


Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas

Maja BRKAN* and Grégory BONNET**

Understanding of the causes and correlations for algorithmic decisions is currently one of the major challenges of computer science, addressed under an umbrella term “explainable AI (XAI)”. Being able to explain an AI-based system may help to make algorithmic decisions more satisfying and acceptable, to better control and update AI-based systems in case of failure, to build more accurate models, and to discover new knowledge directly or indirectly. On the legal side, the question whether the General Data Protection Regulation (GDPR) provides data subjects with the right to explanation in case of automated decision-making has equally been the subject of a heated doctrinal debate. While arguing that the right to explanation in the GDPR should be a result of interpretative analysis of several GDPR provisions jointly, the authors move this debate forward by discussing the technical and legal feasibility of the explanation of algorithmic decisions. Legal limits, in particular the secrecy of algorithms, as well as technical obstacles could potentially obstruct the practical implementation of this right. By adopting an interdisciplinary approach, the authors explore not only whether it is possible to translate the EU legal requirements for an explanation into the actual machine learning decision-making, but also whether those limitations can shape the way the legal right is used in practice.

I. INTRODUCTION

“Algorithm (*noun*). Word used by programmers when they do not want to explain what they did”.¹ This ubiquitous joke has become increasingly popular and notoriously bears a grain of truth. Machine learning algorithms, nowadays pervasively used for various kinds of decision-making, have been traditionally associated with difficulties related to their explanation. Understanding of the causes and correlations for algorithmic decisions is currently one of the major challenges of computer science, addressed under an umbrella

* Associate Professor of EU law, Faculty of Law, Maastricht University (The Netherlands); email: maja.brkan@maastrichtuniversity.nl

** Associate Professor of Artificial Intelligence, Normandy University, UNICAEN, ENSICAEN, CNRS, GREYC (Caen, France); email: gregory.bonnet@unicaen.fr

¹ It is difficult to pinpoint the exact origins and author of this joke, mentioned by numerous online sources; see for example J Alston, “10 Jokes Only Programmers Will Find Funny” (codeslaw, 31 August 2018, with reference to Reddit).

term “explainable AI (XAI)”.² While there is no standard definition of this notion, the core goal of XIA is to make Artificial Intelligence (AI) systems more understandable to humans. Generally, XIA “refers to the movement, initiatives and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept”.³

However, the explainability of algorithmic decisions has attracted not only the attention of computer scientists, but recently also of lawyers. The advent of the General Data Protection Regulation (GDPR)⁴ brought the question of algorithmic transparency and algorithmic explainability to the centre of interdisciplinary discussions about AI-based decision-making. Hitherto, numerous authors have argued that the GDPR establishes the right of data subjects to an explanation of algorithmic decisions.⁵ As the first legal act touching upon this issue, the GDPR seems to go to hand in hand with the XAI in exploring the (im)possibility of providing such explanations. The GDPR is thus becoming increasingly important also for XAI researchers and algorithm developers, since the introduction of the legal requirement for understanding the logic and hence explanation of algorithmic decisions entails also the requirement to guarantee the practical feasibility of such explanations from a computer science perspective.

From a legal perspective, the core incentive to guarantee explanations of algorithmic decisions is to enable the data subject to understand the reasons behind the decision and to prevent discriminatory or otherwise legally non-compliant decisions.⁶ Differently, in terms of computer science, the incentive to gain insight into the algorithms goes far beyond a potential legal requirement to provide explanations for algorithmic decisions. Rather, the incentives relate to both ethical and technical considerations. Being able to explain an AI-based system may help to make algorithmic decisions more satisfying and acceptable, to better control and update AI-based systems in case of failure, to build more accurate models, and to discover new knowledge directly or indirectly.⁷ For example, Lee Sedol was able to defeat AlphaGo in only one game where he used “a very innovative move that even the machine did not anticipate”.⁸ This new knowledge, gained through the use of AlphaGo, has hence benefited the community of players of this game.⁹ However, was this knowledge gained because

² A Holzinger, “From Machine Learning to Explainable Artificial Intelligence” (2018) World Symposium on Digital Intelligence for Systems and Machine 55.

³ A Adadi and M Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence” (2018) 6 IEEE Access 52140.

⁴ 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) [2016] OJ L 119/1.

⁵ See below, note 14.

⁶ M Brkan, “Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond” (2019) 27 International Journal of Law and Information Technology 91, 112, 118; compare also S Wachter et al, “Counterfactual explanations without opening the black-box: automated decisions and the GDPR” (2018) 31 Harvard Journal of Law and Technology 841, pp 843, 863.

⁷ Compare Adadi and Berrada, *supra*, note 3, 52142–52143, who indicate four reasons for the need for XAI: to justify results, reached with algorithmic decisions; to control the system behaviour; to improve models and to gain knowledge. See also T Miller, “Explanation in Artificial Intelligence: Insights from the Social Sciences” (2019) 267 Artificial Intelligence 1, 8, who believes that the core purpose “of explanation is to facilitate *learning*” (emphasis in original).

⁸ PS Yiu, “Peek Into the Future: A Breakdown of the Various Implications of Alphago’s Success Over the Traditional Board Game Go” (2016) 3 Chinese Studies Program Lecture Series 113.

⁹ *ibid.*

the researchers were able to scrutinise AlphaGo's reasoning processes like a glass box? Or was this knowledge gained only through an external (and maybe incorrect) description of what the researchers inferred about AlphaGo's decisions?

These questions illustrate only some of the unresolved issues relating to the explainability of algorithmic decisions. Both technical obstacles as well as legal limits, in particular the secrecy of algorithms, could potentially obstruct the practical implementation of the GDPR's right to explanation. Some authors therefore propose to build and use inherently explainable models.¹⁰ Even though it can be argued that a legal requirement for explainability exists, putting this requirement into practice could potentially be seen as an elusive *fata morgana* that appears on the horizon, yet is in reality intangible and unreachable. Against this backdrop, this paper aims to examine the legal and technical feasibility of explainability of algorithmic decisions and to determine whether the GDPR's right to explanation is a right that can be implemented in practice or rather merely a dead letter¹¹ on paper.

This article is composed of five parts. After the introductory part, the second part seeks to explain the core notions used in this article and the scope of application of the right to explanation from the GDPR. The third part of the article aims to relate different types of AI systems (automated/autonomous and supervised/unsupervised) with the right to explanation. The fourth part explores how the right to explanation can be implemented from both a technical and a legal perspective. The final part concludes the article and offers a schematic overview over different factors that might impact the feasibility to explain algorithmic decisions.

II. SETTING THE SCENE: ALGORITHMIC EXPLAINABILITY BETWEEN LAW AND COMPUTER SCIENCE

1. The requirement for algorithmic explainability in the GDPR

On the legal side, the question whether the GDPR provides data subjects with the right to an explanation in case of automated decision-making has been the subject of a heated debate. In this vein, certain authors firmly oppose such a construction of this right, given that it is not expressly provided by the GDPR,¹² save in one of its recitals.¹³ Other authors claim the opposite: for them, the right to explanation seems to be a real and tangible right, unequivocally provided by the GDPR.¹⁴ Yet others, including one

¹⁰ See C Rudin, "Please Stop Explaining Black Box Models for High-Stakes Decisions", Workshop on Critiquing and Correcting Trends in Machine Learning at the 32nd Conference on Neural Information Processing System 1. The option of "purposefully building interpretable models" is discussed for example also by AD Selbst and S Barocas, "The Intuitive Appeal of Explainable Machines" (2018) 87 *Fordham Law Review* 1085, 1110–1113.

¹¹ The metaphor of dead letter is used also by R Guidotti et al, "A Survey of Methods for Explaining Black Box Models" (2019) 51 *ACM Computing Surveys* 1, p 2.

¹² S Wachter et al, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation" (2017) 7 *International Data Privacy Law* 76.

¹³ Recital 71 GDPR and recital 38 Police Directive.

¹⁴ 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, New York <arxiv.org/pdf/1606.08813v3.pdf> accessed 7 February 2020, 1, claim that the GDPR "effectively create[s] a 'right to explanation'". Compare also Casey, Farhangi and Vogl, who opine that the GDPR creates a "muscular 'right to explanation'"; see B Casey et al, "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise" (2019) 34 *Berkeley Technology Law Journal* 145, p 188.

of the authors of this article,¹⁵ reach the right to explanation through interpretation of various GDPR provisions.¹⁶ Finally, certain scholars claim that the right to explanation is an inappropriate remedy given the difficulties of implementing this right on machine learning algorithms.¹⁷

We do not aim to restate a detailed account of arguments in favour of the existence of the right to explanation in the GDPR,¹⁸ as one of us has provided this analysis elsewhere.¹⁹ To summarise these findings, we submit that the right to explanation in the GDPR should be a result of interpretative analysis of several GDPR provisions jointly. More precisely, we believe that the GDPR provisions on the right to information and access that require the controller to provide the data subject with the “meaningful information about the logic involved” in case of “existence of automated decision-making”,²⁰ should be read together with Article 22 GDPR that regulates automated decision-making, as well as with Recital 71, which epitomises explanation as one of the safeguards in case of such decisions.²¹ One of the core arguments is that the right to explanation enables the data subject to effectively exercise her other safeguards, notably the right to contest the automated decision and to express her point of view.²² In consequence, a clearer understanding of the reasons behind the automated decision can allow the data subject to point out irregularities in decision-making, such as unlawful discrimination.²³

It is important to recall that the GDPR does not provide for a right to explanation for all algorithmic decisions, but only for those that have a legal or significant effect.²⁴ For all other algorithmic decisions, the data subject can exercise other rights provided for by the GDPR, such as the right to information²⁵ or the rights of access,²⁶ but not the right to explanation. For example, in case of a simple targeted advertisement for shoes, the data subject will not have the right to require from the data controller information on

¹⁵ Brkan, *supra*, note 6, pp 110–119.

¹⁶ I Mendoza and LA Bygrave, “The Right not to be Subject to Automated Decisions based on Profiling” in TE Synodinou, *EU Internet Law: Regulation and Enforcement* (Springer 2017) 77, at pp 92–94; AD Selbst and J Powles, “Meaningful information and the right to explanation” (2017) 7 *International Data Privacy Law* 233, p 237 ff.

¹⁷ 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) 16 *Duke Law & Technology Review* 18, p 81.

¹⁸ In the previous analysis, as well as in this article, we focus mostly on the GDPR rather than on the Police Directive, as the latter does not refer to the “meaningful information about the logic involved” within the framework of its right to information and right of access. See Arts 13 and 14 of the Directive (EU) 2016/680 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 by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L 119/89.

¹⁹ Brkan, *supra*, note 6, pp 20–29.

²⁰ Arts 13(2)(f), 14(2)(g) and 15(1)(h) GDPR.

²¹ Brkan, *supra*, note 6, p 22.

²² Mendoza and Bygrave, *supra*, note 11, pp 16–17, rightly observe that the right to contest automated decisions from GDPR is akin to the right of appeal, which is meaningful only if such an appeal is considered on its merits. See also Brkan, *supra*, note 6, p 24.

