Meaningful information and the right to explanation

Andrew D. Selbst* and Julia Powles**

Key Points

- There is no single, neat statutory provision labelled the 'right to explanation' in Europe's new General Data Protection Regulation (GDPR). But nor is such a right illusory.
- Responding to two prominent papers that, in turn, conjure and critique the right to explanation in the context of automated decision-making, we advocate a return to the text of the GDPR.
- Articles 13–15 provide rights to 'meaningful information about the logic involved' in automated decisions. This is a right to explanation, whether one uses the phrase or not.
- The right to explanation should be interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise his or her rights under the GDPR and human rights law.

Introduction

In May 2018, Europe's new General Data Protection Regulation (GDPR) will apply across all European Union Member States.[1] The GDPR is an ambitious, complicated, contested law aimed at making Europe 'fit for the digital age'.[2] Among the law's many provisions are several related to automated decision-making, notably, Article 22 and certain provisions of Articles 13–15.

These provisions, which restrict automated decisions and require associated safeguards, are causing consternation among researchers, lawyers, and others concerned with decisions made by machine learning (ML) or artificial intelligence (AI). ML or AI systems are, among possible modes of decision-making, uniquely in danger of defying any human understanding.[3] Automated decisions without any human intervention or understanding would seem to flout European ideas of autonomy and personhood.[4] Therefore, these provisions exist to provide some meaningful information to data subjects about how their data is used. There is a fierce disagreement over whether these provisions create a data subject's 'right to explanation'. This article seeks to reorient that debate by showing that the plain text of the GDPR supports such a right.

The 'right to explanation' debate has, in part, so captured imaginations because it is knotty, complex, and a non-trivial technical challenge to harness the full power of ML or AI systems while operating with logic interpretable to humans. This issue has drawn immense interest from the technical community.[5] There is also

* Data & Society Research Institute, New York, NY, USA; Yale Information Society Project, New Haven, CT, USA.

** Cornell Tech, New York, NY, USA; New York University, New York, NY, USA; St John's College, University of Cambridge, Cambridge, UK.

The authors wish to thank Jef Ausloos, Solon Barocas, Kiel Brennan-Marquez, Lilian Edwards, Sorelle Friedler, Mireille Hildebrandt, Joris van Hoboken, Orla Lynskey, Viktor Mayer-Schönberger, Frank Pasquale, Ira Rubinstein, Suresh Venkatasubramanian, Sandra Wachter, and Nicolo Zingales for comments on earlier drafts. Selbst and Powles are grateful for the support of the NSF under grants IIS-1633400 and CNS-1704527, respectively.

1 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC [2016] OJ L 119/1.

2 European Commission, Reform of EU Data Protection Rules (2017) <http://ec.europa.eu/justice/data-protection/reform/ index_en.htm> accessed 16 August 2017.

3 J Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 1; C Kuner and others, 'Machine Learning with Personal Data: Is Data Protection Law Smart Enough to Meet the Challenge?' (2017) 7 IDPL 1.

4 M Hildebrandt, 'The Dawn of a Critical Transparency Right for the Profiling Era' in J Bus and others (eds), *Digital Enlightenment Yearbook* (IOS Press, Amsterdam, NL 2012); ML Jones, 'Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood' (2017) 47 Soc Stud Sci 216.

5 See eg R Caruana and others, 'Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission' (2015) Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15) 1721; D Bamman, 'Interpretability in Human-Centered Data Science' (2016) CSCW Workshop on Human-Centered Data Science; M Gleicher, 'A Framework for Considering Comprehensibility in Modeling' (2016) 4 Big Data 4, 75; D Finale and B Kim, 'A Roadmap for a Rigorous Science of Interpretability' (2017) arXiv:1702.08608; E Horvitz, 'On the Meaningful Understanding of the Logic of Automated Decision Making',

rapidly increasing interest from a legal perspective, with a number of scholars beginning to explore both the importance of explanation as a normative value within the ML or AI context,[6] as well as whether there is a requirement for explanation as a matter of positive law.[7]

The legal debate so far has concerned a conception of the right oddly divorced from the legislative text that best seems to support it. The most prominent contributions are two explosive papers out of Oxford, which immediately shaped the public debate.[8] The first paper, by Bryce Goodman and Seth Flaxman, asserts that the GDPR creates a 'right to explanation', but does not elaborate much beyond that point.[9] The second paper, from Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, asserts that no such right presently exists.[10] It does so by unnecessarily constraining the idea of the 'right to explanation', while conceiving of a different 'right to be informed' that amounts to a right to a particular type of explanation. We believe that Wachter and others' response is an overreaction to Goodman and Flaxman that distorts the debate, and that neither paper meaningfully addresses the most relevant provisions supporting such a right—specifically those that create rights to 'meaningful information about the logic involved' in automated decision-making.

This debate appears headed in the wrong direction because it is missing such a major piece. Whether one uses the phrase 'right to explanation' or not, more attention must be paid to the GDPR's express requirements and how they relate to its background goals, and more thought must be given to determining what the legislative text actually means. This article offers a positive conception of the right, located in the text and purpose of the GDPR.

