Show Us the Data: Privacy, Explainability, and Why the Law Can't Have Both*

Thomas D. Grant† *& Damon J. Wischik*§

ABSTRACT

Two rights—the right to privacy and the right to an explanation when automated decision-making affects an individual—are in fundamental conflict. To recognize how they conflict, one needs to understand two things: (1) how machine learning works and (2) how a litigant in a dispute against a data controller will use litigation procedure to demand both explanation and proof of the validity of any explanation proffered.

Machine learning is not based on logic-driven algorithms explicable by examining source code. It is a data-driven process of pattern-matching based on large data sets and statistics. Data subjects affected by machine learning decisions will demand different explanations depending on each subject's relationship to the data controller. Where the data controller and data subject trust one another, the explanation of the decision might be given in general or "accessible" terms. But where they are in conflict and the subject has resorted to adversary procedures, the dynamics of litigation will impel the subject-claimant to demand not only an explanation, but proof that the explanation is accurate and true. This process of demand leads to an intensifying scrutiny of the data that trained the machine. It inevitably will lead to demands to see the training data—not anonymized or partial versions of it—and thus will trespass upon the privacy rights of people from whom the training data was derived.

Various solutions might be attempted, such as data rooms or in camera review. There are reasons to be skeptical of these and, in any event, they remain to be tested. The fundamental conflict, regardless of compliance approaches adopted to mitigate regulatory risk, seems likely to be exposed and played out chiefly in adversarial proceedings in the years ahead.

The Article takes as its focus the European Union's General Data Protection Regulation ("GDPR") because of this regulation's global prominence, but the difficulty it identifies will be presented across national jurisdictions, as many legislatures adopt explainability laws against the backdrop of long-vested privacy rights.

TABLE OF CONTENTS_Toc48924149

* Views and conclusions in this text are those of the authors in their academic capacities alone and do not represent any other individual or institution.

† Fellow, Lauterpacht Centre for International Law, University of Cambridge.

§ Lecturer, Computer Laboratory, University of Cambridge.

INTRODUCTION

Privacy rights and a right to receive explanation of decisions reached by machine both are stipulated in law. Article 8 of the European Convention on

Human Rights[1] includes a broad category of privacy rights.[2] Separately, European Union ("EU") law has for some time protected privacy rights over personal data.[3] As for "explainability," the EU's General Data Protection Regulation ("GDPR"),[4] which was adopted in 2016 and became applicable in 2018,[5] obliges people and institutions that handle and process data to supply laypersons "meaningful information about the logic involved" in automated decision-making.[6] For these provisions to operate as their drafters appear to have intended, any person or organization subject to them must simultaneously protect the privacy of individuals and explain how machines have reached decisions that affect individuals.[7] The two objectives, taken together, are likely not achievable. This Article identifies how the rights to privacy and explainability, especially as embodied in the GDPR, conflict.

To explain the decisions that a machine has reached, one must start by understanding the machine. The GDPR provisions requiring an explanation of automated decision-making would be relatively easy to apply where the machine runs on software that, though perhaps complex, is like an algorithm or logical deduction. Much of the literature and comment concerning "explainability" reflects a premise that this is the machine that the GDPR calls upon "data controllers" to explain.[8] The premise is mistaken because

1 Convention for the Protection of Human Rights and Fundamental Freedoms art. 8, Nov. 4, 1950, 213 U.N.T.S. 221.

2 *E.g.*, S. & Marper v. United Kingdom, 2008-V Eur. Ct. H.R. 167, 193–94 (invalidating the retention of DNA evidence where a criminal defendant is acquitted or a charge is not pressed).

3 *E.g.*, Council Directive 2002/58, art. 1, 2002 O.J. (L 201) 37, 39–40, 42 (EC); *cf.* Council Regulation 2016/679, 2016 O.J. (L 119) 2 (EU) (recognizing a "right to the protection of personal data").

4 Council Regulation 2016/679, *supra* note 3.

5 *Id.* at 87–88.

6 *Id.* arts. 13(2)(f), 14(2)(g) & 15(1)(h), at 41–43; *cf. id.* arts. 21–22, at 14, 45–46 (discussing data subjects' rights, including a right to information, a "right to object," and a "right not to be subject to a decision based solely on automated processing").

7 *See id.* at 41–43, 45–46.

8 *See, e.g.*, Tarleton Gillespie, *The Relevance of Algorithms*, *in* MEDIA TECHNOLOGIES: ESSAYS ON COMMUNICATION, MATERIALITY, AND SOCIETY 167, 192 (Tarleton Gillespie et al. eds., 2014) ("The algorithmic logic . . . depends on the proceduralized choices of a machine, designed by human operators to automate some proxy of human judgment."); ALAN S. GUTTERMAN, BUSINESS TRANSACTIONS SOLUTIONS § 217:146 (June 2020 Update) (ebook) ("[W]ith Artificial Intelligence[,] the machine merely follows a carefully constructed program."); Ashley Deeks, *High-Tech International Law*, 88 GEO. WASH. L. REV. 574, 574 (2020) (describing machine learning as a "[d]ata-driven algorithmic tool[]"); *see also* Christian Chessman, Note, *A "Source" of Error: Computer Code, Criminal Defendants, and the Constitution*, 105 CALIF. L. REV. 179, 184 (2017) ("Evidence produced by computer programs arguably merits additional scrutiny . . . because the complexity of computer programs makes it difficult . . . to detect errors.").

machine learning, which is to say the process behind the present generation of so-called "artificial intelligence," does not operate like algorithmic logic; it operates by learning from data.[9] The conclusions influenced by the premise—including the explainability provisions of the GDPR—need, accordingly, to be reconsidered.

The GDPR, like various legislative innovations at the national level, responds to a general concern that machines are involved in ever more significant ways in reaching decisions that heretofore were made by human beings unaided.[10] A standard account now exists that machines have come to affect decisions of many types.[11] These range from approving an online credit application or targeting an individual for recruitment,[12] to completing a search term on a search engine,[13] writing a financial news article,[14] and calculating how long a convicted felon is to be incarcerated.[15] More than one consideration instigates scrutiny into AI-based decision-making, but the main one is an apprehension that AI-based decisions might discriminate

9 Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [https://perma.cc/32XG-7MQM].

10 Council Regulation 2016/679, *supra* note 3, at 2–3 (calling for the protection of data processed automatically as technology creates new challenges for privacy); *cf.* European Commission Press Release IP/19/4449, General Data Protection Regulation Shows Results, but Work Needs to Continue (July 24, 2019), https://ec.europa.eu/commission/presscorner/detail/en/IP_19_4449 [https://perma.cc/X94H-3N96] (describing the "national data protection laws" adopted by European Union countries and the need to continue monitoring said countries to ensure their laws "remain[] in line with the Regulation").

11 And it is evidently *de rigueur* to give the account, or one much like it, when talking about automated decision-making. *See, e.g.*, Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 636 (2017); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1149 (2017); Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, BIG DATA & SOC'Y, Dec. 2016, at 1, 1.

12 *See* Council Regulation 2016/679, *supra* note 3, at 2–3, 14 (EU) (describing privacy rights for automatic credit application evaluations); County of Riverside v. Perone, No. E037293, 2006 WL 245319, at *1 (Cal. Ct. App. Feb. 2, 2006).

13 *See, e.g.*, Albert v. Google Inc. (No 2), [2015] 1 H.K.L.R.D. 26, 37 (C.F.I.) (describing Google Autocomplete searches); Iowa v. Retterath, 912 N.W.2d 500 (Iowa Ct. App. 2017); Shanta Rangaswami et al., *Analysis of Optimized Association Rule Mining Algorithm Using Genetic Algorithm*, 2014 INT'L J. COMPUTER APPLICATIONS: INT'L CONF. ON INFO. & COMM. TECHS., 12, 12 ("Genetic algorithm is a search heuristic.").

14 *See* Nicholas Diakopoulos, *Accountability in Algorithmic Decision Making*, 59 COMMS. ACM, 56, 56 (2016) ("[A]utomated writing algorithms churn out thousands of corporate earnings articles.").

15 *See* Wisconsin v. Loomis, 881 N.W.2d 749, 755 (Wis. 2016) (using an automated risk assessment tool to "rul[e] out probation"); *see also* Recent Case, Wisconsin v. Loomis, *881 N.W.2d 749 (Wis. 2016)*, 130 HARV. L. REV. 1530, 1530 (2017).

against individuals for invidious reasons.[16] This Article is not a survey of AI-based decision-making across all subject matter, and it is not an assessment of the risks and benefits of AI-based decisions.

Nor is this Article's purpose to supply the final word on the "explainability" provisions of the GDPR, much less to gloss upon the Regulation as a whole. An extensive public discourse is underway concerning "explainability" and the GDPR. Some writers say explainability is regulatory overreach bound to stultify the development of AI.[17] Others say it is indispensable for preserving basic human values in a machine-learning age.[18] Compliance with explainability provisions by organizations that employ AI has entailed significant expenditures of capital, as well as shifts in organizational structure.[19] Compliance with provisions of EU law and the European Convention more generally concerned with privacy has been costly too.[20] Indeed, the legal and compliance professions concerned with the GDPR count large numbers of people broadly describing themselves as privacy specialists.[21] Unsurprisingly, a body of literature has appeared over

16 *See, e.g.*, Robert Bartlett et al., Consumer-Lending Discrimination in the FinTech Era 29 (Nov. 2019) (unpublished manuscript), http://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf [https://perma.cc/H6LM-9WXT] (finding algorithmic bias "in the context of consumer lending").

17 *See, e.g.*, Christopher Kuner et al., *Machine Learning with Personal Data: Is Data Protection Law Smart Enough to Meet the Challenge?*, 7 INT'L DATA PRIVACY L. 1, 1–2 (2017) (expressing an optimistic conclusion about the potential benefits of AI-driven decision-making as corrective for human fallibility and invidious bias that could be threatened by the practical implications of broad regulation).

18 *See, e.g.*, Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841, 843 (2018) (proposing three benefits of explainability); Adrian Weller (Nov. 15, 2018), https://www.bcs.org/content/conEvent/11991 [https://perma.cc/Q8S3-9UVB] (summary).

19 Consider, *e.g.*, IBM's "comprehensive new set of trust and transparency capabilities for AI," Aleksandra Mojsilovic, *Trust and Transparency for AI on the IBM Cloud*, IBM (Sept. 19, 2018), https://www.ibm.com/blogs/research/2018/09/trust-transparency/ [https://perma.cc/R6T5-62KV], and the observation by Allen & Overy LLP that "many companies are finding that there is a lot (for some, too much) to do" in respect of GDPR compliance, *Preparing for the General Data Protection Regulation*, ALLEN & OVERY (June 28, 2018), https://www.allenovery.com/en-gb/global/news-and-insights/publications/preparing-for-the-general-data-protection-regulation [https://perma.cc/6KT8-SR6J].

20 NEIL ROBINSON ET AL., REVIEW OF THE EUROPEAN DATA PROTECTION DIRECTIVE, at viii (2009) (noting "[c]riticisms from within the EU . . . [regarding] the formalities imposed by . . . and the economic costs of compliance and unequal enforcement" of the EU Data Protection Directive of 1995).

21 *See, e.g.*, *Register of EuroPriSe Experts*, EUR. PRIVACY SEAL, https://www.european-privacy-seal.eu/EPS-en/Register-of-experts [https://perma.cc/T2EC-NSE9].

the past several years addressing the impact of explainability (anticipated or realized), and at least as much has appeared on the topic of privacy.[22]

Amidst the immediate practical demands, in particular for regulated parties to demonstrate that they have taken measures to comply with both explainability and privacy rules, a critical problem largely has been ignored: when challenged in a courtroom setting, and subject to long-established rules of evidence, regulated parties are likely to find it difficult or impossible to satisfy both. This Article posits here that the "explainability" provisions of the GDPR are not suited to the machine learning outputs that those provisions require explanations for. The GDPR, including its explainability provisions, belongs to a wider legal framework that promises privacy.[23] But you cannot explain a machine learning output unless you look at the data that trained the machine; this means scrutinizing the data in detail because generalizations, for reasons that this Article articulates,[24] will not do. In a litigation setting, a claimant will demand detail that cannot be disclosed without violating the privacy rights of the data subjects from whom the data came. A proper understanding of machine learning—together with a proper understanding of the legal and forensic setting in which the GDPR will eventually be applied—exposes a paradox: you can have privacy, or you can have explainability, but you cannot have both.

This Article starts with a brief overview of legal texts relevant to privacy and of the provisions of the GDPR relevant to "explainability" (Part I). Much of the effort expended so far on explanation of automated decision has been with a view to building systems that are compliant with best practices or with regulatory directions. However, the further development of a law of explanation will take place not through compliance programs but through the judgments of courts. This Article thus contrasts a compliance approach to explainability with litigation (Part II). Because the type of explanation that one seeks depends not only on the characteristics of the thing to be explained but also on the reasons one seeks the explanation, this Article considers why somebody might ask that a machine learning output be explained (Part III). And, because legislation and public discussion to date too readily addresses machine learning as if it were just another sort of computational process,[25]

22 *See, e.g.*, He Li, Lu Yu & Wu He, *The Impact of GDPR on Global Technology Development*, 22 J. GLOBAL INFO. TECH. MGMT. 1, 1–2 (2019) (exploring the effects of GDPR's privacy and explainability requirements on data controllers).

23 Convention for the Protection of Human Rights and Fundamental Freedoms art. 8, Nov. 4, 1950, 213 U.N.T.S. 221.

24 *See infra* Part IV.

25 *See, e.g.*, Andrea Roth, *Machine Testimony*, 126 YALE L.J. 1972, 1977 n. 16 (2017) (claiming a "need to probe machines' inner workings"); Council Regulation 2016/679, *supra*

this Article explains how machine learning differs from algorithmic programming and how its distinctive characteristics affect the type of explanation that one might give for a machine learning output (Part IV).

Looking more closely at the litigation that explainability will instigate, this Article considers a series of defenses that a data controller might give and the challenges that a data subject who proceeds in court as a claimant against the controller is likely to raise against them (Part V). Finally, this Article suggests that litigants who understand how machine learning works, and who are concerned about invidious effects in machine learning outputs, are unlikely to accept anonymized data (Part VI). Standard methods of privacy compliance thus might well fail to resolve the tension between privacy and explainability, which will challenge governments, courts, and private parties as they implement the new Regulation.

I. THE LEGAL TEXTS

Although the GDPR marks a major addition to EU law, is not entirely a fresh departure. The GDPR belongs to a long-developing body of rules applicable in the EU member states. Rules concerning privacy in particular have formed part of the legal order of Europe for some time.[26] A brief summary of privacy,[27] as reflected in several legal instruments applicable in Europe, sets the stage for present purposes. This Part then turns to the provisions of the GDPR relating to explainability.

A. Privacy

Privacy rights are a fundamental part of the European legal order. Their importance is visible both in the wider European human rights framework and in EU legislation. To start with the wider framework, the Convention for the Protection of Human Rights and Fundamental Freedoms of 1950,[28]

note 3, art. 13(2)(f) (requiring data controllers to give data subjects "meaningful information about the logic involved" in "automated decision-making").

26 *See, e.g.*, Convention for the Protection of Human Rights and Fundamental Freedoms art. 8, Nov. 4, 1950, 213 U.N.T.S. 221.

27 This, perforce, is selective and general only; monograph-length treatments address particular privacy provisions, and so even an introduction to the rules and associated jurisprudence far exceeds the scope of one article. For background, see generally DENIS KELLEHER & KAREN MURRAY, EU DATA PROTECTION LAW (2018); GLORIA GONZÁLEZ-FUSTER, THE EMERGENCE OF PERSONAL DATA PROTECTION AS A FUNDAMENTAL RIGHT OF THE EU (2014); PETER CAREY, DATA PROTECTION: A PRACTICAL GUIDE TO UK AND EU LAW (3d ed. 2009). For an earlier overview at international level, see generally ABRAHAM L. NEWMAN, PROTECTORS OF PRIVACY: REGULATING PERSONAL DATA IN THE GLOBAL ECONOMY (2008).

28 Convention for the Protection of Human Rights and Fundamental Freedoms, Nov. 4, 1950, 213 U.N.T.S. 221.

usually known by its short-form title European Convention on Human Rights, applies to the member states of the EU, to the EU itself, and to the other states of the European area (e.g., Russia, Turkey, Norway).[29] Article 8 of the Convention provides as follows:

> (1) Everyone has the right to respect for his private and family life, his home and his correspondence.
>
> (2) There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.[30]

The second paragraph of Article 8, which indicates exceptions, suggests that the focus of the article is on the conduct of public authorities; indeed, a number of cases applying Article 8 have concerned such conduct. Although special concern arises from the effects of government actions on privacy under the Convention,[31] privacy rights also protect individuals from the conduct of private actors.[32] As to the content of the rights protected, the expression in the first paragraph of a "right to respect for . . . private and family life" has come to be understood to encompass a wide array of rights. A summary of protected rights, as of 2008, was provided by the European Court of Human Rights in *S. & Marper* v. *United Kingdom*,[33] a judgment concerning the retention by police of DNA evidence gathered from persons who were convicted of no crime:

> 66. [T]he concept of "private life" is a broad term not susceptible to exhaustive definition. It covers the physical and psychological integrity of a person. It can therefore embrace multiple aspects of

[29] *Id.* at 223; *Press Country Profile: Russia*, EUR. CT. OF HUM. RTS., https://www.echr.coe.int/Documents/CP_Russia_ENG.pdf [https://perma.cc/F6J4-FS58].

[30] Convention for the Protection of Human Rights and Fundamental Freedoms art. 8, Nov. 4. 1950, 213 U.N.T.S. 221.

[31] *See generally* S. & Marper v. United Kingdom, 2008-V Eur. Ct. H.R. 167 (holding that retention by the authorities of DNA and fingerprints had violated ECHR art. 8); Big Brother Watch v. United Kingdom, App. Nos. 58170/13, 62322/14 and 24960/15, Judgment (Sept. 13, 2018), http://hudoc.echr.coe.int/eng?i=001-186048 [https://perma.cc/VSE6-RGHW] (holding that certain aspects of a bulk data-collection program conducted by the UK's intelligence agency constituted violations of ECHR art. 8).

[32] *See, e.g.*, Satakunnan Markkinapörssi Oy v. Finland, App. No. 931/13, Judgment (June 27, 2017), http://hudoc.echr.coe.int/eng?i=001-175121 [https://perma.cc/7UNC-SW3E] (interpreting privacy rights to allow restraints against a newspaper publishing personal income and assets data).

[33] S. & Marper v. United Kingdom, 2008-V Eur. Ct. H.R. 167.

> the person's physical and social identity. Elements such as . . . gender identification, name and sexual orientation and sexual life fall within the personal sphere protected by Article 8. Beyond a person's name, his or her private and family life may include other means of personal identification and of linking to a family. Information about the person's health is an important element of private life. The Court furthermore considers that an individual's ethnic identity must be regarded as another such element. Article 8 protects, in addition, a right to personal development, and the right to establish and develop relationships with other human beings and the outside world. The concept of private life moreover includes elements relating to a person's right to their image.
>
> 67. The mere storing of data relating to the private life of an individual amounts to an interference within the meaning of Article 8.[34]

Although Article 8 of the European Convention on Human Rights is perhaps the keystone of privacy rights in Europe, a range of more particular rights and obligations regarding privacy are lain down in other regional instruments. For example, the Data Protection Convention of 1981[35] (noted in *Marper*)[36] provides, in Article 6, that "personal data revealing racial origin, political opinions or religious or other beliefs, as well as personal data concerning health or sexual life, may not be processed automatically unless domestic law provides appropriate safeguards. The same shall apply to personal data relating to criminal convictions."[37] At the global level, the Organisation for Economic Co-operation and Development ("OECD") in May 2019 identified "privacy and data protection" among the "[h]uman-centered values" that should be promoted and implemented by actors employing artificial intelligence.[38]

Then there are a number of EU instruments relevant to privacy. For sake of brevity, this Article examines two: Directive 95/46/EC[39] (repealed upon

34 *Id.* at 193–94 (citations omitted).

