DEBATING THE RIGHT TO EXPLANATION

An Autonomy-based Analytical Framework

A most controversial legal tool featured in Europe's General Data Protection Regulation ("GDPR"), the right to explanation ("RTE"), has sparked broad public attention and heated academic debate. Despite controversies and scepticism about its conceptual clarity and practical enforceability, however, the RTE is widely recognised as a promising regulatory instrument that could be utilised to redistribute resources and responsibilities among the main stakeholders of the automated decision making process. Moreover, it holds great promise for empowering data subjects who suffer from information asymmetry and opacity and provides a powerful counterweight to the increasing prevalence of automated decision making in our digitalised everyday life. This article explores the controversies in RTE-related GDPR texts and other related legal instruments. It explores and analyses five aspects of complex issues that shape the RTE's scope, degree, quality, validity and nature. In the second part of this article, the author advances an autonomy-based theory to conceptualise the RTE to reconcile the issues above relating to the legislative texts of the European data protection law, and to clarify three pairs of tensions at the heart of the debate concerning the existence and shape of the RTE: (a) negative *versus* positive autonomy; (b) the instrumental *versus* the intrinsic value of the RTE; and (c) the holistic *versus* segregated approach of interpretation.

Michelle **MIAO**
LLB (East China University of Politics and Law),
LLM (Renmin University), LLM (New York University), DPhil (Oxford);
Associate Professor, Faculty of Law, The Chinese University of Hong Kong.

I. Introduction

1 Algorithmic decision making has become a prevailing characteristic of our current age of big data and artificial intelligence. Artificial intelligence ("AI") models excel at training powerful models from large datasets, often exceeding human-level ability. These data processing tasks can usually be deemed too massive, complex and laborious for human brains only equipped with a finite capacity to

compute. In addition to computational efficiency, compared with fallible humans who are vulnerable to the influences of emotion, impulses, intuitions and biases, computer algorithms provide promising qualities such as fairness, neutrality and objectivity. It has been envisaged that machines may, in specific contexts, overcome "key limitations of human decision-makers and provide us with decisions that are demonstrably fair" so that data subjects affected by human decisions may appeal to and benefit from the algorithmic assessment.[1]

2 Essentially, automated decision making supported by algorithms is the soul of artificial intelligence. From speech censorship on social media platforms to ridesharing, from driving in automated vehicles to predicting future recidivism, algorithms have infiltrated almost the entire universe of contemporary social and economic transactions. During this algorithmic turn, many decisions traditionally solely made by humans, individually or collectively, are now exercised by highly complex and often opaque algorithms. Meanwhile, the perils of automated decision making have taken hold of public discourse among the sceptics and the very tech elites who promoted the growth of the AI industry. Elon Musk, for instance, admitted that he feared that the threat of artificial intelligence has become "a fundamental existential risk for human civilisation".[2]

3 Opacity is at the heart of these concerns. As algorithms take over tasks such as deciding loan granting, credit scoring and insurance qualifications, billions of individuals are left uninformed about how and why a particular automated decision that affects their essential interests was made and how they could adjust their own actions in response to these decisions. These AI algorithms are often opaque because data subjects at the receiving end often have little insight into the process through which AI models arrive at outcomes.[3] Advanced AI models of machine learning and deep neural networks, by nature, focus on *prediction* rather

1 Dimitra Kamarinou *et al*, "Machine Learning with Personal Data" in *Data Protection and Privacy: The Age of Intelligent Machines* (Ronald Leenes *et al* eds) (Hart Publishing, 2017) at p 111.

2 Aatif Sulleyman, "Elon Musk: AI Is A 'Fundamental Existential Risk for Human Civilisation' and Creators Must Slow Down" *Independent* (17 July 2017) <https://www.independent.co.uk/tech/elon-musk-ai-human-civilisation-existential-risk-artificial-intelligence-creator-slow-down-tesla-a7845491.html> (accessed 4 April 2022).

3 Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms" (2016) 3 Big Data & Soc'y 1; Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (2019) 1(5) *Nature Machine Intelligence* 206.

than offer avenues for *understanding*,[4] therefore heightening concerns about opacity.

4 With algorithmic transparency becoming a top concern for public debate and academic critique,[5] recent years saw explainable AI ("XAI") become one of the hottest topics in tech ethics and regulation.[6] However, issues of transparency in algorithmic decision-making have further severe collateral consequences ranging from the erosion of individual autonomy[7] to unfairness. The algorithmic turn accompanies a shift of power and autonomy from individuals to private tech companies and public institutions employing complex mathematical models to shape individual behaviours.[8] Other associated risks and harms include but are not limited to discrimination, unfairness, erosion of privacy and lack of accountability.[9]

---

4 Maithra Raghu & Erica Schmidt, "A Survey of Deep Learning for Scientific Discovery" (26 March 2020) at p 27 <https://arxiv.org/pdf/2003.11755.pdf> (accessed 8 April 2022).

5 Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke L & Tech Rev 18 at 41–43; Joshua Kroll *et al*, "Accountable Algorithms" (2017) 165 U PA L Rev 633; Tal Z Zarsky, "Transparent Predictions" (2013) 4 U Ill L Rev 1503; David Brin, *The Transparent Society: Will Technology Force Us to Choose Between Privacy And Freedom?* (Basic Books, 1998).

6 Derek Doran, "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives" <https://arxiv.org/pdf/1710.00794.pdf> (accessed 12 April 2022); Wojciech Samek *et al*, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer Nature, 2019); Pantelis Linardatos *et al*, "Explainable AI: A Review of Machine Learning Interpretability Methods" (2020) 23(1) *Entropy* 1; Robert R Hoffman *et al*, "Explaining Explanation for 'Explainable AI'" in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (SAGE Publications, 2018).

7 The theory of autonomy is the cornerstone of contemporary political philosophy, moral theory, ethics and law. Meanwhile, it is a highly contested "term of art introduced … to make sense of a tangled net of intuitions, conceptual and empirical issues, and normative claims" to the extent it can be regarded as a conceptual model to justify policies and principles, rather than an overarching, standalone concept. It essentially connotes the sense of the conditions free from external manipulations and coercion as well as the capacity of authentic self-rule. See, in general, Joseph Raz, "Autonomy and Pluralism" in *The Morality of Freedom* (Joseph Raz ed) (Oxford University Press, 1988); John Rawls, *A Theory of Justice* (Harvard University Press, 1971); Gerald Dworkin, *The Theory and Practice of Autonomy* (Cambridge University Press, 1988) at pp 34–47; and Joel Feinberg, "Autonomy" in *The Inner Citadel: Essays on Individual Autonomy* (John Christman ed) (Oxford University Press, 1989) at pp 27–53.

8 See, *eg*, Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown Publishing, 2016); and Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019).

9 See, *eg*, David Lyon, "Configuring the Networked Self: Law, Code and the Play of Everyday Practice" (2015) 65 U Toronto L J 130; and Daniel J Solove, "Data Mining and the Security-Liberty Debate" (2005) 74 U Chi L Rev 343.

