SLAVE TO THE ALGORITHM? WHY A 'RIGHT TO AN EXPLANATION' IS PROBABLY NOT THE REMEDY YOU ARE LOOKING FOR

LILIAN EDWARDS[†] & MICHAEL VEALE[††]

ABSTRACT

Algorithms, particularly machine learning (ML) algorithms, are increasingly important to individuals' lives, but have caused a range of concerns revolving mainly around unfairness, discrimination and opacity. Transparency in the form of a "right to an explanation" has emerged as a compellingly attractive remedy since it intuitively promises to open the algorithmic "black box" to promote challenge, redress, and hopefully heightened accountability. Amidst the general furore over algorithmic bias we describe, any remedy in a storm has looked attractive.

However, we argue that a right to an explanation in the EU General Data Protection Regulation (GDPR) is unlikely to present a complete remedy to algorithmic harms, particularly in some of the core "algorithmic war stories" that have shaped recent attitudes in this domain. Firstly, the law is restrictive, unclear, or even paradoxical concerning when any explanation-related right can be triggered. Secondly, even navigating this, the legal conception of explanations as "meaningful information about the logic of processing" may not be provided by the kind of ML "explanations" computer scientists have developed, partially in response. ML explanations are restricted both by the type of

[†] Professor of Internet Law, Strathclyde Law School, University of Strathclyde, Glasgow, UK [lilian.edwards [at] strath.ac.uk]. Research supported in part 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.

[††] Doctoral candidate, Department of Science, Technology, Engineering and Public Policy (STEaPP), University College London, UK [m.veale [at] ucl.ac.uk]; technical advisor, Red Cross Red Crescent Climate Centre. Michael Veale receives support from the EPSRC, grant number EP/M507970/1 and the World Bank Global Facility for Disaster Reduction and Recovery (GFDRR).

The authors would like to thank Johannes Welbl, Max Van Kleek, Reuben Binns, Giles Lane, and Tristan Henderson, and participants of BILETA 2017 (University of Braga, Portugal), the 2017 Privacy Law Scholars Conference (PLSC) and the 2017 Big Data: New Challenges for Law and Ethics Conference, University of Ljubljana, for their helpful comments.

explanation sought, the dimensionality of the domain and the type of user seeking an explanation. However, "subject-centric" explanations (SCEs) focussing on particular regions of a model around a query show promise for interactive exploration, as do explanation systems based on learning a model from outside rather than taking it apart (pedagogical versus decompositional explanations) in dodging developers' worries of intellectual property or trade secrets disclosure.

Based on our analysis, we fear that the search for a "right to an explanation" in the GDPR may be at best distracting, and at worst nurture a new kind of "transparency fallacy." But all is not lost. We argue that other parts of the GDPR related (i) to the right to erasure ("right to be forgotten") and the right to data portability; and (ii) to privacy by design, Data Protection Impact Assessments and certification and privacy seals, may have the seeds we can use to make algorithms more responsible, explicable, and human-centered.

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. Algorithmic systems are thus increasingly familiar, even vital, in both private, public and domestic sectors of life.

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 an 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

[1] *See, e.g.*, TANIA BUCHER, *The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms*, 20 INFO., COMM. AND SOC'Y 30 (2015) (a qualitative study of how individuals online understand the "algorithms" around them).

mysteriously solve hitherto unassailable problems in society.[2] It seems that we are all now to some extent, "slaves to the algorithm." In his landmark book, Frank Pasquale describes this as "the black box society,"[3] and the issue has become a subject of international attention by regulators, expert bodies, politicians and legislatures.[4]

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 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

---

[2] *See* Randall Munroe, *Here to Help*, XKCD (last visited May 25, 2017), https://xkcd.com/1831/.

[3] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (Harvard University Press 2015).

[4] *See,* e.g., INFORMATION COMMISSIONERS OFFICE (ICO), BIG DATA, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA PROTECTION (2017); EUROPEAN DATA PROTECTION SUPERVISOR (EDPS), MEETING THE CHALLENGES OF BIG DATA: A CALL FOR TRANSPARENCY, USER CONTROL, DATA PROTECTION BY DESIGN AND ACCOUNTABILITY [OPINION 7/2015] (2015). ROYAL SOCIETY, MACHINE LEARNING: THE POWER AND PROMISE OF COMPUTERS THAT LEARN BY EXAMPLE. (2017); Wetenschappelijke Raad voor het Regeringsbeleid [Dutch Scientific Council for Government Policy (WRR)], *Big data in een vrije en veilige samenleving* [Big data in a free and safe society], WRR-RAPPORT 95 (2016); Commons Science and Technology Select Committee, *Algorithms in Decision-Making Inquiry Launched*, UK PARLIAMENT (Feb. 28, 2017)[https://perma.cc/PJX2-XT7X]; NATIONAL SCIENCE AND TECHNOLOGY COUNCIL, PREPARING FOR THE FUTURE OF AI (2016), [https://perma.cc/6CDM-VR3V].

[5] Regulation (EU) 2016/679, of the European Parliament and the Council of 27 April 2016 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), 2016 O.J. (L 119) 1 (hereafter "GDPR").

[6] Bryce Goodman & Seth Flaxman, *EU Regulations on Algorithmic Decision Making and "a Right to an Explanation*," 2016 ICML WORKSHOP ON HUMAN INTERPRETABILITY IN ML (2016).

[7] *See,* e.g., Ian Sample, *AI Watchdog Needed to Regulate Automated Decision-making, Say Experts*, THE GUARDIAN, (Jan. 27, 2017), [https://perma.cc/TW2C-MZWX].

[8] There is a long history of work into explanation facilities, previously referred to as "scrutability" in Web Science. *See,* e.g., Judy Kay, *Scrutable Adaptation: Because We Can and Must*, (2006), *in* ADAPTIVE HYPERMEDIA AND ADAPTIVE WEB-BASED SYSTEMS (V.P. Wade et al. eds., Springer 2006).

[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, 1995 O.J. (L 281) 31 (hereafter "Data Protection Directive" or "DPD"). For earlier discussions concerning what is now referred to as a "right to an

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 indicate.

Restrictions identified within the GDPR's Articles 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; restriction to decisions that produce "legal" or similarly "significant" effects; 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 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

---

explanation," *see* Alfred Kobsa, *Tailoring Privacy to Users' Needs*, *in* USER MODELING, (M. Bauer et al. eds., Springer 2001), doi:10.1007/3-540-44566-8_52; Mireille Hildebrandt, *Profiling and the rule of law*, 1 IDENTITY IN THE INFORMATION SOCIETY 1, 55 (2008), doi:10.1007/s12394-008-0003-1.

[10] Rosemary Jay, *UK Data Protection Act 1998—the Human Rights Context*, 14 INT'L REV. OF LAW, COMPUTERS & TECH. 3, 385, (2000) (doi:10.1080/713673366).

[11] Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INTERNATIONAL DATA PRIVACY LAW 2, 76–99 (2017), doi:10.1093/idpl/ipx005; Mireille Hildebrandt, *The Dawn of a Critical Transparency Right for the Profiling Era*, DIGITAL ENLIGHTENMENT YEARBOOK 2012 (Jacques Bus et al. eds., 2012).

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 generally considered to be adversely affected by algorithmic decision-making.

This exercise also identifies a further problem: data protection (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 practical a right to a "meaningful" explanation is 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.

However, on a more positive note, we observe that explanations may usefully be developed for purposes other than to vindicate data subject rights. Firstly, 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. Secondly, pedagogical explanations (a model-of-a-model), rather than those made by decomposition (explaining it using the innards) may avoid the need to disclose protected 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 Articles 17 and 20 respectively, may in certain cases be as useful, if not more so, than a right to an explanation. However, their application to inferences is still unclear and up for grabs.

Second, we consider several novel provisions in the GDPR which do not 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, to review the accuracy, lack of bias and integrity of a ML system in the round and not simply challenge ML decisions on behalf of individuals. 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.

---

[12] *See*, e.g., Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 1 (2014); Pasquale, *supra* note 3.

[13] *See generally* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, (2017) (on how the FDA might take a watchdog function).

[14] *See generally* Crawford & Schultz, *supra* note 12; Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008) (on how due process might be extended in the Big Data era).

I. ALGORITHMS, AND HOW WE ARE A 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. Such systems 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. These systems—discussed in detail below—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 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

---

[15] A HISTORY OF ALGORITHMS: FROM THE PEBBLE TO THE MICROCHIP (Jean-Luc Chabert et al. eds., 1999) at 2.

[16] HM TREASURY, REVIEW OF QUALITY ASSURANCE OF GOVERNMENT ANALYTICAL MODELS: FINAL REPORT 26 (HM Government, 2013).

[17] A variation on the more legally familiar "computational" turn. *See generally* MIREILLE HILDEBRANDT, PRIVACY, DUE PROCESS AND THE COMPUTATIONAL TURN (Routledge, 2008) (on the legal implications of a variety of data-driven technologies).

[18] *See* RICHARD SUSSKIND, EXPERT SYSTEMS IN LAW (Clarendon Press 1989); JOHN ZELEZNIKOW AND DAN HUNTER, BUILDING INTELLIGENT LEGAL INFORMATION SYSTEMS: REPRESENTATION AND REASONING IN LAW (Kluwer, 1994); *see also* Lilian Edwards & John A.K. Huntley, *Creating a Civil Jurisdiction Adviser*, 1 INFORMATION & COMMUNICATIONS TECHNOLOGY LAW 5 5 (1992), doi:10.1080/13600834.1992.9965640.

explanation facilities.[19] The explanation facilities that *were* developed in the era of expert systems set a high, albeit often overlooked, bar for today's discussions.

A. The Rise of Learning Algorithms

Progress in automated decision-making and decision support systems was initially held back by a lack of large-scale data and algorithmic architectures that could leverage them, restraining systems to the relatively simplistic problems. In recent years, technologies capable of coping with more input data and highly non-linear correlations have been developed, allowing the modelling of 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 (such as the early legal systems noted above) to ML. In rule-based systems, explicitly defined logics turn input variables, such as credit card transaction information, into output variables, such as a flag for fraud. Complex ML algorithms are different: output variables and input variables together are fed into an algorithm theoretically demonstrated to be able to "learn" from data. This process trains a model exhibiting implicit, rather than explicit, logics, usually not optimised for human-understanding as rule-based systems are.[20] The learning algorithms that make this possible are often not blazingly new, many dating from the '70s, '80s and '90s. But we now have comparatively huge volumes of data that can be stored and processed cheaply, such that the performance of and ability to further research 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,[21] such as physical symptoms or 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. Instead, ML systems 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

---

[19] *See,* e.g., John Zeleznikow & Andrew Stranieri, *The Split-up System: Integrating Neural Networks and Rule-based Reasoning in the Legal Domain*, 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] Machine learning techniques that explicitly encode logic are found in natural language processing and in bioinformatics, but not focussed on here.

[21] This vector is like a row in a spreadsheet where columns are characteristics.

cluster might mean.[22] Segmentation by market researchers, for example, would be a relevant field where unsupervised learning might be fruitfully applied since they 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. 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. 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.

ML is the focus of this piece, for several reasons. In our current interconnected, data-driven society, only ML systems have demonstrated the ability to automate difficult or nuanced tasks, such as search, machine vision and voice recognition. As a result, ML systems are fast becoming part of our critical societal infrastructure. Significantly, it would be impractical for many of these decisions to have a "human in the loop"; this is truer still in complex ambient or "smart" environments.