²³ Brkan, *supra*, note 6, 28.

²⁴ Art 22(1) GDPR.

²⁵ Arts 13 and 14 GDPR.

²⁶ Art 15 GDPR.

why she was shown this particular ad for this particular brand, but she will be able to require from the controller information about the data collected about her.²⁷

It is equally important to clarify that the current EU legal regime limits the right to explanation only to decisions based on *personal* data. This is a consequence of the fact that algorithmic decision-making is regulated by the EU data protection legislation²⁸ and not by a more general legislation encompassing broader regulation of algorithmic decisions. However, in computer science, decisions with machine-learning algorithms are not confined only to those based on personal data; rather, these algorithms often deploy either a mixed dataset of personal and non-personal data and sometimes even purely non-personal datasets.²⁹ Non-personal data is either data not relating to natural persons (for example, data concerning companies or environment) or data regarding natural persons that does not allow for their identification (for example, anonymised data).³⁰ In mixed datasets, where personal and non-personal data can be easily separated, the GDPR, including the right to explanation,³¹ applies only to the part containing personal data, whereas the Regulation on the free flow of non-personal data³² applies to the non-personal part of the dataset.³³ If such a separation is not possible, the safeguards from GDPR apply to the entire mixed dataset.³⁴ However, decisions based purely on non-personal data are left in a legal vacuum; not only does the Regulation on the free flow of non-personal data not regulate decision-making based on such data, neither does it provide for the possibility of explanation of decisions based on non-personal data.³⁵

We submit that, in certain circumstances, decisions based on non-personal data would equally necessitate an explanation. The underlying reason is that a legal or similar significant effect on a data subject³⁶ can be produced not only by decisions based on personal data; such an effect can occur also in decisions based on non-personal data.

²⁷ According to the EDPB, targeted advertising does not in principle legally or significantly affect the data subject; see Art 29 data protection working party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, p 22. Critically on WP29's position on targeted advertising, see M Veale and L Edwards, "Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling" (2018) 34 Computer Law & Security Review 398, pp 401–402.

²⁸ See Art 22 GDPR and Art 11 Police Directive. In our analysis, we deliberately limit ourselves only to Art 22 GDPR, even though Art 11 of the Police Directive also regulates [algorithmic] decisions that have an adverse legal effect on data subjects.

²⁹ Hildebrandt warned as early as 2010 that "in the case of profiling the limitation to personal data seriously hinders adequate protection of individuals with regard to" non-personal data; see M Hildebrandt, "Profiling and the Identity of the European Citizen" in M Hildebrandt and S Gutwirth (eds), *Profiling the European Citizen* (Springer 2010) p 320. See also M Hildebrandt and BJ Koops, "The Challenges of Ambient Law and Legal Protection in the Profiling Era" (2010) 73 Modern Law Review 428, at pp 439–440.

³⁰ See in this sense Art 4(1) GDPR in combination with Recital 26 GDPR. Regarding the question of identifiability, see WG Urgessa, "The Protective Capacity of the Criterion of 'Identifiability' under EU Data Protection Law" (2016) 2 European Data Protection Law Review 521.

³¹ We believe it equally problematic that the Police Directive does not encompass the right to explanation.

³² Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union [2018] OJ L 303/59.

³³ Art 2(2) Regulation (EU) 2018/1807.

³⁴ *ibid.*

³⁵ This regulation "aims to ensure the free flow of" non-personal data "within the Union". See Art 1 Regulation on the free flow of non-personal data.

³⁶ Art 22(1) GDPR.

Imagine an autonomous vehicle (AV) collecting purely non-personal data, such as information about the weather, road conditions, distance from other vehicles and tyre grip. On an icy road, the AV miscalculates the braking distance and causes an accident in which a passenger is severely injured. Such an accident undoubtedly has at least significant, if not also legal,³⁷ effects on that passenger. Let us use another example: in precision agriculture, an algorithm might wrongly assess the amount of pesticides needed for agricultural products, thus causing debilitating illness in humans consuming those products. This, too, would lead to a significant effect on these persons. Since none of these decisions is based on the processing of personal data, and hence does not fall under the EU data protection regime, the affected persons do not have the possibility to request the explanation of the reasons behind the decisions. Given a significant impact that certain decisions of this kind can have on natural persons, it would be more appropriate for the legislator to provide for a broader regulation of algorithmic decisions and to ensure a right to explanation whenever persons are legally or significantly affected regardless of the type of data on which the decision is based.

2. Algorithmic explainability: discussing the notions

Before embarking on a discussion about the feasibility of the right to explanation from a computer science and a legal perspective, it is appropriate to provide an elucidation of certain concepts in order to situate the right to explanation in the broader context of AI-based decision-making. In the remainder of this article, instead of “AI-based system”, the term “*artificial agent*” will be used.³⁸ This term is a common metaphor used in computer science to situate software and robots under the same concept, depending on the models, their ability to reason and decide on the action to execute, taking into account different pieces of information. While the GDPR refers to *automated* decision-making,³⁹ AI experts distinguish between *automated*, *autonomous* and *algorithmic* decision-making and processes. While the goal is the same for automation and autonomy, that is, to perform actions without the need for human intervention, they differ considerably in their features. An *automated* process is a software or hardware process which executes a predefined sequence of actions without human intervention.⁴⁰ For example, an automated university selection procedure can function in a way to average all candidates' school marks, then to sort all candidates in decreasing order based on their average mark, and finally to assign admission rank based on this sorting. Therefore, save for machine failures, automated decisions are fully predictable. In this sense, such decisions can be fully explainable, as far as the system's specifications and the situation in which the decision was made are known. For example, the university

³⁷ For example, if the passenger is so badly injured that she loses her legal capacity.

³⁸ S Franklin and A Graesser, “Is it an agent or just a program?: a taxonomy for autonomous agents” (1996) International Workshop on Agent Theories, Architectures, and Languages 1.

³⁹ Art 22 GDPR.

⁴⁰ W Truszkowski et al, “Autonomous and Autonomic Systems with Applications to NASA Intelligent Spacecraft Operations and Exploration Systems” in MG Hinchey (eds), *NASA Monographs in Systems and Software Engineering* (2010) p 10.

admission decision made by an automated system can be explained by the circumstance that the student's average mark was higher than that of the non-admitted candidates.

Differently, *autonomous* decision-making entails that the algorithmic procedure behind the decision is computed by the agent, and relies only on a high-level goal defined by a human. Hence, autonomy emphasises the capacity to compute which decisions must be made in order to satisfy a formalised goal, and therefore *autonomy* is the central notion in the design of artificial agents.⁴¹ For example, making the previous university selection procedure autonomous may consist in training a neural network to identify the best students based on all their school records. Here the combination of variables, thresholds and aggregations is not explicitly defined by a human but computed by the machine. Therefore, an artificial agent with autonomous decision-making capacities⁴² has the capacity "to act according to its own goals, perceptions, internal states, and knowledge, without outside intervention".⁴³ However, autonomy is still bounded: the actions are indeed limited by the amount of information the agent has, by the time available for computation, and by the expressive power⁴⁴ of the language used to specify the decision-making process. Although a human being always specifies the goals themselves or decides how those goals are set, autonomous decisions are more difficult to explain since autonomy allows the machine to decide how to reach the goals, or learn criteria.

Algorithmic decision-making, for its part, is an overarching notion, encompassing both automated and autonomous decision-making. It means that a given decision is made (partly or completely) with the help of an algorithm; this algorithm may be either automated or autonomous and based or not based on AI techniques.

The analysis above demonstrates that the GDPR, by referring only to *automated* decision-making, does not take stock of these definitions. From a computer science perspective, it would be more sensible if this legal act deployed the overarching term *algorithmic*, rather than automated, decision-making. If the GDPR truly covered only automated decisions, the practical implementation of the right to explanation would face very few technical obstacles, since automated decision-making processes are based on predefined step-by-step procedures. However, the examples used to substantiate this decision-making in the GDPR – such as predicting aspects of data subject's performance at work or her economic situation⁴⁵ – fall into the category of autonomous decision-making. Consequently, this legal act in fact seeks to apply to the entire domain of algorithmic decision-making, including decisions made by autonomous agents. The deployment of the latter systems might obstruct the effective application of the law since these systems allow the machine to decide on or learn almost all steps of the decision-making without the *a priori* knowledge of the user.

⁴¹ See Franklin and Graesser, *supra*, note 38, p 2. The authors review several definitions of artificial agents, and many of them use the term "autonomy" or "autonomous".

⁴² Often called an "autonomous agent" for the sake of simplicity.

⁴³ Truskowski et al, *supra*, note 40, p 254.

⁴⁴ In computer science, the expressive power of a language is a means of representing the kind of concepts it can represent. For instance propositional logic is less expressive than first-order logic, and one-layer neural networks are less expressive than multi-layered neural networks.

⁴⁵ Recital 71 GDPR.

This opacity can have an impact not only on the implementation of the legal requirement for explainability, but also on determination of accountability⁴⁶ and liability⁴⁷ for the performance of these systems.

3. The multi-faceted nature of explanations

In order to better understand the meaning of explanations in computer science, it is equally important to discuss the notions of *traces*, *explanations*, *interpretations* and *justifications* as understood in computer science, even though there is a certain degree of disagreement among computer scientists as to the precise meaning of these notions.

All computer programs (thereby artificial agents) can provide execution *traces*, which show what statements are being executed as the program runs.⁴⁸ Those traces can be analysed by a human expert to understand how an agent, being automated or autonomous, made a given decision. For example, a logic-based system can provide a complete formal proof showing how a given decision will allow a given goal to be reached. While this approach can be useful, such traces are difficult to use for non-expert humans. Thus, it might be preferable to rely on another kind of information, called *interpretations*. Interpretations are descriptions of an agent's operations "that can be understood by a human, either through introspection⁴⁹ or through a produced explanation".⁵⁰ Unlike traces, interpretations are tailored to be readable and understandable not only by experts but also by users. Interpretations can be divided

⁴⁶ The literature on algorithmic accountability is abundant. See, for example, A Rosenblat et al, "Algorithmic Accountability" (2014) *The Social, Cultural & Ethical Dimensions of "Big Data"* March 2014 <ssrn.com/abstract=2535540> accessed 7 February 2020; D Keats Citron and F Pasquale, "The Scored Society: Due Process for Automated Predictions" (2014) 89 *Washington Law Review* 1; JA Kroll et al, "Accountable Algorithms" (2017) 165 *University of Pennsylvania Law Review* 637; PT Kim, "Auditing Algorithms for Discrimination" (2017) 166 *University of Pennsylvania Law Review Online* 189; R Binns, "Algorithmic Accountability and Public Reason" (2018) 31 *Philosophy & Technology* 543; M Ananny and K Crawford, "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability" (2018) 20 *New Media & Society* 973; ME Kaminski, "Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability" (2019) 92(6) *Southern California Law Review* 1529; B Casey et al, "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise" (2019) 34 *Berkeley Technology Law Journal* 145; ME Kaminski, "The Right to Explanation, Explained" (2019) 34 *Berkeley Technology Law Journal* 189.