5 Amongst calls for assessing and regulating algorithms,[10] the right to explanation[11] ("RTE") stands out as one of the latest legislative instruments created to respond to these new changes in the "Blackbox Society".[12] There are few more robust and more straightforward power equalisers than knowledge and transparency. Yet, since its academic conception through interpretation of relevant provisions of the primary European data protection law – the Regulation on the Protection of Natural Persons[13] ("GDPR"), the debate surrounding the RTE has not ceased.[14] This article sets out to review, summarise and clarify these controversies pertaining to the interpretation of legislative texts and practicality in utilising the RTE. Drawing from a growing field of RTE and transparency-related legal scholarship, this article will explain why there has been conceptual confusion and offers a sobering analysis of the promises and limits of this new addition to the legal arsenal of regulating AI algorithms. This article address two broad topics. Part II of this article will trace the legislative origin of the RTE, clarifying the main issues attracting debate on the legislative texts upon which the RTE was born. Part III turns to a brief discussion on an autonomy-based analytical framework to understand and reconcile those confusions and

10 Brent Mittelstadt *et al*, "The Ethics of Algorithms: Mapping the Debate" (2016) 3 Big Data & Soc'y 2.

11 This concept has been used broadly to refer to individuals' right to obtain detailed descriptions, justifications and reasoning of automated decisions which significantly impacts their legal, financial and other interests in the context of artificial intelligence and machine learning in particular. See, in general, Bryce Goodman & Seth Flaxman, "European Union Regulations on Algorithmic Decision-making and A 'Right to Explanation'" (2017) 38 *AI Magazine* 50; Mireille Hildebrandt, "The New Imbroglio – Living with Machine Algorithms" in *The Art of Ethics in the Information Society* (Liisa Janssens ed) (Amsterdam University Press, 2016); Bryce Goodman & Seth Flaxman, "European Union Regulations on Algorithmic Decision-making and A 'Right to Explanation'" (2017) 38 *AI Magazine* 50; Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76; and Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking for" (2017–2018) 16 Duke L & Tech Rev 18.

12 See, *eg*, Maayan Perel & Niva Elkin-Koren, "Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement" (2017) 69 Fla L Rev 181; and Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press, 2016).

13 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 (General Data Protection Regulation).

14 For one of the first scholarly elaborations on this subject, please see, *eg*, Bryce Goodman & Seth Flaxman, "European Union Regulations on Algorithmic Decision-making and A 'Right to Explanation'" (2017) 38 *AI Magazine* 50; Mireille Hildebrandt, "The New Imbroglio – Living with Machine Algorithms" in *The Art of Ethics in the Information Society* (Liisa Janssens ed) (Amsterdam University Press, 2016).

controversies in the conceptualisation of the RTE. The conclusion will offer a brief overview of the promises and limits of the RTE and where this new legislative instrument might lead us to in the future.

II. The General Data Protection Regulation – The legislative origin of the right to explanation

6 During the recent decade, the adoption and promulgation of the GDPR sparked intense debates among scholars, media commentators, legal practitioners, *etc*.[15] Most of the debate has centred on a series of provisions found in Chapter 3 of the GDPR that are believed to potentially give rise to the RTE. Although the GDPR was neither the earliest nor the sole legislative source that laid the foundation for a potential RTE, it was the most recent and high-profile legislation that garnered unprecedented intensive academic attention. Goodman and Faxman, among other pioneer scholars in this field, were the first to point out the potential to read the RTE into the text of the GDPR.[16] Their research identified Art 22 and Recital 71 as the legislative basis for the RTE, "whereby a user can ask for an explanation of an algorithmic decision that significantly affects them".[17] Resorting to a liberally-construed RTE in the European data protection law to improve transparency and accountability of automated decision making, however, was met with heavy criticism from scholars who were sceptical about both the legal foundation[18] and enforceability[19] of RTE.

7 Rather than construing the RTE from Art 22 and Recital 71, some turned their focus elsewhere in GDPR to find the legal basis for the RTE. For instance, Sandra *et al* divided explanations of automated decisions

15 See, *eg*, Bryce Goodman & Seth Flaxman, "European Union Regulations on Algorithmic Decision-making and A 'Right to Explanation'" (2017) 38 *AI Magazine* 50; Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76; and Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking for" (2017–2018) 16 Duke L & Tech Rev 18.

16 Bryce Goodman & Seth Flaxman, "European Union Regulations on Algorithmic Decision-making and A 'Right to Explanation'" (2017) 38 *AI Magazine* 50.

17 Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76.

18 Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76.

19 Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke L & Tech Rev 18 at 41–43.

into various categories based on: (a) the timing when explanations are made (*ie*, *ex post versus ex ante* explanations); and (b) the objectives of these explanations (*ie*, explanations of system functionality *versus* specific decisions).[20] They then state that the only legal source for a potential, albeit truncated RTE might be inferred from Art 15(1)(h), which grants "the right to be informed"[21] to data subjects about the logic, significance, envisaged consequences and general functionality of an automated decision-making system rather than specific circumstances and rationales regarding how and why a particular decision has been made. Edwards and Veale, for instance, stop short of making it explicit but suggest a similar approach in that Art 15 (which they term as access rights) is a possible way out for construing the RTE in the GDPR.[22] Still, others disagree with the analysis and insist on inferring the RTE from other parts of the existing text of the GDPR.[23] The growing debate surrounding the RTE, however, has not been limited to the interpretation of legislative texts. Even if the RTE could be established from the GDPR and other legal instruments, its scope and, in particular, efficacy in implementation have been the focus of intense discussion.[24]

8 This section of the article will address the five aspects of the GDPR provisions, which have sparked debate on the source, validity, scope and applicability of the RTE and will also propose a holistic approach to construing the GDPR to allow a flexible reading of the RTE into its text. These five issues, respectively, relate to: (a) the textual source of the RTE; (b) the nature of Art 22 (proscription *versus* entitlement); (c) the degree of automation in the process of decision making ("solely"); (d) the consequence of automated decision making (legal significance); and (e) the quality of explanation ("meaningful").

---

20 Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76.

21 Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76 at 78.

22 Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke L & Tech Rev 18 at 51–54.

23 See, *eg*, Andrew D Selbst & Julia Powles, "Meaningful Information and the Right to Explanation" (2017) 7 Int'l Data Privacy L 233 (arguing Arts 13–15 provide the basis for the RTE).

24 Emre Bayamlıoglu, "The Right to Contest Automated Decisions under the General Data Protection Regulation: Beyond the So-called 'Right to Explanation'" (2021) Reg & Governance 1.

A. The legislative source of the right to explanation: Article 22, Recital 71, or elsewhere?

9 Before venturing into the specific technicality of the RTE, the first hurdle of clarifying the validity of inferring the RTE from the GDPR is the division between Art 22 and Recital 71, both in their respective legal wording and their legislative efficacy. The main reason for claiming that the RTE should not be construed from the GDPR[25] is that a potential RTE can at best be founded on a Recital, not a main text Article of the GDPR, and that the former has no binding legal force. While this clear-cut approach offers a convenient solution to calm the debate, it also may obscure the level of complexity of legislative intentions and how varied and occasionally contradictory those intentions might be, especially when those legislative texts are applied to actual policy contexts and competing interests.