35 Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data art. 6, Jan. 28, 1981, 1496 U.N.T.S. 65.

36 S. & Marper v. United Kingdom, 2008-V Eur. Ct. H.R. 167, 184.

37 Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, *supra* note 35, at 68.

38 OECD Secretary-General, *Recommendation of the Council on Artificial Intelligence*, § 1.2, OECD/Legal/0449 (adopted May 22, 2019).

39 Council Directive 95/46, 1995 O.J. (L 281) (EC).

application of the GDPR)[40] and Directive 2002/58/EC.[41] The former, in its recitals, discusses EU privacy protections:

> [T]he object of the national laws on the processing of personal data is to protect fundamental rights and freedoms, notably the right to privacy, which is recognized both in Article 8 of the European Convention . . . and in the general principles of [EU] law; . . . for that reason, the approximation of those laws must not result in any lessening of the protection they afford but must, on the contrary, seek to ensure a high level of protection . . . .[42]

EU law has thus for some time linked the right to privacy, itself one of the "fundamental rights and freedoms" of the European legal order, to the safeguards that it requires for the processing of personal data.

Article 1(1) of Directive 95/46 provides that "Member States shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data."[43] The Directive also provides, under Article 17(1), for national legislation requiring data controllers[44] to "implement appropriate technical and organizational measures to protect personal data against accidental or unlawful destruction or accidental loss, alteration, unauthorized disclosure or access."[45] Article 17(2) requires data controllers to "choose a processor [of data] providing sufficient guarantees in respect of the technical security measures and organizational measures governing the processing to be carried out."[46] The resultant obligations have been the subject of disputes brought to court.[47]

---

40 Council Regulation 2016/679, *supra* note 3, at 1.

41 Council Directive 2002/58, *supra* note 3.

42 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) 32 (EC). Directive 95/46/EC was repealed on May 25, 2018. Council Regulation 2016/679, *supra* note 3, art. 94(1).

43 Council Directive 95/46, *supra* note 39, art. 1(1).

44 Article 4(7) of the GDPR defines "controller" to mean

> the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.

Council Regulation 2016/679, *supra* note 3, art. 4(7). Article 2(d) of Directive 95/46/EC defined the term the same way. Council Directive 95/46, *supra* note 39, art. 2(d).

45 Council Directive 95/46, *supra* note 39, art. 17(1).

46 *Id.* art. 17(2).

47 *See generally* Case C-210/16, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v. Wirtschaftsakademie Schleswig-Holstein GmbH, Judgment, ¶ 1 (June

As for the 2002 Directive, it identifies as one of its aims "to ensure [in all Member States] an equivalent level of protection of fundamental rights and freedoms, and in particular the right to privacy, with respect to the processing of personal data in the electronic communication sector."[48] Litigation has arisen where the application of the 2002 Directive has been an issue as well.[49] Similar in this respect to the European Convention on Human Rights, EU regulations concerning privacy address the conduct of both government bodies[50] and private parties.[51]

The 2002 Directive, where it applies to matters concerning the protection of fundamental rights and freedoms *vis-à-vis* the processing of personal data and sets out specific obligations with the same objective as the GDPR, would seem to prevail over the latter instrument. It is by no means absolutely clear that a hierarchy operates precisely this way between the two, and the matter has not as yet been clarified in court. It does however seem to follow, from Recital (173) of the GDPR, that the privacy provisions of the 2002 Directive remain intact and, indeed, are to be given some priority: "This Regulation should apply to all matters concerning the protection of fundamental rights and freedoms vis-à-vis the processing of personal data which are not subject to specific obligations with the same objective set out in Directive 2002/58/EC."[52]

5, 2018), http://curia.europa.eu/juris/liste.jsf?num=C-210/16&language=EN [https://perma.cc/Y488-E5F5].

48 Council Directive 2002/58, *supra* note 3, art. 1(1).

49 *E.g.*, Case C-207/16, Ministerio Fiscal v. Spain, Judgment, ¶ 1 (Oct. 2, 2018), http://curia.europa.eu/juris/liste.jsf?num=C-207/16&language=EN [https://perma.cc/X883-SUAC]; Joined Cases C-203/15 & C-698/15, Tele2 Sverige AB v. Post- och telestyrelsen, Judgment, ¶¶ 1, 56 (Dec. 21, 2016), http://curia.europa.eu/juris/liste.jsf?num=C-203/15&language=EN [https://perma.cc/EWE4-FDZM].

50 The cases are numerous which have concerned the retention of personal data by government bodies, for example tax enforcement agencies. *E.g.*, Case C-73/16, Puškár v. Finančné riaditel'stvo Slovenskej republiky, Judgment, ¶ 99 (Sept. 27, 2017), http://curia.europa.eu/juris/liste.jsf?num=C-73/16&language=EN [https://perma.cc/SK58-76KK].

51 A number of cases, for example, have concerned data processing by Facebook and Google (directly and indirectly). *See, e.g.*, Case C-210/16, Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v. Wirtschaftsakademie Schleswig-Holstein GmbH, Judgment, ¶ 15 (June 5, 2018), http://curia.europa.eu/juris/liste.jsf?num=C-210/16&language=EN [https://perma.cc/XXU8-A5UW] (applying EU privacy regulations to Facebook); Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos, Judgment, ¶ 2 (May 13, 2014), http://curia.europa.eu/juris/liste.jsf?num=C-131/12&language=EN [https://perma.cc/RE2P-Z5JX] (applying EU privacy regulations to Google).

52 Council Regulation 2016/679, *supra* note 3, at 31.

It also seems that the protection of personal data in the EU from transfer or disclosure supersedes requests, or commands, from courts in non-EU countries. Article 48 of the GDPR provides that: "Any judgment of a court or tribunal and any decision of an administrative authority of a third country requiring a controller or processor to transfer or disclose personal data may only be recognised or enforceable in any manner if based on an international agreement."[53]

Recital (115) of the GDPR addresses the substance of Article 48:

> (115) Some third countries adopt laws, regulations and other legal acts which purport to directly regulate the processing activities of natural and legal persons under the jurisdiction of the Member States. This may include judgments of courts or tribunals or decisions of administrative authorities in third countries requiring a controller or processor to transfer or disclose personal data, and which are not based on an international agreement, such as a mutual legal assistance treaty, in force between the requesting third country and the Union or a Member State. The extraterritorial application of those laws, regulations and other legal acts may be in breach of international law *and may impede the attainment of the protection of natural persons ensured in the Union by this Regulation*.[54]

This Article's purpose is not to address extraterritorial application, which the Recital identifies as a distinct basis of concern.[55] This Article draws attention instead to the statement that transfer or disclosure of personal data, when asked for or instructed by a non-EU court or other decision-making body, "may impede the attainment of the protection of natural persons."[56] The "protection" concerned includes in particular privacy protection.

From these two recitals and Article 48 of the GDPR, it is seen that, both as a matter of the relation of EU privacy law to other EU law provisions and as a matter of its relation to law in third States, privacy enjoys some measure of priority. As this Article suggests in more detail below, the way machine learning works is going to throw up challenges as courts seek to give that priority effect.

---

53 *Id.* art. 48.

54 *Id.* at 22 (emphasis added).

55 Extraterritorial application of laws is a long-running subject of differences between the EU and third states. *See, e.g.*, Phillippe Bonnecarrère, *What European Response to American Extra-territoriality?*, ROBERT SCHUMAN FOUND. (Apr. 2, 2019), https://www.robert-schuman.eu/en/european-issues/0501-what-european-response-to-american-extraterritoriality [https://perma.cc/6KJM-Z2HT].

56 Council Regulation 2016/679, *supra* note 3, at 22.

Of course, data controllers are expending a great deal of effort, and will continue to, outside of court. Compliance measures to implement obligations—including privacy-related obligations—are a mainstay of data controllers' legal strategy.[57] Much of the effort at compliance is indeed stipulated in the relevant legislation. Among the effects of privacy law in Europe, organizations that handle data have obligations to maintain procedures to mitigate the risk of violation of the privacy rights of individuals. As the Court of Justice of the European Union has indicated, Article 17 of Directive 95/46 provides:

> Member States are to provide that the controller must implement appropriate technical and organisational measures which, having regard to the state of the art and the cost of their implementation, are to ensure a level of security appropriate to the risks represented by the processing and the nature of the data to be protected.[58]

Organizational measures for compliance with privacy rules must be taken by governments as well.[59] Recommendations as to how to implement such organizational measures are a major topic in privacy guidance and literature.[60]

Considering the provisions that this Article has touched upon above and the judgments interpreting and applying them, privacy constitutes a basic constitutional principle in the countries covered by the European Convention on Human Rights. These include all the countries of the EU, which is to say all the countries covered by the GDPR. The precise ramifications of the right to privacy are myriad; the jurisprudence on privacy and the literature addressing the topic are vast.[61] As disputes over privacy suggest, the meaning

---

[57] *See The EU General Data Protection Regulation*, ALLEN & OVERY, http://www.allenovery.com/SiteCollectionDocuments/Radical%20changes%20to%20European%20data%20protection%20legislation.pdf [https://perma.cc/9SW8-XMZU] (observing that the GDPR requires significant effort from "[m]any companies," and "for some, too much").

[58] Case C-553/07, College van burgemeester en wethouders van Rotterdam v. M.E.E. Rijkeboer, 2009 E.C.R. I-03889, ¶ 17; *cf. id.* ¶ 62 (summarizing State arguments on the proportionality of a one-year time limit for data deletion).

[59] *See* Case C-582/14, Breyer v. Bundesrepublik Deutschland, Opinion of Advocate General Campos Sánchez-Bordona, ¶ 41 (May 12, 2016), http://curia.europa.eu/juris/document/document.jsf?docid=178241&doclang=EN [https://perma.cc/4ZGV-R85D] (noting Austria's statement on its compliance obligations).

[60] *See, e.g.*, INFO. COMM'R'S OFFICE, BIG DATA, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA PROTECTION 60, 72, 74, 80, 86, 89, 97 (2017).

[61] For a sense of the scope, see the course bibliography. *Bibliography for Privacy and Data Protection Law*, U. KENT, https://kent.rl.talis.com/lists/87E94F90-5FF8-3C75-35CC-B1E4C43BFF90/bibliography.html?style=nature [https://perma.cc/YGA6-UGQ8]. A number of peer-reviewed journals cover the topic or are dedicated to it. *See, e.g.*, *International*

of the various instruments that refer to privacy is far from completely settled law. These general observations nevertheless are valid: privacy is a legal right of central importance in European law, and its implementation is undertaken through compliance mechanisms in organizations, private and public, and through litigation in court. No doubt the vast majority of resources expended in connection with privacy law are expended on compliance. Litigation, however, plays a crucial role because it is through the settlement of disputes in court that key questions of interpretation and application of any regulation are answered. Substantial questions remain to be answered as to the relation of privacy rules and the GDPR.

B. Explainability

Whereas privacy has been a European legal value for many years,[62] explainability so far has not been imparted much further substance through legal disputes, judgments, or general practice. Moreover, whereas the core legal texts concerning privacy are easy to identify, it is not quite so obvious where in the relevant instrument explainability comes from. The GDPR has no single provision or section titled "explainability." Indeed, except for a single reference to "explanation"—and that in a recital (Recital 71),[63] not the operative part[64]—the regulation does not use any word derived from the verb "to explain." The absence of an express "explainability" provision, together with ambiguities in the adopted text, have given rise to speculation that the regulation in truth makes no provision for "explainability" as such.[65]

Data Privacy Law: About the Journal, OXFORD ACAD., https://academic.oup.com/idpl/pages/About [https://perma.cc/J8BL-PQU4].

62 *See supra* Section I.A.

63 Council Regulation 2016/679, *supra* note 3, at 14.

64 Recitals in a European Union legal instrument may be invoked to interpret operative provisions, but recitals are not an independent basis of legal rights or obligations. Tadas Klimas & Jūratė Vaičiukaitė, *The Law of Recitals in European Community Legislation*, 15 ILSA J. INT'L & COMP. L. 61, 62 (2008). As to specific problems presented by Recital 71 of the GDPR, see Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking for*, 16 DUKE L. & TECH. REV. 18, 49–50 (2017).

65 *See* Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 76 (2017) (expressing doubt over "the legal existence and the feasibility of [a right to explanation]"); Vlad A. Hertza, Note, *Fighting Unfair Classifications in Credit Reporting: Should the United States Adopt GDPR-Inspired Rights in Regulating Consumer Credit?*, 93 N.Y.U. L. REV. 1707, 1732–33, n. 169 (2018) ("[There] has been . . . great debate recently as to whether the GDPR . . . provides for a 'right to explanation' of the reasoning algorithms used in reaching a decision."); *cf.* Edwards & Velae, *supra* note 64, at 50 ("[T]hese . . . seem shaky foundations on which to build a harmonised cross EU right to algorithmic explanation."). This Article notes that the United

The prevailing view, however, is that "explainability" belongs to the principles and rules of data protection of the European Union,[66] a view to which the European Commission lends support, for example in its dealings with the U.S. Department of Commerce.[67] This is also the view of one of the architects of the GDPR.[68]

Several provisions of the GDPR may be considered as embodying a requirement that those data controllers who employ machine learning shall supply an explanation of the results: these are five articles in the operative part, and Recital 71.[69] The five articles are Articles 13–15, which comprise Section 2 ("Information and access to personal data") of Chapter III ("Rights of the data subject"), and Articles 21 and 22, which comprise Section 4 ("Right to object and automated individual decision-making") of Chapter III.[70]

Kingdom Information Commissioner's Office describes the relevant rights as "rights *related to* automated decision making including profiling," alongside six rights specified as "right to be informed," "right of access," "right to rectification," "right to erasure," "right to restrict processing," "right to data portability," and "right to object." INFO. COMM'R'S OFFICE, GUIDE TO THE GENERAL DATA PROTECTION REGULATION (GDPR) 93, 160 (2019).

66 *See, e.g.*, Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 243, 244 (2017) ("[I]n the European Union a sort of 'right to explanation' might be observed within the [GDPR] in the combination of several dispositions.").

67 *See* Federal Trade Commission, Comment Letter to National Telecommunications and Information Administration on Developing the Administration's Approach to Consumer Privacy 2–3 (Nov. 9, 2018), https://www.ftc.gov/system/files/documents/advocacy_documents/ftc-staff-comment-ntia-developing-administrations-approach-consumer-privacy/p195400_ftc_comment_to_ntia_112018.pdf [https://perma.cc/QK6M-9LY4] (recommending a requirement to explain the underlying logic of automated decisions).

68 Paul Nemitz, *Constitutional Democracy and Technology in the Age of Artificial Intelligence*, 376 PHIL. TRANSACTIONS ROYAL SOC'Y A, Nov. 28, 2018, at 1, 13. Paul Nemitz was the Director responsible for Fundamental rights and Union citizenship within the European Commission's Directorate-General for Justice and Consumers. *Paul Nemitz*, COLLEGE OF EUROPE, https://www.coleurope.eu/whoswho/person/paul.nemitz [https://perma.cc/7RZJ-Q6UL]. He is described as "one of the architects of the EU's far-reaching General Data Protection Regulation." Paul Chadwick, *To Regulate AI We Need New Laws, Not Just a Code of Ethics*, GUARDIAN (Oct. 28, 2018, 4:58 PM) https://www.theguardian.com/commentisfree/2018/oct/28/regulate-ai-new-laws-code-of-ethics-technology-power [https://perma.cc/P4B5-BSA5].

69 Council Regulation 2016/679, *supra* note 3, at 14.

70 *Id.* arts. 13–15, 21 & 22.

Articles 13–15 oblige data controllers to supply certain categories of information to data subjects[71] in certain situations.[72] One category of information, expressed in terms common to all three articles, concerns the existence of automated decision-making in the data controller's operations. If the data controller uses "automated decision-making, including profiling, referred to in Article 22(1) and (4)," then, "at least in those cases," the data controller must supply "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[73]

The syntax of these common terms on automated decision-making, and of Articles 13–15 as a whole, at least in their English language version, may leave something to be desired. Be that as it may, it is clear enough that Articles 13–15 are intended to place an obligation on data controllers, under certain situations at any rate, to say something to laypersons—the data subjects—about automated decision-making. The content of what data controllers are obliged to say may be described like this:

(i) whether automated decision-making is being employed;

(ii) "meaningful information about the logic involved";

(iii) meaningful information about "the significance . . . of such processing for the data subject"; and

(iv) meaningful information about "the envisaged consequences of such processing for the data subject."[74]

It could be that the phrase "meaningful information about . . . the significance and the envisaged consequences of" is better read as denoting only one element as to which "meaningful information" must be supplied, not two. Interpreting the phrase, however, one would be expected to give effect to both "significance" and "envisaged consequences." On balance, the four points set out above seem a plausible summary of the provision.

Each of the three articles, Articles 13–15, indicates a different situation in which the data controller is to provide the information described in the common terms. Article 13 indicates the situation in which the data controller has collected personal data from the data subject. It provides that "the [data]

71 A "data subject" is an "identified or identifiable natural person" to whom personal data relates. *Id.* art. 4(1). An "identifiable natural person" is one who "can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." *Id.*

72 *Id.* arts. 13–15.

73 *Id.* arts. 13(2)(f), 14(2)(g) & 15(1)(h).

74 *Id.* art. 13(2)(f).

controller shall, at the time when personal data are obtained, provide the data subject with . . . information necessary to ensure fair and transparent processing" (including the information regarding automated decision-making).[75] Article 14 indicates the situation in which the data controller has *not* obtained personal data from the data subject; it too provides that "the [data] controller shall provide the data subject with . . . information necessary to ensure fair and transparent processing" (including the information regarding automated decision-making).[76] Article 15 provides that "[t]he data subject shall have the right to obtain from the [data] controller confirmation as to whether or not personal data concerning him or her are being processed," and where they are, it further provides that the data subject shall have "access to the personal data" and other information (including, again, the information regarding automated decision-making).[77]

Article 21, "Right to object," specifies situations in which the data subject has a right to object to the processing of personal data concerning him or her.[78] It would seem that "processing" for purposes of Article 21 is a category that encompasses but is broader than "automated decision-making." It thus also would seem that using data for purposes of "automated decision-making" is *a fortiori* covered by the Article 21 right.

Article 22, to which the common terms regarding automated decision-making cross-reference,[79] provides in pertinent part as follows:

> **Article 22**
>
> **Automated individual decision-making, including profiling**
>
> 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
>
> 2. Paragraph 1 shall not apply if the decision:
>
> (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>
> (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>
> (c) is based on the data subject's explicit consent.

75 *Id.* art. 13(2).

76 *Id.* art. 14(2).

77 *Id.* art. 15(1).

78 *See id.* art. 21.

79 *See, e.g.*, *id.* art. 13(2)(f).

. . . .