Background

Legislative provisions at issue

Articles 13(2)(f), 14(2)(g), and 15(1)(h) of the GDPR require data controllers to provide data subjects with information about 'the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject'. Articles 13 and 14 are notification duties imposed on data controllers and Article 15 provides a right to access information throughout processing.

Article 22(1), in turn, elaborates that data subjects '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'. Article 22(2)–(4) specifies limited circumstances where automated decision-making is permitted, and provides for different safeguards so that data subjects can effectively exercise their 'rights and freedoms and legitimate interests'. Most important for this discussion, Article 22(3) provides that where automated decision-making is contractually necessary or consensual, certain safeguards for data subjects must apply, including 'at least the right to obtain human intervention on the part of the controller,

BCLT Privacy Law Forum, 24 March 2017 <https://www.law.berkeley.edu/wp-content/uploads/2017/03/BCLT_Eric_Horvitz_March_ 2017.pdf> accessed 16 August 2017.

6 See eg F Pasquale, Black Box Society: The Secret Algorithms That Control Money and Information (Harvard UP, Boston, USA 2015); M Hildebrandt, 'The New Imbroglio: Living with Machine Algorithms' in L Janssens (ed), *The Art of Ethics in the Information Society* (Amsterdam UP 2016); K Brennan-Marquez, '"Plausible Cause": Explanatory Standards in the Age of Powerful Machines' (2017) 70 Vanderbilt L Rev 1249; AD Selbst, 'A Mild Defense of Our New Machine Overlords' (2017) 70 Vanderbilt L Rev En Banc 87; R Binns, 'Algorithmic Accountability and Public Reason' (2017) Philosophy & Technology, doi:10.1007/s13347-017-0263-5; KJ Strandburg, 'Decision-Making, Machine Learning and the Value of Explanation', The Human Use of Machine Learning: An Interdisciplinary Workshop, 16 December 2016, <http://www.dsi.unive.it/HUML2016/assets/Slides/Talk%202.pdf> accessed 16 August 2017.

7 See eg Hildebrandt (n 4); D Kamarinou, C Millard and J Singh, 'Machine Learning with Personal Data' (2016) ssrn:2865811 forthcoming in R Leenes and others (eds), Data Protection and Privacy: The Age of Intelligent Machines (Hart, Oxford, UK 2017); I Mendoza and LA Bygrave, 'The Right Not to Be Subject to Automated Decisions Based on Profiling' in T Synodinou and others (eds), EU Internet Law: Regulation and Enforcement (Springer, Cham, CH 2017); A Rantanen, 'A Yleinen Tietosuoja – Asetus ja Oikeus Selitykseen Algoritmien Päätöksenteon Logiikasta' (2017) Working Paper, <https://makingmydatareal.files.wordpress.com/2017/03/mmdr_1_20172.pdf> accessed 20 September 2017; L Edwards and M Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" is Probably Not the Remedy You Are Looking For' (2017) 15 Duke L Techno Rev, forthcoming.

8 See eg E Chiel, 'EU Citizens Might Get a "Right to Explanation" About the Decisions Algorithms Make' (*Fusion*, 5 July 2016) <http://fusion.kinja.com/eu-citizens-might-get-a-right-to-explanation-about-the-1793859992> accessed 1 July 2017; C Metz, 'Artificial Intelligence is Setting Up the Internet for a Huge Clash with Europe' (*Wired*, 11 July 2016) <https://www.wired.com/2016/07/artificial-intelligence-setting-internet-huge-clash-europe/> accessed 1 July 2017; I Sample, 'AI Watchdog Needed to Regulate Automated Decision-making, Say Experts' (*The Guardian*, 27 January 2017) <https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions> accessed 1 July 2017; M Burgess, 'Watching Them, Watching Us: Can We Trust Big Tech to Regulate Itself?' (*Creative Review*, April 2017) <https://www.creativereview.co.uk/watching-watching-us/> accessed 1 July 2017.

9 B Goodman and S Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2016) ICML Workshop on Human Interpretability in Machine Learning, arXiv:1606.08813 (v3); (2017) 38 AI Magazine 50.

10 S Wachter, B Mittelstadt, and L Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 IDPL 76.

to express his or her point of view and to contest the decision'.

Non-binding Recital 71 includes a tweak on the safeguards in Article 22(3), by specifying that safeguards for data subjects 'should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, [and] *to obtain an explanation of the decision reached after such assessment and to challenge the decision*' (emphasis added). The omission of the italicised part from Article 22 actually accounts for most of the conflict between Goodman and Flaxman and Wachter and others, and is a point to which we will return. For now, it is sufficient to note that Recital 71 is not meaningless, and has a clear role in assisting interpretation and co-determining positive law. While not enforceable as such, it gives clear support to the effective exercise of data subject rights under Articles 13–15 and 22.