10 Recitals of the GDPR are not meant to establish legal rules alone but provide important and often authoritative insight into the interpretation of binding legal texts.[26] For example, there are multiple ways to interpret the placement of the specific language mandating the RTE (*ie*, "specific information to the data subject" and the right "to obtain an explanation of the decision reached after such assessment") in Recital 71 but not the main text of Art 22.[27] Even among scholars who searched the legislative history of the GDPR to find evidence to clarify legislative intentions and buttress their claims, different conclusions were drawn. For instance, scholars who were RTE-sceptical focused on the fact that the relevant provisions were dropped from the European Parliament proposed draft text, a signal that "legislators intentionally chose to make the right to explanation non-binding by placing it in Recital 71".[28]

---

25 Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76.

26 Margot E Kaminski, "The Right to Explanation, Explained" (2019) 34 Berkeley Tech L J 189 at 193–194; Andrew D Selbst & Julia Powles, "Meaningful Information and the Right to Explanation" (2017) 7 Int'l Data Privacy L 233.

27 While Recital 71 mandates that data subjects enjoy safeguards which should include *specific information to the data subject* and to protect "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" [emphasis added], Art 15(3) merely states that data subjects have 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". See GDPR Art 22(3) and Recital 71.

28 Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76 at 81.

11 Conversely, scholars in support of the RTE observed how the legislative drafts gradually expanded from a single focus on profiling to other types of general automated decision making in the final adopted text.[29] The latter approach of construing legislative intentions connects Art 22 of the GDPR to a deeper "European scepticism towards biases and potentially false decisions that can be taken by automated means if they are not verified by humans".[30] Indeed, if the "long grass" of recitals are where controversial issues and disagreements are dumped out of political expediency and compromises for "throwing up problems of just how binding they are",[31] would the fact that these provisions were retained in the GDPR hold promise that legislators would like to hold on to the RTE to pave the way for future jurisprudential development?

B. The nature and structure of Article 22 – Proscription or entitlement?

12 Out of the multi-faceted issues relating to the textual ambiguity of RTE-related GDPR provisions, the nature of Art 22 stands out as a primary challenge for legal scholars. Article 22(1) grants data subjects "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",[32] followed by qualifications that provide three scenarios (contractual, legal and consensual exceptions) under which the previous paragraph does not apply.[33] The third paragraph clarifies that the qualifications still require the data controller to "implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests", which include the *minimal combination* of the right to obtain human intervention, to express their point of view *and* to contest the decision.[34] Finally, the last paragraph excludes special categories of personal data[35] from the application of para 2 unless otherwise specified "suitable measures to

29 Maja Brkan, "Do Algorithms Rule the World? Algorithmic Decision-making and Data Protection in the Framework of the GDPR and Beyond" (2019) 27 Int'l J L & Info Tech 91 at 96–97.

30 Maja Brkan, "Do Algorithms Rule the World? Algorithmic Decision-making and Data Protection in the Framework of the GDPR and Beyond" (2019) 27 Int'l J L & Info Tech 91 at 96–97.

31 Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke L & Tech Rev 18 at 50.

32 GDPR Art 22(1).

33 GDPR Art 22(2).

34 GDPR Art 22(3).

35 GDPR Art 22(4).

safeguard the data subject's rights and freedoms and legitimate interests are in place".[36]

13 Taken together, the legislative language is elusive and perplexing in at least two ways. First, the multi-layered structure of the provisions in Art 22 is confusing as exceptions are followed by further exceptions, which themselves have more, finer exceptions. The first paragraph of Art 15 – the right not to be subjected to automated decision making – was qualified by an exception that allows the use of automated decision making (Art 22(2)) and a further restriction placed on the previous exception (Art 22(3)) that certain requirements must be met for the use of automated decision making to be permitted. This limited permission is subject to a further exception (Art 22(4): special categories of data), which itself can be lifted on one of various specified conditions (Arts 9(2)(a)–9(2)(g)).

14 Most importantly, the so-called *right* not to be subject to automated decision making, which was created by Art 22(1), is couched in a negative term. The linguistic ambiguity offers several possible avenues for interpretation.[37] Article 22(1) could be understood as granting data subjects a right to object to automated decision making. Alternatively, Art 22(1) could also be read as forbidding data controllers from using automated decision making unless *all three* of the following conditions are met: (a) *at least* one of the Art 15(2) exceptions applies; (b) *all* Art 15(3) requirements are fulfilled; and (c) special categories of data are not processed unless *at least one* of the qualifications under Arts 9(2)(a)–9(2)(g) apply.

15 The first approach of interpretation appears to be conferring rights on data subjects and, therefore, a gesture of empowerment and liberation. Nonetheless, the concern is that it may place an extra burden on individuals whose interests are adversely impacted[38]. This is because one way to understand this is that if Art 22(1) is framed as a right,

---

36 GDPR Art 22(4).

37 Iask Mendoza & Lee A Bygrave, "The Right Not to Be Subject to Automated Decisions Based on Profiling" in *EU Internet Law: Regulation and Enforcement* (Tatiana-Eleni Synodinou *et al* eds) (Springer, 2017); Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76 at 94–96; Maja Brkan, "Do Algorithms Rule the World? Algorithmic Decision-making and Data Protection in the Framework of the GDPR and Beyond" (2019) 27 Int'l J L & Info Tech 91 at 98.

38 See Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76 at 94–96; Maja Brkan, "Do Algorithms Rule the World? Algorithmic Decision-making and Data Protection in the Framework of the GDPR and Beyond" (2019) 27 Int'l J L & Info Tech 91 at 99.

data subjects may need to actively exercise the right to trigger the legal protection under GDPR. Otherwise, the data controller is relieved from the subsequent safeguards and obligations. To activate their rights, data subjects need to be aware of automated decision-making mechanisms and possess and express willingness to lodge a rejection with the data controller as a precondition to activate the right protection.[39] In addition, this mode of interpretation invites more questions, such as under what circumstances should data controllers be obligated to inform data subjects about the use of automated algorithms.

16 In contrast, a second approach that offers protection by default seems to be relieving such hassles for data subjects. Under this approach, the data subject enjoys a passive right and does not need top register his or her objection to enjoy the protections under the GDPR. This protection-by-default route seems to be the preferred approach adopted by the Article 29 Working Party Guidelines ("Article 29 Guidelines"), which broadly interpret the term "right" here as a general prohibition.[40] The Article 29 Guidelines state that Art 22(1) being termed a right does not mean there is a right to be "actively invoked by the data subject". Instead, it should be understood as a blanket prohibition. The prohibition applies independently of any action taken by the data subject.