ML uptake is also driven by business models and political goals, which have led practitioners to seek more "data-driven decisions." 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."[24] These logics of "datafication" have also lead to increasing uptake of ML in areas where the service offering does not

---

[22] Some other types of learning exist, such as semi-supervised learning or reinforcement learning, but they do not consider enough relevant current challenges to be considered here.

[23] TOM MITCHELL, MACHINE LEARNING (McGraw Hill 1997).

[24] Shoshana Zuboff, *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization,* 30 J. INFO. TECH. 1, 75–89 (2015).

necessarily require it, particularly in augmenting existing decisions with ML–based decision-support, in areas such as justice, policing, taxation or food safety.

In this article, we are aware we are speaking across a wide range of very different ML systems—in scope, size, purpose and user—which may raise very different legal, ethical and societal issues, and this may lead to some misleading generalisations. However, at this early stage of the research into ML and the GDPR, a wide scope seems important, and where critical differences arise between private and public sector ML systems, we have tried to make this plain.

B. ML and Society: Issue 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.

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.[25] Once it was commonly thought that machines could not display the biases of people and so would be ideal neutral decision makers.[26] This had considerable influence on some early legal cases involving Google and other online intermediaries and their responsibility (or not) for algorithmic harms.[27] 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 incontrovertible character to which a human decision-maker may attach too much weight, thus abdicating his own responsibilities."[28]

---

[25] See the useful survey in Brent Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3 BIG DATA & SOC'Y 2 (2017), especially at section 7.

[26] *See* Christian Sandvig, *Seeing the Sort: The Aesthetic and Industrial Defence of "the Algorithm,"* 11 MEDIA-N 1 (2015).

[27] 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)*, 21 INT'L. J. L. INFO. TECH. 2, 2 (2013).

[28] *Amended Proposal for a Council Directive on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data*, at 26, COM(92) 422 final—SYN 297 (Oct. 15, 1992).

As Mittelstadt et al. put it, "this belief is unsustainable"[29] given the volume of recent evidence, 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.

Even if, as some argue, "big data" will eventually give us a complete picture of society[30]—we will come to our reservations about this—making decisions based on past data is 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[31]) play a part in decision-making processes.[32] Algorithmic systems trained on past biased data without careful consideration are inherently likely to recreate or even exacerbate discrimination seen in past decision-making. For example, a CV or résumé filtering system based only on past success rates for job applicants will likely encode and replicate some of the biases exhibited by those filtering CVs or awarding positions manually in the past. While some worry that these systems will formalise explicit bias of the developers, the larger concern appears to be that these systems will be indirectly, unintentionally and unknowingly discriminatory.[33]

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, to submit a fraudulent tax or welfare claim, to follow an advert for a pay-day loan, or to 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 seem socially unacceptable. The use of gender—and its recent prohibition—in the pricing of car insurance in the EU serves as a

---

[29] Mittelstadt et al., *supra* note 25, at 25.

[30] VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (Houghton Mifflin Harcourt, 2013).

[31] In relation to the U.K., see Equality Act 2010, c. 15, s. 19; for discussion of US law, compare Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).

[32] For a US discussion in the context of ML discrimination, see Barocas and Selbst, *supra* note 31.

[33] *See* Toon Calders & Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY (Bart Custers et al. eds., Springer 2013).

recent salient example.[34] 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—e.g. 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. Put simply, if the sensitive variable might be predictively useful, and we suspect the remaining variables might contain signals that allow us to predict the variable we omitted, then unwanted discrimination can sneak back in. On rare occasions, this happens explicitly. 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.[35]

More broadly, cases around "redlining" on the internet—"weblining," as it was known nearly 20 years ago[36]—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. 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.[37] Similarly, the online 1-hour-media-delivery service Kozmo received a lawsuit for denying delivery to residents in black neighbourhoods in Washington, DC, which they defended in the media by saying that they were not targeting neighbourhoods based on race, but based on high Internet usage.[38]

---

[34] *See* Case C-236/09, Association belge des Consommateurs Test-Achats ASBL and Others v. Conseil des ministres, 2011 E.C.R. I-00773.

[35] *See* Julia Angwin & Terry Parris Jr., *Facebook Lets Advertisers Exclude Users by Race*, PROPUBLICA (Oct. 28, 2016), https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race. While ProPublica subsequently reported (Feb. 8, 2017) that Facebook amended their dashboard as to "prevent advertisers from using racial categories in ads for housing, employment and credit" and to warn advertisers to comply with the law in other categories, a year later the reporters were still able to successfully place adverts excluding people on the basis of a wide array of inferred characteristics. *See* Julia Angwin et al., *Facebook (Still) Letting Housing Advertisers Exclude Users by Race,* PROPUBLICA (Nov. 21, 2017), https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin.

[36] Marcia Stepanek, *Weblining: Companies are using your personal data to limit your choices—and force you to pay more for products*, BLOOMBERG BUSINESS WEEK, Apr. 3, 2000, at 26.

[37] *Wells Fargo yanks "Community Calculator" service after ACORN lawsuit*, CREDIT UNION TIMES (July 19, 2000), https://perma.cc/XG79-9P74.

[38] Elliot Zaret & Brock N Meeks, *Kozmo's digital dividing lines*, MSNBC (Apr. 11, 2000); Kate Marquess, *Redline may be going online*, 86 ABA J. 8 at 81 (Aug. 2000).

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 reflect a less starkly institutionally racist mass of training data.[39] Racism is surely problematic around the world, yet does not manifest in statistically identical ways everywhere. Countries with deeper or clearer racial cleavages are naturally going to collect deeper or more clearly racist datasets, yet this does not mean that more nuanced issues of racism, particularly in interaction with other variables, does not exist.

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?[40] 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 browser might, even if predictively important, be considered to abuse short-lived, arbitrary correlations, and in doing so, tangibly restrict individuals' autonomy.

In the European data protection regime, fairness is an overarching obligation when data is collected and processed[41] something which is sometimes overshadowed by the focus on lawful grounds for processing. The UK's data protection authority, the Information Commissioner's Office (ICO) published recent guidance on big data analytics which seems to imply that ML systems are not unfair simply because they are "creepy" or produce

---

[39] Note the recent report of the ICO, *supra* note 4, which pays serious attention to issues of fairness and bias but cites only US examples of such despite being a product of the UK regulator. The German autocomplete cases are cited but referred to interestingly, as questions of error or accuracy, rather than discrimination or fairness. *See* Jones, *infra* note 79. *See also* National Science and Technology Council, *supra* note 4, at 30 (specifying that "it is important anyone using AI in the criminal justice system is aware of the limitations of the current data").

[40] *How might your choice of browser affect your job prospects?*, THE ECONOMIST , Apr. 11, 2013 https://www.economist.com/blogs/economist-explains/2013/04/economist-explains-how-browser-affects-job-prospects.

[41] GDPR, *supra* note 5, at art. 5(1)(a).

unexpected results.[42] However, they may be unfair 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.[43] The ICO report argues that unfairness may, on occasion, derive from expectations, where data is used for a reason apparently unconnected with the reason given for its collection,[44] and from lack of transparency, which we discuss in detail below in section I.B.3.

Reliance on past data additionally asks fairness questions that relate to the memory of algorithmic systems—how far back and with which variables should judge people on? Are individuals legally entitled to a *tabula rasa*—a blank slate—after a certain number of years, as is common in some areas of criminal justice?[45] There is a widely-held societal value to being able to "make a fresh start," and technological change can create new challenges to this. Indeed, common institutional frameworks for forgetfulness can be found in bankruptcy law and in credit scoring.[46]

---

[42] ICO, *supra* note 4.

[43] ICO, *supra* note 4 at 20 (referencing Jessica Fleig, *Doctor Locked Out of Women's Changing Room Because Gym Automatically Registered Everyone with Dr Title as Male*, THE MIRROR (Mar. 18, 2015), http://www.mirror.co.uk/news/uk-news/doctor-locked-out-womens-changing-5358594). This is one of very few reports of algorithmic misbehaviour in the ICO report not emanating from the US, and the simplicity of the task means it is very unlikely to be ML based.

[44] *See* the well-known theory of contextual integrity in HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (Stanford Law Books, 2009). In a perfect, data protection compliant world, all purposes for which data is to be used, including re-uses, should be notified in the privacy policy or otherwise. *See* GDPR art. 5(1)(b). However, as discussed *infra* in section V.A, this concept of "notice and choice" is increasingly broken.

[45] 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," discussed *infra* at Section IV.B.1, *See Criminal Record Checks*, NACRO https://perma.cc/GKY4-KHJA.

[46] Jean-François Blanchette & Deborah G. Johnson, *Data Retention and the Panoptic Society: The Social Benefits of Forgetfulness*, 18 THE INFO. SOC'Y 1, 33–45 (2002).

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.[47] 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, Article 4(4) now defines it, profiling is:

> [A]ny 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.

ML is now the favoured way of deriving 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. As Hildebrandt notes, profiling is what all organisms do in relation to their environments, and is "as old as life itself."[48]

For data subjects, privacy concerns here embrace an enormous weight of issues about how data concerning individuals are collected to be bent into profiles, how individuals can control access to and processing of data relating to them, 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 data protection principle that data should be collected for named and specific purposes.[49] Data collected for selling books becomes repurposed as a system to sell advertisements book buyers might like. Connected problems are that "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 Article 5(1)(e) and (c)). These are huge problems at the heart of contemporary data

---

[47] *See generally*, Mireille Hildebrandt, *Defining Profiling: A New Type of Knowledge?*, *in* PROFILING THE EUROPEAN CITIZEN 17 (Mireille Hildebrandt & Serge Gutwirth eds., Springer 2008).

[48] Hildebrandt, *supra* note 9.

[49] GDPR, art. 5(1)(b) (referring to "purpose limitation").

protection law,[50] and we do not seek to review these fully here. We do however want to point out where these issues specifically affect ML.

First, an exceedingly trite point is that data subjects increasingly perceive themselves as having little control over the collection of their personal data that go into profiles. As an example, the most recent Eurobarometer survey on personal data from June 2015 showed 31% of EU citizens as feeling they had no control over the data they provided online and a further 50% feeling they had only partial control.[51] In the GDPR, collection falls under "processing" of data (Article 4(2)) and is theoretically controlled by (inter alia) the need for a lawful ground of processing (Article 6). Most lay people believe consent is the only lawful ground for processing and thus defends their right to autonomous privacy management (though perhaps not in so many words).[52] Yet consent is not the only lawful ground under Article 6. It's quite possible that as much personal data is collected on the grounds of the "legitimate interests" of the controller (at least in the private sector), or on the grounds that the data was necessary to fulfil a contract entered into by the data subject.[53] More importantly, consent has become debased currency given ever-longer standard term privacy policies, "nudging" methods such as screen layout manipulation, and market network effects. It is often described using terms such as "meaningless" or "illusory."[54]

The consent problem is aggravated by the rise of "bastard data," a picturesque term coined by Joe McNamee. He notes that as data is linked and transformed it incentivises new data collection. Thus, data "ha[s] become fertile and ha[s] bastard offspring that create new challenges that go

---

[50] *See* discussion in ARTICLE 29 WORKING PARTY (hereinafter "A29 WP"), OPINION 03/2013 ON PURPOSE LIMITATION, (Apr. 2, 2013); ICO, *supra* note 4 at 11-12; EDPS, *supra* note 4.

[51] EUROPEAN COMMISSION, DATA PROTECTION EUROOBAROMETER, (June 2015), [https://perma.cc/3XLK-VKA6].