⁴⁷ There is currently a discussion about whether the existing Product Liability Directive, which establishes a regime of strict liability (see its Art 1), is appropriate for the liability of autonomous systems. See Council Directive 85/374/EEC of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products [1985] OJ L 210-29; see also European Commission, "Public consultation on the rules on liability of the producer for damage caused by a defective product" <ec.europa.eu/growth/content/public-consultation-rules-liability-producer-damage-caused-defective-product_0_en> accessed 7 February 2020. See also the work of the Expert Group on liability and new technologies <ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupDetail&groupID=3592&NewSearch=1&NewSearch=1> accessed 7 February 2020.

⁴⁸ H Henderson, *Encyclopedia of Computer Science and Technology* (3rd edn, Facts on File 2009) p 61.

⁴⁹ In computer science, introspection is the ability to examine the properties of a program and its inner variables at runtime. Thus, it is a way to produce execution traces.

⁵⁰ O Biran and C Cotton, "Explanations and justifications in machine learning: a survey" (2017) IJCAI Workshop on Explainable Artificial Intelligence 8. This understanding of interpretation is shared by F Doshi-Velez and B Kim, "Towards A Rigorous Science of Interpretable Machine Learning" (2017) ARXIV <arxiv.org/pdf/1702.08608.pdf> accessed 7 February 2020, who defined interpretations as descriptions able "to explain or to present" a machine learning decision model so that it is understandable to a human. See also Guidotti et al supra, note 11, p 5.

into two sub-categories, depending on what is expected by the users or the designers: explanations and justifications.⁵¹

An explanation describes how a given decision was made, whereas a justification describes why a given decision was a “correct”⁵² one.⁵³ In other words, justifications describe why decisions are “reasonable” (providing the meaning of this word has been defined) knowing the applicative domain,⁵⁴ while explanations exhibit causations and correlations that led to the decision.⁵⁵ It is important to remark that the limit between explanations and justifications is blurred. For instance, Fox et al consider that answering the question “why is what you propose to do more efficient/safe/cheap than something else?” can serve as an explanation,⁵⁶ while it is clearly a justification as defined above. Moreover, computer scientists do not agree on a general and formal definition of explanations, preferring to characterise them through questions. For instance, Doshi-Velez et al state that explanations must answer at least one of the following questions:⁵⁷

1. What were the main factors in the decision?
2. Would changing a given factor have changed the decision?
3. Why did two different similar-looking inputs produce different decisions?

However, neither do computer scientists agree on a finite set of questions. For instance Fox et al consider “why didn’t you do something else?” and “why can’t you do that?” as a means to provide explanations.⁵⁸

Therefore, regarding *explanations*, this article rests upon the premise that it is challenging or even impossible to define explanations in the abstract, without taking into account the decision-making model, the kind of data used, the desired understandability, or the kind of questions we ask regarding the decision.⁵⁹ From a computer science perspective, this complexity and multi-faceted nature of explanations led several authors to provide various classifications of explanations,⁶⁰ where the notion of “explanations”

⁵¹ ZC Lipon, “The mythos of model interpretability” (2016) ICML Workshop on Human Interpretability in Machine Learning 96.

⁵² While correctness in computer science means that a given algorithm is correct with respect to a given specification, here correctness means being correct with respect to a given abstract criterion (eg fairness, legality, non-discrimination). As computer science is not a prescriptive discipline, the semantics of correctness is left to the user, the designer or the lawyer.

⁵³ Biran and Cotton, *supra*, note 50, p 8.

⁵⁴ WR Swartout, “XPLAIN: a system for creating and explaining expert consulting programs” (1983) 21 Artificial Intelligence 285, p 287.

⁵⁵ AJ London, “Artificial intelligence and black-box medical decisions: accuracy versus explainability” (2019) 49 Hasting Center Report 15.

⁵⁶ M Fox et al, “Explainable planning” (2017) IJCAI Workshop on Explainable Artificial Intelligence 25, p 26. Here, the terms “efficient”, “safe”, or “cheap” are concrete instance of “correct” (see note 66). Hence, the question asks for a justification.

⁵⁷ F Doshi-Velez et al, “Accountability of AI under law: the role of explanation” (2017) ARXIV <arxiv.org/pdf/1711.01134.pdf> accessed 7 February 2020, 3.

⁵⁸ Fox et al, *supra*, note 56, p 2.

⁵⁹ Guidotti et al observe, for example, that different scientific communities provide a different meaning to explanation. See Guidotti et al *supra*, note 11, p 2.

⁶⁰ See for example Adadi and Berrada, *supra*, note 3, pp 52146–52151; Lipton, *supra*, note 51, pp 15–20; B Mittelstadt et al, “Explaining Explanations in AI” (2019) FAT* 2019 Proceedings <doi.org/10.1145/3287560.3287574> accessed 7 February 2020; D Pedreschi et al, “Open the Black Box Data-Driven Explanation of Black Box Decision Systems” (2018) ARXIV <arxiv.org/pdf/1806.09936.pdf> accessed 7 February 2020, p 4.

has various meanings, mostly depicting different methods to reach an explanation.⁶¹ For example, according to Adadi and Berrada, models can be inherently interpretable (meaning that a model provides its own explanation); the explainability⁶² can be either global or local⁶³ (understanding either the logic of the entire model as opposed to explaining why a model made a specific decision); or explanations can be model-specific or model-agnostic (limited to a specific model or independent of a type of model).⁶⁴ Among model-agnostic explanations are visualisations (exploring patterns in a neural unit),⁶⁵ knowledge extractions (computing an interpretable model that approximates a non-interpretable one),⁶⁶ influence methods (methods that change the input and detect how much the changes impact the performance of a model) and example-based explanations (among which counterfactual explanations⁶⁷ have recently gained popularity).⁶⁸ To that, we could add natural language explanations that explain, in natural language or visually, the relationship between the input and output of the model.⁶⁹

As can be seen from this brief overview of the literature, the methodology as to how explanations are reached often also determines their scope and type. Nevertheless, a common thread shared by all these various approaches towards explanations is a desire to make algorithmic decisions understandable to humans, either experts or non-experts. When humans – and we could add non-experts in particular – desire an explanation for a decision, they in principle seek to understand the “reasons or justifications for that particular outcome, rather than” having a general insight into “inner workings or the logic of reasoning behind the decision-making process”.⁷⁰

⁶¹ Terminologically, in these classifications, the notion of “interpretability” is sometimes juxtaposed to “explainability” and sometimes used instead of the latter notion to denote that the works relate to explainability of machine-learning algorithms. For the latter approach, see for example Adadi and Berrada, *supra*, note 3, 52147. For a former approach, see C Rudin, “Please Stop Explaining Black Box Models for High-Stakes Decisions”, Workshop on Critiquing and Correcting Trends in Machine Learning at the 32nd Conference on Neural Information Processing System 1, pp 1–2. Rudin juxtaposes explainable machine-learning that uses a separate model to replicate the decision-making of a black box, to interpretable machine-learning for models that are “inherently interpretable”, which “provide their own explanations, which are faithful to what the model actually computes”.

⁶² As mentioned *ibid*, Adadi and Berrada, *supra*, note 3, 52147 ff use the notion of “interpretability”, whereas other authors deploy the notion of “explainability” for global and local explanations. This applies also to other types of categories provided by Adadi and Berrada. For instance, the notion of “interpretability” is also used in Edwards and Vale, *supra*, note 18, while the term “local explainability” is used by A Aggarwal et al, “Black Box Fairness Testing of Machine Learning Models” (2019) European Software Engineering Conference and Symposium on the Foundations of Software Engineering 625, p 627. Clearly both terms are interchangeable, as noticed in A Richardson and A Rosenfeld, “A Survey of Interpretability and Explainability in Human-Agent Systems” (2018) 2nd Workshop on Explainable Artificial Intelligence 137, p 137.

⁶³ For example, LIME (Local Interpretable Model-agnostic Explanations), proposed by M Tulio Ribeiro et al, “Why Should I Trust You? Explaining the Predictions of Any Classifier” (2016) 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135, p 1137. On other local explanations, see further section IV of this paper.

⁶⁴ See Adadi and Berrada, *supra*, note 3, pp 52148–52149.

⁶⁵ For example, P Tamagnini et al, “Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations” (2017) ACM Press 1.

⁶⁶ For a survey, see R Andrews et al, “Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks” (1995) 8 Knowledge-Based Systems 373.

⁶⁷ Wachter et al, *supra*, note 6, p 841.

⁶⁸ Adadi and Berrada, *supra*, note 3, pp 52146–52151.

⁶⁹ Mittelstadt et al, *supra*, note 60, p 282. Mittelstadt et al are further referring to Tulio Ribeiro et al, *supra*, note 63, p 1135.

⁷⁰ Adadi and Berrada, *supra*, note 3, p 52142.

Therefore, several authors argue that computer science, by fostering “explanations” as simple models that approximate a more complex one, often misses a true nature of human-like explanations⁷¹ that could be useful for a non-expert user directly affected by an algorithmic decision. Humans are not seeking for a model but rather for contrastive explanations (eg not why a decision D was made, but rather why D was made instead of another decision Q) as well as causal attribution instead of likelihood-based explanations.⁷² We could argue that causality is also one of the main characteristics of the GDPR-based right to explanation because it seems that it is precisely the understanding of causality that would enable the data subject to effectively challenge the grounds for an algorithmic decision. However, while symbolic and Bayesian artificial intelligence proposed many approaches to model and reason on causality,⁷³ the recent advance in machine learning – that leads to the strong resurgence of XIA – is more about learning correlations.⁷⁴ Consequently, as XAI is currently more focused on machine learning, recent explanation methods touch more upon correlations rather than causality.

At the same time, several works⁷⁵ have shown that an explanation is considered futile if the human has to make too great a cognitive effort to interpret it. Therefore, many authors consider that explanations must rely on symbolic approaches,⁷⁶ must be able to be evaluated,⁷⁷ and must help the addressee to understand why a decision was taken to enable her to contest the decision.⁷⁸

Without providing a definition of the notion of explanation *in abstracto*, in the next section of this article, we delve into different possible explanations, depending on the types of artificial agents, their interactions with humans, and the stages of algorithmic decision-making process. Therefore, in the remainder of this article, we discuss how different kind of agents and the legal requirement in GDPR may lead to a greater degree of feasibility for some explanations compared to others.

III. DIFFERENT TYPES OF AGENTS, DIFFERENT TYPES OF EXPLANATIONS

As mentioned above, the nature of artificial agents may have an impact on the nature and feasibility of the explanations that may be provided. In this part of the paper, we analyse this impact more closely. More specifically, in Section III.1, we explain how agents’

⁷¹ Mittelstadt et al, supra, note 60, p 281; Miller, supra, note 7, p 4.