Two caveats

Before getting into the substantive arguments, we want to introduce two caveats to our discussion regarding what this article does not address. First, the practical effect of Article 22(1) and, by association, the referring provisions in Articles 13–15, are a lively matter of dispute. As Wachter and others identify, if the condition 'solely' in 'decisions made solely by automation' is interpreted narrowly, the safeguards and associated requirements of meaningful information will have limited applicability. In particular, Wachter and others suggest even a trivial degree of human involvement could render Article 22(1) inapplicable.[11] As such a reading would render the written provisions nearly purposeless, however, it seems wise to question it.[12] Though these provisions share some language with their predecessors in Articles 12(a) and 15(1) of the Data Protection Directive,[13] there is very little determinative guidance about how they should be interpreted and applied in practice.[14] Moreover, new draft guidelines from the Article 29 Working Party (on behalf of all European data protection authorities) take the position that trivial human involvement will not suffice.[15] Beyond 'solely', Article 22(1) is further conditioned by the requirement that automated decision-making 'produces legal effects concerning [the data subject] or similarly significantly affects him or her'. Given these issues go to the applicability of the right rather than the shape of the right,[16] and will be a matter for future interpretation by legislators, data protection authorities, and courts, we do not consider them further here.

A second important caveat is that our concern here is with the legal requirements, not the technical feasibility of meeting those requirements. If it turns out that the law asks something that certain technologies cannot provide, then those technologies cannot be used without changing or violating the law.[17] That is a possibility that must be accepted in this discussion.

A right to explanation in the GDPR

We believe that a plain reading of Articles 13(2)(f), 14(2)(g), 15(1)(h), and 22 supports a right to explanation. Accordingly, before engaging with the existing scholarly debate, we offer here our interpretation of the text.

Articles 13–15

When an individual is subject to 'a decision based solely on automated processing' that 'produces legal effects . . . or similarly significantly affects him or her', the GDPR creates rights to 'meaningful information about the logic involved'. While it will eventually fall to legislators, data protection authorities, and courts to interpret when particular information may or may not be 'meaningful', we make four observations here. The first we believe must be true, while the remaining three points are clearly open to interpretation. We offer them as arguments for how the right should be treated if it is to have the impact that is suggested by the GDPR's overall trend towards strengthening data protection as a fundamental right.[18] We conclude with our interpretation of the second half of Articles 13(2)(f), 14(2)(g), and 15(1)(h)—

11 ibid 92.

12 Kamarinou, Millard, and Singh (n 7) 11–12; Mendoza and Bygrave (n 7) 87–88; and LA Bygrave, 'Minding the Machine: Art 15 of the EC Data Protection Directive and Automated Profiling' (2001) 17 CLS Rev 17, 20.

13 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] OJ L 281/31.

14 Mendoza and Bygrave (n 7); T Zarsky, 'Incompatible: The GDPR in the Age of Big Data' (2017) 47 Seton Hall L Rev 995, 1016.

15 Article 29 Working Party, Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (2017) <http://ec.europa.eu/newsroom/just/document.cfm?doc_id=47963> accessed 21 October 2017 ('Draft Guidelines'), 10: 'To qualify as human intervention, the [data] 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 available input and output data.' Note that these guidelines were issued while this paper was in press, so their content and pagination may change.

16 See further Bygrave (n 12) 19–20; Mendoza and Bygrave, ibid 88–91; Edwards and Veale (n 7) 19–20; Draft Guidelines, ibid 10–12.

17 Hildebrandt (n 6) 58.

18 O Lynskey, *The Foundations of EU Data Protection Law* (OUP, Oxford, UK 2015).

'the significance and the envisaged consequences of such processing for the data subject'.

First, because Articles 13–15 all relate to the rights of the data subject, meaningful information should be interpreted in relation to the data subject.[19] That is, the information about the logic must be meaningful *to her*, notably, a human and presumably without particular technical expertise.

Second, the test for whether information is meaningful should be functional, pegged to some action the explanation enables in the data subject, such as the right to contest a decision as provided by Article 22(3).[20] Broadly considered, there are two ways of understanding the value of explanation in this context: as an instrumental value or an intrinsic one—a fundamental aspect of autonomy and personhood.[21] Both are important, but the instrumental focus offers a more concrete way to measure whether the explanation is meaningful enough. The intrinsic value of explanations tracks a person's need for free will and control, most familiarly expressed in the desire to avoid living out the plot of a Franz Kafka novel.[22] But under such an approach, it is difficult to even have a discussion about how much or what kind of explanation is required, and the answers will likely be different for different people. A focus on explanation's intrinsic value, therefore, risks weakening the right to explanation as the amorphous concept of autonomy comes into opposition with well established, fiercely defended, and concrete interests in, for example, trade secrecy.[23]

Third, and relatedly, there should be a minimum threshold of functionality for the information provided. That is, the information should be at least meaningful enough to facilitate the data subject's exercise of her rights guaranteed by the GDPR and human rights law. As one example, if an individual receives an explanation of an automated decision, she needs to understand the decision well enough to determine whether she has an actionable discrimination claim. This interpretation is supported by Article 5's requirement that data processing be lawful, fair, and transparent to the data subject, as well as Article 12's emphasis on intelligibility and requirement that '[t]he controller shall facilitate the exercise of data subject rights'.