> 4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.[80]

Article 22 thus provides data subjects with a right not to be subject to certain forms of data processing though that right itself is subject to exceptions as set out in paragraph 2. Article 22, paragraphs 1 and 4, serve, by their cross-references in Articles 13, 14, and 15, to identify certain kinds of "automated decision-making" that trigger the obligation on the data controller to supply "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[81] The phrase "at least in those cases," which modifies the obligatory clause in Articles 13, 14, and 15,[82] suggests that *other* cases, too, besides those identified in Article 22, paragraphs 1 and 4, might entail the same or similar obligation.

Article 9(1) prohibits:

> [p]rocessing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation.[83]

The specification of "data revealing racial or ethnic origin, political opinions, [etc.]" recalls the earlier provisions, including the Data Protection Convention of 1981,[84] addressing privacy as such. A continuity in adopted legal texts thus may be identified from Article 6 of the Data Protection Convention of 1981 and European Court of Human Rights ("ECtHR") judgments applying Article 8 of the European Convention[85] to the GDPR.

Points (a) and (g) of Article 9(2) set out some of the exceptional circumstances in which the categories identified in Article 9(1) may be used notwithstanding the presumptive prohibition against their use. These are where (a) "the data subject has given explicit consent to the processing of

80 *Id.* art. 22.

81 *Id.* arts. 13(2)(f), 14(2)(g) & 15(1)(h).

82 *Id.*

83 *Id.* art. 9(1). The prohibition is subject to ten exceptions specified under Article 9, paragraph (2). *Id.* art. 9(2).

84 Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, *supra* note 35, at 68.

85 *E.g.*, S. & Marper v. United Kingdom, 2008-V Eur. Ct. H.R. 167, 193–94.

those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject" (the subordinate clause here being an exception to the exception—i.e., the availability of the exception may be curtailed by law);[86] and

> (g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.[87]

In short, there are situations in which even the most sensitive personal data may be used in automated processing, but such situations are subject to protections. Article 22(1) indicates a "right not to be subject to . . . automated processing" when "a decision based solely on automated processing . . . produces legal effects concerning" the data subject.[88] It would appear, however, that the right is not limited to decisions producing legal effects. It extends as well to decisions which "similarly significantly affect[]" the data subject,[89] a point to which this Article returns.

In addition to the operative provisions of the GDPR that address automated decision-making and data processing, there is GDPR Recital 71. Recital 71, like the GDPR as a whole, contains no reference to "explainability" as such. Recital 71, as noted, does contain the word "explanation"; it is the only part of the GDPR to use that word:

> The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. . . . In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, *to obtain an explanation of the decision reached after such assessment* and to challenge the decision.[90]

86 Council Regulation 2016/679, *supra* note 3, art. 9(2)(a).

87 *Id.* art. 9(2)(g).

88 *Id.* art. 22(1).

89 *Id.*

90 *Id.* at 14 (emphasis added).

This passage is part of a long recital. Recital 71 runs to 437 words in its English language version.[91] That it bears some relation to Article 22, paragraph 1 is clear. The phrase in Article 22(1) "right not to be subject to a decision based solely on automated processing" relates to the statement in Recital 71 that "[t]he data subject should have the right not to be subject to a decision . . . which is based solely on automated processing."[92] It is clear as well that the "suitable safeguards" to which Recital 71 refers bear some link to a number of the operative provisions of the Regulation concerning safeguards.[93]

Recital 71 relates closely enough as well to Article 13–15, where those provisions require the data controller to provide "meaningful information" to the data subject whose data is involved in automated decision-making.[94] Where a decision "is based solely on automated processing and" it "produces legal effects concerning him or her or similarly significantly affects him or her," Recital 71 says that the data subject has a right to receive from the data controller, *inter alia*, "specific information" and "an explanation of the decision reached."[95] The expression "meaningful information" in the common terms of Article 13–15 would appear to correspond to the expression "specific information" in Recital 71.

Although the existence of a link between Recital 71 and Article 22, paragraph 1 is understandable, how precisely they relate is far from clear. One question is temporal. The recital lays emphasis on "explanation of the decision reached *after* such assessment."[96] The recital, in its emphasis on a decision already reached, suggests that explainability concerns challenges to a decision already reached; it might also suggest that explainability does *not* concern challenges *before* a decision is reached. It thus is not entirely clear, on the text, whether explainability entails a right, preemptively, to require the data controller to refrain from using an automated process. The proper interpretation on this point seems to turn on the meaning of the word "decision" as used in Article 22(1) and accompanying recital. It could be that the right indicated in the clause "[t]he data subject shall have the right not to be subject to a decision based solely on automated processing"[97] is respected if the data controller refrains from taking any action based solely on automated processing that practically affects the data subject. The "decision"

91 *Id.*

92 *Id.* at 14, 46.

93 *E.g.*, *id.* at 14, 37–43.

94 *Id.* at 41–43.

95 *Id.* at 14.

96 *Id.* (emphasis added).

97 *Id.* at 46.

in the provision, on that interpretation, is a decision having practical impact on the data subject, not a "decision" merely in the sense of a result produced by means of an automated process. If "decision," instead, means the latter—i.e., a result produced by means of an automated process, regardless of what the data controller does with the result—then the right articulated in Article 22(1) has a wider compass. It entails that the data controller is obliged to refrain from using the automated process if the data subject has objected.

The "explanation of the decision" to which Recital 71 refers is an explanation sufficient to enable the data subject to "challenge the decision."[98] For one thing, the two concepts are set out together in one sentence, a positioning which suggests that the "explanation" and the "challenge" are not to be read in isolation from one another.[99]

Moreover, a parsimonious interpretation of "explanation" would tend to deny effect to the data subject's right to bring a challenge. The emphasis placed in the complex structure of Articles 9 and 22 upon protecting the data subject from the misuse of a range of sensitive special categories of personal data, plus the general protection against use of *any* personal data,[100] supports the view that the data subject's right to challenge is not to be construed narrowly. This is a legal architecture strongly evincing the purpose of protecting the data subject's rights. As such, it opens the door, on its substantive terms, to a wide range of possible claims by data subjects alleging its breach.

II. COMPLIANCE AND LITIGATION

New regulations spur those whom they regulate to build new processes for compliance.[101] A great deal of valuable comment and analysis has focused on what the GDPR means, including especially its explainability provisions, as regards compliance practice.[102] But compliance considerations are only half the story of a new regulation—the other half is told through litigation and court rulings. Compliance has two primary goals, which are in answer to these two questions: Can I reduce the risk that my system will lead to litigation under the GDPR, and, if it does come to litigation, can I make sure I will not be shown up in court? A further line of questions identifies

98 *See id.* at 14.

99 *See id.*

100 *See id.* arts. 9(1) & 22(1).

101 *Cf.* Mojsilovic, *supra* note 19 (describing IBM's "comprehensive new set of trust and transparency capabilities for AI on the IBM Cloud"); ALLEN & OVERY, *supra* note 57 (observing that the GDPR requires significant effort from "[m]any companies," and "for some, too much").

102 *E.g.*, INFO. COMM'R'S OFFICE, *supra* note 65, at 154, 176.

the goals of compliance more concretely: Can I explain to my CEO that this system is unlikely to lead to litigation under the GDPR, and if it does, that we are likely to win? Can I explain to my insurer that there is minimal risk of a large fine? Can I explain to a regulatory authority that my system is fair, to save me from regulatory action? These are all prospective explanations concerning the *general case*—i.e., with the overall state of a system that the data controller has put in place. Litigation on the other hand is concerned retrospectively with the *specific case*: Can I persuade the court that the specific explanation I gave to the specific data subject who has sued me satisfies the GDPR?

Compliance and litigation thus are closely linked. Compliance involves predicting what might happen in litigation. Where the meaning of terms of a regulatory rule are uncertain, compliance guidelines tend to be more speculative. This is especially true when the rules are new, as is the GDPR, or when the technologies are new, as is modern machine learning. As a body of court practice develops, compliance officers will understand better what the rules mean in practice, and so they will be able to give more precise guidance. Conversely, courts are likely to pay attention to good industry practice and to guidance that experts have developed.

Compliance and litigation are not, however, the same thing. They have different methods, and they have different audiences. To put it crudely, hammering out compliance solutions is not the same thing as preparing for litigation. The former no doubt will consume vast resources[103] and thus present attractive opportunities for professionals who know about machine learning and the GDPR. However, it will be in litigation that the GDPR (and probably any similar legislation in a rule of law society) acquires clearer form.

This Article's focus is litigation. But, given the close connection between compliance and litigation, some further observations are in order about methodologies for compliance and about the role of jurisprudence in defining the compliance landscape. As will be seen, compliance is a process of designing systems to attain certain goals, and compliance is related to litigation in that compliance systems aim to avoid litigation and, in turn, adapt themselves in light of the results of litigation. However, the methodologies of compliance—in essence, those of an engineer working to build a stable and reliable system—are very different from those of litigation—in essence, those of an advocate seeking to win the case.

103 *E.g.*, ALLEN & OVERY, *supra* note 57.

A. The Compliance Approach to Explanation and Accountability

As an illustration of the buzz around compliance and the GDPR, consider this 2019–2020 syllabus for a new master's degree course, "Technology, law and society," at a leading computer science department:

> Data-driven technologies are increasingly the subject of social commentary, political scrutiny and regulatory attention. This module aims to develop a solid understanding of the practical implications these concerns have on systems design and engineering. Areas explored include the legal foundations in data protection (GDPR), privacy, liability, human rights; issues of tech-surveillance and algorithmic accountability; and the related implications for technologies including cloud, machine learning and the IoT.
>
> This course provides students with a practical background regarding how law, policy and societal concerns interact with technology. This is to develop an awareness and consideration of how systems can be designed and engineered to support accountability, be more legally compliant, and generally better for society.[104]

This syllabus highlights that the GDPR has drawn attention to algorithmic accountability, surveillance, and the social and political relevance of data-driven technologies. But the syllabus is interesting for more than its subject matter. Look at the syllabus again, and ask, what *tools* does it propose for addressing the substantive issues? It proposes "systems design and engineering."[105] Implicit here is the proposition that if one uses the proper engineering processes to design a system, then one will comply with the GDPR.

The relationship between the law and systems engineering has been ably developed in recent writings, not least those of Joshua Kroll and his collaborators.[106] They describe the practices of good engineering, such as writing reproducible code and running appropriate audits, to ensure that an engineered system fulfils its design goals.[107] This is, in essence, a compliance approach. It is about readying operators to build systems that have the best chance of avoiding regulatory action or other response by public authority. It is about building systems that behave reliably and soundly, to minimize the chance that a disgruntled data subject makes a complaint about an unfair decision. The systems approach to explainability is central to the

---

104 *Course Pages 2019-20: Technology, Law, and Society*, U. CAMBRIDGE, https://www.cl.cam.ac.uk/teaching/1920/R260/ [https://perma.cc/E8SD-ENUE].

105 *Id.*

106 *See generally* Kroll et al., *supra* note 11.

107 *Id.* at 660–65, 670.

understanding Kroll and others bring to the topic: they are concerned with developing explanations of the overall system[108] that are suitable for a regulatory audience. Management too needs to be given an explanation of a system and of its risk, which it can balance against business objectives.[109] There is much that is useful in the on-going dialogue among academics, regulators, and technical experts who serve regulated parties about compliance. A thoughtful approach to accountability, in the frame of regulations such as the GDPR, is called for, and the demand for the design and engineering of reliable compliance systems is only going to grow as machine learning's applications spread.

In later work, Kroll develops the theme that accountability is much broader than a technological task of implementing a design: "Responsibility and ethics attach not to the specifics of a technical tool, but rather to the ways that tool is used in a sociotechnical context."[110] A system is not engineered in a vacuum. The engineering process is carried out by people with agency who build a system to operate in a given environment. Kroll goes on to say that "explanations must speak to the decisions made during the design of a computer system, as such information is always available and always fulfils the key requirements of a meaningful explanation."[111]

Kroll is not referring to "meaningful explanation" as a term of art in the sense of Articles 13–15 of the GDPR; it would be presumptuous to declare a settled interpretation of substantive legislation before the courts have spoken on the matter. Indeed, any GDPR defense along these lines would quickly fall apart, as this Article discusses in Section V.B. The value of Kroll's explanations, i.e., accounts of how an engineered system *came to be*, is to regulators and to society at large[112] because technology companies should not be allowed to hide their practices behind a veil of technological inscrutability.

108 *Id.* at 633–34.

109 Management may fairly ask what costs compliance approaches impose. In the setting of explainability, Kroll says, "We can distinguish well-governed development processes from unconstrained tinkering." Joshua A. Kroll, *The Fallacy of Inscrutability*, 376 PHIL. TRANSACTIONS ROYAL SOC'Y A, Nov. 28 2018, at 1, 6. What Kroll seems to have in mind here are engineering processes, in which best practices, rules, and ongoing oversight apply to the developer. Not all developers are in an equal position to carry the cost of such "well-governed . . . processes," however. An under-explored area is the fairness of complex regulatory regimes, such as the GDPR.

110 Kroll, *supra* note 109, at 2.

111 *Id.* at 7.

112 Kroll et al., *supra* note 11, at 634.

B. In Litigation, the Data Controller is the Defendant, Not the Judge

A company's compliance department has two faces. One face is inwards, issuing rules to engineers and other employees about how they must conduct themselves in order to meet regulatory requirements, including in particular what sort of explanations they must give to data subjects. The other face is outwards, in court, when they must defend themselves to a judge against a data subject's claims. Litigation is not a compliance operation where reasonable measures or due diligence might suffice, and where the data controller and compliance consultants are the ones making the detailed rules. In litigation, the data controller is the defendant, not the rule-maker, and certainly not the judge.

The GDPR as well as privacy rules preceding it embody requirements that data controllers put and keep in place a range of self-regulatory measures.[113] Even if those requirements did not exist as such, individuals, government bodies, and businesses that hold data have an interest in avoiding claims for breach of regulatory rules such as those contained in the GDPR. Thus, both in order to accord with the law and in order to avoid other breaches of the law, data controllers invest considerable resources in seeking to assure their own compliance.[114] The adoption of the GDPR set in train a significant effort by lawyers and compliance specialists, who sought to prepare institutions to apply its rules. Large volumes of comment and analysis have been published in relation to GDPR compliance.[115]

"Start by getting technical," says one informed writer about GDPR compliance.[116] The writer expands on what it means to "get technical":

> You'll want to get a technical description of the model and the data it was trained on, which you can work off to build your data subject-friendly explanation. . . . You'll also want to get a basic understanding of the data the model is trained on. Where did the

113 *See generally* Council Regulation 2016/679, *supra* note 3; Council Directive 95/46, *supra* note 39.

114 *See, e.g.*, ALLEN & OVERY, *supra* note 57.

115 *E.g.*, INFO. COMM'R'S OFFICE, *supra* note 65; *General Data Protection Regulation*, ORRICK HERRINGTON & SUTCLIFFE LLP, https://www.orrick.com/Practices/GDPR-Readiness [https://perma.cc/G5CC-7XHH]; SHEARMAN & STERLING LLP, THE EU GENERAL DATA PROTECTION REGULATION 1–3 (2017), https://www.shearman.com/-/media/Files/Perspectives/2016/08/GDPR-Briefing-and-Preparation-Checklist-December-2017.pdf?la=en&hash=DB05E549A079B1BDBC739D1CEBD9CE262121515C [https://perma.cc/HFM8-KPQA].

116 Andrew Burt, *Is There a "Right to Explanation" For Machine Learning in the GDPR?*, INT'L ASS'N PRIVACY PROFS. (June 1, 2017), https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr/ [https://perma.cc/8K3K-2FXG].

> data come from, for example? How many features does the model select for? Discuss all of this with your technical experts.[117]

This level of generality sounds like it might suffice. And it might if the data controller's "subject-friendly explanation" is delivered to a data-controller-friendly subject. The data subject, however, is probably not the data controller's friend.

The compliance measures that data controllers implement are self-regulatory responses, in the sense that they are adopted, at least in the typical situation, unilaterally by regulated organizations.[118] However, when a data subject contends that a data controller's compliance measures have failed, it will not suffice for the data controller merely to advert to its own measures. The data subject, if identifying and resorting to a judicial procedure, will challenge the data controller and argue that the failure of the compliance measures constitutes, or has resulted in, an injury to the data subject. When that happens, the data controller is not the judge of her own explanations.

Nor is the data controller the judge of what the GDPR requires. The meaning of each term of the GDPR relevant to explainability will be a potential subject of dispute. True, disputes over the meaning of the

---

117 *Id.*

118 A compliance measure or system might be required under an agreement between the data controller and a regulator, in which case it would not be unilateral action by the former but, instead, a joint approach, or even an imposition by the latter. Compliance systems have been required under a variety of other regulatory regimes. *See, e.g.*, 100Reporters LLC v. U.S. Dep't of Justice, 248 F. Supp. 3d 115, 125 (D.D.C. 2017) ("[A] plea agreement resolving [a] criminal case [against Siemens A.G.] required the company to hire an independent corporate compliance monitor to ensure that Siemens implemented an effective corporate governance system and complied with all applicable laws and regulations."). Stipulations that such compliance systems be monitored by a third party are a noteworthy feature of compliance in such circumstances. *See* Sarah Paul, Olga Greenberg & Andrea Gordon, *How to Avoid and 'Survive' the Dreaded Monitorship*, LAW.COM (JAN. 10, 2020, 2:20 PM), https://www.law.com/newyorklawjournal/2020/01/10/how-to-avoid-and-survive-the-dreaded-monitorship/?slreturn=20200428221824 [https://perma.cc/L3V7-WCWK] (noting the common use of "independent compliance monitor[s]" by the U.S. Department of Justice for "corporate resolutions in Foreign Corrupt Practices Act . . . cases"); Cristie Ford & David Hess, *Can Corporate Monitorships Improve Corporate Compliance?*, 34 J. CORP. L. 679, 682 (2009) (describing the novelty and current use of "corporate monitorships"). However, at the outset, even where elements of a compliance program are more or less precisely stipulated in law, a compliance program is typically the work of the regulated person. Thus, for example, as described by the Export-Import Bank of the United States in its notes concerning the Foreign Corrupt Practices Act (FCPA), 15 U.S.C. § 78dd (2018), "many firms have implemented detailed compliance programs intended to prevent and to detect any improper payments by employees and by third-party agents." *Foreign Corrupt Practices and Other Anti-bribery Measures*, EXPORT-IMPORT BANK OF THE UNITED STATES, https://www.exim.gov/policies/foreign-corrupt-practices-and-other-anti-bribery-measures [https://perma.cc/643E-C5LU].

Regulation may subside in time, as courts clarify disputed provisions. However, even after the meaning of the relevant terms is largely settled, disputes as to whether the data controller has complied with those terms will remain. It is clear under Articles 13–15 that the data controller does not have unlimited discretion as to what scope and content of information to provide. The adjective "meaningful," modifying the term "information" in Articles 13–15,[119] is part of an obligation under law, and it entails a legal standard that the data controller must meet if the data controller is to satisfy the obligation.[120] This much is doubtless the case: When it comes to litigation, the data controller will not be the judge of the meaning of the terms to which it is obliged to comply—and, regardless of their meaning, the data controller will not be the judge of its compliance with those terms.