17 Yet, what if the data subject prefers not to be excluded from automated decision making save for contractual, legal and explicit consent scenarios? Should her free will and self-determination to waive her protection be taken into consideration? Although the Article 29 Guidelines explain that the prohibition means an outright exclusion of fully automated decisions in Art 22(1), if we read all four paragraphs together, we can interpret Art 22 as both a prohibition and a right to the data subjects who have the *opportunity* and *capacity* to object to solely automated decision making barring certain limitations and exceptions. In other words, the Art 22 right is "not merely [a] right to object"[41] contingent on the actions of the data subject. Instead, it is a more complex right that safeguards the autonomy of the data subject. Fully-informed

39 Bygrave's critique was made on the basis of GDPR Art 22's predecessor, Art 15 of the 1995 EC Directive on data protection. See Lee A Bygrave, "Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision Making" in *Algorithmic Regulation* (Karen Yeung & Martin Lodge eds) (Oxford University Press, 2019) at pp 17–18.

40 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 19.

41 Hildebrandt's analysis was based on the draft GDPR, see Mireille Hildebrandt, "The Dawn of a Critical Transparency Right for the Profiling Era" in *Digital Enlightenment Yearbook* (Jacques Bus *et al* eds) (IOS Press, 2012) at p 50.

data subjects are provided with the opportunity to object, but in most scenarios, they are also trusted with the liberty of whether to pursue and exercise this right. That is why Art 22(2) grants explicit consent and contract exceptions to the prohibition and leaves the discretion to the data subject. Thus, this Art 22 right is neither a rigid prohibition nor a superficial right for the defendant to voice her objection but an in-depth, all-rounded right that offers a level playing field to the data subject, and balances the legitimate competing interests of the data controllers and the fundamental rights of the data subject, such as autonomy for self-determination.

C. The degree of automation – Full or nominal?

18 This dichotomy between full and nominal automation, as envisaged by the legislators of the GDPR, lies at the heart of the debate surrounding the scope and extent of the RTE. Article 22(1) states that data subjects have the "right not to be subject to a decision based solely on automated processing, including profiling".[42] The presence of the single word "solely" accounts for most of the ensuing discussions and reflects a certain level of legislative indecision over which algorithmically-made decisions the regulatory net should be cast. The term is important as the prohibition (and/or right) "not to be subject to automated decision making" will only apply to the decision making process and outcomes which are considered "solely" made by automated algorithms. Other partially-automated decisions and human-made decisions are not governed by the terms of Art 22 and its accompanying recitals. Not only does the interpretation of "solely" define the boundary of Art 22, but it also potentially shapes the entire discussion on the RTE.

19 Despite its importance, in practice, it is tremendously challenging to discern to what extent algorithmic decisions are considered entirely made by machines. Given that current AI development has yet to approach singularity, *ie*, when AI can fully replace or supersede humans, almost all automated decision-making processes today still have some degree of human supervision, control and involvement. So how should we interpret the RTE solely in the context of Art 22 and Recital 71?

20 Scholars hold concerns about the possibility that interpreting the wording as "zero human input" could create a loophole via which "even nominal involvement of a human in the decision-making process

42 GDPR Art 22(1).

allows for an otherwise automated mechanism to avoid invoking"[43] protective safeguards for data subjects. An alternative approach interprets "solely" beyond its literal meaning to encompass "substantively" or "meaningfully" instead. In other words, if a person fails to exercise any real influence on the outcome of the decision-making process, then "even if a decision is formally ascribed to a person, it is to be regarded as based solely on automated processing if a person does not actively assess the result of the processing".[44] Conversely, an automated process that merely lends support to a human decision-maker will be discounted from the ambit of Art 22 as humans have the opportunity to shape and influence a substantial or predominant part of the decision-making process.

21 The Article 29 Guidelines seem to side with this latter approach. Although the Guidelines state that "solely" means no human involvement to start with, it later clarifies that the human oversight of the decision must be "meaningful, rather than just a token gesture"[45] to be excluded from the reach of Art 22. Routine maintenance, for instance, does not qualify. The rule of thumb is that human involvement as the human participation "should be carried out by someone who has the authority and competence to change the decision".[46] The Article 29 Guidelines seem to have established a capacity test.

22 These clarifications no doubt helpfully shed light on the issue but still leave open uncertainty concerning the fluid boundaries between humans and machines in automated decision-making processes. Where the line should be precisely drawn remains unclear in the often hybrid collaborative processes. In a recent contractual dispute concerning algorithmic trading of cryptocurrency in Singapore,[47] the court found that human involvement in the trading process relying on automated, "deterministic"[48] software is at best minimal. This seems to be a textbook example of a solely automated decision-making process. Yet the Chief Technical Officer ("CTO") – as someone who has the authority

43 Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76 at 88.

44 Iask Mendoza & Lee A Bygrave, "The Right Not to Be Subject to Automated Decisions Based on Profiling" in *EU Internet Law: Regulation and Enforcement* (Tatiana-Eleni Synodinou *et al* eds) (Springer, 2017) at 87.

45 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 21.

46 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 21.

47 *Quoine Pte Ltd v B2C2 Ltd* [2020] SGCA(I) 2 at [2] and [100].

48 *Quoine Pte Ltd v B2C2 Ltd* [2020] SGCA(I) 2 at [15].

and competence – has the power to alter the outcome of automated decisions. The system allows human intervention to take place before (programming), during and after the decisions are made, rendering the decisions not to be based solely on automated processing. In this case, the CTO, after assessing the process of trading, cancelled the transactions and reversed the credit. Does this make the automated decision neither "solely" nor "meaningfully" made? This is doubtful, but does it matter that the Article 29 Guidelines uses a competence test rather than a contribution test?

23 Consider another example of self-driving cars. SAE International recommends standards of six levels of driving automation, ranging from no driving automation (Level 0) to full driving automation (Level 5).[49] Even when guided by the principles and clarifications made by the Article 29 Guidelines, among the hybrid modes of shared control by human drivers and software (Levels 1 to 4), at which precise level the driver maintains meaningful control of and when he can exercise actual influence on the outcome of the partially-automated process is far from crystal clear. It seems that the GDPR has left considerable room for judicial discretion. How future jurisprudential development will respond to this margin of appreciation (a term borrowed from European human rights law) and clarify to what extent algorithms with "humans in the loop" will be captured under Art 22 remains to be seen.

D. The consequence of automated decision making – Significant or insubstantial?

24 To fall within the purview of Art 22 and Recital 71, in addition to being "solely automated", decisions need to produce legal effects which impact data subjects or create similarly significant effects.[50] Solely-automated decisions which only have an insubstantial or trivial impact on data subjects, therefore, will be excluded from the scope of Art 22 and Recital 71 and affected data subjects will no longer enjoy the protections and safeguards provided by these provisions. Neither Art 22 nor Recital 71 defines these notions, although these texts potentially shape the scope of the RTE. It is challenging to gauge the precise boundaries of these terms without concrete examples illustrating what "legal" and "similarly significant" mean in this specific context. This is perhaps why Recital 71 adds that processes "such as automatic refusal of an online credit application or e-recruiting practices without any human

49 "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles" *SAE International* (30 April 2021) <https://www.sae.org/standards/content/j3016_202104/> (accessed 15 April 2022).