Fourth, the requirement should be interpreted flexibly. Specific rules defining the right methodologically may be too rigid, unnecessarily constraining research and development. For example, one might think that meaningful information should include an explanation of the principal factors that led to a decision.[24] But such a rigid rule may prevent beneficial uses of more complex ML systems such as neural nets, even if they can be usefully explained another way. As we discuss below, Wachter and others are primarily concerned with refereeing between different rigid versions of the right—specifically, whether it refers to explanation of specific decisions or the logic of the system, and whether it applies *ex ante* or *ex post*. We believe such ironclad separations miss the point. Rather, a flexible approach, guided by the functional requirements discussed above, can best effectuate this right while preserving the ability of technologists to innovate in ML and AI. Additional support for a flexible interpretation comes from the different translations of the phrase 'meaningful information'. The German text of the GDPR uses the word 'aussagekräftige', the French text refers to 'informations utiles', and the Dutch version uses 'nuttige informative'. These formulations variously invoke notions of utility, reliability, and understandability. These are related, but not identical concepts, suggesting that a flexible, functional approach will be most appropriate.[25]

Articles 13–15 require that, in addition to meaningful information, the data subject be told the 'significance and the envisaged consequences of such processing for the data subject'. We can see two ways to think about that additional text, both of which bolster the analysis

19 Kamarinou, Millard, and Singh (n 7) 20.

20 See Draft Guidelines (n 15) 16: 'The [data] controller must provide a simple way for the data subject to exercise these rights', and 'The data subject will only be able to challenge a decision or express their view if they fully understand how it has been made and on what basis'.

21 T Lombrozo, 'The Instrumental Value of Explanations' (2011) 6 Philosophy Compass 539. We consider intrinsically valuable explanations, described above, as an *aspect* of autonomy and personhood, distinct from explanations that allow a person to *achieve* autonomy and personhood, which would make the autonomy concern instrumental.

22 DJ Solove, 'Privacy and Power: Computer Databases and Metaphors for Information Privacy' (2001) 53 Stanford L Rev 1393, 1421.

23 See generally, G Malgieri, 'Trade Secrets v Personal Data: A Possible Solution for Balancing Rights' (2016) 6 IDPL 102.

24 See eg J Grimmelmann and D Westreich, 'Incomprehensible Discrimination' (2017) 7 Calif L Rev Online 164, 173. The Draft Guidelines (n 15) 15 provide other typical examples, but stop short of being prescriptive, noting that meaningful information 'will in most cases' require details such as: 'the information used in the automated decision-making process, including the categories of data used in a profile; the source of that information; how any profile used in the automated decision-making process is built, including any statistics used in the analysis; why this profile is relevant to the automated decision-making process; and how it is used for a decision concerning the data subject.' The Draft Guidelines later offer further support for flexible interpretation: 'information about the categories of data that have been or will be used in the profiling or decision making process and why these are considered pertinent will generally be more relevant than providing a complex mathematical explanation about how algorithms or machine-learning work, although the latter should also be provided if this is necessary to allow experts to further verify how the decision-making process works.' ibid 29.

25 See LM Solan, 'Interpreting Multilingual Laws: Some Costs and Benefits' in J Jemielniak and AL Kjaer (eds), *Language and Legal Interpretation in International Law* (OUP, Oxford, UK, forthcoming).

above, but in different ways. One interpretation is that the significance and envisaged consequences constitute information about how the results of the automated processing get used. For example, a data controller that is using automated processing to determine loan provision would first offer meaningful information about the decision-making process itself, and then the 'significance and envisaged consequences' would be the resulting downstream effects—that a loan will or will not be granted and at a certain interest rate.

An alternative interpretation is that the second half of the phrase conditions the first, further refining the right to explanation. This position would hold that it is not meaningful information about the logic of the system in general that is required, but specifically the logic of how the system treats the data subject.[26] That is to say that the meaningful information would be responding to the input data and the processing, rather than laying out an explanation of processing that can apply to any input data, and determining the consequences later.

Between these two interpretations, we are agnostic, and due to space limitations, cannot investigate further. Importantly, though, we believe that the former interpretation would make the envisaged consequences a separate requirement from the right to explanation, and the latter interpretation would incorporate the language. As a result, we make no claims about the ultimate extent of the right to explanation, other than to tie it primarily to meaningful information and to offer the above proposed inferences based on the text and purpose.

Article 22

Our primary focus in reorienting the right to explanation debate is to call attention to the phrase 'meaningful information about the logic involved' in Articles 13–15. Much of the debate to date has glided over these articles, instead seeking to locate or debunk the right in Article 22(3). We flip this approach, and in doing so, observe that Article 22 and Recital 71 support the reading of Articles 13–15 as an independent source of the right.

Under Article 22, all cases of permissible automated decision-making must include 'suitable measures to safeguard the data subject's rights and freedoms and legitimate interests'. Where Article 22(3) applies, it provides a non-exhaustive list of possible safeguards, including 'the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision'. Articles 22(2) and 22(4), which make up the other cases, include no list of examples. Recital 71, which supplements Article 22 as a whole, proposes additional safeguards, including 'specific information to the data subject' and a 'right to ... obtain an explanation of the decision reached after such assessment'.

A right to explanation is, therefore, neither endorsed nor limited by the discussion of safeguards in the text. The only concrete description of the safeguards in the text is their purpose—to safeguard rights, freedoms, and legitimate interests. This is the very same purpose we ascribe to Articles 13–15 under the separate analysis above. This suggests that though a right to explanation cannot be derived from Article 22 itself, Article 22 nonetheless supports the existence of that right derived from Articles 13–15.

The 'right to explanation' debate

We believe the discourse about the right to explanation has gone in an unproductive direction so far. Here we offer a critique of the two most prominent papers in the debate.