It is likely that data controllers, as defendants, will start by echoing the engineering systems approach that prevails in compliance operations. They will supply partial or limited explanations setting out how they built the systems that produced the challenged decision. They will not start by disclosing protected, private information contained in the data that they used to train a machine learning system. That sort of explanation—an explanation that does not involve exposing the data set—indeed might be accepted without challenge or test if it was supplied by an actor who, whether in fact or by law, is trusted. But the data controller, when a data subject litigates against him, is not a trusted source in that sense. In litigation, the claimant will insist upon challenging and testing. A partial or limited explanation, similarly, might satisfy a regulatory requirement setting standards and rules for compliance programs. But again, the claimant in this hypothetical case is asking for the "meaningful information" to which the GDPR entitles her, not for a demonstration that engineering best practices happen to have been followed or a compliance requirement met. As this Article argues below, litigation, unlike the systems approach in compliance, will impel ever more exacting demands to see the data.

C. Court Decisions Will Define the Interpretation of the GDPR's Substantive Rules

There is a further salient difference between compliance and litigation. Compliance operations do not result in binding interpretations of the law;

119 Council Regulation 2016/679, *supra* note 3, arts. 13(2)(f), 14(2)(g), 15(1)(b).

120 Analogously, a treaty under which a State retains discretion "for purposes of scientific research" to kill whales restricts the State's discretion, which it would not have if it provided simply that a State may kill whales, because "whether the killing . . . of whales . . . is for purposes of scientific research cannot depend simply on that State's perception." Whaling in the Antarctic (Austl. v. Japan), Judgment, 2014 I.C.J. 226, ¶ 61 (Mar. 31).

litigation sometimes does. Courts will reach decisions, and these will become the precedents that define how the GDPR is to be interpreted.

Rosemary Jay, the head of the legal office of the UK's data protection regulator for twelve years,[121] in a panel discussion, makes the point that courts have pushed the law forward:

> When you look at something like the GDPR, which is our current new regulation, it's not a done deal, it's not a final frontier. . . . What actually has been created are some mechanisms to move forward, to keep changing, and to keep developing. [The moderator interjects: Because you've got to keep changing, because the technology keeps changing.] Completely and absolutely. And what we have seen as well, the way the courts have dealt with this, has been a willingness by the courts to actually push the law forwards. So if we think about a case called *Costeja*, which was a gentleman who wanted his name taken off search results, his case went through, he succeeded—that brought a huge change. So that was based on a statute, on statutory law, legislation, but the court pushed it forward to take on a new dimension. So I think you don't just look at the law as though it's a static thing, you say "Where's it going to take us . . . ?"[122]

The case to which Jay referred is *Google Inc.* v. *Costeja González*,[123] sometimes called the "right to be forgotten" case.[124] The CJEU in that case placed its interpretation on Directive 95/46/EC (noted above in Part I.A). Google, arguing before the court, sought to narrow the material scope of its obligations under Directive 95/46.[125] The court, in its judgment, did not narrow them. It widened them, considerably. For example, Google was not insulated from its obligation under the Directive because it had gotten the

---

121 *Unreliable Evidence*, BBC, at 1:06–1:20 (Aug. 28, 2019), https://www.bbc.co.uk/sounds/play/m0007wfx [https://perma.cc/24LZ-VGJS].

122 *Id.* at 8:22–9:19.

123 Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos, Judgment (May 13, 2014), http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=12401842.

124 Julia Powles, *What Did the Media Miss with the 'Right to be Forgotten' Coverage?*, GUARDIAN (May 21, 2014, 7:08 AM), https://www.theguardian.com/technology/2014/may/21/what-did-the-media-miss-with-the-right-to-be-forgotten-coverage [https://perma.cc/HFG6-BFA3].

125 Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos, Judgment (May 13, 2014), http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=12401842.

claimant's data from the internet, where it had already been published.[126] Nor was it insulated because it did not alter the personal data.[127] The court interpreted the term "data controller" broadly too:

> [I]t would be contrary not only to the clear wording of that provision but also to its objective—which is to ensure, through a broad definition of the concept of "controller," effective and complete protection of data subjects—to exclude the operator of a search engine from that definition on the ground that it does not exercise control over the personal data published on the web pages of third parties.[128]

The judgment impelled a rapid response by data controllers, including Google, who up to that time had not generally maintained systems to receive and implement requests for the removal of personal data on the scale that the judgment instigated.[129]

D. Litigation Will Expose the Tension Between Explanation and Privacy

The lesson to be drawn from the "right to be forgotten" case just noted above is that jurisprudence affects compliance architecture, sometimes in unexpected ways. There is not yet a body of case law concerning the explainability provisions of the GDPR that clarifies the interpretation of those provisions. The GDPR, like any new legislation, will spawn new disputes. Some of these disputes will turn into legal claims and then proceedings, and courts will be asked to hand down judgements. The GDPR's substantive rules will thus be *tested in an adversarial setting*. The interpretation of the GDPR's rules will not simply be handed down by well-meaning regulators nor by experts from university departments of machine learning: it will be fought over by litigious lawyers. Considering how the litigation process itself might play out reveals a basic tension in the GDPR's provisions. This is a tension between explainability and privacy, and it is inherent in the GDPR. It is a tension that might be smoothed over in a purely technocratic regulatory system, but in a court system with well-established rules of evidence, it will emerge in sharp relief.

* * *

126 *Id.* ¶ 37.

127 *Id.* ¶ 31.

128 *Id.* ¶ 34.

129 *See, e.g.*, Rose Powell, *Google Receives 12,000 Requests to be 'Forgotten' on First Day*, SYDNEY MORNING-HERALD (June 1, 2014, 8:27 AM), https://www.smh.com.au/technology/google-receives-12000-requests-to-be-forgotten-on-first-day-20140601-zru3g.html [https://perma.cc/9VYX-92RV].

The following three Parts consider in detail how litigation under the GDPR's provisions for explainability might proceed. The starting point will be a claim: the data subject will make a claim that the data controller did not provide "meaningful information" about an automated decision which affected her. In Parts III and IV, this Article considers from two points of view what would constitute meaningful information: first from the point of view of what the data subject wants the information for and second from the point of view of the technology of machine learning. Then in Part V, this Article considers the data controller's defenses. This Article explains why disputes over explainability, defenses notwithstanding, seem destined to impel ever more exposure of the data that trained the machine—in short, why one can have explainability, or one can have privacy, but in the machine learning age, it is far from obvious that one can have both.

III. THE DATA SUBJECT'S RIGHT TO A MEANINGFUL EXPLANATION: WHAT IS IT FOR?

The GDPR specifies rights *of the data subject*,[130] which means a serious source of legal risk will be private claims. This is as opposed to, e.g., filing requirements, where the main source of legal risk would be action by a public regulator. This Article is not chiefly concerned here with investigations by a public regulator, which well may have a policy of seeking settlement, though to be sure public regulators too sometimes litigate. The concern here is with litigation and all the idiosyncrasies of the many potential claimants who might sue. This is not an area of regulation in which the sole, or even the main, source of legal risk will be action by a public regulator. The source of possible legal challenges is much more widespread.

The right that concerns this Article here is that stated in Articles 13–15: for systems with automated decision-making, the data controller must provide "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."[131] The wide scope of the "envisaged consequences" is made clear in Article 22(1), which is expressly incorporated by reference into Articles 13–15:[132] they comprise not only any legal effects that automated decision-making might exercise upon the subject, but also any other consequence that "similarly significantly affects him or her."[133]

130 Council Regulation 2016/679, *supra* note 3, at 1.

131 *Id.* arts. 13(2)(f), 14(2)(g) & 15(1)(h).

132 *Id.*

133 *Id.* art. 22(1). That the inclusion of such a wider class may have consequences is illustrated by the *a contrario* situation—the situation where a legal provision is limited to *legal* effects or interests. Article 6(2)(b) of the UN Convention on Jurisdictional Immunities

Assuming that the legal remedies available for breach of the GDPR cover all the rights and obligations established by the GDPR,[134] then the remedies that a claimant might pursue are for a large class of injuries.

An explanation that is useless to the data subject could hardly be described as "meaningful information." To understand the scope of the term, it is therefore helpful to consider *how a data subject might wish to use an explanation* from the perspective of envisaged consequences of automatic decision-making. An ongoing dispute between Uber and four Uber drivers provides a useful illustration. The drivers have requested information from Uber and did not receive a satisfactory response, so they threatened action under Article 15 of the GDPR.[135] According to their press release,[136] they have requested:[137] information about Uber's system for maintaining profiles of drivers, e.g., labelling drivers with tags for "inappropriate behaviour" or "missed eta"; an "explanation of how personal data is used" for dispatching trip requests to drivers; data about individual trip ratings by passengers (Uber will suspend drivers if their average ratings fall too low, but drivers do not know which trips resulted in which ratings, so they might be "suspended and fired at will without due process, right of an appeal or even an adequate explanation"); and "[t]he complete set of GPS trace data" for those drivers, including data about when they were "logged on [to the platform] and

of States and Their Property concerned that situation. G.A. Res. 59/38 (Dec. 16, 2004); *see also* Belhaj v. Straw, Rahmatullah (No 1) v. Ministry of Defence [2017] UKSC 3 [26] ("[A]cademic commentators have concluded that any uncertainty in [article 6(2)(b)'s] scope should be addressed by recognizing that 'interests should be limited to a claim for which there is some legal foundation and not merely to some political or moral concern of the State in the proceedings.'" (citation omitted)); Belhaj v. Straw [2014] EWCA (Civ) 1394 [45] (Eng.) (limiting Article 6(2)(b) to "'interests' of states to legal interests").

134 The right to an effective judicial remedy against a controller or processor is provided in Article 79 of the GDPR. Under Article 79(1), the remedy extends to "where [the data subject] considers that his or her rights under this Regulation have been infringed as a result of the processing of his or her personal data in non-compliance with this Regulation." Council Regulation 2016/679, *supra* note 3, art. 79(1). The limitive terms there are "as a result of the processing" and "non-compliance." The latter ("non-compliance") is not much of a limit, because whether there has been non-compliance will be the subject of litigation in most cases; it would only be exceptionally, if at all, that absence of non-compliance would be determinable as a preliminary matter. There is no limitation as to particular "rights under this Regulation."

135 Dena Tyrrell, *Uber Again Fails to Provide Drivers Access to Their Personal Data*, WORKER INFO EXCHANGE (May 7, 2019), https://workerinfoexchange.org/index.php/2019/08/29/uber-again-fails-to-provide-drivers-access-to-their-personal-data/ [https://perma.cc/PB5S-RDPJ].

136 *Id.*

137 The Uber drivers are also demanding data specific to them, and the dispute also touches on the regulatory question of whether the drivers are employees or contractors, but for present purposes this Article is only concerned with the aspect of explanation. *Id.*

waiting for work [or] en route to" a pickup (such information could be useful for calculations about minimum wage).[138] Two broad motives can be read into this request for explanation: to seek guidance and to seek redress.[139] This Article now considers each, and then considers a third societal impetus behind explanation: validating decisions for purposes of legitimacy in light of general public values.

A. Explanations that Give Guidance for Future Conduct

A layperson might seek an explanation in order to obtain guidance about how to behave.[140] For example, drivers might want to know how Uber's profiling system uses driver profiles, so that they can take steps to improve the quality and value of the trips that are offered to them. Or, they might want to know which specific passengers gave them low ratings, so they can make an informed guess about the specific types of driver behavior that are likely to result in low ratings. In its *Guidelines on Automated Individual Decision-making*, the Article 29 Data Protection Working Party[141] gives another example of an insurance company giving tips to drivers, as an illustration of "meaningful information about . . . the significance and envisaged consequences."[142]

There are three assumptions behind this type of explaining. The first is that the data subject accepts that the decision-making process is legitimate, and that the onus is therefore on them to change their conduct. The second is that there is scope for the data subject to change their future conduct. The

138 *Id.*

139 *See* Wachter et al., *supra* note 18, at 843. Wachter, Mittelstadt, and Russel also categorize explanations "according to the specific goal or action they are intended to support." *Id.* They list seeking guidance, seeking redress, and "inform[ing] and help[ing] the subject understand why a particular decision was reached" as benefits of explanations. *Id.* This Article has not included this last motivation here, because it is unclear what action it is intended to support.

140 *Id.*

141 The Working Party was the independent advisory group set up in accordance with Article 29 of Directive 95/46/EC (Oct. 24, 1995) for the purpose, *inter alia*, of "mak[ing] recommendations on all matters relating to the protection of persons with regard to the processing of personal data in the Community." Council Directive 95/46, *supra* note 39, arts. 29, 30(3). It was "composed of a representative of the supervisory authority or authorities designated by each Member State and of a representative of the authority or authorities established for the [EU] institutions and bodies, and of a representative of the Commission." *Id.* art. 29(2). The Working Party functioned until the entry into force of the GDPR, at which time the European Data Protection Board established under the GDPR took its place. *See* Council Regulation 2016/679, *supra* note 3, arts. 68–76, 94.

142 Article 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-making and Profiling for Purpose of Regulation 2016/679*, at 26 (adopted Oct. 3, 2017; revised Feb. 6, 2018).

conduct in question might be repeated behavior, such as a driver who continues to work for Uber, or it might be further actions that follow on from the original decision, such as a house buyer who resubmits a mortgage application but for a lower value. The third is that the explanation must give enough information for the data subject to decide how to change their conduct.

Conversely, if the explanation does not allow the data subject to decide how to change their conduct, then it cannot be considered "meaningful information" in the sense of Articles 13–15. The precise meaning of "meaningful information" remains open to dispute (as one may expect of key phrases in new and as yet largely untested legislation). However, it is reasonable to say that the purpose of explanation under GDPR is not abstract; it is for a concrete purpose. Information that is useless to the concrete purpose for which it is provided is not "meaningful."

To illustrate the extremes of what might be considered meaningful information, here is a thought experiment. Suppose there is a layperson who wants to obtain a license to run a fish and chip shop, and an algorithm denies her the license. Suppose that circumstances mean that she is unable to change her application in any meaningful way. One might think in such a case that an explanation would not be useful. However, suppose that the decision-making process depends in some convoluted way on other circumstances outside her control, such as the price of North Sea fish futures and predictions about weather and footfall. A useful explanation in this case might be: "The decision is to all intents and purposes random, and the probability of the answer Yes is 25%." (Many machine learning algorithms are in fact probabilistic in nature, and the probability value might in fact be produced as an intermediate stage of the algorithm). In this case she might decide it is worth the effort of resubmitting the application as-is. Or she might be told "The probability of Yes is 0.001%," in which case she might decide not to resubmit. The point of this thought experiment is to stress the point that meaningful information about a decision-making algorithm does not entail step by step details of the operation of the algorithm, but it does require sufficient information to guide the data subject in their future conduct.

B. Explanations for the Purpose of Challenging a Decision

Recital 71 says that a data subject has the right "to obtain an explanation of the decision reached . . . and to challenge the decision."[143] As this Article

143 Council Regulation 2016/679, *supra* note 3, at 14.

noted above, the exact relation of these two clauses might well be disputed.[144] It should be common ground that the right to challenge a decision must not be illusory. Considering, moreover, that such a right to challenge and the right to obtain an explanation are closely related, this Article posits that adequacy of the explanation must be judged, at least in part, by whether it supports the data subject's right to challenge a decision in a given situation. A reasonable interpretation of Recital 71 follows from this: the data subject has the right to obtain an explanation, so she has meaningful information for the purposes of challenging the decision. To interpret the right to an explanation in this sense, three situations in which a challenge might be made must be understood. This Article highlights three purposes.

First, the data subject might believe that the decision was a mistake in the sense of a technical failure, for example a result of erroneous data entry. This might be difficult as a practical matter to address, but it is conceptually straightforward: something went wrong with the machine, and therefore the output is not valid. In this case, "meaningful information" would mean enough evidence to give confidence that the decision-making system is functioning as expected at each stage of processing, or, if it is not, where it has gone wrong.

Second, the data subject might believe that the decision was made on the basis of invidious grounds. This is a more subtle problem, and from the standpoint of public policy it is a more serious one. Public policy in many countries, and as reflected in some treaty provisions at the international level, lays down proscriptions, such as those against racial or gender discrimination.[145] For decisions made by elementary algorithms, it may be possible to detect invidious grounds by inspecting the algorithm's code: for example, the line of code "if gender=female then fishandchip_license_status=DENIED" is clear evidence of invidious grounds. But for more sophisticated algorithms, especially for machine learning algorithms, it is far harder to detect invidious grounds for a decision. It is almost inconceivable that the problem could be detected by inspecting the algorithm's logic, as implausible as being able to detect sexist patterns of

144 Wachter et al., *supra* note 18, at 874 ("An explicit link is not established in the GDPR between the right to explanation and the right to contest, wherein the former would provide information necessary to exercise the latter."). Whether or not courts read Recital 71's *implicit* link into the "meaningful information" of Articles 13–15 remains to be seen.

145 *E.g.*, Council Regulation 2016/679, *supra* note 3, art. 9 (prohibiting "[p]rocessing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation"); *see also id.* at 14 (prohibiting many kinds of "profiling").

thought by inspecting the wiring of neurons in a human brain. In Part IV, this Article discusses further what form of explanation might be usable for detecting invidious grounds in machine learning algorithms and argues that it is likely to consist of data about other data subjects.

The third situation in which a challenge might be made is when the algorithm is legitimate but inaccurate. For example, consider a mortgage decision algorithm that does not take account of the amount of money requested. Such an algorithm would be foolish on the part of the bank, and so one would not expect to see it in practice. However, it is not intrinsically problematic from the point of view of public policy. In this situation it is easy to see that this particular feature of the input data is relevant for the decision, but in practice there may be a multitude of possible features and it can be very hard to detect which are relevant, and so an algorithm might well have lacunae. The machine learning procedure attempts to differentiate among relevant features and irrelevant accidents of the training dataset. If it errs by including too many features then it tends to perform poorly on novel cases, and if it errs by including too few features, then it has lacunae (such as the hypothetical omission in the mortgage decision process). The challenge here is known as the "bias-variance tradeoff." It is a grey area of machine learning, full of heuristics and good practices but still an area of active research.[146] It, too, is likely to be the locus of disputes about explanation. As with the second case discussed above, the meaningful information needed to explain the decision will consist of data about other data subjects.

Considering these situations in which explanation might be demanded, the data controller may find itself faced with a range of unpalatable outcomes. If the data controller discloses a mistake (as in the first situation) or if it discloses a failure to give a relevant feature appropriate weight (as in the third situation), then the data controller exposes itself to allegations of incompetence. Shareholder complaints might ensue against the company, or the professional reputation of the engineers or programmers might be impugned. If the data controller discloses that the decision was reached on invidious grounds—e.g., sex discrimination—then it opens itself to major complaints and liabilities on that basis. This is a tension that can be expected to surface as more challenges are brought against data controllers who use machine learning systems.

146 Even experts make mistakes. *See* TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING, at viii (2d ed. 2009) (rewriting the first edition's discussion of accuracy and noting its "discussion of error-rate estimation in Chapter 7 was sloppy, as we did not clearly differentiate the notions of conditional error rates (conditional on the training set) and unconditional rates").

C. *Explanation as a Source of Legitimacy*

Society at large—the public—will wish to be assured that an automated decision-making system is fair and trustworthy. The data controller equally will wish that its system be accepted as legitimate.

Explanation can confer legitimacy in two senses. First, the data controller might give an *account of the system* for the system's design and engineering such that the public is satisfied that the system is fair. This is the compliance concept of explanation, as described in Section II.A above. Second, the data controller can build a system that gives an *explanation for each decision* it makes. A system will gain acceptance in society if it provides explanations for each of its decisions, acceptable to the data subjects concerned. Acceptance of the decision-making system for this purpose need not be pleasure with a result; acquiescence in the result will do.