[52] The UK's Information Commissioner recently addressed the prevalence of these beliefs head on with a series of "GDPR Myths" blogs addressing what she refers to, tongue-in-cheek, as "alternative facts." *See* Elizabeth Denham, *Consent is Not the 'Silver Bullet' for GDPR Compliance*, INFORMATION COMMISSIONER'S OFFICE NEWS BLOG, (Aug. 16, 2017), https://iconewsblog.org.uk/2017/08/16/consent-is-not-the-silver-bullet-for-gdpr-compliance/.

[53] GDPR, art. 6. The public sector has separate public interest grounds for processing. Policing and national security are exempted from the GDPR, but covered in a connected directive.

[54] For a full discussion of the illusory nature of consent in the Internet world, see Lilian Edwards, *Privacy, Law, Code and Social Networking Sites*, *in* RESEARCH HANDBOOK ON GOVERNANCE OF THE INTERNET (Ian Brown ed., Edward Elgar, 2013) at 323; Rikke Frank Joergensen, *The Unbearable Lightness of User Consent*, 3 INTERNET POL'Y REV. 4 (2014); Brendan Van Alsenoy et al., *Privacy notices versus informational self-determination: Minding the gap*, 28 INT'L REV. OF LAW, COMPUTERS & TECHNOLOGY 2 185–203 (2014).

far beyond what society previously (and, unfortunately, still) consider[] to be 'privacy.'"[55] Many of these offspring are profiles produced by ML systems. Typically, data about people, which are personal, are transformed into data which have often been seen as non-personal and therefore fall outside the scope of data protection law, perhaps simply because the data subject name or other obvious identifier has been removed.[56] 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 data protection control. In recent times, the anonymity argument has been parried on grounds of potential for re-identification.[57] This has become especially crucial in the emerging ambient environment deriving from the Internet of Things (IoT), which collects data that on first glance looks mundane, but can be used with relative ease to discover granular, intimate insights. 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, can be re-connected to individuals, and identities established from it.[58] In practice, this has meant that the day-to-day actions that individuals undertake, especially in "smart" environments,[59] leave trails of potentially sensitive latent personal data in the hands of controllers who may be

---

[55] Joe McNamee, *Is Privacy Still Relevant in a World of Bastard data?*, EDRI EDITORIAL, (Mar. 9, 2016), https://edri.org/enditorial-is-privacy-still-relevant-in-a-world-of-bastard-data.

[56] "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 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." *Id.* at 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, e.g. it is a form of "privacy by design" and is excluded from mandatory security breach notification. *Id.* at art. 25,

[57] *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010). In Europe, see A 29 WP, OPINION 05/2014 ON ANONYMISATION TECHNIQUES (2014).

[58] *See* Yves-Alexandre de Montjoye et al., *Unique in the crowd: The privacy bounds of human mobility*, 3 SCIENTIFIC REPORTS (2013); MICHAEL VEALE, DATA MANAGEMENT AND USE: CASE STUDIES OF TECHNOLOGIES AND GOVERNANCE (The Royal Society and the British Academy 2017).

[59] *See* 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*, 1 EUR. DATA PROT. L. REV. 28, 28–58 (2016).

difficult to identify.[60] If controllers are not identifiable, data subjects may not be able to effectively exercise the data protection rights we discuss in sections II and V 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 data protection law and possibly not in equality law.[61]

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,[62] 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."[63] 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, importantly, control over) the relevant group profile. A decade later, the GDPR refines this argument by asserting that if a profile can be used to target or "single me out"[64]—for example, to deny me access to luxury services or to discriminate about what price I can buy goods at—then the

---

[60] On the practical difficulties for data subjects to identify data controllers, *see* Max Van Kleek et al., *Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps*, *in* CHI'17 (2017), doi:10.1145/3025453.3025556.

[61] Brent Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, __ PHILOS. TECHNOL doi:10.1007/s13347-017-0253-7. *See also* Anton Vedder, *KDD: the Challenge to Individualism*, 1 ETHICS & INFO. TECH. 4, 275 (1999); Alessandro Mantelero, *From GroupPrivacy to Collective Privacy: Towards a New Dimension of Privacy and Data Protection in the Big Data Era*, *in* GROUP PRIVACY: NEW CHALLENGES OF DATA TECHNOLOGIES (Linnet Taylor et al. (eds.), Springer 2017).

[62] Vedder, *supra* note 61.

[63] Hildebrandt, *supra* note 59.

[64] *See* GDPR, recital 26.

profile is my personal data as it relates to me and makes me identifiable.[65] 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.[66]

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.[67] In European data protection law, special categories of data (known as "sensitive personal 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] Protected characteristics in other domains or jurisdictions often differ—in US privacy law, no general concept of sensitive data applies but there highly regulated statutory privacy regimes for health, financial and children's data.[69]

A relevant and well-publicised "war story" involves the American supermarket Target profiling its customers to find out which were likely to be pregnant so relevant offers could then be targeted at them. A magazine piece, now urban legend, claimed a teenage daughter was targeted with pregnancy related offers before her father with whom she lived knew about her condition.[70] In data protection law, if consent is used as the lawful ground for processing of special categories of data, that consent must be

---

[65] 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 *infra* Sections IV.B.1, IV.B.2.

[66] GDPR recital 26. See discussion of the status of recitals below, in addition to Klimas and Vaicuikaite, *infra* note 132. This approach to personal data has however been championed by the A29 WP for many years. *See*, e.g., ARTICLE 29 WORKING PARTY, OPINION 4/2007 ON THE CONCEPT OF PERSONAL DATA 01248/07/EN WP 136, at 13.

[67] *See* Emannuel Benoist, *Collecting Data for the Profiling of Web Users*, *in* PROFILING THE EUROPEAN CITIZEN 169, 177 (Mireille Hildebrandt & Serge Gutwirth eds., Springer 2008).

[68] GDPR, art. 9; *see also* Case C-101/01, Bodil Lindqvist, 2003 E.C.R. 1-12971.

[69] *See* HIPAA (Health Insurance Portability and Accountability Act of 1996), Pub. L. No. 104-191, 110 Stat. 1936; Sarbanes–Oxley Act of 2002 (also known as the "Public Company Accounting Reform and Investor Protection Act"), Pub. L. No. 107-204, 116 Stat. 745; COPPA (Children's Online Privacy Protection Act of 1998), 78 C.F.R. § 4008 (2013).

[70] Charles Duhigg, *How Companies Learn Your Secrets*, The NEW YORK TIMES MAGAZINE (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html; Kashmir Hill, *How Target Figured out a Teen Girl Was Pregnant Before Her Father Did*, FORBES (Feb. 16, 2012), https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.

"explicit."[71] But 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, unless 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? Should explicit consent be given where personal data is gathered from public social media posts using "legitimate grounds"[73] and transformed into data about political preferences which is "sensitive" data in the GDPR (Article 9(1))?[74] What

[71] This was a significant safeguard in an analogue world—consent would often be taken with especial care, such as in written form. Online, it is of limited protection, at best pushing controllers to more consumer-friendly consent collection such as opt-in rather than opt-out. [71] GDPR, art. 9(2)(a). Other grounds are available but notable for commercial data controllers is that processing cannot be justified by "legitimate interests" of the controller, nor because it was necessary for the performance of a contract between data subject and controller – the two prevalent grounds for processing used in the commercial world. For executive and judicial processing of special data, the main grounds are art. 9(2)(c) (emergency health situations), (f) (regarding legal claims or defences or judicial action) and (g) (substantial public interest).

[72] GDPR, Article 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 a recent collective statement of a number of EU Data Protection Authorities (DPAs) (*see Common Statement by the Contact Group of the Data Protection Authorities of The Netherlands, France, Spain, Hamburg and Belgium*, CNIL (May 16, 2017), https://www.cnil.fr/fr/node/23602) announcing a number of privacy breaches by Facebook, one issue is that company "uses sensitive personal data from users without their explicit consent. For example, data relating to sexual preferences were used to show targeted advertisements." (noted specifically by the *Autoriteit Persoonsgegevens*, the DPA of the Netherlands). It is not said if that data was created algorithmically or existed as a user input.

[73] GDPR, art. 6 (1)(f). Note that these interests may be 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 supra* note 71).

[74] This is likely part of the data collection and processing to produce political targeted advertisements pushed out via Facebook, allegedly undertaken in the UK and elsewhere by companies such as Cambridge Analytica. *See* Jamie Doward, *Watchdog to Launch Inquiry into Misuse of Data in Politics*, The GUARDIAN, Mar. 4, 2017, https://www.theguardian.com/technology/2017/mar/04/cambridge-analytics-data-brexit-trump. This area of the law is highly sensitive, given concerns about recent elections and referenda, and is under investigation by the UK's ICO. As noted within, the article itself is currently the subject of legal dispute as of mid-May 2017!

about when ordinary data collected via a wearable like a Fitbit is transformed into health data used to reassess insurance premiums?[75]

3. Opacity and Transparency

Users have long been disturbed by the idea that machines might make decisions for them, which they could not understand or countermand; 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 (for example credit scoring, job applications, speeding camera tickets) to the emergent, fictional worlds of technology (such as 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).[76] SARs empowered 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, Article 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 discuss later,[77] fused and extended into the so-called "right to be forgotten" in the GDPR, which succeeded the DPD in 2016. Although the US lacked an omnibus notion of data protection laws, similar rights emerged in relation to credit scoring in the Fair Credit Reporting Act 1970.[78]

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. In Europe, this led to a specific though rather under-used right in the DPD (Article 15) to stop a decision being made solely on the basis of automated processing.[79] Data subjects had a right to obtain human

---

[75] This 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 and the case of "Unfit Bits," see Olga Khazan, *How to Fake Your Workout*, THE ATLANTIC (Sep. 28, 2015), https://www.theatlantic.com/health/archive/2015/09/unfit-bits/407644/.

[76] *See* DPD, art. 12 (GDPR, art. 15).

[77] *See infra* Section IV.B.1.

[78] *See* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV.1,16 (2014).

[79] This has been interpreted to imply that European systems are more interested in the human dignity of data subjects than the US system. *See* Meg Leta Jones, *Right to a Human in the*

intervention (a "human in the loop"), in order to express their point of view but this right did not contain an express right to an explanation.[80] This right was updated in the GDPR to extend to a more general concept of profiling.[81] As Citron and Pasquale[82] 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."[83] 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, "[t]he "default" of governmental action should be transparency."[84] The opposite is more or less true of private action, where secrecy, including commercial or trade secrecy (and autonomy of business practices[85]) and

---

Loop: Political Constructions of Computer Automation & Personhood from Data Banks to Algorithms, 47 SOC. STUD. OF SCI. 216, 217 (2017).

[80] *But see supra* note 75.

[81] *See* GDPR, art. 4(4) ("'Profiling means 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 [...] performance at work, economic situation, health, personal preferences, interests, reliability or behaviour, location or movements."). Profiling may be achieved through means other than by ML. *See supra* Section I.A.1.

[82] Citron & Pasquale, *supra* note 78, at 16.

[83] LOUIS BRANDEIS, OTHER PEOPLE'S MONEY, AND HOW BANKERS USE IT 62 (National Home Library Foundation, 1933).

[84] Tal Zarsky, *Transparency in Data Mining: From Theory to Practice*, DISCRIMINATION AND PRIVACY IN THE INFO. SOC'Y 301, 310 (2013).

