around applications of machine learning in personal finance by financial regulators.[222]"

We have already noted that sectors are likely to have specific needs for explanation and that a sectoral approach might be fostered by certification. In a world apparently scrambling to create as many new bodies as possible for various types of oversight of

---

[221] European Commission https://ec.europa.eu/digital-single-market/en/news/public-consultation-building-european-data-economy .

[222] This submission is primarily drawn from the recent report: The Royal Society, *Machine learning: the power and promise of computers that learn by example* (2017).

AI, ML and algorithmic decision making in embodied forms such as robots[223], it is worth keeping a sector–specific, purpose-driven sentiment in mind.

As we have already noted, many of the problems with algorithms are more problems for groups than for individuals. Remedies aimed at empowering or protecting groups — remedies such as "an FDA for algorithms" or a "supercomplaint" system to empower third party organisations, or a European-style ombudsman body — may be more useful things to consider and reinvent than struggling to transform the individual rights paradigm of DP.

Finally, this work has been a true (and sometimes heated) interdisciplinary collaboration between (reductively) a ML specialist and a DP lawyer. Any attempts to increase the transparency or explicability of ML systems, and indeed, in general to better harness them to social good, will not function effectively without this kind of interdisciplinary work. We need to consider algorithms in the sociotechnical context within which they work. We will, as Mireille Hildebrandt describes, "have to involve cognitive scientists, computer engineers, lawyers, designers of interfaces and experts in human-computer interaction with a clear understanding of what is at stake in terms of democracy and the rule of law".[224]

We thus end with a reiteration of the common plea for collegiate work not only across different legal jurisdictions and across different disciplines, but also between academics and practitioners. In relation to applied domains in particular, we fear that the situation is becoming more adversarial than collaborative, and that colleagues risk burning bridges with the very practitioner communities they should be working with, rather than against. Only with exemplary, trans-disciplinary collaboration can we hope not just to enslave the algorithm, but to purpose them towards legitimate societal ends.

Acknowledgements

Lilian Edwards's work was supported by the Arts and Humanities Research Council (AHRC) centre CREATe, and the Engineering and Physical Sciences Research Council (EPSRC) Digital Economy Hub Horizon at University of Nottingham, grant number EP/G065802/1; Michael Veale is supported by the EPSRC, grant number EP/M507970/1.

---

[223] See Commons Science and Technology Committee (2016) *The Big Data Dilemma*, UK Parliament; and Commons Science and Technology Committee (2016) *Robotics and artificial intelligence*, UK Parliament; The Conservative Party, *The Conservative and Unionist Party 2017 Manifesto* at 79; European Parliament, *Report with recommendations to the Commission on Civil Law Rules on Robotics*, 2015/2103(INL).

[224] Hildebrandt (n 11) at 54.