"Society at large" is not a data subject in the terms of the GDPR, and so it does not have the right to an explanation in either sense. However, if society at large, in the form of pressure groups, trades unions, or other organized bodies, does not accept that the decisions made by an automated system are fair and trustworthy, then it can channel its concern through individual data subjects. For example, the Uber drivers referred to above are being supported by Worker Info Exchange,[147] which describes itself as "a non profit organisation dedicated to helping workers access and gain insight from data collected from them at work usually by smartphone."[148] If society at large is not satisfied with the explanations it has been given, it may prompt concrete instances of litigation under the GDPR.

The value of explanation as a source of legitimacy is well recognized in another type of decision-making system: courts and other organs that apply the law. Explanations, as expressed in a judge's ruling, demonstrate that the outcome is not arbitrary or, worse, misshapen by malign influence at variance with the substantive rules of the legal system. This reason for giving reasons indeed has been directly associated with "legitimacy," which is what Lord Chief Justice Hewart seems to have had in mind when he admonished that "justice should not only be done, but should manifestly and undoubtedly be seen to be done."[149] What will not do is a widespread belief that judicial

147 Tyrrell, *supra* note 135.

148 WORKER INFO EXCHANGE, https://workerinfoexchange.org/ [https://perma.cc/SN3L-3ZA2].

149 R. v. Sussex Justices [1924] 1 KB 256 at 259 (Eng.); *see also* R. v. R.E.M. [2008] 3 S.C.R. 3, para. 11 (Can.) (discussing the functions of "reasons for judgment in a criminal trial"); *cf.* Liteky v. United States, 510 U.S. 540, 565 (1994) (Kennedy, J., concurring) ("In matters of ethics, appearance and reality often converge as one.").

decisions are unfair or otherwise invalid. Judicial explanations address society at large. The explanation serves the integrity of the decision-making process as a whole, not just particular parties testing the validity of a given decision. This is also a reason for explanation that jurists have long identified.[150]

In jurisprudence, explanations that are *accounts of the system* are the domain of scholars and academics, and they complement the *explanations for each decision* written by judges. Both types of explanation play a role in giving legitimacy to the legal system.

D. All of the Above

This Article has considered several reasons for wanting a meaningful explanation: in order to guide future conduct, to challenge individual decisions, and to challenge the legitimacy of the overall decision-making system.

Any party bringing a claim is likely to do so for a mixture of reasons, and the data subject will not know in advance which reason or reasons are involved. In the dispute between Uber drivers and Uber, for example, the drivers have simply asked for more information about how the dispatch algorithm makes decisions based on driver profiles.[151] Based on what Uber reveals, the drivers will decide whether it is a fair algorithm, and it will be up to them to adjust their conduct: whether there are particular instances where they were treated unfairly and for which they should seek redress, or whether there are grounds for claiming the algorithm has systematic deficiencies and must be discarded or fixed.

Such ambiguity is characteristic of a decision that lacks a meaningful explanation. A first step in litigation will be for the claimant to ask for enough information to discern which, if any, of the situations above is relevant. In other words, "We suspect that the decisions might be unfair, but the explanations we've had so far aren't good enough for us to be sure, so we'll go to court and demand a better explanation. If the explanation is convincing, at least we'll know what's going on. If not, we will challenge the decision."

150 *See, e.g.*, HERBERT BROOM, CONSTITUTIONAL LAW VIEWED IN RELATION TO COMMON LAW 147–48 (George L. Denman ed., 2d ed. 1885) (stating that reasons are "due to the suitors and to the community at large").

151 Tyrrell, *supra* note 135.

E. An Analogy with Legal Explanation

There is an analogy between explaining an algorithm's decision, and a judge explaining a court's decision: both types of explanation can confer legitimacy and maintain public confidence in the decision-making system, as suggested in Section III.C. "[J]ustice should not only be done, but should . . . be seen to be done."[152]

The analogy between machine learning algorithms for decision-making and legal decision-making can be extended further. A data subject may seek an explanation in order to guide their future conduct, as discussed in Section III.A. Likewise, one of the reasons that courts explain the decisions they reach is that the decisions, if stated without reasons, do not give the parties much information. The decisions, standing on their own, are not sufficient to guide the parties in adapting their behavior to the rules that the decisions have applied.[153] The parties whom the decisions specifically address seek the information in order to adapt their behavior to accord with the law or to achieve lawful results under regulatory regimes that concern them.

A data subject may seek an explanation for the purpose of contesting the validity of an algorithmic decision, either to seek remedy for a particular decision or to challenge the algorithm as a whole, as discussed in Section III.B. Likewise, in legal decision-making, there is an interest that the rules of public policy be faithfully observed, and a judge's explanation can provide a basis for challenging the decision. The parties might be concerned that a perfectly sound law has been misapplied; they might also be concerned that the law is unsound and thus needs to be re-considered entirely.

Bystanders—i.e., persons not specifically addressed by a given decision in a dispute or a regulatory process—seek the information too because they might be subject to the rules that the decision has interpreted and applied. If they are given it, then they can use the information contained in the court's reasoning to plan their future conduct. Likewise, as this Article discussed in

152 *Sussex Justices*, 1 KB at 259.

153 It is accepted, for example, in the jurisprudence of the International Court of Justice, that reasons given for a dispositive holding in a judgment may give meaning to the holding not self-evident in its terms alone. *See, e.g.*, Request for Interpretation of the Judgment of 15 June 1962 in the Case Concerning the Temple of Preah Vihear (Cambodia v. Thai.), Judgment, 2013 I.C.J. 281, ¶¶ 48–49, 68 (Nov. 11) ("In determining the meaning and scope of the operative clause of the original Judgment, the Court . . . will have regard to the reasoning of that Judgment to the extent that it sheds light on the proper interpretation of the operative clause."); BROOM, *supra* note 150, at 147–48 ("A public statement of the reasons for a judgment . . . is essential to the establishment of fixed intelligible rules.").

Section III.C, society at large has an interest in explanations of algorithmic decisions.[154]

Thus, with machine learning results and legal decisions alike, the people whom the results and decisions specifically address and society at large have reasons to know the reasons. A technical mistake by a court may give rise to a right of appeal.[155] A decision reached on grounds that violate public policy[156] all the more may lead a party to seek adjustment through some mechanism of control.[157] And, at least in systems with judicial review of legislative acts, the law being applied might need to be struck down. In a similar fashion, when deciding whether to accept a machine learning output or whether to change his behavior in response to the output, the layperson benefits from understanding how the output was produced. An explanation of a machine learning output serves a purpose like that of a reasoned judicial decision. The explanation supplies information for shaping behavior; information that people might use to challenge a specific output, and; information that people might use to cast doubt on the automated process itself. Explanation of a machine learning output shares these purposes with judicial reasoning.

IV. THE FORM OF A MEANINGFUL EXPLANATION

The GDPR states that a data subject has the right to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" in situations where the data controller uses automated decision-making.[158] In Part III, this Article considered what would constitute meaningful information in light of the consequences of such decision-making. This Article now turns to the first half of the phrase, "meaningful information about the logic involved." This

154 *Cf.* BROOM, *supra* note 150, at 147–48 (promoting the public benefits of articulated reasoning).

155 *See R.E.M*, 3 S.C.R. para. 11 ("Reasons permit effective appellate review."); *Sussex Justices*, 1 KB at 256.

156 Thus, the exceptional possibility allowed under Article V(2)(b) of the Convention on the Recognition and Enforcement of Foreign Arbitral Awards of 1958, June 10, 1958, 330 U.N.T.S. 3, that an award may be denied recognition or enforcement where granting it "would be contrary to the public policy of [the] country" where recognition or enforcement is sought. *See* Mitsubishi Motors Corp. v. Soler Chrysler-Plymouth, Inc., 473 U.S. 614, 638 (1985).

157 A phrase denoting the range of possible procedures for challenge and correction, whether formally of appeal or of other types of challenge, such as nullification, non-recognition, non-enforcement. It also has been used to denote extra-legal mechanisms, such as in social or engineering systems. W. MICHAEL REISMAN, SYSTEMS OF CONTROL IN INTERNATIONAL ADJUDICATION AND ARBITRATION: BREAKDOWN AND REPAIR 1–3 (1992).

158 Council Regulation 2016/679, *supra* note 3, arts. 13(2)(f), 14(2)(g) & 15(1)(h).

Article argues that, as the decisions are made by machine learning algorithms, "meaningful information" might well entail disclosure of the data that trained the machine.

Explaining an output from a computer is widely described as explaining an operation of logic. To take an example, which would be trivial if it were isolated, an article on machine learning on the website of the International Association of Privacy Professionals refers to "logic" ten times (e.g., "the logic and significance of machine learning systems").[159] The notion that "explainability" means explaining the "logic" behind machine learning is a recurring one in the literature.[160] And it is not restricted to academia or think tanks. The European Commission, in its communications with the other main national regulator, the U.S. Department of Commerce, recommends "considering a requirement to explain *the underlying logic* of automated decisions."[161] It will not have escaped notice that the common terms of Articles 13–15 of the GDPR call for information about "the *logic* involved."[162]

At least as relevant here, the explainability provisions repeatedly use the word "processing." Processing implies that a fixed mechanism exists, such that, the data controller feeds a question into the mechanism, and the mechanism gives a result—and that this is all there is to it. Articles 13–15 use the term "processing" like this.[163] Article 9(2) (a) and (g) use it too.[164] Viewing a machine learning system which has already been trained, and limiting one's consideration to a discrete episode in which the data controller uses that system to obtain an answer to a question regarding a data subject, one might just about justify saying that this is all there is to it. With one's attention focused on the discrete episode, especially on the question posed and answer produced, it looks like "processing" in the way a traditional logic-based software system "processes": ask the machine a question and get an answer. This, however, is not a promising start to analysis, where what calls for analysis is a legal obligation that somebody explain *machine*

159 Burt, *supra* note 116.

160 *See, e.g.*, Suzanne Rodway, *Just How Fair Will Processing Notices Need to Be Under the GDPR?*, 16 PRIVACY & DATA PROTECTION, Jan./Feb. 2016, at 16.

161 Letter from European Commission, Directorate-General Justice and Consumers, to National Telecommunications and Information Administration (NTIA), U.S. Department of Commerce (Nov. 9, 2018) (on file with The George Washington Law Review) l(emphasis added).

162 Council Regulation 2016/679, *supra* note 3, arts. 13(2)(f), 14(2)(g) & 15(1)(h) (emphasis added).

163 *Id.*

164 *Id.* art. 9(2).

learning output. It ignores the fundamental difference between the machine learning process and logic-based systems.

A. Machine Learning Is Not (Just) Logic

Legal writers addressing automated decision-making have suggested that machine outputs are necessarily explicable by reference to source code.[165] This is an accurate description of classic algorithms, such as those for sorting or path finding,[166] which can be understood entirely based on the logic of their source code. There is a school of thought that believes this sort of thinking, the classic practice of "how computer scientists build and evaluate software,"[167] should be the basis for artificial intelligence—but this school has been eclipsed by machine learning. Will Knight, in the MIT Technology Review, explains the difference:

> From the outset, there were two schools of thought regarding how understandable, or explainable, AI ought to be. Many thought it made the most sense to build machines that reasoned according to rules and logic, making their inner workings transparent to anyone who cared to examine some code. Others felt that intelligence would more easily emerge if machines took inspiration from biology, and learned by observing and experiencing. This meant turning computer programming on its head. Instead of a programmer writing the commands to solve a problem, the program generates its own algorithm based on example data and a desired output. The machine-learning techniques that would later evolve into today's most powerful AI systems followed the latter path: the machine essentially programs itself.[168]

A typical modern machine learning algorithm is an algorithm with source code, but to see it just as an algorithm is entirely inadequate for explaining it. This is so on two counts. First, the initial phase of a machine learning algorithm is to process a training dataset, from which it computes "knowledge," represented as millions of weight parameters specifying the strengths of connections in a neural network. The source code alone does not reveal what output will be produced by a new input; for that we also need to know the weight parameters it has learnt—or, instead, we could be given

165 *E.g.*, Roth, *supra* note 25, at 1977, 1981–82, 1995, 2025 n.16.

166 *See, e.g.*, ROBERT SEDGEWICK & KEVIN WAYNE, ALGORITHMS 288–302, 535–37 (4th ed. 2011) (discussing "the [common] sorting algorithm . . . *quicksort*" and path-finding algorithms).

167 Kroll et al., *supra* note 11, at 642. Those writers do, however, acknowledge that "source code alone teaches a reviewer very little, since the code only exposes the machine learning method used and not the data-driven decision rule." *Id.* at 638.

168 Knight, *supra* note 9.

access to the training dataset. Machine learning is not just source code; it is source code plus data.

An algorithmic view of machine learning is inadequate in a second way: the source code and the weight parameters together constitute a reductionist *description* of the machine learning algorithm's decisions, but they are not an *explanation*. They are as inadequate as would be an explanation of "love" or "justice" through synapse biochemistry and a neuron connection map. A useful explanation, responsive to the goals listed in Part III, needs a different level of abstraction. What then is the right level?

B. The Value of Data

Alexander Wissner-Gross neatly illustrated the practical importance of data by compiling a list of machine learning breakthroughs (speech recognition, image classification, etc.) together with the date when the relevant algorithm was proposed and the date when the first high-quality training dataset became available.[169] He observed that "the average elapsed

169 Alexander Wissner-Gross, *Datasets Over Algorithms*, EDGE (2016), https://www.edge.org/response-detail/26587 [https://perma.cc/J26K-JYNT]. Wissner-Gross provides a full list of breakthrough examples:

> A review of the timing of the most publicized AI advances over the past thirty years suggests a provocative explanation: perhaps many major AI breakthroughs have actually been constrained by the availability of high-quality training datasets, and not by algorithmic advances. For example, in 1994 the achievement of human-level spontaneous speech recognition relied on a variant of a hidden Markov model algorithm initially published ten years earlier, but used a dataset of spoken Wall Street Journal articles and other texts made available only three years earlier. In 1997, when IBM's Deep Blue defeated Garry Kasparov to become the world's top chess player, its core NegaScout planning algorithm was fourteen years old, whereas its key dataset of 700,000 Grandmaster chess games (known as the "The Extended Book") was only six years old. In 2005, Google software achieved breakthrough performance at Arabic- and Chinese-to-English translation based on a variant of a statistical machine translation algorithm published seventeen years earlier, but used a dataset with more than 1.8 trillion tokens from Google Web and News pages gathered the same year. In 2011, IBM's Watson became the world Jeopardy! champion using a variant of the mixture-of-experts algorithm published twenty years earlier, but utilized a dataset of 8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg updated one year prior. In 2014, Google's GoogLeNet software achieved near-human performance at object classification using a variant of the convolutional neural network algorithm proposed twenty-five years earlier, but was trained on the ImageNet corpus of approximately 1.5 million labeled images and 1,000 object categories first made available only four years earlier. Finally, in 2015, Google DeepMind announced its software had achieved human parity in playing twenty-nine Atari games by learning general control from video using a variant of the Q-learning algorithm published twenty-three years earlier, but the variant was trained on the Arcade Learning Environment dataset of over fifty Atari games made available only two years earlier. *Id.*

time between key algorithm proposals and corresponding advances was about eighteen years, whereas the average elapsed time between key dataset availabilities and corresponding advances was less than three years, or about six times faster."[170] So breakthroughs may come many years after the algorithm was proposed, but they generally come within a few years of the dataset, "suggesting that datasets might have been limiting factors in the advances."[171]

Despite the centrality of data, algorithmic thinking has had consequences for how people talk about explainability. For example, some have said that intellectual property rights in software might conflict with explainability. Wachter, Mittelstadt, and Floridi (who are skeptical overall about the explainability provisions of the GDPR) refer in particular to copyright of software.[172] It may be that those writers have identified a real obstacle to implementing explainability. However, to talk about intellectual property in an algorithm is to place weight on algorithmic logic and thus, to miss the distinctive characteristic of machine learning. Perhaps the intellectual property rights of data controllers will frustrate data subjects seeking to learn more about the algorithm that embodies the training procedure or the neural network structure, but disclosure of an algorithm would be, at most, half an explanation. If you want more than that, then you need to see the data that trained the machine. The data that trained the machine is indispensable to explaining the output of the machine. Arguably, curated datasets are the valuable assets for machine learning. The intellectual property of source code, while no doubt valued, is unlikely to tell the whole story.

C. Explanations Grounded in Data

Even before the recent boom[173] in machine learning, some regulators recognized that automated decision-making entails data and that to understand how automated decision-making works, one must engage in "rigorous assessment of data quality and relevance."[174]

---

170 *Id.*

171 *Id.*

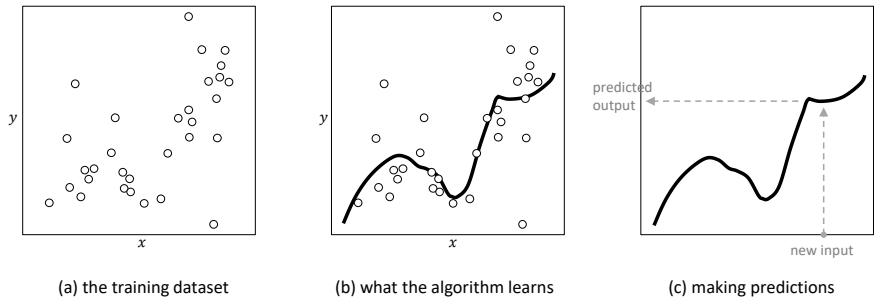
172 Wachter et al., *supra* note 65, at 93; *cf.* Roth, *supra* note 25, at 2028 (discussing disclosure of algorithmic source code during legal challenges).

173 T.S., *Why Artificial Intelligence Is Enjoying a Renaissance*, ECONOMIST (July 15, 2016), https://www.economist.com/the-economist-explains/2016/07/15/why-artificial-intelligence-is-enjoying-a-renaissance [https://perma.cc/32FB-A7YH] ("When a deep-learning system won an annual image-recognition contest in 2012, vastly outperforming rival systems, people both within the academic community and beyond sat up and took notice.").

174 BD. OF GOVERNORS OF THE FED. RESERVE SYS., SR 11-7, GUIDANCE ON MODEL RISK MANAGEMENT 2–3 (Apr. 4, 2011).

More recently, organizations concerned with privacy and the use of machine learning in surveillance programs have placed emphasis on data as well. For example, Access Now, a public interest group that made submissions as a third party in *Big Brother Watch v. United Kingdom*,[175] referred to machine learning as a process through which "mathematical algorithms could draw inferences from collections of data."[176] This is a fair description of the process. The Court took no view on it. Some commentators also recognized that machine learning involves inferences from data. In a short article published a year before the GDPR entered into application, a policy analyst in a Brussels-based think tank suggested that data's role in "algorithmic decision" could be an insurmountable hurdle to explanation. He said, "Often, the challenge of explaining an algorithmic decision comes not from the complexity of the algorithm, but the difficulty of giving meaning to the data it draws on."[177] Whether or not the "challenge of explaining" can be overcome, statements of the challenge like these are closer to the mark than many others because they recognize the centrality of data to machine learning.

FIGURE 1. THE ROLE OF TRAINING DATA IN MACHINE-LEARNING DECISIONS



175 Big Brother Watch v. United Kingdom, App. Nos. 58170/13, 62322/14 & 24960/15, Judgment (Sept. 13, 2018), http://hudoc.echr.coe.int/eng?i=001-186048 [https://perma.cc/7QEC-H3C2].