50 GDPR Art 22(1) and Recital 71.

intervention" are meant to fall within the scope of these provisions. The Article 29 Guidelines go even further to provide non-exclusive lists of a number of scenarios that produce "legal effects", including contractual cancellation, immigration and citizenship, and social benefits; "similarly significant effects" include automated decisions which affect the financial circumstances of data subjects, access to health services, employment and education opportunities.[51]

25 Interestingly, beyond these unusual suspects, the Article 29 Guidelines also cover targeted online advertising based on profiling and price discrimination, especially if these algorithmic decisions concern minority groups or vulnerable adults. For targeted advertising, intrusive advertising algorithms which track personal data and explore the vulnerabilities of the data subjects may fall within the scope of these provisions. The Article 29 Guidelines offer a hypothetical example of algorithms obtaining knowledge about the financial vulnerabilities of the data subject and using this personal weakness to promote advertisements for high-interest loans. The person is highly likely to sign up for those high-interest loans, therefore incurring further debts and ending up in exacerbated financial hardship.[52] More severe concerns might arise if the advertising algorithms target the medical conditions and cognitive vulnerabilities of data subjects, resulting in irreparable damage to the victims, such as inappropriate treatment[53] and severe emotional distress.[54] Similarly, price discrimination based on personal data and profiling, which results in individual consumers being barred from specific products and services, could also count as "similarly significant" effects on data subjects.[55]

---

51 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at pp 21–22.

52 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 22.

53 Yun Wang, "Baidu's Ad Business May Crack Under Student's Cancer Death" *Forbes* (5 May 2016) <https://www.forbes.com/sites/ywang/2016/05/05/baidus-ad-business-heads-for-shakeout-after-college-student-death/?sh=3e7a23bd2f22> (accessed 15 April 2022).

54 Emma Fletcher, "Social Media A Gold Mine for Scammers in 2021" *Federal Trade Commission* (25 January 2022) <https://www.ftc.gov/news-events/data-visualizations/data-spotlight/2022/01/social-media-gold-mine-scammers-2021> (accessed 15 April 2022).

55 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 22.

26 These most updated legislative clarifications might reflect awakening public sensibilities not only in Europe but worldwide toward promoting data autonomy, self-determination, privacy and algorithm accountability when large sections of private and societal transactions are mined, predicted and exploited by data processors who use AI algorithms to gain financial profit and power. The ambiguous and open-ended nature of the text provides golden opportunities for jurists and lawyers to develop future jurisprudence and expand the non-exhaustive examples through courts.

E. The quality of explanation – Formalistic or meaningful or specific?

27 The last core issue concerns the quality of explanations that data processors are required to provide to data subjects. Here, we move beyond issues of validity, scope or consequence to enquire about what constitutes an acceptable (at the minimum) and sound (in ideal scenarios) explanation. The conventional wisdom holds that there has been a trade-off between complexity and explainability of AI algorithms, and the growing complexity of cutting-edge AI technology results in a diminished prospect of deciphering algorithmic decisions in human-understandable language.[56] Yet, some scholars reject that algorithmic transparency needs to be obtained at the cost of its performance and function.[57] Therefore, the complexity of the algorithms is no excuse for opacity.

28 Furthermore, technical explainability itself does not necessarily lead to sound explanations, much less transparency. While the latest AI technologies, such as unsupervised machine learning, are not explanation-friendly, the quality of explanation is not merely an issue of technical capacity. Dumping obscure technical jargon and flooding data subjects with marginally-relevant data can hardly constitute a legitimate explanation. The quality of explanations is contingent on real-world issues such as finite resources, competing interests, and to what extent regulatory and legal efforts are made to ensure data subjects can enjoy a decent level of transparency in the context of automated decision making.

29 Within the GDPR, standards for the quality of explanations may be inferred from the overarching provisions regarding the transparency requirements, which mandate that the "controller shall take appropriate measures to provide any information referred to in Arts 13 and 14 and

---

56 See, *eg*, Tal Z Zarsky, "Transparent Predictions" (2013) U Ill L Rev 1503 at 1548.

57 See, *eg*, Cynthia Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (2019) 1(5) *Nature Machine Intelligence* 206.

any communication under Arts 15 to 22 and 34 relating to processing to the data subject in a *concise, transparent, intelligible and easily accessible* form, using *clear and plain language*"[58] [emphasis added]. Similarly, the Explanatory Report of the modernised version of the Council of Europe's Convention 108 stipulates that, as a general requirement of the transparency of the data process, data controllers should present information to the data subject that is "*easily accessible, legible, understandable and adapted to* the relevant data subjects"[59] [emphasis added].

30 We can also find traces of the quality standards in provisions regarding the notification duties of data controllers and rights of data subjects. Articles 13(2)(f), 14(2)(g) and 15(1)(h) of the GDPR share the identical textual language, which requires data controllers to provide data subjects with information about "the existence of automated decision-making, including profiling … 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" [emphasis added]. No doubt, the restrictive modifier of "meaningful" is the useful textual phrase that sets the benchmark for assessing whether the notification duties imposed on data controllers (Arts 13 and 14) are fulfilled and to what extent a right to access information throughout data processing is provided to the data subject with satisfaction (Art 15). Recital 71 embraced a similar degree of linguistic brevity, stipulating that the data processors need to enable safeguards "which should include *specific* information to the data subject and the right to … obtain an explanation of the decision reached after such assessment".[60] Although the term "specific" has not been used to directly modify or constrain "explanation" in this clause, it is safe to interpret that data subjects should have access to information with a considerable degree of specificity for explanations about what, how, why and by whom particular automated decisions affecting their rights and interests are made.

31 Based on these legislative requirements, scholars have teased out various criteria to assess how data controllers communicate with data subjects about algorithmic decision making, such as comprehensibility, comprehensiveness and actionability.[61] In other words, legitimate

58 GDPR Art 12(1).

59 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (28 January 1981), Eur TS No 108 (entry into force 1 October 1985) ("Convention 108"), Explanatory Report, para 68.

60 GDPR Recital 71, para 1.

61 See, *eg*, Margot E Kaminski, "The Right to Explanation, Explained" (2019) 34 Berkeley Tech L J 189 at 213 (stating that explanations about algorithmic decision-making should be, simultaneously, understandable (or "legible"), meaningful and actionable).