⁷² *ibid.*

⁷³ Symbolic artificial intelligence uses logic-based model, see for instance JY Halpen, “Axiomatizing Causal Reasoning” (2000) 12 Journal of Artificial Intelligence Research 317. Bayesian artificial intelligence is interested in modelling probabilistic relationships between events, see J Williamson, *Bayesian Nets and Causality* (Oxford University Press 2005).

⁷⁴ G Marcus, “Deep Learning: A Critical Appraisal” (2018) ARXIV <arxiv.org/pdf/1801.00631.pdf> accessed 7 February 2020, p 20.

⁷⁵ The reader may refer to S Gregor and I Benbasat, “Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice” (1999) 23 MIS Quarterly 497, p 530 for a complete survey.

⁷⁶ D Pedreschi et al, “Meaningful Explanations of Black Box AI Decision Systems” (2019) 33rd AAAI Conference on Artificial Intelligence 9780, p 9782.

⁷⁷ ST Mueller et al, “Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI” (2019) DARPA Report ARXIV <arxiv.org/ftp/arxiv/papers/1902/1902.01876.pdf> accessed 7 February 2020, p 95.

⁷⁸ Wachter et al, supra, note 6, p 843.

behaviours can be described; in Section III.2 we distinguish different types of artificial agents with respect to how they interact with a human. Finally, in Section III.3, we demonstrate that a decision-making process is composed of different steps, which call for different kinds of explanation.

1. Explainability as a question of point of view: external and internal descriptions

The nature of explanations that may be provided about an artificial agent depends on what we know about it. Therefore, it can be claimed that explanations always depend on a point of view. The process of engineering an artificial agent involves a *designer* who designs it. Once the agent is designed, its behaviour can be observed by an *observer*. The agent can then be described by a set of properties that are understandable by the observer, by the designer, and sometimes by the agent itself. The notion of observer is particularly important, since some properties that may be observed do not have any practical and direct implementation in the agent as built by the designer. Thus, an artificial agent can be described from two perspectives: an external description and an internal description.⁷⁹ On the one hand, the *external description* is the description of the artificial agent based on a set of properties used by the observer.⁸⁰ This description is objective in the sense that it is built from the functioning of the agent without knowing what the designer had in mind and how the agent is really implemented. On the other hand, the *internal description* is the description of the agent based on a set of properties used by the designer, which may for instance be described by the source code of the agent, the fine-tuned parameters within the algorithms, or the weights learned by a neural network.⁸¹ This description is subjective because it relies on how the designer intended to implement the properties.

It is important to note that the internal and external descriptions of an artificial agent do not necessarily match. It is indeed possible for an observer to explain the behaviour of an artificial agent in a way that is completely different from how its internal architecture behaves in reality. This difference can be illustrated by a well-known example from the AI field. In 1948, the cybernetician William Grey Walter built a couple of tortoise robots, named Elmer and Elsie.⁸² Those robots were embedded with light sensors and a small light placed on the front that switched on when no light is detected. Their behaviour was based on a simple automaton that made them walk towards light if detected, and wandering randomly in search for light in other cases. When placed in front of a mirror, the tortoise robots were attracted by the reflection of their own light, making the light continuously switch off and on. The alternation of light switching on and off made the robots shake as their motors alternated quickly

⁷⁹ J Sichman et al, "When Can Knowledge-Based Systems be Called Agents?" (1992) Brazilian Symposium on Artificial Intelligence 1, p 8; C Carabelea et al, "Autonomy in Multi-Agent System: A Classification Attempt" (2003) International Conference on Agents and Computational Autonomy 103, pp 107–109.

⁸⁰ D Kinny et al, "A Methodology and Modelling Technique for Systems of BDI-Agents" (1996) 1038 Lecture Notes in Artificial Intelligence volume 56, p 58.

⁸¹ *ibid.*, p 60.

⁸² L Barrett, *Beyond the Brain: How Body and Environment Shape Animals and Human Minds* (Princeton University Press 2011) p 47.

Table 1: Explanations based on external and internal descriptions

	Explanation based on external description	Explanation based on internal description
Advantages	Relies on observed behaviour only Takes into account side-effects	More accurate
Drawbacks	Not necessarily correct	Needs access to source code or parameters

between moving towards the mirror and wandering randomly.⁸³ Walter noticed that the robots appeared to be “dancing” when “recognizing themselves in the mirror”.⁸⁴ It shows, as stated by Louise Barrett, that “without the knowledge of how [the robots] were put together, it would be natural to attribute more internal complexity than was warranted, as well as to overestimate the cognitive demands of the tasks that they could perform, like ‘recognizing’ themselves in the mirror”.⁸⁵ Consequently, from a right to explanation perspective, it is of the utmost importance to know whether the explanations provided about the behaviour of a given system are based on internal or external descriptions.

Grounding explanations on external descriptions means that it is only necessary to observe the behaviour of the system, and to analyse it afterwards. This is particularly relevant in cases where the internal description cannot be accessed, for instance if the source code of the assessed system is patented. Moreover, explanations based on external descriptions allow unintended behaviours or side effects to be better taken into consideration. Indeed, external descriptions rely on properties that hold in the eye of the observer, and those properties may have been unconsidered by the designer. For example, let us consider a medical assistant agent deciding what kind of medication a patient needs. Its decision may be explained based on the internal criteria implemented within the agent (eg the agent proposed this medication because it minimises the risks for the patient given her health state). This explanation is based on an internal description. However, from an external point of view, an explanation may be grounded on other criteria that were not implemented in the agent and that the agent was not intended to consider (eg the agent proposed this medication because it minimises the hospital’s financial cost).

The example above shows that explanations based on external descriptions have also drawbacks. Indeed, explaining a decision is not just about giving an explanation, but it is also about giving a *correct* explanation. Human beings might be misled by these external descriptions – in particular end users who do not have computer literacy. In this sense, explanations based on internal description can be more accurate as they are based on elements that are actually implemented in the assessed system. Table 1 summarises the advantages and the drawbacks of both types of descriptions in the perspective of explaining an algorithmic decision.

⁸³ *ibid.*

⁸⁴ *ibid.*

⁸⁵ *ibid.*

2. Interactions as explainability factors: supervised and unsupervised agents

The explainability of an agent, that is the capacity to explain its actions, can be further impacted depending on whether the decision is made by a supervised or by an unsupervised agent. In the case of a *supervised* agent, the human being is an *operator* who interacts with the agent (eg a robotic agent such as a drone) to make it achieve its functions. An example of a supervised agent is given by Coppin and Legras:⁸⁶ a military operator supervises a bio-inspired⁸⁷ unmanned aerial vehicle (UAV) swarm in order to perform various missions. In this case, the operator is trained to use the agents through setting high-level goals, eg commanding the swarm to watch or avoid an area, or to intercept a target, or directly taking the control of the UAV. Since the operator is trained to understand how the agents behave, it may be easier for her to explain the decision taken by the agent. However, as highlighted by Bathaee,⁸⁸ this task might nevertheless be challenging, because correctly explaining a supervised agent's decision requires the correct assessment to be made of how the operator influenced the agent's decision.

Another degree of complexity is added when the supervised agent is also an autonomous agent. Indeed, several types of supervised agents – called agents with *adaptive autonomy*, with *adjustable autonomy* or agents in *mixed initiative* – may be considered according to how the artificial agent, the human operator or both entities can take control of the artificial agent's behaviour.⁸⁹ Adaptive autonomy means that the artificial agent can take the control from the operator, whereas adjustable autonomy means that the operator can take control whenever she wants.⁹⁰ Finally, mixed initiative means that both the operator and the artificial agent can take over authority.⁹¹ Consequently, in terms of explainability, it is first important to know who (either the artificial agent, or the human operator) had control, in order to explain a given decision.

Differently, in case of an *unsupervised* agent, the human that interacts with the agent is a *user*. She uses the functions of the agent (eg a search agent on the Internet), mostly not knowing how they are implemented. An example of unsupervised agents are socially interactive agents such as conversation bots (eg Apple's Siri) where the user makes requests to an artificial agent in natural language.⁹² Since the users interacting with the agent are not trained to supervise the decisions taken by the agent, the "level" of explanations that an untrained user is capable of understanding may be different to the "level" understood by a trained operator. The level of expertise of the user is

⁸⁶ G Coppin and F Legras, "Controlling Swarms of Unmanned Vehicles through User-Centered Commands" (2012) AAAI Fall Symposium 21, p 25.

⁸⁷ Bio-inspired swarm algorithms are distributed (and often decentralised) algorithms inspired from insects' biological and social processes. Each agent in the swarm is an automated process with simple rules which are biased by the other agents' decisions. From a global point of view, the whole swarm is an autonomous system with a complex behaviour that emerges from each agent's local decisions.

⁸⁸ Y Bathaee, "The Artificial Intelligence Black-Box and the Failure of Intent and Causation" (2018) 31 Harvard Journal of Law and Technology 889, p 933.

⁸⁹ B Hardin and MA Goodrich, "On Using Mixed-Initiative Control: a Perspective for Managing Large-Scale Robotic Teams" (2009) International Conference on Human-Robot Interaction 166, pp 167–168.

⁹⁰ *ibid.*

⁹¹ *ibid.*

⁹² T Fong et al, "A Survey of Socially Interactive Robots" (2003) 42 Robotics and Autonomous Systems 143, p 145.

relevant also with regard to previously analysed external and internal descriptions. Indeed, it is easier to assume that an operator can better understand an explanation based on an internal description than a non-expert user.

3. Explainability in different stages of a decision-making process

As shown in Section II.3, explanations are multi-faceted. Considering the nature of the objects to be explained, the literature focuses on two kinds of explanations: global explanations that explain the whole system, and local explanations that explain a given decision.⁹³ However, this distinction does not take into account that the algorithmic decision-making process is not a monolithic process. Each decision involves several steps, ranging from data acquisition to the final decision, and, in the context of logic-based expert systems, it has been highlighted that an explanation showing the reasons behind each step of the decision-making process is more acceptable for users.⁹⁴ Traditionally, a decision-making process of an artificial agent is divided into four steps:

1. the agent perceives *raw data* through its sensors;
2. the agent uses a *situation awareness* module to transform those data into a *situation representation* in which it must make a decision;
3. the agent uses a *decision module* to compute with respect to its *situation representation* which *decision* is the best according to its *goals*;
4. the agent *acts* according to the decision.

Evidently, if we desire a *complete* explanation of a given decision, we need an explanation for each step: first, an explanation of how the agent assesses the situation with the perceived data; second, explanation of factors impacting the decision with respect to its goal; third, explanation of each factor involved in the decision-making process; and fourth, explanation of the final “product” of the decision. Explaining only step 4 may not be sufficient, as each step depends on the previous one. An explanation may require an understanding of why given data leads the artificial agent to a given decision. For example, let us consider a university admission system that uses an artificial agent to decide whether a given student is allowed to enrol, based on her academic dossier. Suppose the following explanation is given at the decision module step: “the artificial agent decided to not enrol the student as her chances of academic success are lower than the other students, and her Humanities’ marks are below a critical threshold”. Here, the different criteria used to weigh the decision are given: the student’s chances of academic success, and a threshold on the marks. However, we can wonder on what elements the student’s chances of academic success and this threshold are based on? In this case, a complete explanation would help to understand how the artificial agent assessed the student.