176 *Id.* ¶¶ 4, 296.

177 Nick Wallace, *EU's Right to Explanation: A Harmful Restriction on Artificial Intelligence*, TECHZONE360 (Jan. 25, 2017), https://www.techzone360.com/topics/techzone/articles/2017/01/25/429101-eus-right-explanation-harmful-restriction-artificial-intelligence.htm [https://perma.cc/8KQY-PQC6].

For present purposes, this Article assumes, as the GDPR must, that "the difficulty of giving meaning to the data" is a technically achievable task.[178] But because machine learning is an operation of induction from data, not an operation of deduction grounded on logic, the explanation of a machine learning output will not be like the explanation of a mathematical derivation. It will be an explanation of how a particular training data set influenced the machine learning output. More specifically, as Judea Pearl[179] explained in an interview with Quanta Magazine, explanations of data's effect on outputs will describe an exercise in "curve fitting":[180]

> As [Judea Pearl] sees it, the state of the art in artificial intelligence today is merely a souped-up version of what machines could already do a generation ago: find hidden regularities in a large set of data. "All the impressive achievements of deep learning amount to just curve fitting," he said recently. . . . The way you talk about curve fitting, it sounds like you're not very impressed with machine learning [remarks the interviewer]. "No, I'm very impressed, because we did not expect that so many problems could be solved by pure curve fitting. It turns out they can."[181]

To illustrate what Pearl means, consider an algorithm for making predictions based on simple two-dimensional data, shown in Figure 1 above. The training data set is a collection of (x,y) points. Perhaps x represents a person's income and y represents the profit made from that person. A machine learning algorithm might process this training dataset and obtain a fitted curve, stored as a collection of parameter values. When a new person comes along with a new x value, it can predict the profit to be made by simply reading off the corresponding y value on the curve. To explain the predicted profit, it is nonsensical to report the parameter values and the algorithm for reading off a y value. Instead, the only sensible way to explain the output is by saying "the training dataset had several points with x values similar to the new person's x, and the predicted y value is close to their average." For a complete explanation, we might also provide those cases from the training dataset. With that data in hand, the data subject can see not only the fitted curve used to predict her outcome; she also sees the data used to arrive at the curve.

178 *But see id.* (suggesting that correlations between data points will be inexplicable).

179 Judea Pearl "won the [2011] Turing Award, computer science's highest honor." Kevin Hartnett, *To Build Truly Intelligent Machines, Teach Them Cause and Effect*, QUANTA MAG. (May 15, 2018), https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/ [https://perma.cc/2XEE-GRPT].

180 *Id.*

181 *Id.*

This does not suggest that explanation is a simple matter of "report the most similar cases in the training dataset." Instead, that explanation should be centered around the problem of how the training dataset relates to the decision at hand. Explanation is an exercise in data science, not an exercise in algorithmics.

There are numerous complications here. For example, how is it decided which points in the training dataset are similar? How well did the training procedure manage to fit the data? How should the data science reasoning be presented in a way that is useful to the audience? Most importantly, how should these explanations be oriented towards the explanatory needs of the data subject listed in Part III? None of these questions is answered in the GDPR. It therefore seems inevitable that they will be contested in litigation, because you need answers to them if you are to make practical use of the GDPR's concept of explainability.

C. Explanations that Meet the Needs of the Data Subject

What sort of explanation grounded in data would meet the needs of a data subject? More precisely, how does the "curve fitting" view of explanation relate to the requirements for a "meaningful explanation" discussed in Part III?

One reason for seeking an explanation is to give guidance for future conduct, as discussed in Section III.A. The explanation would have to give enough information for the data subject to decide how to change their conduct, if at all. Formally, the data subject needs to be able to answer a hypothetical question, "What would the decision be if my conduct was *X*?," where *X* ranges over the actions that the data subject might reasonably take.

In the language of curve fitting, the hypothetical question is expressed as "What is the shape of the fitted curve?" There are many forms in which an answer might be given. The algorithmic form is to tell the data subject about the "logic involved," i.e. the code of the machine learning system plus its parameter values. The data subject would then, in principle, be able to re-run the algorithm for various *X*, to see what the decision would be, and thus learn how to change their conduct. Very few data subjects would be able to benefit from such an explanation. Another form of explanation might consist in providing the data subject with the decision that would be made for each action *X* in a reasonable range; this has been described as "counterfactual explanation."[182] It might even consist in providing them with a simpler

[182] Wachter et al., *supra* note 18, at 844, 854 (presenting counterfactuals of this type as "a novel type of explanation of automated decisions that overcomes many challenges

statistical model or a rule of thumb that approximates the true behavior of the decision-making algorithm;[183] such an approximate model might in some cases be more helpful to the data subject.

A second reason for seeking an explanation is for the purposes of challenging a decision, for instance to claim that it is based on invidious grounds, as discussed in Section III.B. The question here is "Why did the algorithm make the decision it made?" In the language of curve fitting, this is expressed as "Why does the fitted curve have this value at this point?" In systems based on machine learning, the only possible answer is "because of related points in the training dataset."[184] It does not make sense to explain an individual outcome by reference only to that individual. The only way to explain it is by describing what and how the machine learnt by analyzing its training dataset. Thus, "meaningful information" for detecting invidious grounds is likely to consist of data about other data subjects, because the machine learning algorithm's training dataset will be a compilation of data about other data subjects. This Article returns to this characteristic of the machine learning process, because its implications may be significant for how disputes over explainability unfold.

A third reason for seeking explanations is to confer legitimacy on a decision-making system. As discussed in Section III.C, this calls for explanations that give an account of the system, and relates to compliance. For example, a regulator might judge a system according to "a definition of fairness in which similarly situated people are given similar treatment," [185] i.e., a definition of fairness that can be expressed as "Do we approve of the shape of the fitted curve?" Compliance and accountability are thoroughly dealt with elsewhere,[186] and because this Article's focus is litigation, this Article does not explore this further.

V. THE DATA CONTROLLER'S DEFENSES

associated with algorithmic interpretability and accountability," and proposing machine learning techniques for generating meaningful counterfactuals).

183 Advait Sarkar, *Interactive Analytical Modeling*, U. CAMBRIDGE, May 2018, at 1, 51 (referring to the simpler statistical model as a "metamodel," and describing it as an "explanatory metaphor[]" for the underlying machine learning model).

184 Wachter, Mittelstadt, and Russell suggest that "[c]ounterfactuals offer a solution and support for contesting decisions." Wachter et al., *supra* note 18, at 878. They explain how counterfactual explanation helps when the decision is being contested on grounds of technical error, but it is doubtful that such an explanation would be of use if the concern is invidious grounds. *Id.*

185 Kroll et al., *supra* note 11, at 685.

186 *See, e.g.*, *id.*

(AND WHY THEY PROBABLY WILL NOT WORK)

The data subject, unhappy with a decision made by an automatic system, will begin by asking the data controller for an explanation. She suspects that the decision might be unfair, but the explanations she has been given did not contain enough information for her to be sure of this. She begins proceedings, and will start her case under the available procedures.[187] Her substantive claims will include a claim that the data controller did not "explain" an automated decision which affected her—meaning that the data controller did not explain it in a manner that borne out, she will challenge the decision or seek other legal remedies.[188]

The data controller will of course defend. The data controller's defense will involve some form of explanation of how the machine learning system functioned. Perhaps it will include a more or less baffling description by an expert about the software code that is involved, with a more or less general acknowledgement that a training dataset was involved in training the system. The data subject (the complainant or plaintiff in this hypothetical scenario) will of course introduce her own experts. She will draw pointed attention to the distinguishing characteristic of machine learning—its reliance on the training dataset. It is hard to imagine any explanation getting past the data subject's lawyer that does not acknowledge that a training dataset plays a role in machine learning.

The data controller's defense might try to give a high-level description of the training dataset: "Our algorithm was trained on data about how people use public buses in Sheffield." Data subject's counsel will demand to know more because to explain why a machine reached a given result, one must

187 *See* Council Regulation 2016/679, *supra* note 3, art. 79(2) (providing a "[r]ight to an effective judicial remedy against a controller or processor" and establishing choice of venue rules); *id.* art. 82(6) (describing "[c]ourt proceedings for exercising the right to receive compensation"); *id.* at 27 (discussing jurisdiction and choice of venue in recitals); *cf. id.* art. 77 (providing a "[r]ight to lodge a complaint with a supervisory authority").

188 *See id.* art. 82 (providing a "[r]ight to compensation and liability"); *id.* art. 83 (describing administrative fines); *id.* at 27–28 (describing administrative, criminal, and compensatory remedies); *cf. id.* art. 78 (establishing a "right to an effective judicial remedy against a legally binding decision of a supervisory authority"); *id.* art. 84 (providing penalties "for infringements which are not subject to administrative fines pursuant to Article 83"). Questions have been asked about the extent to which the GDPR protects individual rights, if at all. *See generally* Bart van der Sloot, *Do Data Protection Rules Protect the Individual and Should They? An Assessment of the Proposed General Data Protection Regulation*, 4 INT'L DATA PRIVACY L. 307 (2014). On balance, the Regulation indicates both individual rights and procedures for challenging alleged breaches thereof. It is not this Article's purpose to examine the specific causes of action that claimants will invoke under the GDPR. This Article further acknowledges that the manner in which specific claims procedures interact with the causes of action must be considered closely, if one is to have a full picture.

scrutinize the training data. In particular, the data subject, who suspects that the cause of the result is some invidious influence in breach of a rule of public law, will demand to know more. Even if regulations suggest, as a formal matter, that it suffices to simply say "we used data about how people use the buses," counsel will still demand more. Counsel will insinuate that the training dataset was gathered in a manner which, if not calculated to do harm to the data subject, negligently brought about an unfair, unjust, and debilitating impact on her. The training dataset, data subject's counsel will say, was a stacked deck. Data controller's counsel will act offended, dismissive, or both. Data controller's counsel will assure, probably with expert witnesses as back-up and documentary materials, that nobody does datasets quite as conscientiously as this particular data controller—or at least that this data controller got its data from an industry-respected data source whose standards are beyond question. Maybe the data controller will give an account of some of the very conscientious methods for data collection that were employed to ensure that the training data set did not manifest some invidious bias or other deleterious effect. And so on.

The plaintiff will have none of it. The plaintiff, no matter the response the data controller gives, will ask for more. True, a well-managed courtroom lets such an interrogation go only so far. Past a certain point, the judge will cut off certain inquiries as irrelevant. But the object of the inquiry here has a logical end-point, and that end-point is beyond doubt as relevant as any evidence a plaintiff might seek. This colorful sketch makes the point that the dynamics of adversary procedure,[189] when combined with the nature of machine learning, invite the demand "show us the data!"

In reply to each demand in court, the data controller will proffer defenses. There are few limits to the creativity of legal counsel (defense or claimant). One nevertheless may anticipate certain lines of argument that data controllers will consider. First, the data controller will try to knock out the explainability claim by saying that the situation is not covered by the explainability obligation, there having been a human involved in the decision. This defense is to assert that, if a decision is not based *solely* on automated processing, then it is excluded from the explainability obligation. Alternatively, the data controller will say that it has followed all appropriate

189 Significant differences exist among different legal systems when it comes to the management of trial proceedings. The observations this Article makes are particularly salient to the common law systems that prevail in much of the English-speaking world. However, writers have noted that the so-called inquisitorial systems are much more adversarial than simplified accounts would give. *See, e.g.*, J.A. Jolowicz, *Adversarial and Inquisitorial Models of Civil Procedure*, 52 INT'L & COMP. L.Q. 281, 281 (2003) (arguing that there are no "purely inquisitorial" systems).

regulatory guidelines and compliance, both in explaining the decision-making system as a whole, and in explaining each decision. Another defense will be for the data controller to say that, contrary to plaintiff's claim, it indeed provided a meaningful explanation. This defense will involve the parties in an exchange over both fact and law: the factual parameters of the explanation given will be challenged, and so too will the meaning of the term "meaningful explanation." Finally, the data controller will plead some combination of impossibility or proportionality, saying, in effect, that to provide more information is either technically infeasible or out of proportion to the legal interests of the claimant concerned. This Article turns now to consider each of these defenses—and the way each is likely to fail.

A. "We Didn't Have to Explain, Because There's a Human Involved."

Article 22(1) of the GDPR offers a degree of comfort to data controllers, where it apparently excludes from the explainability requirement "a decision based *solely* on automated processing";[190] the modifier "solely" is apparently a limitive term. That term might be read to suggest that a decision reached with even minimal human involvement is not a decision requiring explanation under the explainability provisions of the GDPR. However, any limitive effect here is not to be read too far.

In its Guidelines on Automated Individual Decision-Making ("Guidelines"), the Article 29 Data Protection Working Party advised that "[t]he [data] controller cannot avoid the Article 22 provisions by fabricating human involvement."[191] The Working Party's concern that "human involvement" might be fabricated identifies the matter as an important one. It also identifies the duty of explanation to be more than a procedural requirement. If it were only a matter of human involvement in form, and not in substance, then there would be no point in drawing special attention to it in this way. The better understanding is that the data controller does not escape the potential for claims simply by involving a human in decision-making in some superficial way. The Guidelines then discuss the substance of the concept of "human involvement":

> To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture. It should be carried out by someone who has the authority and competence to change the decision. As part of the analysis, they should consider all the relevant data.[192]

As this Article argued in Part III, the term "meaningful" needs to be interpreted in the light of envisaged consequences. It is of no benefit to the

190 Council Regulation 2016/679, *supra* note 3, art. 22(1) (emphasis added).

191 Article 29 Data Protection Working Party, *supra* note 142, at 21.

192 *Id.*

data subject merely to *have* a human in the decision-making loop. The point of human oversight of decision-making must surely be that humans by their very nature can be asked for explanations of their decisions—if the machine will not explain, then let there be a human overseer who will.

So now consider the defense "the decision in question was not based solely on automated decision-making, hence no explanation is required." The data controller, making that defense, would assert that human beings have had meaningful involvement in the decision and that they "consider[ed] all the relevant data"[193] as part of their involvement. Plaintiff would then ask (1) for an explanation of the human's decision, perhaps so it can be challenged, and (2) for evidence that the human considered all relevant data.

This Article is concerned with automated decision-making systems based on machine learning. Suppose that the human overseer is aware of the output that the machine gave with respect to the particular data subject concerned. In this situation, it is not at all clear that the data controller satisfies its obligations in respect of automated decision making simply by disclosing the data that the human directly handled (i.e., data that the human read himself). There is also the data that was involved in the machine learning process which gave an output that the human saw. Before announcing the decision, the human thus will have handled certain data, but he will also have considered the machine's output—which, as set out in Part IV above, is meaningfully explained only by considering the data that trained the machine. The plaintiff would assert that the training dataset is therefore relevant to the human's decision and would insist on seeing it for the purpose of challenging that decision. This matter only arises with machine learning. Simple algorithmic decision-making is embodied purely in code and is not based on training data.

Plaintiff might furthermore argue that the grounds for considering the data behind the machine learning output are strengthened where the machine learning output and the decision which the data controller alleges was human-made are in substance the same. The claimant's concern would be that the human is a façade and the real decision was the automated one. That the two results are identical would be circumstantial evidence of the façade. Even if the human were a meaningful participant in the decision, determining that he really was, assuming that the *Guidelines* advice is followed, requires considering "all the relevant data," and that would mean the data behind the machine output as well as the data evaluated by the human directly. Or so a claimant-data subject very likely will say.

The defense that there was meaningful human involvement in decision-making might well be a complete defense to explainability; that will be clarified as courts and tribunals interpret that part of the explainability rule. But to prevail in the defense, the data controller must do more than merely

193 *Id.*

assert it. Where human involvement was accompanied by a machine learning output, the training dataset enters the picture. The adversary process in court is likely to test the defense with escalating demands for disclosure of the training dataset.

There is, perhaps, a situation in which the demand for data exposure might be quelled, even where there had been machine learning involvement. Suppose the company has a procedure whereby a decision is reached through machine learning, and the data subject can request human reconsideration. Under the company's procedure, the human reconsideration takes place in complete isolation of the machine learning output—i.e., the human who performs the reconsideration does not know what the machine output was.[194] The human proceeds to give an entirely independent decision. The data subject-claimant would likely want rigorous proof that the human indeed was isolated from the machine. It could be that, if the data controller-defendant showed that there had indeed been such isolation, then this would be a situation in which the data behind the machine would no longer matter—and thus no longer be susceptible to exposure. However, even this defense might raise a problem for the data controller. If an isolated human reached a separate decision, then that decision would have been based on a much (much) smaller data set than the decision reached by machine; surely the human decisionmaker will not have considered the hundreds of thousands, or millions, of elements in the training data set. The data subject-claimant might ask whether her rights now are being disposed of by a crude and inferior decision process, performed simply to get her out of the way. The data controller, therefore, even here would be in a potentially awkward position. Not to mention, in order to get that far, the data controller would need to have maintained as part of its compliance system a redundant set of human decisionmakers ignorant of the machine-based branch of the company's decision-making operations.

B. "We Gave a Meaningful Explanation of Our Decision-making System."

A data controller is obliged to disclose whether automated decision-making is being employed, and if it is, then it must also supply "meaningful information about the logic involved."[195] A data controller might publish such information as part of its terms and conditions in an effort to be open and transparent: it might describe the overall operation of the system, perhaps also including evidence of fairness. As to the form of the meaningful information, perhaps the data controller has taken to heart the advice of

194 Query: does it matter whether the human knows that there was such an output? Whether the human knew is a question likely to be asked. A tenacious claimant will seek to claw the explainability requirement back in, and if the human did know, then that knowledge will give some purchase to the claw.

195 Council Regulation 2016/679, *supra* note 3, art. 13.

Kroll: "explanations must speak to the decisions made during the design of a computer system, as such information is always available and always fulfils the key requirements of a meaningful explanation."[196] The data controller's defense might argue that the data controller has provided a full and complete explanation of the decision-making system. Indeed, the data controller might find an expert witness who says, following Kroll, that an explanation of the machine's output for a plaintiff's specific decision would not be meaningful and would potentially be misleading:

> Explanation is not an unalloyed good, both because it is only useful when it properly engages the context of the tool being explained and because explanations, at a technical level, do not necessarily provide understanding or improve the interpretability of a particular technical tool. Rather, explanations tend to unpack the mechanism of a tool, focusing narrowly on the way the tool operated at the expense of contextualizing that operation. Explanations risk being 'just-so' stories, which lend false credence to an incorrect construct.[197]

It certainly may serve the purposes of a compliance designer to ask "how measurement of a system beyond understanding of its internals and its design can help to defeat inscrutability."[198] This is of a piece with skepticism about "explanations [which] tend to unpack the mechanism of a tool, focusing narrowly on the way the tool operated."[199] The compliance approach elevates the vantage point.

But, in court, the "bigger picture" defense is what you argue when the details hurt your case. The other side will not fail to see what you are doing. This Article is not saying this as a rhetorical device to criticize the compliance systems approach; the purpose is to draw attention to the very different direction that the inquiry into explanation is likely to take when the GDPR's explainability provisions are invoked in a claims process. The two approaches are operating in very different socio-technological settings. Thus, although each may have something to say to the other, neither gives a complete answer to the problems that are thrown up in the places where the other operates.