(cont'd on the next page)

explanations need to be understandable by both experts and laymen, have sufficient coverage and depth and can serve as the basis for further remedial actions by the data subjects such as contests and requests for corrections. Others see the gaps in the requirements regarding the quality of explanations scattered in different parts of the GDPR.[62] The meaningfulness test in Arts 13 to 15, which requires data controllers to provide information *ex ante* pertaining to systemic function of the algorithms to data subjects, fulfils the comprehensiveness requirement but is not tailored to the specific needs of individual data subjects in terms of comprehensiveness and actionability. The standard of specificity in Recital 71 (and Art 22), which may help pierce the veil of opacity for particular cases, meets the needs of individual data subjects whose interests and rights are affected by automated decisions. Yet, it loses sight of the overall picture and can hardly be seen as satisfying the comprehensiveness requirement. This leads scholars to propose new standards to assess the quality of information, such as a legibility test.[63] Ideally, a good explanation combines the best of the two worlds, falling in "the sweet spot between comprehensibility and full detail".[64]

32 The Article 29 Working Party, through various guidelines, has partially filled some of the loopholes by specifying, for instance, the general transparency requirements that information is "intelligible" does not only mean to be comprehensible but also tailored to the actual audience.[65] In the context of the right to be informed, the Article 29 Guidelines articulate that meaningfulness means the relevant information must be "sufficiently comprehensive"[66] so that data subjects can understand "the rationale behind, or the criteria relied on in reaching the decision",[67] but this does not necessarily entail "a complex explanation of the algorithms

62 Gianclaudio Malgieri & Giovanni Comande, "Why A Right to Legibility of Automated Decision-making Exists in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 243.

63 Gianclaudio Malgieri & Giovanni Comande, "Why A Right to Legibility of Automated Decision-making Exists in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 243.

64 Robert R Hoffman *et al*, "Explaining Explanation for 'Explainable AI'" in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (SAGE Publications, 2018) at p 198.

65 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at pp 7–8.

66 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 25.

67 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 26.

used or disclosure of the full algorithm".[68] But how comprehensive should the information be to satisfy this sufficiency threshold? Using examples of credit scoring and insurance premium calculation, the Guidelines recommend that data controllers provide detailed information in two main aspects: (a) the input end; and (b) output end of the automated decisions.

33 The former part of the sufficiency threshold could include main characteristics considered in reaching the decision (what information is included?), the source of this information (how the information is obtained, through voluntary provision by the data subject, public records or conduct data?) and the relevance (why certain personal information is relevant to calculate the credit score of data subjects?).[69] In the context of the right to access, in addition to illuminating the factors taken into account during the decision-making process, their respective "weight" on an aggregate level is also useful for the data subject to challenge the decision.[70] For the output end, to illustrate the significance and envisaged consequences of insurance premium calculation based on customers' driving behaviours, the processors ought to explain what dangerous habits (such as fast acceleration and last-minute braking driving) may result in certain consequences (higher insurance payments) and compare fictional drivers with and without these habits.[71]

34 In sum, for transparency rights enshrined in Arts 13 to 15, the Guidelines relieved data controllers from the obligation to disclose "a complex mathematical explanation about how algorithms or machine-learning works". Instead, to pass the meaningfulness test, the controller should provide details as to what, by whom, why and how personal input data is used in automated decision making in clear and comprehensive ways.[72] It seems that the quality requirement of explanations demands higher standards than offering mere *ex ante*

68 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 25.

69 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at pp 25–26.

70 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 27.

71 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 26.

72 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 31.

counterfactual[73] or pedagogical[74] explanations of the functionality, which do not necessitate the opening of algorithmic black boxes but impose a lower level of obligations on data controllers than full, *ex post* disclosure of the source code, mathematical models and the algorithm's components and parameters in individual cases.[75] Once again, both the GDPR and the Article 29 Guidelines provide sufficient wiggle room for future jurisprudence to define the quality of explanations.

III. Conceptualising the right to explanation as an instrument to express individual autonomy and dignity

35 No doubt, the RTE is a promising new legal mechanism for promoting fairness, accountability and transparency in contemporary societies saturated with complex algorithms and opaque automated decisions which affect the interests and rights of digital citizens. With Europe pioneering the making of laws in this field, other parts of the world are faced with similar tensions among individual rights, commercial incentives and public interests. The repercussions of the European legislation, therefore, have and will continue to influence future legislative progress in many other jurisdictions, including common law countries outside of Europe, such as Singapore. This is precisely why it will be beneficial to explore what would be the best approach to unpack and reconcile the confusions highlighted above in the GDPR texts. This article proposes a holistic approach based on the notion of individual autonomy to examine these contested issues. Such a framework will: (a) clarify the negative and positive dimensions of the autonomy of data subjects; (b) elucidate utilitarian and categorical perspectives relating to the RTE; and (c) explain why a holistic approach to interpretation will help resolve conclusions concerning the RTE.

36 In general, the RTE could be regarded as a salient example of GDPR architects' awareness of the categorical (rather than purely consequential) importance of the human-in-the-loop approach. Fully understanding why certain decisions critically affect human life is made in a meaningful way to define the self-identity of human beings. As scholars observed, "reading between the lines of these explanatory statements, we can discern not just fear about humans letting machines make mistakes but a concern to uphold human dignity by ensuring that

73 Sandra Wachter *et al*, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR" (2018) 31 Harv J L & Tech 841.

74 Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking for" (2017–2018) 16 Duke L & Tech Rev 18 at 65.

75 See, *eg*, Tal Z Zarsky, "Transparent Predictions" (2013) U Ill L Rev 1503.

humans (and not their 'data shadows') maintain the primary role in 'constituting' themselves".[76] In other words, control of one's own data is an intrinsic ingredient of his or her dignity and autonomy. This dimension of individual autonomy and self-identity goes beyond concerns about the utility, fairness and accuracy of automated decision making.

37 Understood in this way, the RTE is not merely of instrumental value but a legal construction reflecting the fundamental principles enshrined in the GDPR – transparency and accountability.[77] An autonomy-based theory with a special focus on the impacted individual data subjects offers the most comprehensive and "powerful arguments for transparency and accountability of the entire process of data processing".[78] Such a holistic analytical approach to interpreting RTE as an expression of intrinsic values of individual autonomy will also help clarify many of the competing interpretations outlined in Part II.

38 As analysed above, both the structure and wording of Art 22 of the GDPR are far from being straightforward. This leads scholars to criticise the "shaky foundations" for a "right to algorithmic explanation"[79] and question the necessity of inserting a second right – RTE – after already providing a primary right to exclude fully automated decisions.[80] So, is the RTE merely a product of legislative ambivalence and redundancy? Not necessarily. The multiplicity of rights provided in Art 22 might express different legislative intents concerning both *negative* and *positive* autonomy of data subjects. As discussed above, Art 22(1) is best understood as a prohibition or a negative right to ban wholly automated decisions. Data subjects are automatically protected, "whether or not the data subject takes action regarding the processing of their personal data".[81] The Article 29 Guidelines explain that this unique legislative technique

76 Iask Mendoza & Lee A Bygrave, "The Right Not to Be Subject to Automated Decisions Based on Profiling" in *EU Internet Law: Regulation and Enforcement* (Tatiana-Eleni Synodinou *et al* eds) (Springer, 2017) at p 84.

77 GDPR Art 5(1).

78 Tal Z Zarsky, "Transparent Predictions" (2013) U Ill L Rev 1503 at 1550.

79 Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking for" (2017–2018) 16 Duke L & Tech Rev 18 at 50.

80 Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking for" (2017–2018) 16 Duke L & Tech Rev 18 at 49.

81 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 19.