[85] Freedom to conduct a business is a fundamental right in the EU. *See* art. 15, Charter of Fundamental Rights of the European Union (CFEU) 2012/C 326/02.

protection of IP rights, are de facto the norm. Data protection law in fact seems quite odd when looked at from outside the informational privacy ghetto, as it is one of the few bodies of law that applies a general principle of transparency[86] even-handedly to private and public sector controllers, with more exceptions for the latter than the former in terms of policing[87] and national security.[88] But disclosure of personal data to its subject, from both public and private data controllers, is of course justified at root in Europe by the fundamental nature of privacy as a human right, sometimes extended to a separate right to DP.[89]

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[90] 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[91]—e.g. the introduction in the GDPR of mandatory security breach notification,[92] or the US EPA Toxics Release Inventory.[93] Disclosures may

---

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

[87] *See* GDPR art. 2(2)(d). *But see supra* note 53.

[88] Despite these exceptions, European countries have traditionally been more transparent than the US in the development of ML systems used for judicial/penal decision support. ML systems in Europe are often developed in-house, rather than privately procured and subject to proprietary secrecy. *See* A COMPENDIUM OF RESEARCH AND ANALYSIS ON THE OFFENDER ASSESSMENT SYSTEM (OASYS) (Robin Moore ed., Ministry of Justice Analytical Series, 2015); Nikolaj Tollenaar et al., *StatRec —Performance, Validation and Preservability of a Static Risk Prediction Instrument*, 129 BULL. SOC. METHODOLOGY 25 (2016) (discussing published UK and Dutch predictive recidivism models).

[89] *See* ECHR, art. 8; CFEU arts. 7 and 8.

[90] An ancillary question relates to 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.

[91] *See* Zarsky, *supra* note 84, at 311.

[92] GDPR arts. 33 and 34.

[93] Madhu Khanna et al., *Toxics Release Information: A Policy Tool for Environmental Protection*, 36 J. ENV'T ECON. & MGMT. 243 (1998); Archon Fung and Dara O'Rourke, *Reinventing Environmental Regulation from the Grassroots Up: Explaining and Expanding the Success of the Toxics Release Inventory*, 25 ENVIRO. MGMT. 115, 116 (2000).

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).

Despite the sometimes almost unthinking association of transparency and accountability, the two are not synonymous.[94] 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.[95] 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.[96] 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.[97] 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.[98]

Notwithstanding these doubts, and turning to ML systems, transparency rights remain 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.[99] Similar ideas pervade the many calls for reinstating due process in algorithmic decision making,[100] for respecting the right to a "human in the loop" as an aspect of human dignity[101] and for introducing "information accountability" in the form of "policy awareness"

---

[94] *See*, e.g., PASQUALE, *supra* note 3, at 212 (rejecting the idea that transparency has created any real effects on or accountability of the financial sector post-crash and recession).

[95] Mark Bovens, *Analysing and Assessing Accountability: A Conceptual Framework*, 13 EUROPEAN LAW JOURNAL 447, 450 (2007).

[96] *See* Helen Margetts, *Transparency and Digital Government*, *in* TRANSPARENCY: THE KEY TO BETTER GOVERNANCE?, 197, 200 (C. Hood & D. A. Heald, eds., Oxford University Press 2006).

[97] *See* Onora O'Neill, *Transparency and the Ethics of Communication*, *in* Hood & Heald, *supra* note 96, at 75–90.

[98] David A. Heald, *Varieties of Transparency*, *in* Hood & Heald, *supra* note 96, at 30.

[99] Hildebrandt, *supra* note 9, at 66.

[100] *See*, e.g. Crawford & Schultz, *supra* note 12, at 95; Citron, *supra* note 14, at 1254.

[101] *See* Jones, *supra* note 79, at 217.

which will "make bad acts visible to all concerned"; [102] or varied similar ideas of "algorithmic accountability."[103]

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 I.B.2 above, or another recent case of outrage affecting a group, when Google wrongly categorized black people in its Photos app as gorillas.[104]

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*[105] 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 González, 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 González'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.[106]

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,

---

[102] Daniel J. Weitzner et al., *Information Accountability*, 51 COMMUNICATIONS OF THE ACM 6 82,84 (2008).

[103] Maayan Perel & Nova Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STAN. TECH. L. REV. 473, 478 (2016); NICHOLAS DIAKOPOULOS, ALGORITHMIC ACCOUNTABILITY REPORTING: ON THE INVESTIGATION OF BLACK BOXES (Tow Centre for Digital Journalism, 2013). For a rejection of rights of transparency as the answer to algorithmic accountability, see Joshua Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 638 (2017).

[104] *See* Sophie Curtis, *Google Photos Labels Black People as 'Gorillas,'* THE TELEGRAPH (May 4, 2017, 11:20 AM), http://www.telegraph.co.uk/technology/google/11710136/Google-Photos-assigns-gorilla-tag-to-photos-of-black-people.html.

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

[106] For further detail, see Kohl, *supra* note 27, at 192; Jones, *supra* note 79, at 216.

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 in the face of institutionalised power or money,[107] 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.[108]

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—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.[109] 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). 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 IV.A, and to the question of what types of explanation in what circumstances may actually be useful, and to whom (section III). First however we consider the recent legal debate on whether a "right to an explanation" of algorithmic decisions does indeed exist in EU data protection law.

---

[107] *See* PASQUALE, *supra* note 3, at 212.

[108] *See* Bruce Schneier, *The Myth of the 'Transparent Society*,' WIRED, (Mar. 6, 2008, 12:00 PM), https://www.wired.com/2008/03/securitymatters-0306/.

[109] *See* Kroll et al., *supra* note 103, at 654. For a further discussion on "gaming," see *infra* Section III.C.1.

II. SEEKING A RIGHT TO AN EXPLANATION IN EUROPEAN DATA PROTECTION LAW

In 2016, to the surprise of some EU data protection lawyers, and to considerable global attention, Goodman and Flaxman asserted in a short paper that the GDPR contained a "right to an explanation" of algorithmic decision making.[110] As Wachter et al. have comprehensively pointed out, the truth is not quite that simple.[111] 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 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.

A. GDPR, Article 22: Automated Individual Decision-Making

Our starting point is Article 15 of the now-replaced DPD, which was originally aimed at protecting users from unsupervised automated decision making. This rather odd provision[112] was mainly overlooked by lawyers and commentators by reason of non-significance and few saw the potential it had towards algorithmic opacity. It is clear that Article 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, Article 22 which provides:

> [T]he 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*.[113]

Importantly, Article 22, like Article 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

---

[110] Goodman & Flaxman, *supra* note 6, at 2.

[111] Wachter et al., *supra* note 11, at 72.

[112] *See* Izak Mendoza & Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, *in* EU INTERNET LAW: REGULATION AND ENFORCEMENT 2 (Tatiani Synodinou et al. eds., Springer, forthcoming 2017) (describing 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 Harbour); Lee A. Bygrave, *Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling*, 17 COMPUTER L. & SECURITY REP., 17 (2001). *See also* Hildebrandt, *supra* note 9, at 65.

[113] GDPR, art. 22(1) (emphasis added).

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).[114]

Even after this there are a number of hurdles to get over. First, Article 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[115]—and indeed in a great deal of these cases—for example involving victims of crimes or accidents—full automation seems inappropriate or far off. Article 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 *ProPublica* in 2016.[116] 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 data protection authorities are currently worrying at this point.[117] Human involvement can also be rendered nominal by "automation bias," a psychological phenomenon where humans either over or under-rely on decision support systems.[118] The Dutch Scientific Council for Government Policy in early 2016 specifically recommended that more attention be paid to "semi-automated decision-making" in the GDPR, in relation to profiling.[119]

---

[114] Mendoza & Bygrave, *supra* note 112, at 13 (arguing that DPD arts. 13-15 and 22 suggest that there is a right to be informed that automated decision is being made).

[115] CABINET OFFICE, DATA SCIENCE ETHICAL FRAMEWORK (HM Government, May 2016), https://www.gov.uk/government/publications/data-science-ethical-framework. This specifically advises human oversight in non-trivial problems, even where autonomous systems are possible.

[116] *See* Julia Angwin et al., *Machine Bias*, PROPUBLICA, (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Lack of effective supervision has led to bans in some states. *See* Mitch Smith, *In Wisconsin, a Backlash Against Using Data to Foretell Defendant's Futures*, THE NEW YORK TIMES (Jun. 22, 2016).

[117] The UK ICO at the time of writing recently concluded consulting on this point: see ICO, FEEDBACK REQUEST – PROFILING AND AUTOMATED DECISION-MAKING [v 1.0 ,2017/04/06] (2017) at 20, https://ico.org.uk/media/2013894/ico-feedback-request-profiling-and-automated-decision-making.pdf (asking "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?").

[118] *See*, e.g., Linda J Skitka et al., *Accountability and Automation Bias*, 52 INT'L J. HUMAN-COMPUTER STUD. 4, 4 (2000).

[119] WRR, *supra* note 4 at 142.

Second, Article 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 Article 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.[120]

Second, even if we posit that algorithmic "output" and human "decision" may be conflated in Article 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.

1. Article 22 in the Context of "Algorithmic War Stories"

Consider two well-known and influential early examples of "algorithms gone bad." In 2013, Latanya Sweeney, a security researcher at Harvard University, investigated the delivery of targeted adverts by Google AdSense using a sample of racially associated names.[121] She found statistically significant discrimination in advert 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 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

---

[120] HEATHER DOUGLAS, SCIENCE, POLICY AND THE VALUE-FREE IDEAL (University of Pittsburgh Press 2009). We return to the issue of "decisions" and ML below.

[121] Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 COMMUNICATIONS OF THE ACM 44, at 44 (2013).

beyond the scope of her research to know what was happening in the "inner workings of Google AdSense,"[122] 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 for the word "Jew." Google (in stark contrast to its more recent attitudes)[123] 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. It was stated that "[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."[124] 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 some 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 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)[125] but did the

---

[122] Sweeney contemplates advertisers providing 'black sounding names' themselves for targeting, or auto-adjustment of Google's algorithm based on distribution of 'hits.'

[123] Google has rethought its approach to such cases, especially after unfavourable press reports. *See* Samuel Gibbs, *Google Alters Search Autocomplete to Remove 'Are Jews Evil' Suggestion*, THE GUARDIAN (Dec. 5, 2016, 10:00 AM), https://www.theguardian.com/technology/2016/dec/05/google-alters-search-autocomplete-remove-are-jews-evil-suggestion. Now, a "quality rating" downgrades pages rather than removes them: interesting considering issues raised later regarding altering ML models using the "right to be forgotten." *See Google Launches New Effort to Flag Upsetting or Offensive Content in Search*, SEARCHENGINE WATCH (Mar. 14, 2017, 1:00 PM), http://searchengineland.com/google-flag-upsetting-offensive-content-271119.

[124] *See Google in Controversy Over Top-Ranking for Anti-Jewish Site*, SEARCHENGINE WATCH (Apr. 24, 2004), https://searchenginewatch.com/sew/news/2065217/google-in-controversy-over-top-ranking-for-anti-jewish-site.

[125] See Mendoza & Bygrave, *supra* note 112 at 10 (suggesting that 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 (see *supra* note 81) includes in its definition in art. 4(4) the evaluation of "personal aspects" of a person including their

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 data protection 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[126] 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 the GDPR discussing 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 discusses,[127] 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, and quite likely given the evidence quoted above, the searcher might not themselves be of the class affected.[128]

1. Re-enter the "Right to an Explanation"?

The ban on automated decision-making in Article 22(1) 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

---

"personal preferences." This sounds a lot like targeted advertising, though see below on whether that decision would be "significant.").

[126] Mendoza & Bygrave, *supra* note 112 at 12.

[127] *See* Kohl, *supra* note 27.