In litigation, the plaintiff will point to Recital 71, which calls for "specific information to the data subject" and "an explanation of the decision reached *after* such assessment."[200] It plainly calls for an explanation of the specific decision, in a form that gives meaningful guidance to the data subject, as described in Parts III and IV above. An explanation of "the

196 Kroll, *supra* note 109, at 7.

197 *Id.* at 3.

198 *Id.* at 7.

199 *Id.* at 3.

200 Council Regulation 2016/679, *supra* note 3, at 14 (emphasis added).

decisions made during the design of a computer system"[201] does not supply what is called for.

C. "Our Explanations Follow Regulatory Guidelines."

In the UK, the GDPR has been incorporated into domestic law through the Data Protection Act of 2018.[202] This Act gives powers to a public body called the Information Commissioner's Office ("ICO"), which is charged with developing guidelines and conducting investigations, and which has powers to issue requests for information and to levy fines.[203] The defense might point to official guidance from the ICO, and say that it complied with the letter of that guidance, and argue that the explanation is therefore adequate.

However, to the extent that the regulator's guidelines can be satisfied *pro forma*, those guidelines are open to challenge in court. Explanations that are explanations in form only, but which are not of substantive use to the data subject for the purposes outlined in Part III, can be challenged on the grounds that they are not meaningful. The GDPR does not give a legislative role to regulators: a national regulator can issue guidelines that circumscribe the enforcement actions it will take,[204] but it cannot issue guidelines that constrain how courts will interpret the GDPR in private claims, and it is not up to regulators to limit what is meant by "meaningful information." Perhaps for this reason, the UK's ICO in its compliance guidelines has avoided definitive language like "If you explain X, Y, and Z, you will be safe," which suggests *pro forma* compliance standards, and has instead used suggestive language that indicates it is offering its *opinion* about the interpretation of the GDPR.[205]

An interesting illustration of the hesitancy to provide definite answers in such a regulatory environment is supplied by privacy guidance for churches. For example, the ICO hesitantly rebuts the myth that churches cannot ask for Christmas prayers for named parishioners who are ill or sick: "If this is something that the parishioner concerned might reasonably expect and welcome and the church can justify processing their health data, then it

201 Kroll, *supra* note 109, at 7.

202 *See* Data Protection Act 2018, c. 3, § 22 (Eng.).

203 *Id.* §§ 115–16.

204 *E.g.*, INFO. COMM'R'S OFFICE, REGULATORY ACTION POLICY, https://ico.org.uk/media/about-the-ico/documents/2259467/regulatory-action-policy.pdf (last visited Sept. 6, 2020).

205 *See, e.g.*, *id.* at 18 ("Although *not an exhaustive* list, this *could* include, for example . . . .") (emphasis added).

is unlikely to be breaching the law."[206] One church newsletter takes a more cautious line:

> Catholic Insurance Service have given instructions to the Diocese regarding the new GDPR legislation which is now in force. It means, for instance, that without the written consent of those who are ill (they and they alone can give it) we are no longer able to publish their names in the newsletter.[207]

The insurance firm has perhaps noticed that the ICO used suggestive words "reasonably" and "unlikely" rather than definitive terms, and has chosen instead its own more cautious interpretation of the GDPR.

Another church organization in the United Kingdom offers similarly cautious guidance:

> If a church holds personal data either on a computer or in a paper-based filing system, it must follow the rules set out in the Data Protection Act 2018 and the GDPR. This leaflet explains what this means for churches. It should however only be taken as general guidance and should not be used as a substitute for obtaining legal advice. At the end of the leaflet we have provided a checklist for churches to work through. If churches have questions that fall outside the scope of this leaflet then we would advise that you contact the Information Commissioner's Office for their advice.[208]

Thus, the national church organization—in this example, the Baptist Union—of many other denominations have published similar guidance[209]—approaches the matter with caution. Its interpretation is only that. And its message to individual churches? Get yourself a lawyer. Wariness has permeated society over the litigations to come.

As of 2020, the legal standard to identify what constitutes sufficiency under the GDPR for purposes of "meaningful information" is undefined (or at best imprecisely defined). As that standard is clarified through dispute settlement practice and official guidance, it might be clarified in a way that is permissive to the data controller. Even under a vague or woolly standard however, the data controller's task would still not be easy. Disputes will still

---

206 Steve Wood, *Blog: Sleigh-ing the Christmas GDPR Myths*, INFO. COMM'R'S OFFICE, https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2018/12/sleigh-ing-the-christmas-gdpr-myths/ [https://perma.cc/RUF7-QFGS].

207 *Newsletter for 25th August 21st Sunday in Ordinary Time 2019*, OUR LADY ASSUMPTION & ENG. MARTYRS (Cambridge), Aug. 15, 2019, at 2.

208 *Guideline Leaflet L13: Data Protection*, BAPTIST UNION OF GR. BRIT., https://www.baptist.org.uk/Publisher/File.aspx?ID=111329&view=browser [https://perma.cc/P4SE-U3PH].

209 *See, e.g.*, *General Data Protection Regulation (GDPR)*, DIOCESE OF LONDON, https://www.london.anglican.org/kb/data-protection/ [https://perma.cc/S765-7DK9]. *See generally GDPR*, IKNOW CHURCH, https://www.iknowchurch.co.uk/gdpr [https://perma.cc/W2YM-X33J] (providing advice to churches generally on the GDPR).

arise in specific cases as to whether a data controller has satisfied the standard. Whatever the standard is, information that merely satisfies the standard in form—*prima facie* compliance—will not satisfy the claimant-data subject in litigation. It is to be expected that a careful data controller will give an explanation that adheres in form to whatever standard has emerged in respect of "meaningful information." The standard might emerge from regulators' guidance, industry best practice, or a combination of both. The accordance of an explanation with formal requirements is unlikely to bring the matter to a close.

D. "We Gave a Meaningful Explanation of the Decision."

A data subject who suspects that the decision of a machine learning system has breached her rights will seek an explanation of that decision. If the explanation does not satisfy her, she may bring a claim that the explanation provided was not adequate. As discussed in Section III.D above, the first step in contesting a decision is likely asking for a better explanation of the decision. The data controller-defendant's counsel probably will reply that the data controller has already provided an explanation, and that it constituted "meaningful information" in the sense of Articles 13–15 of the GDPR.

Various commentaries on explainability suggest that an explanation is adequate if it gives a basic account of how the automated decision was reached.[210] Some go so far as to say that, if the account is *not* basic, then it will not be satisfactory, because data subjects will not understand.[211] Such confidence in basic accounts is misplaced. Data controllers facing data subjects in litigation are likely to find that simplified or basic accounts of how a machine learning system has come to give a particular output will not be accepted by data subjects—or their legal counsel—as sufficient. Practically any account will attract challenge—up until an account that exposes the data that trained the machine. To see why a "better" explanation, but one that still falls short of disclosing that data, will not bring a dispute to conclusion, it is necessary to think about how the litigation is likely to proceed.

When a data controller is a respondent in litigation, and the claimant-data subject challenges the explanation that the data controller has given in purported fulfilment of its obligation of explainability, the data controller does not unilaterally judge whether the information that it has given amounts

210 *See, e.g.*, Kroll, *supra* note 109, at 7.

211 *Cf. id.* at 3 ("It may in many cases be unnecessary to understand the precise mechanisms of an algorithmic system, just as we do not understand how humans make decisions, so long as we describe the outlines of the system's interaction with the world.").

to "meaningful information."[212] Disputes over explainability that go to litigation, even at a future time when the question of defining the legal standard of "meaningful information" will have been answered, thus will include disputes over whether a data controller has satisfied explainability. Data subject's counsel will demand that the explanation be scrutinized. It is hard to see how scrutiny could be refused. The explanation will likely be scrutinized in two respects: the factual question of whether it is accurate, and the legal question of whether it is meaningful.

Although the common terms of Articles 13–15 do not explicitly say that the data controller, when providing "meaningful information," must tell the truth, it would be bizarre to read a rule that requires an explanation as meaning that the party subject to the rule is free to give an explanation that is either inaccurate or false. It is implicit that "explanation," whatever its degree of specificity or precision, shall be *accurate* explanation. Herein lies the crux of a problem described in Section IV.C above: to judge whether an explanation of a machine learning output is accurate in the relevant sense, without being shown the data behind the output, is impossible. The legal logic of an explainability dispute demands accuracy of explanation. Accuracy of explanation can be tested only with precise explanation, and precision, because of the character of machine learning, means precision about the data.

There is as yet no binding definition of "meaningful information" tested in actual disputes concerning the GDPR.[213] The first generation of disputes, now arising, will likely involve, among other things, skirmishes over precisely what constitutes "meaningful information." It seems clear that supplying mere "information" will not be enough. The common terms of Articles 13–15 of the GDPR call for something more than mere "information." The standard of "meaningful information" in the GDPR would be devoid of purpose if it required nothing more than proffering words

212 *See* Whaling in the Antarctic (Austl. v. Japan), Judgment, 2014 I.C.J. 226 (Mar. 31).

213 Cases concerning other topics have considered the expression "meaningful information" as employed in other EU texts. *See, e.g.*, Case 73/84, Denkavit Futtermittel GmbH v. Land Nordrhein-Westfalen, 1985 E.C.R. 1013 (discussing the meaning of a meaningful information provision in a compound feed directive); Case T-61/99, Adriatica di Navigazione SpA v. Comm'n of the European Cmtys., 2003 E.C.R. II-5349, ¶ 34 (acknowledging meaningful information requirement in market notice definition); *cf.* Case T-442/12, Changmao Biochemical Eng'g Co. v. Council of the European Union, Judgment ¶¶ 129, 137 (June 1, 2017), http://curia.europa.eu/juris/celex.jsf?celex=62012TJ0442&lang1=en&type=TXT&ancre= [https://perma.cc/U6KR-5EE6] (describing an applicant's plea that she had been deprived "meaningful information on the method used for calculating normal value for DL tartaric acid").

that are an explanation in form. The standard of "meaningful information" must itself be meaningful. In a litigation setting, the claimant-data subject will insist on having evidence that the data controller has satisfied it.

The rules of evidence, in the adversarial proceedings that some claimants will insist on running through to a final judgement, equip the claimant to place demands on the data controller. The claimant will demand evidence of how the data controller trained the machine. No doubt the data controller will object to the relevance of such evidence. The data controller will object that whatever explanations it already has given are all that are needed to satisfy the explainability provisions of the GDPR. It is doubtful, however, that attempts to lessen the immediate burden of explanation will protect the data controller from demands by claimants to see the data that trained the machine.

E. "Any Further Explanation Would be Impossible or Disproportionate."

It is likely that data controllers, attempting to fend off demands for more information about automated decisions, will invoke the proportionality of compliance burden to compliance benefit. This is similar to the defense that says "we gave a meaningful explanation" but merits separate treatment because its weakness is fundamental.

Principles of risk mitigation and proportionality for some time have given data controllers a degree of safe harbor under privacy law.[214] Absolute compliance has seldom been the goal in regulatory regimes—it has not been in EU privacy law—and so, a data controller might argue, it will not be in respect of explainability either. The problem is, anything less than a review of the data used to train the machine will be challenged by a claimant-data subject as falling short of a real explanation. The data subject will challenge the sufficiency of the information that the data controller has supplied.

In setting forth a proportionality defense in response to an explainability claim, the data controller likely would refer to Recital 62 of the GDPR. Recital 62 addresses exceptions to "the obligation to provide information."[215]

---

214 *See* Article 29 Data Protection Working Party, *Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks*, at 2 (May 30, 2014) (noting that a "scalable and proportionate approach to compliance," a "risk-based approach," was "well known under the current Directive 95/46/EC," which was repealed on May 25, 2018, in particular its Articles 8, 17, and 20); *see generally* Charlotte Bagger Tranberg, *Proportionality and Data Protection in the Case Law of the European Court of Justice*, 1 INT'L DATA PRIVACY L. 239 (2011) (discussing the importance and history of proportionality). The principle that compliance measures need not be out of proportion to risk is visible in international privacy guidelines as well. *See* THE OECD PRIVACY FRAMEWORK ¶ 18 (2013).

215 Council Regulation 2016/679, *supra* note 3, at 12.

One of the exceptions is that "the provision of information to the data subject proves to be impossible or would involve a disproportionate effort."[216] This is a proportionality defense, which, as this Article notes above, is available in respect of data privacy failures. The difficulty is that, when an explanation is sought for a machine learning output, the training set must be examined to explain the output and validate the output's explanation.[217]

There is a fundamental problem with a proportionality defense against a claim for explanation: an explanation is valid or it is not; a proportionality test cannot achieve validity.

It is accordingly difficult to see how an exception for "disproportionate effect" would be applied to explaining a machine learning output: without the training data, the output is not explained, and without disclosure of the data, any explanation of the output is not susceptible to effective scrutiny. This is not a situation in which it will do to "sort of explain," "kind of explain," or "explain in a manner not imposing a disproportionate effort on the explainer." "Explanation" of a machine learning output, if the explanation is to be tested for its truthfulness or accuracy, must expose the data that trained the machine. These are disputes in which (or in some of which) data subject-claimants will allege invidious intent. Even where they allege mere mistake or misadventure, they will do so in an adversary process. They will seek validation of whatever explanation they are given, and it is hard to see a principled way for a court to refuse. Critical here is to be clear about the validation at issue: there is validation of the machine learning process that produced a decision, and there is validation of the explanation of that process. The present challenge entails the second validation.

There may be ways to implement a compliance obligation that gets the proportion right between cost and regulatory benefit, e.g., assuring a high degree of privacy protection while accepting the harm of occasional privacy failure; there is no "proportionate" version of a truthful explanation. Unless the exception of "disproportionate effort" under Recital 62 is allowed to supersede the rule of explanation in the common terms of Articles 13–15, it is hard to see how the exception would apply. It is an exception that, if applied in respect of explanation, would swallow the rule.

---

216 *Id.*

217 Indrė Žliobaitė and Bart Custers, in a paper published in 2016, argue strongly that, even for its own internal purposes, a data controller probably will not arrive at a reliable indication of whether it has respected the rules against invidious discrimination unless it considers sensitive personal data. Indrė Žliobaitė & Bart Custers, *Using Sensitive Personal Data May be Necessary for Avoiding Discrimination in Data-Driven Decision Models*, 24 ARTIFICIAL INTELLIGENCE & L. 183, 183–84 (2016).

VI. DOES ANONYMIZATION OFFER A WAY OUT?

In seeming response to the conflict between privacy and explainability, the GDPR provides for certain "[s]afeguards and derogations relating to processing [of data] for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes."[218] Article 89, paragraph 1, provides as follows:

> Processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes, shall be subject to appropriate safeguards, in accordance with this Regulation, for the rights and freedoms of the data subject. Those safeguards shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation. Those measures may include pseudonymisation provided that those purposes can be fulfilled in that manner. Where those purposes can be fulfilled by further processing which does not permit or no longer permits the identification of data subjects, those purposes shall be fulfilled in that manner.[219]

It is no surprise that the GDPR contains an exception for anonymization of data. The exception acknowledges that situations arise that call for the retention of personal data, and it reflects the practice of present-day data controllers to address privacy concerns by taking steps to obscure the identity of people whose data they control.[220] It is not clear, however, whether Article 89, paragraph 1 was drafted with the situation in mind where a claimant has challenged an explanation of a machine learning result that directly concerns a legal right or interest of the claimant.

At the heart of Article 89, paragraph 1 is the concept of "appropriate safeguards." The premise behind Article 89, paragraph 1 is that, in the current state of the art, personal data can be stored in a manner that conceals such information as would be needed to identify the individual to whom the data relate. No doubt, at some level of concealment data would not be sufficient to identify the people to whom they relate. For example, a list of blood types with no other information would not suffice to attribute a particular item on the list to a particular individual. Data sets that contain more information, however, present the risk of de-anonymization even if an individual's data is not immediately visible to an untrained eye.

Article 4, paragraph (1) of the GDPR defines personal data to be "any information relating to an identified *or identifiable* natural person ('data

---

218 Council Regulation 2016/679, *supra* note 3, art. 89.

219 *Id.*

220 *Cf.* INFO. COMM'R'S OFFICE, *supra* note 65, at 30, 42 (encouraging data controllers to "periodically review the data you hold, and erase or anonymise it when you no longer need it").

subject')."[221] By "identifiable natural person," Article 4 means "one who can be identified, *directly or indirectly*, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."[222] Here, the GDPR acknowledges that identification of an individual is possible through indirect means. The strides that have been made in identifying individuals, notwithstanding sophisticated anonymization of data,[223] suggest the significance of this point.

The drafters of the GDPR implicitly acknowledged, albeit in the recitals, that what passes for anonymization today might not tomorrow. Recital 26 addresses this problem:

> To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments.[224]

Advances in technology will likely challenge the effectiveness of anonymization. "Technological developments" already have demonstrated that anonymization is not necessarily a robust protection of privacy. In 2007, Arvind Narayanan and Vitaly Shmatikov demonstrated that the Netflix Prize dataset, though anonymized, did not protect the privacy of the data subjects whose data it contained.[225] They concluded that "the adversary with a small amount of background knowledge about an individual can use it to identify, with high probability, this individual's record in [that] anonymized dataset and to learn all anonymously released information about him or her, including sensitive attributes."[226]

Much of the focus on privacy protection has been on this sort of solution: anonymize the data set, and the problem is thought to be solved.[227] True, technical advances undo the anonymization strategies that might have

---

221 Council Regulation 2016/679, *supra* note 3, art. 4(1) (emphasis added).

222 *Id.* (emphasis added).

223 *See, e.g.*, Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 INST. ELECTRICAL & ELECTRONICS ENGINEERS SYMP. ON SECURITY & PRIVACY, Spring 2008, at 111, 111.

224 Council Regulation 2016/679, *supra* note 3, at 5.

225 Bruce Schneier, *Why 'Anonymous' Data Sometimes Isn't*, WIRED (Dec. 12, 2007, 9:00 PM), https://www.wired.com/2007/12/why-anonymous-data-sometimes-isnt/ [https://perma.cc/Q37J-7UCG].

226 Narayanan & Shmatikov, *supra* note 223, at 111.

227 *Cf.* INFO. COMM'R'S OFFICE, *supra* note 65, at 30, 42 (encouraging data controllers to "periodically review the data you hold, and erase or anonymise it when you no longer need it").

worked yesterday, and so they no longer work today. A widespread view holds that further technical advances will solve the problem.[228]

Even if data controllers develop a foolproof method for anonymizing data, however, the challenges here involve legal rules as well as technology. As this Article has suggested already with respect to the term "meaningful information," some of the most important legal rules remain untested. The exceptions set out in Article 89 are among the untested rules. It is open to question how far those exceptions are to be extended. Article 89's exceptions involve data controllers who have a right to hold and use data under certain conditions and for certain purposes. It is not clear that those conditions or purposes include exposing data in a legal proceeding for purposes of testing the sufficiency of an explanation in the Articles 13–15 sense.