⁹³ Guidotti et al supra, note 11, p 13. We note that Guidotti et al use the term “outcome explanations”, whose meaning is the same as Adadi and Berrada, supra, note 3, p 52148 “local interpretability” and Mittelstadt et al, supra, note 84, p 280 “local explanations”.

⁹⁴ RL Ye and PE Johnson, “The Impact of Explanations Facilities on User Acceptance of Expert Systems Advice” (1995) 19 MIS Quarterly 157, p 172.

IV. THE RIGHT TO EXPLANATION: TECHNICAL AND LEGAL FEASIBILITY

In the previous sections, we discussed certain difficulties that an explanation of algorithmic decision-making processes might face both from computer science as well as legal perspectives. We demonstrated that any explanation should also take into account different types of agents and different steps of the decision-making process. In this section, we discuss whether these different types of explanations are feasible from a technical and a legal perspective. We review different technical approaches aimed at providing explanations and we investigate their legal feasibility, focusing in particular on intellectual property obstacles and the secrecy of algorithms.

1. Technical feasibility

Many of the surveys of XAI techniques focus on explaining machine learning systems, without putting the focus on the general perspective of artificial agents.⁹⁵ They distinguish *external systems* (also called post-hoc or black-box approaches), which are able to analyse an artificial agent and *reflexive systems* (also called intrinsic, interpretable, constructive or by design approaches), which allow artificial agents to provide explanations by themselves.⁹⁶ While each approach has its own advantages, they all have drawbacks: they are not fit for all types of artificial agents and cannot provide all kinds of explanations. Therefore, this section is not structured around different kinds of explanation but rather based on what is needed to compute those explanations. Such a structure highlights the technical requirements for each kind of explanation.

a. External explanation systems

External⁹⁷ explanation systems are devoted to analysing an artificial agent and proposing explanations. Such an approach has several advantages. First, it makes the explanation system distinct from the artificial agent, allowing the use of the same explanation system for different agents. Second, it can be used on agents that are unable to provide high-level representations of their behaviours. Indeed, such explanation systems can be used both on automated systems and autonomous artificial agents, whatever their inner decision-making processes are (eg logic-based rules, machine learning techniques). Two approaches can be considered, depending on whether the explanation system has

⁹⁵ This claim is supported by the systematic literature review provided in S Anjomshoae et al, "Explainable Agents and Robots: Results from a Systematic Literature Review" (2019) 18th International Conference on Autonomous Agents and Multiagent Systems 1078, p 1078. The authors highlight the lack of literature surveys on "explainable agency", meaning explaining the decisions of artificial agents.

⁹⁶ See Adadi and Berrada, supra, note 3, p 52151; Biran and Cotton, supra, note 50, p 3; Pedreschi et al, supra, note 76, p 9782; WJ Murdoch et al, "Interpretable Machine Learning: Definitions, Methods, and Applications" (2019) ARXIV <arxiv.org/pdf/1901.04592.pdf> accessed 7 February 2020, p 3.

⁹⁷ Here the term "external" is not directly related to the notion of "external description" given in Section III.1. The adjective "external" is therefore used in two different settings. The term "external" here refers to the fact that the explanation system is distinct from the explained system. An external explanation system is software that analyses agents to explain their decisions, while an internal (or reflexive) explanation system is a component embedded within the agents. Furthermore, external white-box approaches provide explanations based on internal descriptions. External black-box approaches probe explanations based on external descriptions.

access to an internal description of the agent:⁹⁸ the white-box approach, where analysis of the code or specification of the agent is possible, and black-box approach, where there is no knowledge of the agent's code.

1. Verification and validation techniques may be used as a *white-box testing technique* which inspects the agents in order to infer general properties such as invariance (what always happens) and liveness (what will happen).⁹⁹ The main advantage of this approach, called *property checking*,¹⁰⁰ is being able to provide pre-computed external descriptions for the users (eg a given banking decision-support system will never grant a loan to somebody with less than a given annual income). As stated by Fey and Drechsler, “these properties summarize the functionality of a design in a different way and thus explain the behaviour”.¹⁰¹ However, such approaches can only deal with agents whose internal description is known, and rely either on source code analysis or (more conveniently) formal specifications. For example, these approaches are not appropriate to explain the decisions of artificial agents based on neural networks, as the most important part in those agents' decisions are based on learned parameters, and not the algorithm. Moreover, those external white-box approaches often suffer from a high computational complexity.¹⁰²
2. *Black-box testing systems* may be used to probe the artificial agent by simulating different inputs and observing the results,¹⁰³ in order to infer either *local explanations*, or *counter-factual faithfulness*, or *decision trees*. Local explanation consists in computing a simpler decision model locally by sampling possible inputs.¹⁰⁴ This model is then used to highlight the main factors of the original decision-making process. For example, one major black-box algorithm explaining the predictions of a classifier is LIME¹⁰⁵

⁹⁸ C Castelluccia and D Le Métayer, “Understanding Algorithmic Decision-Making: Opportunities and Challenges” European Parliamentary Research Service Report PE 624.261, p 5; see also G Friedrich and M Zanker, “A Taxonomy for Generating Explanations in Recommender Systems” (2011) AAI AI Magazine 90, p 92. The distinction between a white-box and a black-box approach also exists in software engineering concerning program evaluation: “black-box evaluation” means the evaluator does not know how the program works, and “clear-box evaluation” means the inner component of the program are known. See M Scriven, “The Fine Line Between Evaluation and Explanation” (1999) 9 Research on Social Work Practice 521, p 523 (first published in (1994) 15 Evaluation Practice 75).

⁹⁹ G Frazer et al, “Testing with Model-Checkers: a Survey” (2007) SNA technical report (SNA-TR-2007P2-04) p 6.

¹⁰⁰ Property checking, or model checking, is an established approach to check exhaustively whether a model satisfies a given property. See Frazer et al, supra, note 99; R Jhala and R Majumdar, “Software Model Checking” (2009) 41 ACM Computing Survey (article no 21). Initially developed for symbolic systems, computer scientists try to adapt property checking to neural networks, see for instance S Wang et al, “Efficient Formal Safety Analysis of Neural Networks” (2018) 31 Advances in Neural Information Processing Systems 6367.

¹⁰¹ G Fey and R Drechsler, “Self-Explaining Digital Systems – Some Technical Steps” (2019) 22nd Workshop of Methods and Description Languages for Modelling and Verification of Circuits and Systems 2.

¹⁰² See Frazer et al, supra, note 99, p 4; property checking reasons on what will or can happen in the future, or until something else happens by using temporal logics. However, most used temporal logics are difficult to compute. For a survey, see P Schnoebelen, “The Complexity of Temporal Logic Model Checking” (2002) 4 Advances in Modal Logic 1, p 20.

¹⁰³ Doshi-Velez et al, supra, note 57, p 8.

¹⁰⁴ X Zhou and H Lin, “Local Sensitivity Analysis” in X Zhou et al (eds), *Encyclopedia of GIS* (Springer 2017).

¹⁰⁵ Tulio Ribeiro et al, supra, note 63, p 1137.

(Local Interpretable Model-agnostic Explanations) and its variants SPLIME,¹⁰⁶ DLIME¹⁰⁷ or LEAFAGE.¹⁰⁸ While those algorithms rely on machine learning techniques, other algorithms relying on principal component analysis¹⁰⁹ or Shapley value computation¹¹⁰ exist. Counter-factual faithfulness (or counter-factual explanations) consist in computing how an input would have needed to be changed in order to produce a desired result. It can then be used to show the impact of a given factor on the decision. From a computational point of view, apart from fixing a desired output, known black-box algorithms to compute counter-factual explanations consist in probing the inputs. For instance, the Growing Sphere¹¹¹ method samples the inputs around the input for which a counterfactual explanation must be computed, while Wachter et al¹¹² sample the inputs with an optimisation procedure that allow to find a counterfactual explanation which minimises the difference with the input to be explained.

Another approach, *decision decompilation*, consists in building logic-based programs from the input and results of the probed artificial agents, such as a system that generates easily explainable decision trees,¹¹³ or LISP-like¹¹⁴ programs, which interpret how traces change with time.¹¹⁵ For example, Guidotti et al proposed LORE (Local Rule-based Explanations), which provide logic-based local explanations with a decision rule associated to a set of counterfactual rules.¹¹⁶ In general, black-box testing is a very interesting approach to deal with a broad range of agents. For instance, it is very well adapted to provide explanations for neural networks. However, it only provides high-level links between inputs and outputs, namely an external description. It sees the decision-making as a whole and cannot consider explicitly each step in the process. Moreover, it is difficult for such approaches to

¹⁰⁶ *ibid*, p 1138.

¹⁰⁷ MR Zafar and NM Khan, “DLIME: A Deterministic Local Interpretable Model-Agnostic Explanations Approach for Computer-Aided Diagnosis Systems” (2019) ACM SIGKDD Workshop on Explainable AI/ML for Accountability, Fairness and Transparency, p 2.

¹⁰⁸ J Adhikari et al, “LEAFAGE: Example-based and Feature importance-based Explanations for Black-box ML models” (2019) ARXIV <arxiv.org/pdf/1812.09044.pdf> accessed 7 February 2020, p 2.

¹⁰⁹ C Brinton, “A framework for explanation of machine learning decisions” (2017) IJCAI Workshop on Explainable Artificial Intelligence 14, p 14.

¹¹⁰ SM Lundberg et al, “Explainable AI for Trees: From Local Explanations to Global Understanding” (2019) ARXIV <arxiv.org/pdf/1905.04610.pdf> accessed 7 February 2020, p 21.

¹¹¹ T Laugel et al, “Comparison-Based Inverse Classification for Interpretability in Machine Learning” (2018) 853 Communications in Computer and Information 100, pp 105, 106.

¹¹² Wachter et al, *supra*, note 6, p 854.

¹¹³ S Singh et al, “Programs as Black-Box Explanations” (2016) NIPS Workshop on Interpretable Machine Learning in Complex Systems <arxiv.org/pdf/1611.07579.pdf> accessed 7 February 2020.

¹¹⁴ LISP (LISt Processing) is a family of programming languages characterised by a fully parenthesised prefix notation. For instance, “A and B” is denoted “(and A B)”. This family of language is well-adapted to represent mathematical expressions. See K Hinsien, “The Promises of Functional Programming” (2009) 11 Computing in Science & Engineering 86.

¹¹⁵ S Penkov and S Ramamoorthy, “Using program induction to interpret transition system dynamics” (2017) ICML Workshop on Human Interpretability in Machine Learning 22, p 24.