[128] This compares to European cases of algorithmic defamation, where Google autocomplete suggested particular names was falsely associated with reputation-harming terms. Yet there the causal connection between algorithm and data subject harm seems more obvious. See discussion in Jones, *supra* note 79.

entering a contract, authorized by law or, most crucially, based on explicit consent (Article 22(2)).[129] In these cases Article 22 does not prevent automated decision-making 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."[130]

Recital 71 then mentions all of the above safeguards but *also* adds a further, explicit "right to an explanation." Is this therefore another route to a "right to an explanation" in Article 22? This seems paradoxical. Article 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 Article 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.[131] So it may be possible to read a "right to an explanation" into these cases as well, and indeed given that we are not pointed to Article 22(3) with its contradictory list of safeguards, this might indicate that we could rely more heavily on the full extent of the Recital 71 list.

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

---

[129] 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 in Section II.A.1.

[130] GDPR, art. 22(3) .

[131] Note also that paragraph 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.

to be interpretative of the main text rather than creating free standing extra obligations.[132] 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 et al. argue that the history of Article 22 in the preliminary drafts indicates a deliberate omission of a "right to an explanation" from the main text of Article 22, not an accidental or ambiguous omission[133] 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.[134]

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 a "special" category of data[135] (in this case, race)? If so, it worth noting that Article 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 Article 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,"[136] 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.

---

[132] *See* Tadas Klimas & Jurate Vaiciukaite, *The Law of Recitals in European Community Legislation*, 15 ILSA J. OF INT'L AND COMP. LAW 1 61, 92 (2008). Recitals in EU law can be perplexing and is at core politicised. They lack "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."

[133] Wachter et al., *supra* note 11, at 9–11.

[134] 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.

[135] GDPR, art. 9.

[136] *See supra*, Section I.B.2 and Mittelstadt et al., *supra* note 25.

Finally looking at the primary remedy Article 22 provides, a "human in the loop," how valuable is it truly? 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 through their outputs using both real and hypothetical queries. In such circumstances, an intuitive response is likely to be correct and this is something machines do badly. But typically,[137] the types of ML algorithms that are highly multidimensional make "decisions" with which humans will struggle as much as, if 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"[138]—humans are as likely to make spur of the moment decisions as reasoned ones. For these reasons, Kamarinou et al. 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."[139] In such an event they recommend it might be better, not to have the "appeal" from machine to human which Article 22 implies, but to have the reverse.[140]

B. GDPR, Article 15: A Way Forward?

A right which might be more usefully employed to get a transparent explanation of a ML system is *not* part of Article 22 but rather in Article 15, a provision not specially related to automated decision making. 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 Article 22(1) and (4)," access to "meaningful information

---

[137] *See infra* Section III.B.

[138] The trolley problem is an ethical thought experiment often applied to autonomous vehicles. Imagine a runaway train careering towards a group of people unable to avoid it. A bystander could flip the lever to send to train to fewer, equally helpless people—but in doing so would determine who lives and dies. What should they do? This is interesting, but we note here that when faced with stressful situations with many factors in the real world, the challenge is both psychological and ethical. *See generally* PHILIPPA FOOT, THE PROBLEM OF ABORTION AND THE DOCTRINE OF THE DOUBLE EFFECT IN VIRTUES AND VICES (Basil Blackwell, 1978).

[139] *See* Calders & Žliobaitė, *supra* note 33; *see also* FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACHINE LEARNING, WWW.FATML.ORG.

[140] Dimitra Kamarinou et al., *Machine Learning with Personal Data*, QUEEN MARY SCHOOL OF LAW LEGAL STUDIES RESEARCH PAPER NO. 247/2016, https://ssrn.com/abstract=2865811.

about the logic involved, as well as the significance and the envisaged consequences of such processing."[141]

As noted above, this version of the "right to an explanation" is not new, but has existed in the DPD since 1995.[142] While this may seem a more straightforward way to get to such a right than via Article 22, it has its own problems.

A first issue is timing. Wachter et al. suggest that Article 15 "subject access rights" should be contrasted with the "information rights" of the GDPR. Articles 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 (Article 13), or from another party (Article 14). This information is reminiscent of that required to inform consumers before entering, say, distance selling contracts. In contrast, Article 15 refers to rights of "access" to data held by a data controller. This seems to imply data has been collected and processing has begun or taken place. From this Wachter et al. argue that the information rights under Articles 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."[143]

In the case of Article 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, i.e. "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.[144] However,

---

[141] GDPR, art. 15(1)(h)).

[142] DPD, art. 12(a).

[143] Wachter et al., *supra* note 11 at 78.

[144] Wachter et al., *supra* note 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 technologically and restricting art. 15(h) to *ex ante* explanations seems against the purpose indicated by the recital.

whether such an explanation can be "meaningful" in substance is another story as will be discussed below in section III.

Secondly, Article 15(h) has a carve out in the 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."[145] 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."[146] Several other factors also give us hope for overcoming this significant barrier. First, as we discuss below,[147] some explanation systems which build a model-of-a-model need not necessarily infringe IP rights. Secondly, the EU Trade Secrets Directive, the provisions of which must be adopted by member states by June 2018, specifically notes in Recital 35 that the directive "should not affect the rights and obligations laid down" in the DPD,[148] going on to specifically name the right of access—although we caution, as previously, that the status of recitals is murky at best.

Finally, it has been suggested that the text of Article 15(h)'s "right to meaningful information" is just as restricted as any remedy derived from Article 22, given it refers to "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4)."[149] We disagree. It seems quite possible to view the reference to Article 22 as merely instantiating one form of automated decision making, not excluding others, which are e.g., achieved partially but not solely by automation. Furthermore, Article 15(h) says the right to "meaningful information" refers "at least" to these types of automated decision making. This seems to logically imply a wider scope. Given the dearth of European case law on the matter, it is hard to say this was a settled matter in the DPD.

Next, drawing on literatures from computer science and elsewhere, we turn to some of the practical opportunities and challenges implementing any similar right to "meaningful information about the logic involved" to that Article 15 potentially provides.

---

[145] GDPR, recital 63.

[146] GDPR, recital 63 [emphasis added].

[147] *See infra* Section III.C.2.

[148] Recital 35, Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, 2016 O.J. (L 157) 1.

[149] GDPR, art. 15(2)(h).

III. IMPLEMENTING A RIGHT TO AN EXPLANATION

Explanations and the demand for them in machine learning systems are not new, although emphasis has more recently turned to explanations for the decision-subject, rather than the user of the decision-support tool. Computer scientists have been long concerned that neural networks "afford an end user little or no insight into either the process by which they have arrived at a given result,"[150] and that "people should be able to scrutinise their user model and to determine what is being personalised and how."[151] ML explanations are not just good for decision subjects 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 at least 10% of the time on new unseen data. In the real world, this is usually worsened by the changing nature of tasks, the world and the phenomena ML systems are often expected to accurately model.[152] 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 feedback can help to both ensure system performance and support varying notions of quality.[153]

Here, our focus is however mainly on decision subjects (data subjects, in data protection parlance), who, as discussed above, might display an array of overlapping reasons for wanting an explanation. Below we discuss what types of explanation are possible (and what they might substantially provide to decision-subjects), and consider in what situations and for who an explanation of an ML system may be difficult, limited or impossible. Finally, 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 (model-of-model) rather than decompositional (i.e., explain by taking apart) explanations as a way to avoid perceived IP and trade secrets restraints on ML algorithms.

---

[150] Alan B. Tickle et al., *The Truth Will Come to Light: Directions and Challenges in Extracting the Knowledge Embedded Within Trained Artificial Neural Networks*, 9 IEEE TRANSACTIONS ON NEURAL NETWORKS 1057, 1057 (1998). *See also* Zelznikow and Stranieri, *supra* note 19.

[151] Kay, *supra* note 8 at 18.

[152] Joao Gama et al., *A Survey on Concept Drift Adaptation*, 1 ACM COMPUTING SURVEYS 1 (2013).

[153] 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* Nava Tintarev & Judith Masthoff, *Explaining Recommendations: Design and Evaluation*, *in* RECOMMENDER SYSTEMS HANDBOOK (Francesco Ricci et al. eds., Springer 2015).

A. Types of Explanation: Model-Centric Versus Subject-Centric Explanations

We can broadly discern two categories of explanations that might be feasible. The first centers on the model itself. In this, we include logics that might be generally applicable to many decision subjects as well as motivations, context, variables, and performance behind the model and the decision process. The second concerns particular predictions of interest, which may or may not lead to 'decisions.' Here, even in complex models, some information can often be provided about 'why' a particular prediction was made—although this information has its limits.

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 some aspects of 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 disciplines than clarity. We extend the focus from the computational behaviour of a model to consider the motivations and context behind this model in action. As Singh et al. note, machine learning is part of a process, and the dimensions of 'explanation' that relate to the broader context are important and should not be ignored.[154] MCEs provide one set of information to everyone, but there are limitations on how detailed, practical and relevant—and thus, how "meaningful"[155]—such an explanation can be alone.

Information provided with an MCE approach 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, the provenance of such data, 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;

---

[154] Jatinder Singh et al., *Responsibility & Machine Learning: Part of a Process* (Oct. 27, 2016), https://ssrn.com/abstract=2860048

[155] GDPR art. 15(h).

- *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.[156] Information about the logics, which might be provided in the form of cryptographic assurances,[157] 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 much "meaningful information" for individual data subjects, we are probably going to need to look beyond MCEs alone.

2. Subject-Centric Explanations (SCEs)

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 real or fictitious or exploratory. As a result (and somewhat contrary to the approach of Wachter et al.), they are theoretically possible to give before or after a "decision" as discussed in the sense of data protection, if access to the model is provided. 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 rapidly become known as 'black boxes' in the media. Despite this, only considering certain relevant parts of them at any one time might allow for more useful explanations.

To better understand this, we introduce a concept from computer science: the "curse of dimensionality." 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 variables you have in your data. As you increase the dimensions (i.e., the number of variables) the number of ways that all potential values of them can be combined grows

---

[156] Kroll et al., *supra* note 103.

[157] Only limited prior work has demonstrated the feasibility of verifying certain types of ML systems with cryptographic methods for any purpose. *See* George Danezis et al., *Private Client-Side Profiling with Random Forests and Hidden Markov Models*, PRIVACY ENHANCING TECHNOLOGIES (Springer 2012).

exponentially. It is this dynamic which makes the data especially complex to comprehend. Layered onto this, models which mix arbitrary combinations of variables in multiple different ways in parallel, interdependent ways, means that the complexity of the data by its extent is compounded by the complexity of the procedures used to analyse it. Explaining everything in one go, as MCEs try to, quickly becomes unwieldy.

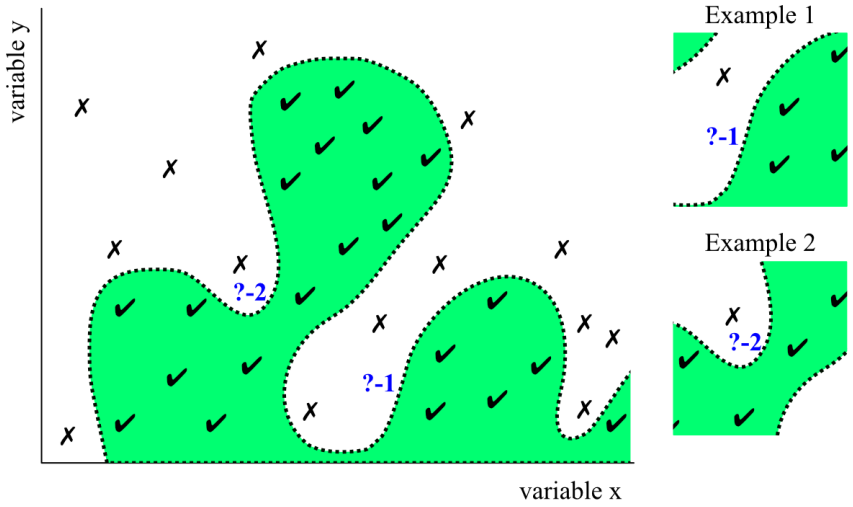

Figure 1 Subject-centred explanations in practice. The dotted line represents a machine learned decision boundary, where the ticks are classified one way, and the crosses another. This is a highly simplified, illustrative diagram that consciously omits uncertainty, or points misclassified during training in order to illustrate a broader point.