The original claim

Goodman and Flaxman's paper, an explainer for a technical audience, reads the provisions cited as creating a right to explanation. The section of the article making this argument is rather short and the argument is not fleshed out. In the original conference version of their paper,[27] Goodman and Flaxman based their argument of the existence of a right on Recital 71: 'Although [Article 22] does not elaborate ... beyond "the right to obtain human intervention", the GDPR recitals state that a data subject has the right to "an explanation of the decision reached after [algorithmic] assessment" '.[28] Goodman and Flaxman later changed the argument, omitting Recital 71 entirely, and deriving the right from Articles 13–14, presumably because recitals are inherently non-binding and, therefore, cannot impose a 'requirement'.[29] But they did not clearly express how the right would work, and they do not discuss Article 15's identical language at all.[30]

26 This is slightly different from the point above that the information must be meaningful *to* the data subject. The point above allows for general information as long as the data subject can make use of it, whereas this interpretation would require a tailored explanation every time.

27 B Goodman and S Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2016) ICML Workshop on Human Interpretability in Machine Learning 26, arXiv:1606.08813 (v1).

28 ibid 28.

29 Goodman and Flaxman (n 9) 50, 55.

30 ibid.

Notably, though Goodman and Flaxman seem to envision a fairly narrow version of the right, which would provide for explanations of a specific decision about a person, they do not expressly argue—outside the abstract, at least—that the right is actually that narrow. They argue that explanation means the 'trained model can be articulated and understood by a human' and state that 'any adequate explanation would, *at a minimum*, provide an account of how input features relate to predictions . . .' (emphasis added).[31] While Goodman and Flaxman do not do nearly enough to defend their claim of a right to explanation, they also do not make strong claims about the scope of such a right.

The response

Wachter and others were ready with a response. But as we elaborate in the following discussion, the cure is arguably worse than the disease. Wachter and others' paper is problematic for two broad reasons. First, the paper makes irresponsible rhetorical moves regarding the 'right to explanation'. It gives the impression of a strong argument against a right to explanation, while in reality it defines the right to explanation down, attempts (and fails) to debunk the narrowed version, and offers a new 'right' that should have been included in any fair definition of a 'right to explanation' in the first place. It is pushing back on a proxy, an envisioned phrase that the authors see as meaning something different than its ordinary meaning. As a result, it claims to debunk something it does not actually try to fairly interpret. This is not only disingenuous but dangerous, as it invites less scrupulous or more time-pressed advocates to cite the paper for the proposition that there is no right to explanation, which is not even what the paper argues in substance.

Secondly, the argument itself, limited though it is, relies on an artificial analytical framework, which in turn relies on unstated legal and technical assumptions. The conclusions Wachter and others reach almost entirely fall out of the framework they create. Worryingly, Wachter and others treat the framework as fact and other scholars have begun relying on it as well.[32] But this framework is merely an intellectual abstraction, not a necessary part of either the law or technology, and it does not work.

Wachter and others' analytical construct includes two distinctions. One is *system functionality* versus *specific decisions*. They define the former as the 'logic, significance, envisaged consequences and general functionality' of a system and the latter as reasons for a specific outcome, 'e.g. the weighting of features, machine-defined case-specific decision rules, information about reference or profile groups'. The other distinction is *ex ante* versus *ex post* explanations—that is, explanations provided before or after automated decision-making. When analysed carefully, this strict construct leads to certain conclusions about the scope of data controllers' duties under Articles 13–15 that render them much weaker, and conflict with our interpretation discussed above. We do not believe the drafters of the GDPR intended such a result, but whether one agrees or not, we aim here to illustrate the unfounded assumptions and unsettling implications of such an analytical frame.

The proxy 'right to explanation'

The article makes a number of conflicting assertions about the 'right to explanation'. The title makes the sweeping claim that the 'right to explanation of automated decision-making does not exist'. In the introduction, the article tempers that claim, instead arguing against a 'right to explanation of specific automated decisions of the type currently imagined elsewhere in public discourse'.[33] Whether they intend to challenge the concept of the right itself or a more limited version 'imagined elsewhere' is not always clear. Where they explain their analytical frame, Wachter and others argue that the four states—an *ex ante* or *ex post* explanation of *system functionality* or *specific decisions*—are the full scope of 'what one may mean by an "explanation" of automated decision-making'.[34] It would, therefore, seem to follow logically that a 'right to explanation' could encompass any or all of them. Thus, as they consistently argue that a 'right to explanation' does not exist, it would seem they are making a quite broad claim indeed.

But they are not. Wachter and others introduce a proposed 'right to be informed', 'a limited *right to explanation* of the functionality of automated decision-making systems' (emphasis added).[35] They find in the law a right to a particular kind of explanation in the very same paper that argues—repeatedly—that a 'right to explanation' does not exist. They can only do this because they equate the 'right to explanation . . . elsewhere in public discourse', or 'as popularly proposed', with a right to *ex post* explanations of *specific decisions*. But if those two ideas were conflated in the public discourse, then the right thing is to suggest that explanation is

31 ibid; (n 26) 29.
32 See eg Edwards and Veale (n 7).
33 Wachter and others (n 10) 78.
34 ibid 78.
35 ibid 96.

more complicated than that, not to artificially narrow the concept into a straw man to knock down.