"reinforces the idea of the data subject having control over their personal data, which is in line with the fundamental principles of the GDPR".[82]

39 Here, the notion of control over one's own data, which is firmly rooted in European legislation and jurisprudence, could be regarded as an extension of the autonomy of data subjects.[83] In particular, this focus on control and authorship reflects the *positive* dimension of individual autonomy. Individual autonomy does not merely describe a condition that individuals are free from manipulation and coercion by external forces. Autonomous individuals are also the masters of their own lives, activities and data. Free control and use of one's own data is, therefore, an important step towards one's fulfilment of self-governance. In this sense, individual autonomy is more than the freedom of being left alone or independence from interruption. Its positive front requires awareness and capacity to exercise deliberate choices, if not actual realisation of these options. Data control is a central plank of self-authorship and self-governance in a digital age. Data autonomy is the first step towards addressing the power asymmetry between data subjects and controllers created by the rise of new forms of digital economy feeding on individual data: the platformisation of society, big tech dominance and rule by artificial intelligence.

40 Along this line of reasoning, an autonomy-based theoretical framework could provide clarity to various debatable issues highlighted in the above section. For instance, it sheds light on the debates surrounding both the degree of automation and the quality of explanation, which are seemingly distinguishable yet closely interrelated. In contrast with the degree of automation, which deals with when the RTE is invoked, the quality of explanation concerns how and to what extent the RTE should be applied to a specific case scenario. An autonomy-based theoretical framework can offer useful guidance to the debate about the degree of automation by switching the focus from the hybrid formality of human–machine collaboration to the impact of such hybridity on individual autonomy. Similarly, an autonomy-centred approach assesses the quality of explanation on the basis of the perception and action of the data subject.

41 First, on the degree of autonomation, if the human involvement in automated decision making is partial but merely nominal, from the perspective of maximising individual autonomy, the RTE should

82 *Guidelines on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679* (Article 29 Data Protection Working Party, 6 February 2018) at p 20.

83 Tal Z Zarsky, "Transparent Predictions" (2013) U Ill L Rev 1503 at 1541.

be offered to data subjects. This is because a nominal involvement of human activity in decision-making processes that are predominantly performed by machines will not ameliorate the obscurity of the mechanisms and processes dominated by the automation. These processes, although theoretically speaking are not fully automated, are different from traditional human decision-making processes. Data subjects, as autonomous human beings, remain subject to a transparency deficit which substantially undermines their capacity to comprehend, act on and challenge such decisions. An autonomy-based test, instead of hinging on the technical measurement of labour division between human and machine, simply asks: would this mode of human–machine collaboration render the data subject no access to information about the logic, functionality and consequences that they would have access to if the decision had been fully made by a human? If the answer is yes, then no matter what the appearance of human input is – nominal or formalistically partial – as far as the data subject's capacity to rationally comprehend, control and govern his or her own actions and activities under the impact of automated decisions is compromised by opacity, the RTE will serve as a useful legal remedy to restore human autonomy.

42 Adopting this approach to re-examine the high frequency trading software case in Singapore above, we will see that it does not matter whether the CTO had the competence or actually got involved in manipulating the algorithms. What solely matters is whether data subjects who are financially impacted possess the capacity to understand, think critically and challenge decisions made as such by the trading algorithms. If the court is convinced that neither the data subject nor any reasonable person in his or her position would be able to understand how the decisions were reached, despite the fact that CTO can make *ex post* interventions, the decisions would qualify as "solely" automated decisions under the wording of GDPR Art 22. Similarly, let us imagine a hypothetical road accident involving a SAE Level 3 self-driving car. A human driver who was behind the wheel failed to intervene before the collision and therefore was accused of negligence by not paying sufficient attention to the road conditions and failing to take control of the vehicle to apply emergency braking. The autonomous vehicle ("AV") company alleged they should not be held liable as SAE Level 3 cars are technically immature and demand supporting human drivers to take charge during emergency. Yet, the human plaintiff may argue that how the AV failed to detect a moving human object is unclear and therefore human agents were unable to make rational appraisal of the situation; *ie*, how and when to intervene. Higher levels of transparency could thus be achieved by granting them the RTE.

43 The concept of individual autonomy is associated with qualities of critical reflection and knowledge of one's own interests.[84] As Thomas Scanlon puts it, as an autonomous individual, a person "must see himself as sovereign in deciding what to believe and in weighing competing reasons for action".[85] In order to truly achieve the status of autonomy, an individual subject, therefore, needs not only the entitlement to the RTE but also the capacity to meaningfully exercise his or her right in at least two interrelated ways: (a) obtain *awareness* of the logic, mechanisms, consequences and envisaged consequences of automated decisions on his or her interests; and (b) gain the *capacity* of performing rational calculations on and critical reflections of his or her past, current and future courses of actions based on such awareness. Whether the explanation provided to the data subject is sufficiently meaningful and specific depends on whether it meets the two-pronged criteria to enable the data subject to make informed, rational and critical decisions.

44 Take credit rating, which has a significant impact on data subjects' mortgage application, as an example. A third-party credit agency computes a credit score based on which the data controller (a bank or a real estate developer) makes decisions which either approve or deny a certain mortgage application. How could the data controller and the credit rating agency ensure the quality of the explanations they provide is meaningful? Under the autonomy approach, the awareness of the consequences could be interpreted broadly to include not only the success or failure of this mortgage application, but also other derivative consequences such as secondary uses of the data: how might the automated decision making affect the data subject in his or her future application for credit cards, loans and jobs? Envisaged consequences may uncover information as to how the data subject may improve his or her success rate of application in the future. Will this current rejection have a detrimental impact on her future mortgage application? Will this affect her capacity to act as a guarantor in the future? In terms of logic and mechanisms, the controllers may disclose the sources of the data: what data will the credit rating agency acquire from the bank to calculate the score? What factors are taken into consideration when such calculation is made, and what are the weighting and schemes of computation? They need to inform the data subject in a clear, comprehensible and comprehensive way how the credit was scored without necessarily disclosing the full algorithm or source code.

84 Gerald Dworkin, *The Theory and Practice of Autonomy* (Cambridge University Press, 1988) at p 6.

85 Thomas Scanlon, "A Theory of Freedom of Expression" (1972) 1 *Philosophy and Public Affairs* 204 at 215.

45 The goal here is not to enable the data subject to replicate every step of the decision-making process but to simply form the capacity to make rational decisions with current and future decisions. Using the capacity test, even if the data controllers are willing to share all proprietary information regarding the source codes or algorithmic models of the credit scoring software, dumping such information on the data subject without helping the data subject to unpack and understand the implications and mechanisms as a lay person, will not help the data subject to make informed decisions and critically reflect on how to take actions in a responsible, rational manner. Such a "lazy" but "generous" approach to explanation, therefore, does not meet the autonomy-based criteria.