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 taking only a slice of the system, it can become considerably more interpretable. Take the simple (and simplified) example above in Figure 1. The dotted line represents a machine learned decision boundary over two input variables. Using this boundary, we can classify points into ticks and crosses, which might, for example, be acceptance of an application for a financial product. Giving out the whole model in a useful form will be challenging, especially since usually there are more than the two or three dimensions we can grasp visually with relative ease. Yet zooming in to a particular point: such as why were ?-1 and ?-2 rejected, might be easier to explain. ?-1 is easier in many ways—it seems that they have to just increase variable x to switch the decision, and this is helped a little if they also increase variable y at the same time, and hindered if they don't. Yet ?-2 is slightly trickier. They could increase or reduce variable *y*, or increase or reduce variable *x*. This

seems pretty unsatisfactory—the individual is likely to wonder (in an MCE fashion) why the model was shaped to have these odd 'pockets' they could be stuck in, anyway. What is clear is that models can be explained in terms of one or two things an individual could change. In other cases, they can only be framed in terms of many variables, which change in inconsistent and non-linear ways.

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?[158] (Where do I have to move in Figure 1 to be classified differently?)
- *Case-based* subject-centric explanations: which data records used to train this model are most similar to mine?[159] (Who are the ticks and crosses nearest to me?)
- *Demographic-based* subject-centric explanations: what are the characteristics of individuals who received similar treatment to me?[160] (Who, more broadly, was similarly classified?)
- *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? (How many ticks and crosses nearer me were misclassified during training? Am I a difficult case?)

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, SCEs are considerably more linked

---

[158] Wojciech Samek et al., *Evaluating the Visualization of What a Deep Neural Network Has Learned*, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS 1, 7 (2016); Marco Tulio Ribeiro et al. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1135–44 (2016).

[159] DONAL DOYLE ET AL., A REVIEW OF EXPLANATION AND EXPLANATION IN CASE-BASED REASONING (Department of Computer Science, Trinity College, Dublin, 2003).

[160] Liliana Adrissono et al., *Intrigue: Personalized Recommendation of Tourist Attractions for Desktop and Handheld Devices*, 17 APPLIED ARTIFICIAL INTELLIGENCE 687, 696 (2003); Tintarev & Masthoff, *supra* note 153.

to communities of interface design and human-computer interaction than to communities concerned with engineering issues, such as those building the cryptographic assurances that a system did what it was expected to, discussed above.

B. Barriers to Explanations

MCEs and SCEs are far from perfect solutions, let alone easy ones to operationalize in many cases. Here, we present an in-exhaustive overview of two issues in this field: one that relates to the domain of decision-making, and one that relates to the interaction between who needs an explanation and how solid that explanation is likely to be for them.

1. 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 often have only a few input variables that are combined in relatively straightforward ways, such as increasing or decreasing relationships. 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, colour channels in pixels are treated as individual inputs. While we struggle to read a table full of numbers at a glance, when an image is meaningful, the brain can process thousands of pixels at once 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.

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, claims to have over 100,000 variables held on every user that feed into ML modelling.[161] Many of these will not be clear variables like "age," but more abstract ways you interact with the webpage, such as how long you take to click, the time you spend reading, or even text

[161] KUN LIU, DEVELOPING WEB-SCALE ML AT LINKEDIN—FROM SOUP TO NUTS, PRESENTED AT THE NIPS SOFTWARE ENGINEERING FOR MACHINE LEARNING (Dec. 13, 2014).

you type in a text box but later delete without posting.[162] 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? Do we even possess the mental vocabulary of categories and concepts to grasp the important aspects in the data?

2. 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.[163] 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.

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 some outcomes are more difficult to predict than others for ML systems, given their relative individual complexity.[164] 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.[165] ML practitioners expect this kind of dynamic within the data they use—the common technique of *boosting* relies on learning more from cases previously misclassified.

---

[162] Sauvik Das & Adam Kramer, *Self-Censorship on Facebook*, PROCEEDINGS OF THE SEVENTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA 120, 120 (2013).

[163] JAMES C. SCOTT, SEEING LIKE A STATE (Yale University Press 1998); SALLY ENGLE MERRY, THE SEDUCTIONS OF QUANTIFICATION (University of Chicago Press 2016).

[164] Gary M. Weiss, *Mining with Rare Cases*, *in* DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK, 747–57 (Oded Maimon & Lior Rokach eds., Springer 2009).

[165] Taeho Jo & Nathalie Japkowicz, *Class Imbalances Versus Small Disjuncts*, 6 ACM SIGKDD EXPLORATIONS NEWSLETTER 40 (2004).

Why might this challenge meaningful explanations? SCEs practically focus on taking the model you have, selecting a certain region 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?[166] 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.

B. Opportunities for Better Explanations

Better explanations are possible, although it may involve rethinking how we make and use them. We highlight two promising avenues. The first centers on allowing users to interactively explore algorithmic systems, which can enable individuals to develop good and trustworthy mental models of the systems they use and are subject to. The second rests on another insight—you do not have to have access to the innards of a model to attempt to explain it, but can instead wrap a simpler model around it and use that as the explanation facility.

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? The best explanations of complex systems seem to be "exploratory," using subject-centric inputs. Experimental tests have found that interfaces that provided SCEs rather than MCEs appeared far more effective at helping users complete tasks, even where the experiment was constructed so that unusually, the same amount of information was provided by both.[167] For users, it seems that when done well, SCEs are more appealing, convenient and compelling. Here, explanation facilities might

---

[166] Tickle et al., *supra* note 150, at 1058.

[167] Dianne C. Berry & Donald E. Broadbent, *Explanation and Verbalization in a Computer-Assisted Search Task*, Q.J. EXPERIMENTAL PSYCHOL. SEC. A 585, 596 (1987).

allow decision subjects to build more effective and relevant mental models, build justified trust and work better with algorithmic systems.[168]

Drawing on the literature on human–computer interaction (HCI), SCEs can be thought of as "seams" in the design of a ML system.[169] Seam*less* 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.[170] Yet "seam*ful*" algorithmic systems, 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.[171] 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.[172] "Seamful" systems might help restore what Mireille Hildebrandt terms "double contingency"—the mutual ability to anticipate, or "counter-profile" how an agent is being "read" by another, so she can change her own actions in response.[173]

Some SCEs already let individuals hypothetically explore the logics happening around their own data points. Tools already exist to let you "try out" what your credit score might be online, including through filling a questionnaire, or signing into these using your data profile (for example, by authorising a 'soft' check on your credit file, or potentially one day, by

---

[168] 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* Maayan Perel (Filmar) & Niva Elkin-Koren, *Black Box Tinkering: Beyond Transparency in Algorithmic Enforcement*, FLORIDA LAW REVIEW (forthcoming 2017). Some authors suggest facilities help acceptance of decisions. *See* Henriette Cramer et al., *The Effects of Transparency on Trust in and Acceptance of a Content-Based Art Recommender*, 18 USER MODELING AND USER-ADAPTED INTERACTION 455 (2008); while others consider trust building overall. *See contra* A Busone et al., *The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems, in* ICHI '15, 160–169 (2015).

[169] Matthew Chalmers & Ian McColl, *Seamful and Seamless Design in Ubiquitous Computing, in* WORKSHOP AT THE CROSSROADS: THE INTERACTION OF HCI AND SYSTEMS ISSUES IN UBICOMP (2003).

[170] *See* BRUNO LATOUR, PANDORA'S HOPE: ESSAYS ON THE REALITY OF SCIENCE STUDIES 304 (Harvard University Press 1999).

[171] Kevin Hamilton et al., *A Path to Understanding the Effects of Algorithm Awareness*, CHI '14 633 (2014).

[172] Motahhare Eslami et al., *First I "Like" It, Then I Hide It: Folk Theories of Social Feeds*, CHI' 16 (2016).

[173] Hildebrandt, *supra* note 59.

giving access to your social media API). More advanced approaches might let a data subject see how the system might make decisions concerning other users, thus taking the user out of their own "filter bubbles."

Unfortunately, it will be easier to build SCEs that let you explore the logics around yourself rather than around others, because simulating the inputs of others convincingly is hard. Some 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.[174] 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."[175] Individuals everywhere will find it harder to "fake" a new persona without changing their lifestyle, 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," but one that we believe to be overstated, is that such querying might be used to "game the system." In many cases, this is more unlikely or less consequential than often assumed. In private sector systems such as targeted advertising deriving from social media information, users anecdotally do often try to "game" or self-optimize 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 is patchy at best.[176] If this is true then it seems unlikely that prisoners will change their characteristics just 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

---

[174] Aniko Hannak et al., *Measuring Price Discrimination and Steering on E-Commerce Web Sites*, *in* PROCEEDINGS OF THE 2014 CONFERENCE ON INTERNET MEASUREMENT CONFERENCE (2014).

[175] Sam Jones, *The Spy Who Liked Me: Britain's Changing Secret Service*, FINANCIAL TIMES (Sep. 29, 2016), https://www.ft.com/content/b239dc22-855c-11e6-a29c-6e7d9515ad15.

[176] *See*, e.g., Daniel S. Nagin, *Deterrence in the Twenty-First Century*, 42 CRIME AND JUSTICE 199, 201 (2013).

these courses, or behaving well, would not be useful or transformative in and of itself.

For important decisions, we might question if a system that "works" but can so easily be gamed is not a system which is already too fragilely reliant on obfuscation to achieve its policy goals. If all that is preventing misuse is 'keeping the lid' on the logic, then it is not a great stretch to assume some individuals or organizations, likely assisted by money and power, have already pried the lid open further than others. In particular, researchers have demonstrated that with significant financial resources, there is a feasibility of "model stealing" i.e. reverse engineering models such as those Google and Amazon offer as-a-service via APIs.[177] It might also be questioned if a system is only based on "entrenched" factors that are costly or impossible to change or hide (e.g. race), is this really a fair system?[178]

2. Explaining Black Boxes Without Opening Them

As we have seen, the way that ML systems optimize for performance usually comes at the expense of internal interpretability. 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 to a non-expert. Since then researchers have generally seen explanation as an entirely separate optimization challenge—*decoupling* algorithmic reasoning from algorithmic explanation.[179]

There are two main styles of decoupled algorithmic explanations.[180] They differ from MCEs and SCES, which concern the *focus* of explanation, to consider the way in which that explanation (MCE or SCE) is decided. The first type is the *decompositional* explanation, which attempts 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. This requires access to the bulk of the model structure itself. Some types of machine learning, like regression, are decomposable by design, and commonly used to explain phenomena in social sciences. Others can be made more decomposable with relatively little effort—random forest models can be trained to also produce "variable

---

[177] Florian Tramèr et al., *Stealing Machine Learning Models via Prediction APIs*, *in* USENIX SECURITY SYMPOSIUM, AUSTIN, TX, USA, AUGUST 11 2016 (2016).

[178] On the danger of inequalities and gaming, *see generally* Jane Bambauer & Tal Zarsky, *The Algorithm Game* [draft manuscript, on file with authors].