This rhetorical gamesmanship is irresponsible and could have disastrous real-world effects. As described above, the GDPR clearly mandates 'meaningful information about the logic' of decisions to which Article 22 applies. If 'meaningful' is to have any substance, that appears on its face to be a move in the direction of explanation of *some* type—and all parties in this debate, including Wachter and others, seem to agree on that point. But it is almost impossible to imagine that the brief filed by the defendant in the first case to construe this provision will not immediately cite an article titled 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' to try to evade or dilute that requirement. Moreover, 'a right to be informed' sounds weaker than 'a right to explanation' or even 'a right to be meaningfully informed'—even though the authors expressly argue that it is a right to explanation (of *system functionality*). That is to say, Wachter and others go out of their way to name their proposed right in a way that happens to downplay its significance even on their own terms. There is no good reason to do so.

As we analyse the substance of Wachter and others' argument in the next section, we treat their paper as arguing for a 'right to explanation' of *system functionality*, but not *specific decisions*.

The consequences of the analytical frame

The *system functionality—specific decision* distinction

Almost all of Wachter and others' conclusions come from their analytical frame. But the framework collapses upon examination. Start with the *system functionality—specific decision* distinction. The idea is broadly useful, but they build the argument on the idea that requiring explanation of a *specific decision* is a heavier burden than an explanation of *system functionality*. This is not obviously true in all or even most cases, and appears to be based on a misunderstanding of the technology.

ML systems find patterns in training data, and build models based on those patterns. New inputs are then shown to the model, and the model returns a result.[36] That result can take a number of forms, including, among other things, a classification of the input or the probability that the inputs are classified a certain way. Though some models output probability distributions, the system itself is generally[37] deterministic; that is, given the same inputs to the same model, the same output will result, unless the model changes.

As a result, in many systems, a *complete* system-level explanation tells you everything about specific cases.[38] Even though ML models are complex and often inscrutable, and probability may be involved in the creation of a model, the determinism makes them predictable. Therefore, where it is possible to generate an explanation for *system functionality*, it should be possible to generate an explanation of *specific decisions* given the input data. Wachter and others seem to treat ML systems as inherently probabilistic, suggesting that 'the use of complex probabilistic analytics' is a hindrance to explanation of *specific decisions*, even where it does not similarly hinder explanations of *system functionality*. There is no reason in general for this to be so.

Moreover, the items that they list in their definitions betray a misunderstanding of the realistic separation of the categories. As they define it, explanations of *system functionality* include among other things, 'the system's requirements specification, decision trees, pre-defined models, criteria, and classification structures' and *specific decisions* include 'the weighting of features, machine-defined case-specific decision rules, information about reference or profile groups'. But there is no such thing as 'case-specific decision rules'—the model is the model, and the rules that constitute the model decide all cases. The same for the weighting of features. It is true that the input data is needed to determine the *specific decision*, but once the model is built and the input data is known, the logic determines the outcome.

Now, Wachter and others do not argue that the GDPR requires a complete explanation of *system functionality*, and neither do we. But the GDPR does require 'meaningful information about the logic involved'. Given that Wachter and others' analysis states that the GDPR requires explanation of *system functionality* but not *specific decisions*, it also implies that it is possible to provide meaningful information about *system functionality* that does not give the data subject meaningful information about *specific decisions*.[39]

36 P Flach, Machine Learning: The Art and Science of Algorithms that Make Sense of Data (CUP, Cambridge, UK, 2012) 11.

37 We use the word 'generally' here not to be loose with our claims, but rather to be more precise. We cannot be sure this is true for every system. As far as we know, current machine learning systems create models that are deterministic in the end, but the field is rapidly developing, and it is conceivable that there might be a use case that develops for which randomised outputs are useful.

38 Horvitz (n 5).

39 We want to be clear here, that ML systems as currently built are often not explainable from either a *specific decision* or *system functionality* standpoint. Developing tools for explanation is an active area of computer science research. See ZC Lipton, 'The Mythos of Model Interpretability' (2016) ICML Workshop on Human Interpretability in Machine Learning, arXiv:1606.0490; MT Ribeiro, S Singh, and C Guestrin, 'Why Should I Trust You?: Explaining the Predictions of Any

What would such a distinction look like? Is it possible to construct a meaningful explanation that includes information from the first list and not the second? That is hard to say, but it clearly depends on one's definition of 'meaningful'. Would it be meaningful to provide a list of all the features in the model, without enough information about how they are weighted to make an individual determination? It would perhaps allow a data subject to know what features were examined and enable a degree of error checking, but it is not much of an explanation. How about the system's 'decision trees'? It is hard to see how a decision tree could be meaningful to the average person but fail to inform about a *specific decision*.

Wachter and others might instead be thinking of outliers. But this still does not solve their problem. One possible type of *system functionality* explanation is black-box testing to determine the average importance of individual feature types to the output.[40] Whether this would be meaningful information depends on the data subject it is given to. Such an analysis explains a good amount to the typical data subject, but nothing at all to outliers. That is, the same explanation might be meaningful to some data subjects but not others. Ultimately, however, this still does not cleave *system functionality* from *specific decisions*, because the right to meaningful information is held by the individual data subject.