¹¹⁶ R Guidotti et al, “Local Rule-Based Explanations of Black Box Decision Systems” (2018) ARXIV <arxiv.org/pdf/1805.10820.pdf> accessed 7 February 2020, pp 1–10; Pedreschi et al, *supra*, note 76, p 9782.

distinguish a controlled usage of randomness¹¹⁷ in a decision-making process from simple chaotic behaviours.

b. Reflexive explanation systems

A second way to provide explanations is to use artificial agents embedded with reflexive systems. These are agents which are designed for a given goal but which can also reason on their own behaviour and thus can produce formalised explanations.¹¹⁸ This approach is complementary to external explanation systems as it produces explanations based on a fine-grained knowledge of the agent's internal description, and is clearly fitted to this latter. To this end, artificial agents embedded with a reflexive explanation system satisfy a *reflexivity* property, meaning their internal description must be able to reason by itself and produce external descriptions. Two approaches can be considered, depending on whether the reflexivity is implemented within the agent's reasoning module or is implemented in addition to the reasoning module:

1. Reflexive systems added to the reasoning module are often systems which translate the agent's traces into natural language, providing an *ex ante* explanation through *interpretable traces*.¹¹⁹ More sophisticated systems try to highlight causal links between the decisions and the data the agent used when the decision was made. For example, some works propose an explanation module which inspects the agent's hierarchy of tasks (eg a sequence of actions "to achieve A, the agent needs to achieve B, then C") and explains a given action through the underlying goal and the current beliefs at the time the decision was taken (eg "the agent decided B in order to achieve A because B is needed to achieve A and the agent believed that B was possible").¹²⁰ Others propose a similar system for translating optimal policies¹²¹ in human-interpretable sentences: the artificial agent can communicate not only about data it uses to decide but also about likelihoods of possible outcomes or expected rewards.¹²²
2. Similarly, as in the previous approach, reflexive systems built within the agent's reasoning module aim to translate the agent's traces into natural language. Unlike the previous approach, built-in reflexive systems aim to produce explanations during the reasoning process and not *ex ante*. To this end, the explanation system must be interlinked with the reasoning module and *provide arguments*. For example, Briggs and Scheutz proposed a deontic logic-based

¹¹⁷ For instance, in some game-theoretical settings where artificial agents must face uncertainty, the "best" decision can be computed as a mixed strategy, namely a probability distribution over the actions to choose. The agents act randomly according to this distribution. As it is a stochastic decision, only observing few decisions hardly allows to distinguish between an agent applying a mixed strategy or a random behaviour.

¹¹⁸ B Moulin et al, "Explanations and Argumentation Capabilities: Towards Creation of More Persuasive Agents" (2002) 17 Artificial Intelligence Review 169, p 171.

¹¹⁹ MG Core et al, "Building Explainable Artificial Intelligence Systems" (2006) AAAI Conference on Artificial Intelligence 1766, p 1768.

¹²⁰ M Harbers et al, "Design and Evaluation of Explainable BDI Agent" (2010) International Conference on Web Intelligence and Intelligent Agent Technology 125, p 127.

¹²¹ For an explanation of the notion of policies in computer science, see LP Kaelbling et al, "Planning and Acting in Partially Observable Stochastic Domains" (1998) 101 Artificial Intelligence 99.

¹²² N Wang et al, "The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams" (2016) International Conference on Autonomous Agents and Multi-Agent Systems 997, p 999.

explanation module for the special case where an artificial agent can decide to reject a human user request.¹²³ General reasons to reject the request are hard-wired in the reasoning process (eg in the context of the autonomous university admission agent: “the agent cannot enrol more than a given number of students”) and concrete explanations are generated when those rules are violated (eg “your academic dossier ranks you on 211th position while we can only enrol 200 students”). In the context of machine learning, Ding¹²⁴ and Rudin¹²⁵ advocate the use of neural logic networks whose combination of weights explicitly represent logical rules. Another way to interlink explanations and reasoning is to use formal argumentation.¹²⁶ This technique represents *pro* and *contra* arguments about a set of possible decisions and computes acceptable decisions with respect to different semantics (eg a decision can be made if – and only if – there always exists an argument against all arguments against the decision). All arguments form a hierarchical structure where users may search for more detailed explanations. For example, Čyras et al proposed to explain decisions, determined by humans or by machines indifferently, by modelling argumentative disputes between two fictitious disputants arguing, respectively, for or against the decision in need of explanation.¹²⁷

c. Classification of different explanation systems

In conclusion, both self-explainable artificial agents, either built on top or built within, as well as external explanation systems, either white-box testing or black-box testing approaches, have advantages and drawbacks, summarised in Table 2.

As shown by Table 2, if technical constraints prevent the use of external white-box explanation systems or reflexive modules, then the right to explanation is limited to external black-box approaches. Unfortunately, the external black-box approaches cannot explain the different steps of a decision-making process. Indeed, only the output, namely the final step of the decision-making process, is explained in terms of inputs. How the inputs are used to produce internal representations is outside the scope of the explanations, as well as the fact the decision may have been influenced by human supervision in case of supervised systems.

The other types of explanation approaches also have their limits. Both reflexive approaches are limited to agents specially designed for such approaches. Moreover, those agents are best fit when they rely on logic-based mechanism. Consequently, while it seems interesting to design such an agent from a scientific point of view, it

¹²³ G Briggs and M Scheutz, “Sorry I Can’t Do That: Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions” (2015) AAAI Fall Symposium 32, p 33.

¹²⁴ L Ding, “Human Knowledge in Constructing AI Systems – Neural Logic Networks Approach towards an Explainable AI” (2018) 126 *Procedia Computer Science* 1561, p 1563.

¹²⁵ Rudin, *supra*, note 10, p 8.

¹²⁶ X Fan and F Toni, “On Computing Explanations in Argumentation” (2015) AAAI Conference on Artificial Intelligence 1496, p 1496; B Liao et al, “Representation, Justification and Explanation in a Value Driven Agent: An Argumentation-Based Approach” (2018) ARXIV <arxiv.org/pdf/1812.05362.pdf> accessed 7 February 2020; Moulin et al, *supra*, note 118.

¹²⁷ K Čyras et al, “Explanations by Arbitrated Argumentative Dispute” (2019) 127 *Expert Systems with Applications* 141.

Table 2: Advantages and drawbacks of different explanation systems

Approach	Subtype	Advantages	Drawbacks
External	White-box	Each step can be explained Provides internal descriptions	Needs source code or formal specifications Can only pre-compute explanations
External	Black-box	No source code or formal specifications needed Fits to a broad set of agents	Provides only external descriptions Views the decision as a whole
Reflexive	On top	Each step can be explained Provides internal descriptions	Needs agents designed for it Simple translation of traces
Reflexive	Within	Each step can be explained Provides internal descriptions	Needs agents designed for it Only fits to logic-based agents

seems difficult to oblige designers to comply with such constraints. Lastly, the external white-box approach, while not relying on an agent especially designed for it, suffers from two limitations: it needs access to the source code or a formal specification of the agent's behaviour; and it does not provide explanations in itself but only shows some general properties that may be used for explanation afterwards.

Finally, not all approaches are tailored to deal with multi-agent systems. Such systems are composed of several (even hundreds of) artificial agents which may be heterogeneous in their design, and their global behaviour is the result of the local interactions between all agents.¹²⁸ A concrete example of such a system is a web service composition system: a global service (eg a trip management service) is an agent that uses several small services, also agents, dynamically chosen with respect to some quality metrics (eg cost or response time).¹²⁹ To use another example, a network of autonomous vehicles is also a multi-agent system. Due to the distributed and decentralised properties of such systems, providing proofs and analysis is more difficult. Moreover, those systems are generally open, meaning that agents are able to join or leave the system. Such agents can be heterogeneous (in architecture, designers and owners) and can act on behalf of different users. In this context, several new questions are raised. How to provide explanations while some part of the decision is heavily influenced by the decisions made by the other agents? Assuming the agents can provide explanations, how to use and combine them to provide a global explanation? How to deal with agents unable to provide explanations? Those questions currently remain open in the context of the future massive interconnected multi-agent systems.

2. Legal feasibility

As stated at the beginning of this article, numerous scholars have argued that the GDPR establishes or at least should establish the right to explanation.¹³⁰ Recognising the legal

¹²⁸ See JP Müller and K Fisher, "Application Impact of Multi-agent Systems and Technologies: A Survey" (2014) *Agent-Oriented Software Engineering* 27 for a review of the industrial applications of multi-agent systems.

¹²⁹ QZ Sheng et al, "Web Service Composition: A Decade's Overview" (2014) 280 *Information Sciences* 218.

¹³⁰ See supra, note 15.

right to explanation of algorithmic decisions might be seen as a panacea that clearly and easily enables an in-depth understanding of such decisions. However, the ease of such a claim could be compared to an elusive *fata morgana*. Not only does the existence of the right to explanation not guarantee its smooth practical implementation, due to concerns related to technical feasibility, as explained in the previous section; the effective understanding of reasons behind an algorithmic decision might also be obstructed by legal obstacles. Specifically, trade secrets or other confidential information can stand in the way of algorithmic transparency. Below we examine the legal factors that can impact the effective exercise of the right to explanation.

a. Trade secrets: a genuine obstacle?

Trade secrets,¹³¹ which are distinct from intellectual property rights,¹³² can potentially stand in the way of effective exercise of the right to explanation.¹³³ As stipulated by the Trade Secrets Directive,¹³⁴ a trade secret is information which is secret, has commercial value due to its secrecy and has been kept secret by reasonable steps of the information holder,¹³⁵ such as an undisclosed know-how and business information.¹³⁶ Algorithms can certainly fall within this definition¹³⁷ and have been effectively covered by trade secrets in practice. Indeed, national case law has already recognised such protection¹³⁸ and some Member States specifically offered the possibility of acquiring trade secrets over technology when transposing the Trade Secrets Directive into national law.¹³⁹ On a global scale, the Amazon recommendation

¹³¹ We do not address here the issue of overlap between trade secrets and business secrets, as this question exceeds the scope of this article. The EU General Court indeed seems to make the distinction between the two concepts; see T-643/13, *Rogesa v Commission*, ECLI:EU:T:2018:423, paras 90, 101. Moreover, the EDPS opines that trade secrets differ from business secrets; see Opinion of the European Data Protection Supervisor on the proposal for a directive of the European Parliament and of the Council on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, 12 March 2014 <edps.europa.eu/sites/edp/files/publication/14-03-12_trade_secrets_en.pdf> accessed 7 February 2020, paras 16–17.

¹³² Recital 39 Trade Secrets Directive.

¹³³ Wachter et al, *supra*, note 12, p 87 ff; G Malgieri and G Comandé, “Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation” (2017) 7 *International Data Privacy Law* 243, pp 262–264; See also M Brkan, “AI-Supported Decision-Making under the General Data Protection Regulation” (2017a) Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, London, 12–16 June 2017 <dl.acm.org/citation.cfm?id=3086513> accessed 7 February 2020, pp 3, 6; M Brkan, “Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond” (2017b), available at SSRN <ssrn.com/abstract=3124901> accessed 7 February 2020, pp 21–22. The cross-references used below refer to the published version of this paper (2017a) rather than the paper in the SSRN eLibrary (2017b).