[179] Michael R. Wick & William B. Thompson, *Reconstructive Expert System Explanation*, 54 ARTIFICIAL INTELLIGENCE 33, 35 (1992). This corresponds to the "naïve" approach Kroll et al. talk about of merely dumping source code, inputs and outputs (*supra* note 103).

[180] Combinations are also possible. See Tickle et al., *supra* note 150.

importance scores" alongside the model.[181] Decomposing others, particularly when the innards are complex as they are in deep learning systems, requires extra methods—a hot research area.[182]

On the other end of the spectrum, *pedagogical* systems, 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.[183] Pedagogical systems have the great advantage of demanding a much lower level of model access and are thus less likely to run into the IP or trade secrecy barriers embedded in Article 15(h) (see section II.B 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.[184]

IV. SEEKING BETTER REMEDIES THAN EXPLANATIONS WITH THE GDPR

A. Avoiding a "Transparency Fallacy"

Above, we have seen a large number of difficulties, as well as some opportunities, around providing meaningful explanations in ML systems. This leads us in this section to stop and consider if "the game is worth the candle": if meaningful information about the logic of ML is so hard to provide, how sure are we that explanations are actually an effective remedy and if so, to achieve what? In section I.B.3, we already began to explore a little sceptically the notion of transparency as a remedy, drawing on

---

[181] Durham Police's recidivism system publicises these measures, for example. *See* Sheena Urwin, Presentation at TRILCon '17: Algorithms in Durham Constabulary custody suites—How accurate is accurate? (May 3, 2017).

[182] *See*, e.g., George Montavon et al., *Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition*, 65 PATTERN RECOGNITION 211 (2017).

[183] For examples of pedagogical systems, *see* Ribeiro et al., *supra* note 158; Anupam Datta et al., *Algorithmic Transparency via Quantitative Input Influence, in* TRANSPARENT DATA MINING FOR BIG AND SMALL DATA (Tania Cerquitelli et al. eds. Springer 2017).

[184] Cynthia Dwork, *Differential Privacy: A Survey of Results, in* INTERNATIONAL CONFERENCE ON THEORY AND APPLICATIONS OF MODELS OF COMPUTATION (Springer 2008).

historical experience from the financial crash and from freedom of information laws. A useful warning can also be taken about so-called remedies or safeguards that may simply not work by considering the history of consent in information privacy.

Privacy scholars are already very 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."[185] Why is this? 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, understand or truly consider 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.[186] Consent as an online institution in fact arguably no longer provides any semblance of informational self- determination but merely legitimises the extraction of personal data from unwitting data subjects. As behavioural economics have taught us, 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, consent is manipulated by those, such as employers or insurers, who can exert pressures that render "free" consent imaginary. To posit in a rather utopian way that consent can be given once to a data controller in a free and informed way, will require constant vigilance 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 via consent in the world of online dependence

---

[185] *See* discussion and references *supra* Section I.B.2 and *supra* note 54.

[186] 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 (Article 7(1)), that terms relating to consent in user contracts must be distinguishable from other matters, and written in "clear and plain language" (Article 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 (Article 7(4)). It is submitted however that these changes are not major, and that much will depend on the willingness of EU member state data protection 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 (*supra* note 72) is certainly an interesting first shot over the bows.

and "bastard data."[187] As a result, it is now beyond trite to talk about a "notice and choice fallacy."[188]

Relying on individual rights to explanation as the means for users to take control of ML systems risks creating a similar "transparency fallacy."[189] 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 III 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 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.

Ananny and Crawford recount the numerous ways in which transparency "as a method to see, understand and govern complex systems"[190]—both in the past, and now in the time of algorithmic ML systems—is not only limited but at times misleading and unhelpful. Inter alia, they note that transparency can support "neoliberal models of agency,"[191] placing a tremendous burden on individuals both to seek out information about a system, interpret it, and determine its significance, only then to find out they have little power to change things anyway, being "disconnected from power."[192] Liberal democracy in the past has taught us "the feeling that seeing something may lead to control over it"[193] but in fact in its search for a technical solution, dependence on transparency may occlude the true problems which rest in societal power relations and institutions as much as the software tools employed.

B. Beyond Explanation Rights: Making Fuller Use of the GDPR to Better Control Algorithms

We now consider if in the stampede to find a legally enforceable right to an explanation, other new user rights and tools in the GDPR have been given undeservedly little attention. We first explore two main rights: the right to erasure (colloquially often called "right to be forgotten") in Article 17, and the right to data portability in Article 20, before turning to look at the proposed supporting environment for enforcement the GDPR

---

[187] *See* McNamee, *supra* note 55.

[188] *See* full discussion in Edwards, *supra* note 59.

[189] *See* Heald, *supra* note 98 (discussing the notion of a "transparency illusion").

[190] *See* Mike Annany & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, NEW MEDIA & SOCIETY 1, 5.

[191] *Id.* at 7.

[192] *Id.* at 6.

[193] *Id.* at 3.

establishes using a varied range of instruments, such as Data Protection Impact Assessments (DPIAs) and privacy seals.

1. GDPR, Article 17: The Right to Erasure ("Right to be Forgotten")

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."[194] This is not an unrestricted right. Erasure can be obtained on one of various grounds,[195] including: that the data is 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 sixteen years old. Most usefully, erasure can be sought if the data was being used to profile the data subject and had been collected lawfully but without her consent.[196] The right can conceivably be repelled by the controller on "compelling legitimate grounds" and there are other exceptions, including to safeguard freedom of expression and the historical record.[197]

In the context of ML, we believe a data subject might usefully seek erasure as a remedy in at least three main circumstances.

1. Seeking Erasure of Personal Data "Concerning" a Data Subject

First, 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.

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 *explicitly provided data* used as inputs to an ML system (e.g. name, age, medical history) but does she have the right to erase *observed* data about her behaviour and movements both in the real and virtual world? This is important—ML systems such as those run by Facebook or LinkedIn make heavy use of observed behaviour—for example the type of links clicked on

---

[194] The right to erasure ("right to be forgotten") in GDPR, arts. 17 and 18 (restriction of processing) emerged after the landmark CJEU case, *Google Spain*, *supra*, note 105, and it is both wider in effect and more specified than the rule elaborated in that case out of the DPD.

[195] GDPR, art. 17(1).

[196] 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).

[197] 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 (Article 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.

on-site, the photos viewed, the pages "liked"; or, in the real world, the location and movement as tracked by GPS. While this implicitly provided data, should arguably 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 Article 17 ever contemplated its use for such purposes.

Perhaps most importantly in relation to ML, what about the *inferences* that are made by the system when the data subject's inputs are used as query? These seem what a user 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 Article 29 Working Party (the body of DPA representatives that advises the Commission), in the context of the right to portability[198] have already issued guidance that the inferences of a system is *not* the data of the subject but "belongs" to the system that generated it.[199] It is not yet clear if this approach would be advised regarding the right to erase, though it logically might, as the two rights (GDPR Articles 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. One example is the "right to be forgotten," as vindicated in *Google Spain*.[200]

2. Seeking Withdrawal of Personal Data From a Model: "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 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.

---

[198] *See infra* Section IV.B.2.

[199] *See* A29 WP, *infra* note 218.

[200] *Google Spain*, *supra* note 105.

3. Machine "Unlearning" Take 2: Erasing Models as Themselves Personal Data

Third, 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 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,[201] which is a right over the arrangement of data in a certain system, personal or otherwise, rather than the data itself.

Yet 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.[202] 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.[203] In this type of attack, individual records can be recovered from a model with high probability. Indeed, some applications of ML specifically utilise this characteristic to try and improve or better understand data compression techniques.[204]

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

---

[201] *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 to cover "models," at least in the UK. In the US it seems to have been accepted in some courts that ML algorithms such as search algorithms are protected as trade secrets. See especially Viacom Intern. Inc. v. YouTube Inc., 253 F.R.D. 256, 259-60 (S.D.N.Y. 2008).

[202] *See* Weiss, *supra* note 164.

[203] *See* the literature on model inversion attacks, including Matt Fredrikson et al. *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, *in* CCS' 15, OCTOBER 12–16 2015, DENVER, COLORADO, USA (2016). For attacks against machine learning, *see generally* Ling Huang et al., *Adversarial ML*, AISEC '11, 43–58 (2011). For counter methods, *see* Ananad D. Sarwate et al., *Signal Processing and ML with Differential Privacy: Algorithms and Challenges for Continuous Data*, 86 IEEE SIGNAL PROCESS. MAG. Sep., 86–94 (2013); Ji Zhanglong et al., *Differential Privacy and ML: A Survey and Review* (2014).

[204] GEORGE TODERICI ET AL., FULL RESOLUTION IMAGE COMPRESSION WITH RECURRENT NEURAL NETWORKS (2016).

model and, especially, revising the features of that model.[205] 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.[206]

4. Model Trading and the Right to Erasure

A rising business model involves the trading, publishing of or access to 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, 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 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.

Article 17, section 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."[207]

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

---

[205] In relation to the importance of feature engineering, *see* Pedro Domingos, *A Few Useful Things to Know About ML*, 55 COMMUN. ACM 10, 1, 4 (2012). 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.

[206] Yinzhi Cao & Junfeng Yang, *Towards Making Systems Forget with Machine Unlearning*, *in* SP'15 PROCEEDINGS OF THE 2015 IEEE SYMPOSIUM ON SECURITY AND PRIVACY, MAY 17–21, 463–480 (2015).

[207] GDPR, recital 66.

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, the GDPR 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."[208] 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.

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."[209] Data portability is conceptually a sibling right to Article 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.[210] 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.[211] Data portability is aimed at explicitly 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.[212]

In the context of ML, it is possible to imagine Article 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 of what are sometimes known as Personal Data Containers (PDCs). Using these technologies, personal data does not have to be shared to secure desired services from giants such as Google or Facebook. These companies do not use this 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

---

[208] GDPR, art. 17(2).

[209] GDPR, art. 20(1).

[210] GDPR, art. 20(3). But note the right to erasure covers data "concerning" a data subject rather than as here, merely "provided" by the data subject. This is considerably more restrictive.

[211] GDPR, art. 20(2).

[212] It is also often said that art. 20 is intended to be more of an atypical competition remedy than a privacy remedy. *See* A Diker Vanberg and MB Ünver, "The right to data portability in the GDPR and EU competition law: odd couple or dynamic duo?" 8 EUR. J. OF LAW AND TECH. 1, (2017). *See also* the UK's voluntary *midata* scheme which preceded art. 20.

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.[213] Proponents of data containers, which encompass research projects such as DataBox and Hub of all Things (HaT),[214] 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.[215] Article 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."[216] Using Article 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 Article 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.[217] 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. Sometimes, as 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.

---

[213] *See* LARRY LESSIG, CODE 2.0. 111 (Basic Books, 2006); *see also* visions of this in the marketing literature, such as ALAN MITCHELL, RIGHT SIDE UP 6 (HarperCollins, 2002).

[214] *See* discussion in Lachlan Urquhart et al., *Realising the Right to Data Portability for the Internet of Things* (March 15, 2017), doi:10.2139/ssrn.2933448.

[215] Richard Mortier et al., *The Personal Container or Your Life in Bits*, DIGITAL FUTURES '10, OCTOBER 11–12, 2010, NOTTINGHAM, UK (2010).

[216] Hildebrandt, *supra* note 59.

[217] *See infra* Section V.