46 On a related note, the RTE applies when the blanket prohibition is lifted but contractual, consensual and legal exceptions are present.[86] In these scenarios, however, the GDPR is "constitutionally sceptical" whether consent could indeed keep "humans in the loop".[87] Given the considerable asymmetry of information, resources and capacity between data subjects and data controllers, consent can often be made in an uninformed state while the parties lack symmetrical access to information and resources. Uninformed and partially voluntary consent by knowledge and resource-depleted data subjects is nothing but a hollow façade to justify automated decisions engineered by powerful data controllers. Such consent is not conducive to enhance the individual autonomy of the data subject. This is why the GDPR supplies a secondary, positive RTE even after data subjects have expressed their consent or acquiesced in being subject to automated data processing through contractual agreements. The RTE is an extra layer of protection offered to data subjects. It is a novel legal counterweight, or at its minimum, a reasonable first step toward individual empowerment in the context of data protection.

47 The negative nature of the general prohibition clause forbids public and private bodies from infringing upon the rights and interests of individual data subjects by deploying a purely automated decision-making mechanism. In comparison, the RTE is a positive right that imposes obligations on data controllers. It empowers the data subject to exercise positive actions, *ie*, to request and receive information concerning why and how fully automated decisions are made in his or her individual case. If the prohibition serves to cut off the stream of personal data flowing from the weak actors to the powerful, the RTE reverses the flow of the stream so that it flows from the powerful to the

86 GDPR Art 22.

87 Chris Jay Hoofnagle *et al*, "The European Union General Data Protection Regulation: What It Is and What It Means" (2019) 28 Info & Comm Tech L 65 at 68.

weak actors. This combination of passive prohibition and active RTE is a response to what scholars describe as the threat of algorithmic profiling and automated decision making to "an intricate combination of negative and positive freedom".[88]

48 Such an analytical framework will also help differentiate the *instrumental* and the *intrinsic* aspects of the RTE. The impact of the RTE will be significantly enhanced if it could be implemented to rebalance real-world asymmetrical institutional power relations in algorithmic data collection, mining and analysis. This ideal scenario is precisely why the RTE has been aspired to not only enhance the transparency of automated data processing and decision making but also the accountability of the data controllers to individuals (government to individual and/or business to individual). The fact that this legal institution is embedded in and reflects existing political, social and economic fabrics of societies, not *vice versa*, is not a solid rationale for questioning the value or doubting the necessity[89] of the RTE. Interestingly, as scholars observe, the RTE often serves as a precondition to the exercise of other transparency-related rights, such as contest, correction and erasure. In order to substantiate a right to contest automated decision making, for instance, the data subject needs to understand how and why the decision has been made to start with. This suggests other venues for the utility of the RTE to take effect. The RTE is therefore a gateway right to put in place and compel data controllers to use appropriate technical resources and organisational measures to safeguard the interests and rights of data subjects. It is a precondition for data controllers to fulfil their legal and ethical accountability in an age of algorithmic dominance.

49 Last but not least, framing the RTE in a *holistic* analytical framework centred around individual autonomy could also free the RTE from an "unnecessarily rigid"[90] textual approach of interpretation which is often preoccupied with *segregated* single articles. For instance, some prominent accounts of the RTE refuse to acknowledge the existence of an *ex post* "general" right to explanation of system functionality and specific

88 Mireille Hildebrandt, "The Dawn of a Critical Transparency Right for the Profiling Era" in *Digital Enlightenment Yearbook 2012* (Jacques Bus *et al* eds) (IOS Press, 2012) at p 47.

89 Lilian Edwards & Michael Veale, "Slave to the Algorithm: Why A Right to An Explanation Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke L & Tech Rev 18.

90 Andrew D Selbst & Julia Powles, "Meaningful Information and the Right to Explanation" (2017) 7 Int'l Data Privacy L 233.

decisions in the GDPR[91] as it focuses on each of the main legislation components which might give rise to the RTE – Art 22, Recital 71, Art 15 and Arts 13 to 14.[92] In addition, the "hard law" of the main text of the GDPR is understood to be separated from its surrounding "soft law", including, but is not limited to, the Recitals, Article 29 Guidelines, and other guidelines, conventions and authoritative documents used to aid the interpretation and enforcement of the GDPR. Under this approach, it is easy to lose sight of the forest because too much attention is paid to the trees. A foreseeable conclusion drawn with this segregated approach of interpretation is that there is a lack of sufficient ground for the RTE as each of the clauses alone could not provide an adequate legal basis for this right. However, if these provisions are read in conjunction, the conclusion would be quite the contrary.

IV. Conclusion

50 The RTE has been the focus of academic debate and public discourse in the past decade since its formal inception in the GDPR. This article provides an overview of the five main issues at the heart of the debate: (a) the legislative source of the RTE; (b) the nature and structure of Art 22; (c) the degree of autonomation; (d) the consequence of automated decision making; and (e) the quality of explanations. Taken together, these issues concern the validity, nature, scope, implications and quality of the right to explain why and how fully-automated decisions affecting particular individual data subjects are made. Each of the above issues is rife with controversy and competing interpretations offered by scholars, media commentators, policymakers and the like.

51 It is further argued that at the heart of the debate are ambivalence and disagreements with regard to the theoretical rationale underpinning the existence and shape of the RTE. This article proposes an autonomy-based theoretical framework to unpack and reconcile the complex and seemingly contradictory interpretations of the law. Such an analytical framework will clarify three pairs of tensions in the understating of the RTE: (a) the negative *versus* positive dimensions of autonomy and rights; (b) the intrinsic *versus* instrumental value of the RTE; and (c) a holistic *versus* segregated approach of interpretation.

---

91 Sandra Wachter *et al*, "Why A Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation" (2017) 7 Int'l Data Privacy L 76 at 90.

92 Iask Mendoza & Lee A Bygrave, "The Right Not to Be Subject to Automated Decisions Based on Profiling" in *EU Internet Law: Regulation and Enforcement* (Tatiana-Eleni Synodinou *et al* eds) (Springer, 2017) at p 93.

52 Paradoxically, the legislative ambiguity and uncertainty in the GDPR can be understood both as difficulties as well as opportunities. There are realistic chances and plenty of wiggle room for scholars to inject a dose of autonomy-based analysis into the GDPR. This will also allow courts to adopt a flexible and holistic approach to interpreting relevant provisions with an aim to reinforce the transparency of the process and outcome of automated decision making, enhance the accountability of data controllers and restore a level playing field between the often outpowered individual data subjects and the resourceful institutional data controllers in the future. At the moment, despite jurisprudence on this new legal instrument remaining incredibly thin, there is hope that social, technological and political changes in the future will further incentivise impacted individuals to make fuller use of this right to vindicate machine-made decisions, which significantly affects their social, financial, educational and medical well-being. On the legislative plane, we can already see the widening of interests in and the gradual expansion of RTE-related provisions into non-European countries, including, most recently, Canada and China. These new countries which opened their legal systems to the RTE will be the new frontiers of debate and negotiation concerning the RTE. Regardless of how impactful and appealing this right will be in practice in Europe and elsewhere, it might be safe to assume that the legislative formula and practical implications of the RTE will be contingent on the particular social, legal, economic and political context of each country.