Yet exposing the data is what claimants will demand, because they will demand a meaningful test of the data controller's explanation. No doubt a legislator (or even a judge) could extend exceptions such as those set out in Article 89 so as to protect a data controller whom a claimant has called upon to prove the accuracy and truthfulness of its explanations. To extend the exceptions in this way, one would need clearly to articulate the limit of the exceptions. One would need a detailed plan prescribing the mechanisms within which a process of proof would take place. Formulating such a plan is not in itself an insurmountable task; modern courts have prescribed structural remedies which require governments to change administrative, judicial, and other public institutions.[229] Such remedies take time to achieve,

228 These advances are being pursued through substantial work on differential privacy. Differential privacy is a mathematical guarantee on identifiability of individuals from a dataset. (PLEASE PROVIDE SOURCE) Informally, a set of statistics computed from a dataset is said to satisfy differential privacy if a person analysing them cannot determine whether or not a particular individual's data was used in computing those statistics. (PLEASE PROVIDE SOURCE) The work on differential privacy by Cynthia Dwork is of particular note. Dwork and her co-authors were awarded the 2017 Gödel Prize for their 2006 paper that introduced the concept. *2017 Gödel Prize*, EUR. ASS'N FOR THEORETICAL COMPUTER SCI., https://www.eatcs.org/index.php/component/content/article/1-news/2450-2017-godel-prize [https://perma.cc/CE5Y-KC4Z]; Cynthia Dwork et al., *Calibrating Noise to Sensitivity in Private Data Analysis*, THIRD THEORY CRYPTOGRAPHY CONF., Mar. 2006, at 265, 265. The work has stood the test of time: in the workshop on privacy in machine learning held as part of the 2019 NeurIPS conference, 25 out of the 42 accepted papers concerned differential privacy. *Privacy in Machine Learning*, PRIML'19, https://priml-workshop.github.io/priml2019/ [https://perma.cc/Y5DM-E8TJ]. The general thrust of those 25 NeurIPS papers is to offer novel and useful differentially-private statistics for particular classes of data. In other words, researchers know what they have to do to guarantee privacy, they just need clever mathematics to achieve it.

229 *See, e.g.*, Xenides-Arestis v. Turkey, App. No. 46347/99, Judgment, ¶ 40 (Mar. 22, 2006), http://hudoc.echr.coe.int/eng?i=001-71800 [https://perma.cc/ER9T-38VS] (requiring that "the respondent State must introduce a remedy which secures the effective protection of the rights [at issue] in relation to the present applicant as well as in respect of all similar applications pending before the Court"); Broniowski v. Poland, 2005-IX Eur. Ct. H.R. 1, 7–8 (acknowledging a previous decision "[finding] that [a] violation had originated in a systemic

and they are contested along the way. Until such a structural remedy was put in place, however, the problem of the fallibility of anonymization would remain.

The UK Information Commissioner's Office, the UK independent public body having regulatory responsibilities under, *inter alia*, the GDPR,[230] has addressed anonymization. It acknowledges studies that have cast doubt on the anonymity of "anonymised" data,[231] and it acknowledges that "[i]t may not be possible to establish with absolute certainty that an individual cannot be identified from a particular dataset, taken together with other data that may exist elsewhere."[232] It then suggests that risk mitigation, not the assurance of privacy, is the approach to be taken.[233] This Article has noted that the law takes a risk-based approach to privacy protection; compliance measures in respect of privacy need not be disproportionate to the risk of compliance breach. This Article has posited that such an approach is not applicable to the explanation of a machine learning result, where the complaining party alleges that the training data embedded some invidious (and illegal or tortious) bias. The complaining party is likely in that case to insist on a detailed evaluation of the data that trained the machine. In view of how machine learning works, nothing less than a detailed evaluation will produce a meaningful (and validated) explanation. In view of how adversary proceedings work, proof that the data set really is what the data controller says it is will be demanded.

It is likely that, if data from a large dataset used in machine learning is opened for evaluation, the time will come when another data subject will complain that his personal data came to light, by which he suffered injury—and that this circumstance supports a finding that some party, whether the data controller or another, is liable to make reparation. The precise contour of a liability claim will depend on the law in the particular case, including procedural rules in the jurisdiction in which the claim is brought, as well as

problem connected with the malfunctioning of domestic legislation and practice caused by the failure to set up an effective mechanism to implement the 'right to credit' of Bug River claimants"). Steps also have been taken to protect privacy in particular, such as the privacy of minor defendants in criminal trials. *See* T. v. United Kingdom, App. No. 24724/94, Judgment, ¶ 121 (Dec. 16, 1999), http://hudoc.echr.coe.int/eng?i=001-58593 [https://perma.cc/8B92-SHHB].

230 The U.K. Information Commissioner's Office ("ICO") also has responsibilities under the Data Protection Act 2018 ("DPA"), the Freedom of Information Act 2000 ("FOIA"), and various other statutes. *See* INFO. COMM'R'S OFFICE, MANAGEMENT AGREEMENT 2018–2021 38–39 (2018), https://ico.org.uk/media/about-the-ico/documents/2259800/management-agreement-2018-2021.pdf [https://perma.cc/EKK3-BKZS].

231 INFO. COMM'R'S OFFICE, *supra* note 60, ¶ 133.

232 *Id.* ¶ 134.

233 *Id.* ¶¶ 134–35.

on the facts of the case. However, claims of this general contour seem a plausible result of the tension between privacy and explainability.[234]

The inadequacy of anonymization is fundamental. When it comes to an adversarial legal procedure, a party to the dispute is likely to reject anonymization in any form. A claimant-data subject, alleging that a data controller has employed an automated decision-making process to the detriment of the subject—and in breach of a rule like that against gender discrimination[235]—will pursue a legal remedy against a data controller. In pursuing the remedy, the claimant is likely to challenge the very act of anonymization regardless of the technical proficiency with which it was performed. To anonymize the data is, by definition, to obscure the data. To obscure the data is to obscure how the decision was reached.

This will not be the first time that the legal system has struggled to reconcile such competing interests in a decision-making process. In the United States, controversy has grown in recent years over anonymous juries, a matter canvassed in a recent prize paper at Cornell.[236] An anonymous jury is a departure from the "usual case" in which "the parties know the names, addresses, and occupations of potential jurors, as well as those of any spouses."[237] It is used especially in cases involving allegations of terrorism or organized crime, where prosecutors have argued that the identity of the jurors needs to be concealed for their safety, or in order to make it more

234 Such claims would not be the first under EU law in which privacy rights were pitted against another public order value. In *European Commission v. Bavarian Lager Co.*, "Regulation (EC) No 45/2001 . . . on the protection of individuals with regard to the processing of personal data" restricted the disclosure of the minutes of a meeting at the Commission. Case C-28/08 P, European Comm'n v. Bavarian Lager Co., 2010 E.C.R. I-6112, ¶ 3. The meeting had been attended by certain regulators and others, but not by the Bavarian Lager Co., though the meeting concerned the Company's grievance in respect of measures that effectively restricted the sale of German beer in the UK. *See id.* at I-6126–28. The respondent invoked Regulation No. 1049/2001, Council Regulation 1049/2001, 2001 O.J. (L 145) 43 (EC), for public access to EU documents and sought the minutes of the meeting; the Commission refused to disclose the names of the participants at the meeting, on grounds that this would breach their rights under Regulation No. 45/2001, *Bavarian Lager*, 2010 E.C.R. at I-6130. The Court sided with the Commission. *Id.* at I-6147. As a writer commenting on the case observed, the Court thus "[put] the right to data protection in conflict with the freedom of information rules." ORLA LYNSKEY, THE FOUNDATIONS OF EU DATA PROTECTION LAW 126 (2015). This Article notes Lynskey's point that "data protection" and "privacy" are not precisely the same concept. *Id.* at 89. That they are congeners, however, is evident in their concurrent treatment in a range of situations. *See id.*

235 *See, e.g.*, Council Directive 2006/54, *supra* note 3, art. 14 (prohibiting gender discrimination affecting "access to employment").

236 *See generally* Leonardo Mangat, Note, *A Jury of Your [Redacted]: The Rise and Implications of Anonymous Juries*, 103 Cornell L. Rev. 1621 (2018).

237 United States v. Ross, 33 F.3d 1507, 1519 n.22 (11th Cir. 1994).

difficult for them to be suborned.[238] A central objection to anonymous juries is that the parties are unable to scrutinize the jurors to learn if they harbor invidious prejudice.[239] Appellate courts in some U.S. states have determined the use of anonymous juries at trial to be unsound on constitutional grounds.[240]

The difficulty with an anonymous jury is that, if you do not know the juror, then you will not know whether the juror is appropriate for your case. Anonymity will render any right you might have to challenge the juror ineffective.[241] Similarly, if you do not know the data that trained the machine learning system, you will not know whether the decision the system gave is subject to some impermissible bias in that data. Nor will you be able to satisfy yourself whether the explanation the data controller supplied you of the machine learning decision is valid. Your right to challenge the decision thus will be as illusory as the right to challenge the anonymous juror.

The difficulty with explainability is the demand that a claimant will place on the data controller in litigation. A claimant will demand, first, an explanation adequate to serve the two purposes of explanation that this Article has identified: supplying a guide for future behavior and demonstrating that the decision or output was reached on grounds that are neither mistaken nor invidious. And a claimant will demand, second, information sufficient to establish whether the explanation is accurate and truthful. It is hard to see how claimants will accept occluded, altered, or otherwise incomplete data, when the gravamen of a claim is that in the data lurks an erroneous or invidious ground in breach of right—and that the data controller is, either willfully or by error, concealing this. Doubts exist as to whether anonymization will work even on its own terms, because it is getting easier to identify people even when their data has been anonymized.[242] Even if anonymization does work as a technological function, it is doubtful that it will be acceptable to litigants challenging a machine learning decision. To assert that anonymization is effective at concealing whose data was used to train the machine might give comfort to the people from whom the data came. It gives no comfort to the claimant who has a right to an explanation of an output produced by a machine trained on that data.

238 *See* Mangat, *supra* note 236, at 2, 18 (describing how safety and harassment concerns drove the use of anonymous juries in a variety of cases, including those involving terrorism and organized crime). Anonymous juries seem to have begun in a federal prosecution in New York in 1977, spreading to other states "in the 1990s." Christopher Keleher, *The Repercussions of Anonymous Juries*, 44 U.S.F. L. REV. 531, 534–39 (2010).

239 Mangat, *supra* note 236, at 1639.

240 *See* Keleher, *supra* note 238, at 543–46 (describing state appellate cases where courts reversed decisions by anonymous juries due to concerns about anonymity's effect on defendants' constitutional rights).

241 *See* Mangat, *supra* note 236, at 1639 ("[P]eremptory challenges are . . . hampered when an anonymous jury is empaneled.").

242 Narayanan & Shmatikov, *supra* note 223.

Another approach, and perhaps a more promising one, would be some form of *in camera* review of datasets. Such an approach might use protected facilities for the examination of the data,[243] supported by protective orders or the like from the competent court.[244] Facilities are in widespread use already for the purpose of showing data to restricted audiences under conditions that prevent or impede disclosure of the data to other parties. "Data rooms" are a common feature of intellectual property-intensive industries like biotech, where potential investors need to satisfy themselves that they understand what they might be investing in, and the intellectual property owners need to satisfy themselves that their secrets are safe.[245] The European Commission has used such facilities in administrative procedures.[246] But an analogy from such settings is precarious: a data set containing millions of data points will involve the privacy rights of many individuals; the individuals have not necessarily consented to the review of their data by judges or other assessors in the litigation that will call for such review. What will be demanded is a path to reviewing the data that does not trespass upon the privacy rights of the people from whom the data came. Such a path is likely to prove elusive.

243 Writers have begun to suggest a range of possible approaches to controlling and limiting access to sensitive data. *See generally* Jenn Rolnick Borchetta, *Curbing Collateral Punishment in the Big Data Age: How Lawyers and Advocates Can Use Criminal Record Sealing Statutes to Protect Privacy and the Presumption of Innocence*, 98 B.U. L. REV. 915 (2018). Most, if not all, proposed approaches are chiefly administrative, even when they might involve adversarial procedures. *See, e.g.*, *id.* at 927 (discussing sealing statutes). The difficulty is that, in the adversarial procedure, litigants will call for exposure of the data and thus test the effectiveness of any mechanism that limits access to data. For some thoughtful suggestions about how adversarial procedure—in the criminal setting—is likely to entail demands for exposure of how computers work. *See generally* Chessman, *supra* note 8. Chessman's emphasis on source code gives too little attention to the data-driven approach that characterizes machine learning, but his Note is valuable for what it says about adversarial procedure and computer-generated evidence. *See id.*

244 *See* Uniloc 2017 LLC v. Microsoft Corp., No. 8:18-CV-02053-AG (JDEx), 2019 WL 451345, at *1, 5 (C.D. Cal. Feb. 5, 2019) (granting a protective order in respect of "protected data," including on grounds of GDPR protections).

245 Data rooms are employed in a range of settings where one party has grounds for scrutinizing a set of assets and the holders of the assets have grounds for protecting confidentiality. *See, e.g.*, Davey v. Money [2018] EWHC (Ch) 766 [176] (Eng.) (describing use of "an electronic data room for prospective purchasers").

246 *See* Case T-194/13, United Parcel Services, Inc. v. European Comm'n, 2017 EUR-Lex-CELEX LEXIS ¶ 15 (Mar. 7, 2017) (describing Commission's use of a data room for "external legal counsel to examine . . . confidential extracts from internal documents"). There are also situations in which privacy is said to be at odds with other values, such as national security, when surveillance programs entail the gathering of information about individuals. Closed judgments and other forms of special review have been applied when claimants have called for scrutiny of surveillance programs. *See, e.g.*, Privacy Int'l v. Sec'y of State for Foreign & Commonwealth Affairs [2018] UKIPTrib 15_110_CH 1 [2]–[3], [6].

CONCLUSION

The GDPR requires explanations of certain machine learning results. But it does not say what explanations constitute adequate ones. To say that, one needs to understand how machine learning works and one needs to appreciate why and in what setting explanations are needed.

This Article has suggested how a layperson might use an explanation of a machine learning result. The layperson might use an explanation as a guide to modify his behavior in an effort to get a better result. The data controller fails if she does not supply an explanation that enables the layperson to use the explanation.

Another way the layperson might use an explanation is to test whether the machine learning decision has infringed any of the various public values that the law protects, such as equality of treatment. The concern behind explainability, thus, is more than the fact that the machine gives answers that a layperson might respond with changes in conduct—answers to questions such as "how much money is to be awarded" or "should the applicant get a license to sell fish and chips?" The explanation is also about serving public values in general. Explanations will be demanded in order to validate—or to invalidate—decisions that data controllers have reached with the aid of machine learning.

Just as much as one needs to understand why a data subject is asking for explanations, one needs to understand the procedural setting in which a data subject is asking. The procedural setting will affect how the data subject asks and what precisely the data subject will ask for. It is not necessarily in a spirit of collaboration, or in sedate, deliberative settings, that questions will be asked about machine learning decisions; they will be fought over in court.[247] One of the two main purposes for seeking an explanation of a machine learning output will be to contest the validity of the process behind the output. A data controller might address the data subject with a form of words

247 This Article has not aimed here to canvass the emerging docket of GDPR-related cases, which is of worldwide scope. This Article notes that parties in litigation in U.S. courts have invoked the privacy provisions of the GDPR in both claims and defenses; the results are mixed, at best. *See, e.g.*, Corel Software, LLC v. Microsoft Corp., No. 2:15-cv-00528-JNP-PMW, 2018 WL 4855268, at *1 (D. Utah Oct. 5, 2018) (denying Microsoft's defense that production of Telemetry Data would "raise[] tension with Microsoft's [GDPR] obligations"); Finjan, Inc. v. Zscaler, Inc., No. 17-cv-06946-JST (KAW), 2019 WL 618554, at *3 (N.D. Cal. Feb. 14, 2019) (rejecting defendant's claim that GDPR protected emails and ordering production); D'Amico Dry D.A.C. v. Nikka Fin., Inc., No. CA 18-0284-KD-MU, 2018 WL 5116094, at *4 (S.D. Ala. Oct. 19, 2018) (rejecting in most part defendant's claim that the GDPR required a protective order for videotaped testimony where defendant was "aware his deposition [wa]s to be videotaped").

that seem to explain the process, but the data subject will not accept an explanation by the data controller as *ipse dixit*. Although different legal systems take different approaches, it is a characteristic of adversary trial procedure that litigants seek to impugn one another's assertions and their defenses.

There is a great deal already written on algorithmic explainability, some of it dealing with machine learning and much of it addressed to compliance practitioners. But litigation pushes in a different direction to the systems design and engineering approach that compliance practitioners and theorists are at home with. Control, recording, containment, systematization, and oversight—these are concepts of the engineering culture behind compliance operations. They aim for reliability and stabilization. They are not concepts that will be helpful when defending against the claim: "You didn't provide me with a meaningful explanation of your machine learning system's decision."

This Article has suggested that the litigation that likely will result from disputes over explainability will pose challenges to privacy. The UK Information Commissioner's Office Discussion Paper on *Big Data, Artificial Intelligence, Machine Learning and Data Protection* concluded in 2017 that "while data protection can be challenging in a big data context, the benefits will not be achieved at the expense of data privacy rights."[248] This Article takes this statement to mean, by extension, that the Office believes that "data privacy rights" can be preserved when a controller explains a machine learning output, including when she is under scrutiny in litigation. This Article is not confident that a clear path exists under present legal and institutional arrangements to achieve that result. Nor is it likely that anonymization, a technical measure, will solve the problem because anonymization obscures the data and it is the data that a litigant will demand to see when challenging a machine learning output. Claimants will not accept anonymized data assurances just as as criminal defendants would not accept a prosecutor's assurance that a secret jury really consists of unbiased peers.

The impetus toward exposing the data that trained the machine, thus, does not just come from the technologist, who understands that vague or general accounts of the data obscure the basis of the output. It also comes from the data subject, who understands that such accounts do not supply the information she needs as a guide to future conduct and as a test to validate the explanation.

248 INFO. COMM'R'S OFFICE, *supra* note 60, ¶ 3.

The difficulties that this Article has outlined might arise in any number of jurisdictions that employ adversarial trial procedures because the GDPR has already influenced legislation far afield. The California Consumer Privacy Act ("CCPA") of 2018[249] is an example.[250] It exceeds the scope of the present article to review globally the adopted rules that address machine learning. The difficulties should be kept in mind as lawmakers, regulators, lawyers, and regulated parties consider the implications of legislation now in force and the rules yet to come.

This Article's purpose has been to draw attention to the difficulties by considering the GDPR in light of the technology that it is intended to regulate—and in light of the challenges which claimants will make when they demand that data controllers explain their decisions. European law today contains provisions to protect privacy and provisions to assure that those who use machine learning explain its outputs to those whom those outputs affect. It is not self-evident, in the present state of the law and practice, how both of those objectives will be fulfilled. A period of trial-and-error, of experimentation with methods for compliance, and of dispute and litigation, lies ahead.

The text of the GDPR does not contain all the answers to the questions it raises about machine learning. Explainability will involve contests over the scope and meaning of "meaningful information." It further will involve demands for meaningful testing of the data that have trained machine learning systems. Meaningful testing, in turn, will entail exposure of datasets. And exposure of datasets will present new challenges for the right to privacy and the institutions entrusted with its protection.

249 Assemb. B. 375, 2017–2018 Reg. Sess. (Cal. 2018); *see, e.g.*, Sarah Hospelhorn, *California Consumer Privacy Act (CCPA) vs. GDPR*, VARONIS (updated June 17, 2020), https://www.varonis.com/blog/ccpa-vs-gdpr/ [https://perma.cc/HHX3-58M9] (comparing the GDPR with the California Consumer Privacy Act).

250 *Cf.* Hertza, *supra* note 65, at 1730 ("The EU's newly-adopted GDPR is a good source of inspiration for reforms to the FCRA and ECOA.").