Secondly, from a legal perspective, Article 20, much like Article 22, is hedged around with what often seem capricious restrictions. It only applies to data the subject "provides." There seems to be no clear consensus on whether this covers just the explicit data a person provides (e.g. name, hobbies, photos etc. on Facebook); metadata supplied 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.[218] Furthermore, Article 20 only applies to data provided by "consent"[219]—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.[220] Lastly, it is worth emphasising this right only covers data which was being processed by "automated means"[221]—though not, as in Article 22, "s*olely*" automated means, which sets it up as a fundamentally more useful provision concerning algorithms-as-decision-support, rather as decision-makers.

V. BEYOND INDIVIDUAL RIGHTS IN THE GDPR: PRIVACY BY DESIGN

The General Data Protection Regulation discussion so far has revolved around rights given to individual data subjects. Although section I.B above demonstrates that algorithms create societal harms, such as discrimination against racial or minority groups, a focus on data protection remedies makes an individual's rights approach inevitable. Data protection 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).[222]

This means that even if the rights we have discussed above—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, as we noted in our section IV.A on "notice and choice" and transparency fallacies. This is even truer perhaps in the EU where consumers are on the whole far less prepared and empowered to litigate than in the US. The UK and many

---

[218] *See* A29WP: A29 WP, GUIDELINES ON THE RIGHT TO DATA PORTABILITY, 16/EN. WP 242 (Dec. 13, 2016).

[219] GDPR, art. 20(1)(a).

[220] This bizarre choice can only be explained by thinking of art. 20 as a solution to promote competition by allowing data subjects to make active choices to retrieve their voluntarily posted data from social networks.

[221] GDPR, art. 20(1)(b).

[222] Lilian Edwards & Edina Harbinja, *Protecting Post-Mortem Privacy: Reconsidering the Privacy Interests of the Deceased in a Digital World*, 32 CARDOZO ARTICLES & ENT. LAW J. 102, 113 (2013).

other EU nations have no generic system of class actions. Although this has been viewed as a problem for many years, attempts to solve it on an EU wide basis have repeatedly stalled. Individuals are further hampered in meaningfully attaining civil justice by a general prejudice against contingency lawyering combined with dwindling levels of civil legal aid. Some options are emerging in the GDPR to provide a semblance of class action remedies, such as mandating third party bodies to act in court around data protection issues on a data subjects' behalf, or, with specific derogations by member states, for third party bodies to act without being mandated on behalf of a particular sector. However, bodies that are not mandated by a data subject have no ability to claim compensation under DP law, leaving them still far from a US-style class action, and their utility is still to be seen.[223]

The data protection regime contemplates that individual 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[224] (its Data Protection 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?

A. "Big Data Due Process" and Neutral Data Arbiters

From a European perspective, it is interesting to observe how the predominant North American legal literature has tried to solve the problems of algorithmic governance without the low-hanging fruit of a data protection–based "right to an explanation." One notable bank of literature explores the idea of "big data due process." Crawford and Schultz,[225] drawing on early work by Citron,[226] 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

---

[223] *See* GDPR, art. 80.

[224] *See* GDPR, art. 51.

[225] Crawford & Schultz, *supra* note 12 at 123.

[226] Citron, *supra* note 14.

systems in the governmental sector. Citron's work argues[227] for a number of radical adaptations to conventional due process which might include:

- extra education about the biases and fallacies of automation for government agencies using automated systems;[228]
- 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";[229]
- agencies to be required to regularly test systems for bias and other errors;[230]
- 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.[231]

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.[232] 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."[233]

Seen through European eyes, two problems quickly emerge. One, the EU data protection regime applies to private and public sector alike and

---

[227] 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, *supra* note 14 at 1284).

[228] *Id.* at 1306.

[229] *Id.* at 1307.

[230] *Id.* at 1310.

[231] *Id.* at 1305.

[232] Crawford & Schultz, *supra* note 12 at 124.

[233] *See* Tutt, *supra* note 13. Other suggestions for algorithmic audit are usefully compiled by Brent Mittelstadt, *Auditing for Transparency in Content Personalization Systems*, 10 INTERNATIONAL JOURNAL OF COMMUNICATION 4994 (2016).

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.

B. Data Protection Impact Assessment 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, starting from the beginning of the process of design rather than "tacking privacy on at the end!." They recognize 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.[234] 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);[235]
- every public authority and every "large scale" private sector controller and any controller processing "special" categories of data[236] (sensitive personal data) must appoint a Data Protection Officer (DPO);[237]

DPIAs especially have tremendous implications for ML design. PIAs (as they were formerly known) have traditionally been voluntary measures, in

[234] GDPR, art. 25.
[235] GDPR, art. 35.
[236] GDPR, art. 9.
[237] GDPR, art. 37.

practice largely taken up by public bodies bound to compliance and audit, such as health trusts. Attempts to expand their take up in Europe into areas like radio-frequency identification (RFID)[238] and the Internet of Things[239] by the private sector have in the main been unsuccessful. However, the new Article 35 is compulsory, not voluntary, and its definitions of "high risk" technologies are almost certain to capture many if not most ML systems. The GDPR 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."[240]

This is almost identical to the formulation used in Article 22 around automated decision-making. The ICO[241] note firmly that "potential privacy risks" have already been identified with "the use of inferred data and predictive analytics." Accordingly, they provide a draft DPIA for big data analytics.[242] It seems clear that, despite the uncertainty of the "high risk" threshold, DPIAs are quite likely to become the required norm for algorithmic systems, especially where sensitive personal data, such as race or political opinion, is processed on a "large scale."[243]

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.[244] Given the fines that can also be levied against non-compliant controllers under the GDPR (in the worst cases, up to

---

[238] *See* EUROPEAN COMMISSION, PRIVACY AND DATA PROTECTION IMPACT ASSESSMENT FRAMEWORK FOR RFID APPLICATIONS (Jan. 12, 2011), http://cordis.europa.eu/fp7/ict/enet/documents/rfid-pia-framework-final.pdf.

[239] *See* The Data Protection Impact Assessment Template supported by Commission Recommendation 2014/724/EU, Smart Grid task Force 2012-14, Expert Group 2: Regulatory Recommendations for Privacy, Data Protection and Cybersecurity in the Smart Grid Environment, Data Protection Impact Assessment Template for Smart Grid and Smart Metering Systems (Mar. 18, 2014).

[240] GDPR, art. 35(3)(a).

[241] ICO, *supra* note 4.

[242] ICO, *supra* note 4, annex 1.

[243] GDPR, art. 35(3)(b). *See also* DATA PROTECTION WORKING PARTY, *Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is "Likely to Result in a High Risk" for the Purposes of Regulation 2016/679*, Art. 29, WP 248 (Apr. 4, 2017). Judging by this guidance, almost every ML system seems likely to require a DPIA.

[244] GDPR, art. 36(2).

4% of global turnover[245]) this is potentially a very effective method to tame unfair ML systems.[246] Binns describes this as a kind of regulatory "triage."[247]

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.[248] In the UK, a tender has already been advertised by the ICO for a certification authority to run a UK privacy seal,[249] 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:

1. certification of the algorithm as a software object by
    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);
2. 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,"[250] for example, might be provided by such provisions, perhaps as a form of alternate dispute resolution.

---

[245] GDPR, art. 83. Maximum fines are the higher of €10m or 2% of global turnover for less severe transgressions, €20m or 4% for more severe ones.

[246] 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* Lilian Edwards et al., *From Privacy Impact Assessment to Social Impact Assessment*, *in* 2016 IEEE SECURITY AND PRIVACY WORKSHOPS (SPW), 53–57 (2016), doi:10.1109/SPW.2016.19.

[247] Reuben Binns, *Data Protection Impact Assessments: A Meta-Regulatory Approach*, 7 INTERNATIONAL DATA PRIVACY LAW 22, 28 (2017).

[248] For an early analysis, *see* Rowena Rodrigues et al., *Developing a Privacy Seal Scheme (That Works)*, 3 INTERNATIONAL DATA PRIVACY LAW 100 (2013).

[249] Gemma Farmer, *What's the Latest on the ICO Privacy Seals?*, INFO. COMMISSIONER'S OFF.: BLOG (Aug. 28, 2015), https://iconewsblog.org.uk/2015/08/28/whats-the-latest-on-the-ico-privacy-seals/.

[250] Citron, *supra* note 14 at 1254–58.

In these cases, not only could fairness and discrimination issues be considered in the standards to certify against,[251] 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 have noted in this paper, explanations and their effectiveness differ strongly by type, domain, and the user seeking explanation, 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[252] 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 the EU–US data agreement *Safe Harbor* alone,[253] 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.[254]

Another issue is that DPIAs, PbD, certification and the general principle of "accountability"[255] in the GDPR bring with them a real danger of formalistic 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.

CONCLUSION

Algorithms, particularly of the ML variety, are increasingly used to make decisions about individuals' lives but have caused a range of

---

[251] Issues of algorithmic fairness are specifically discussed in GDPR, recital 71. Tristan Henderson in private correspondence has suggested that a certifying authority might well under art. 42 be given the power to require explanation facilities, thus side stepping the Article 22/15(h) debate.

[252] GDPR, arts. 40 and 41.

[253] *See* Schrems v. Data Prot. Comm'r of Ir., Case C-362/14, 6 October 2015.

[254] TrustE and similar privacy seals failed to meet European privacy standards. *See* Andrew Charlesworth, *Data Privacy in Cyber Space*, LAW AND THE INTERNET (2000).

[255] GDPR, art. 5(2). This may lead to a new world of form-filling for data controllers.

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 I, 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 conclusion. First, the law is restrictive on when any explanation-related right can be triggered, and in many places is unclear, or even seems paradoxical. Secondly, 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, "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 et al. argue, "*after* an automated decision has taken place," [256] 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 wording of provisions like the GDPR Article 15(h), the search for a legally enforceable 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."[257] So, as our last exercise, we turned our focus to the other legal rights of the GDPR which might aid those impacted adversely by ML systems. We noted with caution some possible uses of the GDPR's "right to erasure" and the "right to data portability" to "slave" the algorithm, but found that, like the "right to an explanation," these rely too much on individual rights for what are too often group harms.

---

[256] Wachter et al., *supra* note 11.

[257] *See supra* Section V.

However, radically, in section IV we found that some of the new tools in the GDPR, in particular the mandatory 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[258] seems eminently more appealing than grimly pursuing against the odds a "meaningful" version of the interior of a black box.

A. 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,[259] identification systems that think only light skinned people are beautiful[260] 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 overarching 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?

Furthermore, what about ML systems that mainly deal with non-personal data? Should they be excluded from any data protection 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.[261] On the other hand, the limitation of scope to personal data could

---

[258] *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.

[259] *See passim* the glorious *Trolley problem memes* page at www.facebook.com/TrolleyProblemMemes/.

[260] *See* Dave Neal, *FaceApp Sorry for Suggesting that Light Skin is 'Hotter' than Dark Skin*, THE INQUIRER (Apr. 25, 2017), https://www.theinquirer.net/inquirer/news/3008961/faceapp-sorry-for-suggesting-that-light-skin-is-hotter-than-dark-skin.

[261] European Commission, *Public Consultation on Building the European Data Economy* (Jan. 10, 2017), https://ec.europa.eu/digital-single-market/en/news/public-consultation-building-european-data-economy.

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.[262]

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 AI, ML and algorithmic decision making in embodied forms such as robots,[263] 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

---

[262] This submission is primarily drawn from The Royal Society, *supra* note 4.

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

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 data protection lawyer and an ML specialist. 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."[264]

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 continuing trans-disciplinary collaboration can we hope not just to enslave the algorithm, but to create a more legitimate, more comprehensible and in the end more useful algorithmically-mediated society.

[264] Hildebrandt, *supra* note 9 at 54.