The only other instance we can imagine of a system that is able to provide an explanation of *system functionality* but not *specific decisions* is one that has a randomised output. Given Wachter and others' discussion of probability, that appears to be their model. Putting aside the atypicality of such a model, in such a case *specific decisions* are not explainable because there is nothing to explain. The *system functionality* sets up a random number, then a virtual die is rolled. That the ultimate outcome is the result of a random process does not add anything to the meaningful information provided by the explanation of *system functionality*. If it was meaningful, as we have defined it, it remains so; if it was not, it remains so. The takeaway from this discussion is that if one accepts the instrumental definition of 'meaningful' we put forth earlier in the article, it is hard to see how to construct a meaningful explanation of *system functionality* that fails to offer meaningful information about *specific decisions*, because by the definition, a data subject would need enough explanation to vindicate her rights.

We consider one other possibility. Perhaps their understanding corresponds to a thin understanding of 'meaningful', akin to an autonomy placebo. This is why we distinguish between the autonomy rationale for explanation and the instrumental one. If 'meaningful' just means 'enough information so that people do not feel like they entirely lack control', then it is conceivable that Wachter and others' separation can exist and comply with the GDPR because people differ wildly on such a subjective metric. In this case a basic explanation of the *system functionality* akin to 'we examine features X, Y, and Z and determine an outcome' could, in theory, suffice. But neither an instrumental nor a more robust autonomy rationale would permit the distinction they draw. And such an interpretation seems to go against the purpose of the GDPR to ensure more robust rights for data subjects.

The timing distinction

The *ex ante—ex post* distinction similarly falls apart. Wachter and others' main purpose in making this distinction is to state that *ex ante* explanations of *specific decisions* are impossible because a decision must be reached before it can be explained. But the discussion above demonstrates the falsity of that claim. As soon as *system functionality* can be explained, *specific decisions* can be as well. Again, this is because most models are deterministic once created. To determine the outcome, it is true that input data must be run through the model, which means that data is processed. But the *ex ante—ex post* distinction the authors draw is not before and after processing, but rather before and after a decision. Because *specific decisions* can be revealed following data processing, without requiring an ultimate decision, the timing distinction does not make sense; explanations of both *specific decisions* and *system functionality* are available before issuing an ultimate decision.

Even worse, because the explanation right is derived from Articles 13–15, the person with a right to the explanation is the data subject herself, and she has the input data (or has access to it via the very same articles). Therefore, it would be theoretically possible to provide

Classifier' (2016) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135; A Henelius and others, 'A Peek Into the Black Box: Exploring Classifiers by Randomization' (2014) 28 Data Mining and Knowledge Discovery 1503; A Datta, S Sen, and Y Zick, 'Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems' (2016) Proceedings of the 2016 IEEE Symposium on Security and Privacy 598; P Adler and others, 'Auditing Black-Box Models for Indirect Influence' (2016) Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM) 1. Under any version of a right to explanation, including Wachter and others' 'right to be informed', these systems as designed will likely run into legal trouble. But in this discussion, we are interested only in whether Wachter and others' treatment of the *difference* between *system functionality* and *specific decisions* is justified. To analyse that, we must assume a case in which explanation is possible for one and not the other.

40 See Datta, Sen, and Zick, ibid.

a system explanation to the data subject whereby *she herself* could process her data and determine her answer, so even if processing were the timing touchstone, this distinction would not work. (Though, such a feat would likely require disclosing the overall model, and trade secrecy could get in the way. More on that later.)

Finally, there is only a limited legal basis for the timing distinction that Wachter and others seek to draw. None of the relevant articles refers to decision timing explicitly. Articles 13–14 refer to notification duties of data controllers and Article 15 speaks to the access rights of the data subject. The only reference to a pre- or post-decision timing element comes from Recital 71, which informs Article 22, not Articles 13–15. Article 13 does require notification 'at the time when personal data are obtained' but once a data controller has the personal data, if they are able to comply with Article 13(2)(f) with an explanation of *system functionality* then they can also comply with a *specific decision*, as discussed above. This conclusion is reinforced by the Draft Guidelines, which state that information to be provided under Article 15(1)(h) should already have been provided under Articles 13(2)(f) and 14(2)(g).[41]

The resulting legal analysis

With that background, we can evaluate Wachter and others' legal arguments. First, responding to their narrow (*ex post*, *specific decision*) reading of the 'right to explanation', they note that such a right is articulated only in Recital 71, but not Article 22, making it non-binding. They further argue that such a configuration was an intentional choice to remove the right to explanation from Article 22(3), seeking support in the inconclusive[42] drafting history of various versions of the GDPR text. Regardless of the merits of this argument, it is not an argument against a right to explanation in general, but only their cramped version of it. Moreover, both Wachter and others, and Goodman and Flaxman, ignore the positive value of Recital 71.

Next, Wachter and others turn to Articles 13–14. They argue that because Articles 13–14 impose notification duties at the time data is collected, they cannot possibly require *ex post* explanation, and thus cannot require explanations of *specific decisions*. Again, this is an argument only against their proxy right, and they concede that an explanation of *system functionality* might be required. The discussion above also demonstrates both that their timing argument is incorrect and meaningful explanation of *system functionality* enables the necessary inferences about *specific decisions*.