¹³⁴ Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure [2016] OJ L 157/1.

¹³⁵ Art 2(1) Trade Secrets Directive. This definition essentially follows the one from Art 39(2) of the Agreement on Trade-related Aspects of Intellectual Property Rights. A similar definition from TRIPS is mentioned also by Brkan (2017a), *supra*, note 133, p 6.

¹³⁶ The Trade Secrets Directive expressly mentions undisclosed know-how and business information in its title.

¹³⁷ See for example also M Maggolino, “EU Trade Secrets Law and Algorithmic Transparency” (2019), available at SSRN <ssrn.com/abstract=3363178> accessed 7 February 2020, p 5 ff; Brkan (2017a), *supra*, note 133, p 6.

¹³⁸ As pointed out by Wachter et al, *supra*, note 12, p 87 ff, national authorities have already recognised trade secrets as a defence against providing a full explanation as to how the algorithm is functioning. However, from their analysis, it is not entirely clear whether the trade secrets in question protected the software, the algorithm or the code. They seem to use these three notions interchangeably (p 87).

system, the Instagram algorithm for publication diffusion or Google's search algorithms are among the most well-known examples of trade secrets.¹⁴⁰ In fact, the core algorithm of Google's search engine, the PageRank algorithm, based on a randomised procedure to diagonalise a matrix which represents the relationships between web pages, is widely known and has been used in several other algorithms, such as an e-reputation algorithm called EigenTrust.¹⁴¹ However, Google's precise modalities to determine the relationships between pages, the optimisations built into the search system and the parameters used to detect such manipulations are not revealed.¹⁴² For example, it is unknown how different criteria are weighted, such as the number of links, the traffic on the pages or the structure of the pages' source code.

The importance of trade secrets and the resulting tension with data subject's rights to access and information is also recognised by the GDPR. Article 23 GDPR allows for these rights to be limited for the protection of "the rights and freedoms of others"¹⁴³ and one of the recitals specifically requires that the right to access should not "adversely affect" the trade secrets of the controller.¹⁴⁴ For its part, the Trade Secrets Directive allows for a suspension of a trade secret "for the purpose of protecting a legitimate interest recognised by Union or national law";¹⁴⁵ such a legitimate interest could potentially be providing an explanation of an algorithmic decision to a data subject.¹⁴⁶ However, a trade secret should also not affect data subjects' rights, in particular the right of access to the "personal data being processed".¹⁴⁷ From a legal perspective, it therefore remains rather unclear which set of rules should take precedence in case of conflict of trade secrets with data subjects' rights. Malgieri and Comandé argue, based on a systematic interpretative method, that data protection rights should take precedence over trade secret rules.¹⁴⁸ However, such a solution might not correspond to the purpose-based analysis; if GDPR always prevailed over trade secrets, the latter could never be protected when providing an explanation of an algorithmic decision to the data subject.

In this contribution, we would like to offer an alternative solution to the legal conundrum of prevalence of the two sets of rules. The core of our argument is that, even though an algorithm is protected by a trade secret, explaining the decision based

¹³⁹ For example, Belgium, Bulgaria, Estonia and Greece; European Union Intellectual Property Office, *The Baseline of Trade Secrets Litigation in the EU Member States* (EUIPO, 2018, doi: 10.2814/19869) pp 16, 26, 32, 33, 77, 95.

¹⁴⁰ The secrecy of the Google search algorithm is pointed out also by G Noto La Diega, "Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information" (2018) JIPITEC 3, para 34.

¹⁴¹ SD Kamvar et al, "The eigentrust algorithm for reputation management in P2P networks" (2013) 12th International Conference on World Wide Web 640, p 641.

¹⁴² See for example also Maggiolino, *supra*, note 137, p 7.

¹⁴³ Art 23(1)(i).

¹⁴⁴ Recital 63 GDPR. See further on this issue Malgieri and Comandé, *supra*, note 133, pp 262–264. They rightly observe (at p263) that the limitation extends only to the right of access and not also to the right to information which equally requires the revelation of the "meaningful information about the logic involved" in the decision-making.

¹⁴⁵ Art 5(d) Trade Secrets Directive.

¹⁴⁶ Brkan (2017a), *supra*, note 133, p 6.

¹⁴⁷ Recital 35 Trade Secrets Directive.

¹⁴⁸ Malgieri and Comandé, *supra*, note 133, pp 263–264. See also G Malgieri, "Trade Secrets v Personal Data: a possible solution for balancing rights" (2016) 6 International Data Privacy Law 102, pp 104–105.

on this algorithm does not necessarily need to encroach upon this secret.¹⁴⁹ First, we believe that the extent to which a trade secret and the underlying functioning of an algorithm are revealed largely depend on the scope of the explanation. If the explanation requires the revelation of the entire underlying logic of the algorithm, the trade secret would obviously be revealed. If, on the other hand, the explanation aims to divulge, for example, only the main factor influencing a decision, the trade secret would remain untouched.¹⁵⁰ In cases where an algorithm is protected by the trade secret, the right to explanation (albeit in its more limited version) could be safeguarded through the balancing of different interests.

Second, revealing trade secrets on algorithms would depend also on who probes and tests the algorithm to reach its explanation. If the company itself holding the trade secret on the algorithm provides the data subject with a textual explanation as to the factors impacting the decision, this might not mean that the trade secret would be unlawfully revealed. Moreover, if an algorithm protected by a trade secret is probed by a court, such probing should equally be allowed as long as the further non-disclosure of a trade secret by anyone involved in legal proceedings is respected.¹⁵¹ For example, in a procedure before the court to determine the existence of discrimination by an algorithm protected by a trade secret, the latter could be revealed provided that anyone involved in the procedure keeps this information confidential.

Third, the explanation could be further obtained through reverse engineering of a protected algorithm in the public domain.¹⁵² According to the Trade Secrets Directive, if a product has “been made available to the public” and the revelation of trade secret is a consequence of “observation, study, disassembly or testing” of this product, the trade secret is considered to have been acquired lawfully.¹⁵³ By way of example of Google’s search algorithm, it is legally allowed to reverse engineer this algorithm, even though technically this is a difficult endeavour given frequent changes to the algorithm.¹⁵⁴ Furthermore, this provision of the Trade Secrets Directive equally seems to allow for “black-box testing” or, in other words, “external explanation systems” based on the black-box approach as presented in section IV.1 of this article. Black-box testing means that another algorithm will probe the tested algorithm in order to build a model of the latter, and use this model to provide an explanation. For example, an algorithm can be probed to identify the influence of a given variable¹⁵⁵ or to try to identify whether the outcome is discriminatory.

¹⁴⁹ Brkan (2017a), *supra*, note 133, p 6.

¹⁵⁰ Compare Wachter et al, *supra*, note 6, pp 871, 883.

¹⁵¹ Art 9 Trade Secrets Directive; this provision can be invoked in “legal proceedings relating to the *unlawful acquisition, use or disclosure* of a trade secret” (emphasis added). It can be argued that a procedure where the court probes an algorithm protected with a trade secret in order to reveal this secret falls within this category.

¹⁵² See also Maggiolino, *supra*, note 137, p 11.

¹⁵³ Art 3(1)(b) Trade Secrets Directive. See also Maggiolino, *supra*, note 137, 13.

¹⁵⁴ M Martinez, “Why You Cannot Reverse Engineer Google’s Algorithm” (*SEU Theory*, 8 January 2011) <www.seo-theory.com/why-you-cannot-reverse-engineer-googles-algorithm/> accessed 7 February 2020.

¹⁵⁵ See for example A Datta et al, “Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems” (2016) IEEE Symposium on Security and Privacy, 22–26 May 2016, pp 598–617, <ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7546525> accessed 25 February 2020.

b. Classified information

Classified information on the level of the EU¹⁵⁶ or its Member States can potentially be one of the biggest obstacles for algorithmic transparency. The EU distinguishes between different levels of classified information, ranging from “EU top secret” to “EU restricted”.¹⁵⁷ The level of classification depends on the degree of harm that revealing of such information can cause to the interests of the EU or its Member States.¹⁵⁸ In certain cases, it is in the interest of the public authorities not to reveal exactly why a certain decision was taken.¹⁵⁹ For example, if military forces use algorithms to plan a sensitive and targeted military operation, and if the modalities of this operation constitute classified information, it could be justified that such algorithms not be revealed to the public. It is important, however, that labelling of a particular information as classified is not abused with the intention to keep the algorithms intransparent. It is expected that the modalities of algorithmic decision-making would need to be part of a broader content of classified information. In other words, it is difficult to see how an algorithm by itself could be designated as “classified” without the classified nature of the information for which it is being used.

The desire of public authorities to keep algorithms secret should not, therefore, be equated with the classified nature of information of which these algorithms form part. For example, the government in the Netherlands has been using secret algorithms to profile its citizens and categorise them on the basis of the risk they might pose, in particular the risk for tax fraud and allowance fraud.¹⁶⁰ The secrecy of these algorithms has been widely debated in the media and even led to a legal action against the Dutch government for non-compliance with privacy rights.¹⁶¹ Consequently, the Dutch court declared this risk indication system, known under the abbreviation SyRI, as incompatible with the right to privacy.¹⁶² Similarly, the French algorithm for university admission remained (unjustifiably) secret until a high school association managed to obtain the information about its modalities.¹⁶³ However, to our knowledge, the use of these algorithms did not seem to fall under “classified” information. Therefore, contrary to the classified information, these “unclassified”

¹⁵⁶ On classified information in the EU, see for example V Abazi, *Official Secrets and Oversight in the EU: Law and Practices of Classified Information* (Oxford University Press 2019); D Curtin, “Overseeing Secrets in the EU: A Democratic Perspective” (2014) 52(3) *Journal of Common Market Studies* 684; D Galloway, “Classifying Secrets in the EU” (2014) 52(3) *Journal of Common Market Studies* 668.

¹⁵⁷ See Art 2(2) of the Council Decision 2013/488/EU of 23 September 2013 on the security rules for protecting EU classified information [2013] OJ L 274/1. For an in-depth analysis of questions relating to classified information in the EU, see Abazi, *supra*, note 156.

¹⁵⁸ For example, revealing “EU top secret” information “could cause exceptionally grave prejudice to the essential interests of” EU or its Member States; Art 2(2) of the Council Decision 2013/488/EU.

¹⁵⁹ Brkan (2017a), *supra*, note 133, p 6. For example, the Dutch tax authorities and customs authorities rely heavily on automated decision-making. We are grateful to Professor Sjir Nijssen for this insight.

¹⁶⁰ See for example J Henley and R Booth, “Welfare surveillance system violates human rights, Dutch court rules” (The Guardian, 5 February 2020) <www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules> accessed 25 February 2020; see also website <bijvoorbbaatverdacht.nl/> accessed 7 February 2020.