Finally, Wachter and others turn to Article 15. They note that because Article 15 relates to requests for information during or after data processing, it lacks the timing problem they have ascribed to Articles 13–14. Nonetheless, they argue that because the relevant language is identical in all three Articles, there is no substantive difference, and Article 15 cannot require explanations (of *specific decisions*) either. But the timing problem drove their initial analysis of what meaningful information is in Articles 13–14, so bootstrapping that analysis to Article 15 because the language is the same makes little sense. Wachter and others then turn to an extended analysis of German and Austrian data protection law under the Directive, which contained a similar, but not quite equivalent right. But this analysis, though informative and welcome for its civil law focus, is not conclusive, as it relies on non-binding interpretations by just two Member States of underutilised provisions of the Directive, now replaced with a law designed to enhance data subject rights.

Crucially, because Wachter and others are so wedded to their analytic framework, at no point in their legal discussion do they ever engage with the text of Articles 13(2)(f), 14(2)(g), and 15(1)(h) on its own terms. While in its previous incarnation, Article 12(a) of the Directive required that data subjects have access to '*knowledge of the logic involved* in any automatic processing' (emphasis added), Articles 13–15 require information on 'the existence of automated decision-making [and] *meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*' (emphasis added). The difference is clear—the GDPR specifically added a requirement of meaningful information, as well as information about the significance and consequences for an individual. At no point do Wachter and others

41 Draft Guidelines (n 15) 15. The Draft Guidelines later offer some support for Wachter and others' analysis separating *specific decisions* from *system functionality*: 'Article 15 implies a more general form of oversight, rather than a right to an explanation of a particular decision' ibid 24. As we have discussed above, however, we do not believe that explanations of system functionality are so easily separable from specific decisions, and that ultimately, the touchstone for the type of explanation required will revert to the meaningfulness requirement. See ibid 29 (n 23).

42 Wachter and others' argument on this point centres on a mid-negotiation proposal that would have expressly included 'the right to … an explanation of the decision reached after [human] assessment' in art 22(3), and which was subsequently rejected: (n 10) 81. In general, many proposals were made and rejected during the negotiations (see eg M Vermeulen, 'Regulating Profiling in the European Data Protection Regulation: An Interim Insight into the Drafting of Article 20' (2013) EMSOC Working Paper, ssrn:2382787), and we must be careful drawing too much inference from them. The GDPR's starting point is the Directive, so it is inconclusive in itself if the negotiators failed to reach a consensus in adding to the Directive text. It is also conceivable that the lack of direct correspondence between art 22(3) and recital 71 is because certain of those rights—in particular, to information and to explanation—were replicated elsewhere, such as in arts 12–15.

substantively analyse what 'meaningful information about the logic' means.

Wachter and others also fail to consider the context provided by other substantive changes that broadly enhance data subject rights against automated decisions. For example, there is no equivalent in the Directive to the GDPR's requirement in Article 22(3) of 'the right to obtain human intervention on the part of the controller ... and to contest the decision'.[43] There is a new focus on profiling,[44] as a subset of the broader issue of automated decision-making, in order to improve data subjects' ability to deal with these activities. Enhancements to Articles 5 and 12 reinforce the GDPR's emphasis on meaningful transparency and accountability, in a way that is useful, intelligible, and actionable to the data subject. Wachter and others' analysis omits any substantive discussion of these textual changes and shifts in emphasis, which will bear on the interpretations of the text. At one point in their paper, Wachter and others claim that '[t]he GDPR appears to offer less protection to data subjects concerning explanations' than the Directive.[45] That is simply an unfathomable reading, and any analysis that leads to it should be considered suspect.

The last point worth discussing relates to Wachter and others' analysis of trade secrecy and their right to explanation of *system functionality*. They argue that the right they propose 'could be heavily curtailed to protect the controller's interests (eg trade secrets, intellectual property)'.[46] But the argument is unpersuasive. The precedents that Wachter and others rely on are Member State interpretations of the Directive. Not only was the Directive expressly repealed by the GDPR, but in considering the relationship between explanation and trade secrets, the textual changes in the GDPR alter the balance in favour of explanation. And given Wachter and others' dismissal of Recital 71 as a decisive feature in rejecting the right to explanation, it is striking that, in the absence of any supporting Articles in the GDPR, they draw on Recitals 47 and 63 in making such a claim. Recitals 47 and 63 merely recognise that data controllers may have relevant rights and interests, but they also expressly provide for the overriding rights of data subjects, especially in relation to information to be provided by the data controller. There is no justification for treating trade secret restrictions as axiomatic under an entirely new law with a new emphasis.

Conclusion

Articles 13–15 provide rights to 'meaningful information about the logic involved' in automated decisions. We think it makes sense to call this a right to explanation, but that point is less important than the substance of the right itself. We believe that the right to explanation should be interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise his or her rights under the GDPR and human rights law.

doi:10.1093/idpl/ipx022

43 Mendoza and Bygrave (n 7) *contra* Bygrave (n 12).

44 See Hildebrandt (n 4); Mendoza and Bygrave, ibid.

45 Wachter and others (n 10) 89.

46 ibid 90.