¹⁶¹ See for example H Teunis, “Staat gedagvaard om “ondoorzichtige” risicoanalyses burgers” (rtlnieuws, 27 March 2018) <www.rtlnews.nl/tech/artikel/4128921/staat-gedagvaard-om-ondoorzichtige-risicoanalyses-burgers> accessed 7 February 2020.

¹⁶² See the judgment of the District Court of the Hague in the SyRI case, 5 February 2020, C-09-550982 HA ZA 18-388, ECLI:NL:RBDHA:2020:865, <uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:865> accessed 25 February 2020.

algorithms may be revealed. In responding to the question whether they should be revealed, different interests have to be balanced: the right to privacy and data protection, the prevention of fraud, or – in the case of France – the secrecy of deliberation for university admission.

c. Patents and copyrights¹⁶⁴

It is submitted that patent protection does not stand in the way of effective exercise of the right to explanation. Since patent law is not harmonised in the EU, save for a few exceptions,¹⁶⁵ this legal domain remains to be covered by the European Patent Convention (EPC)¹⁶⁶ to which all EU Member States are parties. The EPC allows for patenting of “computer-implemented inventions”,¹⁶⁷ provided that the software is new, involves an inventive step and allows for industrial application.¹⁶⁸ However, such patentability does not extend to algorithms where the inventive step criterion would be fulfilled only if the algorithm possessed “technical features”.¹⁶⁹ In this regard, the European Patent Office (EPO) confirmed that an algorithm is essentially a mathematical method which should not be considered an invention, except if it serves a technical purpose which would give it a technical character.¹⁷⁰ This understanding is in line with the general definition of an algorithm, which can be seen as “a set of steps to accomplish a task that is described precisely enough that a computer can run it”.¹⁷¹ The nature of a mathematical method of an algorithm could be portrayed with an example of a procedure to build a house or a more complex procedure to train a neural network. Based on this line of reasoning, the EPO for example refused to award patents for an algorithm for automatic document classification,¹⁷² an optimisation

¹⁶³ See for example C Chausson, “France: How a high school association finally obtained a source code” (Open Source Observatory, 24 April 2017) <joinup.ec.europa.eu/news/france-how-high-school-ass> accessed 7 February 2020.

¹⁶⁴ This part is an extended version of Brkan (2017a), supra, note 133, p 6.

¹⁶⁵ Directive 98/44/EC of the European Parliament and of the Council of 6 July 1998 on the legal protection of biotechnological inventions [1998] OJ L 213/13. Moreover, Regulation 1257/2012 gives a patent granted by EPO a unitary effect in Member States participating in enhanced cooperation; see Regulation (EU) No 1257/2012 of the European Parliament and of the Council of 17 December 2012 implementing enhanced cooperation in the area of the creation of unitary patent protection [2012] OJ L 361/1. See related to language arrangements also Council Regulation (EU) No 1260/2012 of 17 December 2012 implementing enhanced cooperation in the area of the creation of unitary patent protection with regard to the applicable translation arrangements [2012] OJ L 361/89.

¹⁶⁶ Convention on the Grant of European Patents (European Patent Convention) of 5 October 1973, available at <www.epo.org/law-practice/legal-texts/html/epc/2016/e/ma1.html> accessed 7 February 2020.

¹⁶⁷ See Art 52(2)(c) of the European Patent Convention in combination with Guidelines for Examination, point 3.6 Programs for computers, available at <www.epo.org/law-practice/legal-texts/html/guidelines/e/g_ii_3_6.htm> accessed 7 February 2020.

¹⁶⁸ See Art 52(1) of the European Patent Convention; Brkan (2017a), supra, note 133, p 5.

¹⁶⁹ According to the European Patent Office, non-technical features should be “ignored in assessing novelty and inventive step”; see European Patent Office, “Case Law of the Boards of Appeal”, point 9.1, <www.epo.org/law-practice/legal-texts/html/caselaw/2016/e/clr_i_d_9_1.htm> accessed 7 February 2020. See also Noto La Diega, supra, note 140, para 40.

¹⁷⁰ Decision of the EPO Board of Appeal, T 1784/06 (Classification method/COMPTEL) of 21.9.2012, ECLI:EP:BA:2012:T178406.20120921, point 3.1.1. See also Noto La Diega, supra, note 140, para 40.

¹⁷¹ TH Coormen, *Algorithms Unlocked* (MIT Press 2013) p 1.

¹⁷² Decision of the EPO Board of Appeal, T 1358/09 (Classification/BDGB ENTERPRISE SOFTWARE) of 21.11.2014, ECLI:EP:BA:2014:T135809.20141121. See also Noto La Diega, supra, note 140, para 40.

algorithm determining a travel route in navigation system,¹⁷³ and an algorithm aimed at finding relationships between items.¹⁷⁴ According to the EPO, in order to patent a computer algorithm, the “programmer must have had technical considerations beyond ‘merely’ finding a computer algorithm”.¹⁷⁵ Even though Edsger W Dijkstra is considered to have “invented” the algorithm nowadays used in most¹⁷⁶ of the modern GPS systems,¹⁷⁷ this “invention” still does not fulfil the criteria of patentability because the algorithm remains a mathematical procedure. In consequence, since algorithms as mathematical methods are currently not patentable under the EPC,¹⁷⁸ the patent protection cannot truly create obstacles for understanding and explaining of algorithmic decision-making.

The absence of copyright protection of algorithms in the EU leads to a similar result.¹⁷⁹ While the Directive on the legal protection of computer programs¹⁸⁰ expressly allows for copyright protection of those programs, this is not the case for algorithms underpinning them. In line with the TRIPS agreement¹⁸¹ and the WIPO copyright treaty,¹⁸² the said Directive protects only “the expression of a computer program” and not “ideas and principles” behind those programs.¹⁸³ Therefore, if algorithms comprise such ideas and principles, they cannot enjoy copyright protection,¹⁸⁴ an approach which is largely in line with some other non-European jurisdictions.¹⁸⁵ Even though the Court of Justice of the EU (CJEU) has not yet expressly decided on this issue, its *obiter dictum* in *SAS Institute*¹⁸⁶ seems to have implicitly confirmed this reasoning. By

¹⁷³ Decision of the EPO Board of Appeal, T 2035/11 (Navigation system/BEACON NAVIGATION) of 25.7.2014, ECLI:EP:BA:2014:T203511.20140725.

¹⁷⁴ Decision of the EPO Board of Appeal, T 0306/10 (Relationship discovery/YAHOO!) of 4.2.2015, ECLI:EP:BA:2015:T030610.20150204.

¹⁷⁵ See Opinion of EPO G 0003/08 (Programs for computers) of 12.5.2010, ECLI:EP:BA:2010:G000308.20100512, point 13.5; Brkan (2017a), *supra*, note 133, p 6.

¹⁷⁶ To be precise, Dijkstra’s algorithm is the best algorithm to compute shortest paths on a map. However, on large maps, the algorithm is very slow. Real path planning software generally uses both a generalisation of Dijkstra’s algorithm (named A*) which can be faster if correctly parametrised, and compact and efficient representations of the maps (for instance contraction hierarchies). Nevertheless, Dijkstra’s algorithm is the basis of all path-planning algorithms. See for example D Delling et al, “Engineering Route Planning Algorithms” (2009) 5515 Lecture Notes in Computer Sciences 117, p 118.

¹⁷⁷ *ibid.*

¹⁷⁸ This might be different in the US; see Noto La Diega, *supra*, note 140, para 37.

¹⁷⁹ Brkan (2017a), *supra*, note 133, p 6.

¹⁸⁰ Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L 111/16.

¹⁸¹ Art 9(2) of the Agreement on Trade-Related Aspects of Intellectual Property Rights prevents copyright protection of “ideas, procedures, methods of operation or mathematical concepts”.

¹⁸² See Art 2 of the World Intellectual Property Organization Copyright Treaty.

¹⁸³ Recital 11 Directive 2009/24.

¹⁸⁴ Recital 11 Directive 2009/24. Compare also Noto La Diega, *above*, note 164, para 37.

¹⁸⁵ Brkan (2017a), *supra*, note 133, p 6. For example, in Japan, copyright protection of algorithms is not allowed; see DS Karjala, “Japanese Courts Interpret the ‘Algorithm’ Limitation on the Copyright Protection of Programs” (1991) 31 *Jurimetrics Journal* 233; for US scholarship on this issue (as well as patents on algorithms) see RH Stern, “On Defining the Concept of Infringement of Intellectual Property Rights in Algorithms and Other Abstract Computer-Related Ideas” (1995) 23 *AIPLA Quarterly Journal* 401; J Swinson, “Copyright or Patent or Both: An Algorithmic Approach to Computer Software Protection” (1991) 5 *Harvard Journal of Law & Technology* 145.

¹⁸⁶ Case C-406/10, *SAS Institute*, ECLI:EU:C:2012:259. For a discussion of this case, see Noto La Diega, *supra*, note 140, paras 34, 42.

referring to the equivalent recital of the directive previously in force¹⁸⁷ and pointing out that the copyright protection extends only to the source or object code of the program, the CJEU not only confirmed that algorithms cannot be subject to copyright protection, but also that the algorithm and the code need to be clearly distinguished.¹⁸⁸ Despite that, some national jurisprudence nevertheless seems to recognise this possibility, as demonstrated by the recent decisions of an Italian administrative court referring to a copyright protection of algorithms.¹⁸⁹

Within the system of copyright protection, it is important to distinguish algorithms from source code and software. A source code is a particular implementation of an algorithm in a particular programming language, whereas an algorithm is a “reliable definable procedure for solving a problem” which is independent of the language and the implementation.¹⁹⁰ Software is the “final product” of the software engineering process, ie the functionalities through an executable program and the packaging, such as the name of the software, the data used as parameters, the appearance of the user interface and the documentation.¹⁹¹ The distinction between these concepts can be illustrated with an example of a facial recognition software based on neural networks technology. The procedure used to train the data structure within the neural network is the algorithm (for instance a back-propagation algorithm¹⁹² which plays the same role as Dijkstra’s algorithm for pathfinding systems¹⁹³). The particular implementation of both the data structure and the procedure (for example in C++ language) is the source code. All that remains is the software, such as the data used to train the network, the particular values of the neurons once the network has been trained, the interface to interact with a user or the documentation. Therefore, a software and a source code can be copyrighted, but not an algorithm that is merely an abstract idea underpinning the software and the source code.

Furthermore, even if an algorithm forms part of a copyrighted computer program, this does not mean that the algorithm itself is protected by copyright and that its functioning cannot be revealed. The Directive on the legal protection of computer programs allows the user of a computer program “to observe, study or test the functioning of the program in order to determine the ideas and principles”¹⁹⁴ – that is, also algorithms – behind the program. In terms of the right to explanation, a user of a computer program could

¹⁸⁷ That is, Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs [1991] OJ L 122/42.

¹⁸⁸ *SAS Institute*, paras 32–34.

¹⁸⁹ Regional Administrative Court in Lazio, case No 3742/2017, 21 March 2017, *CISL, UIL, SNALS v MUIR*; Regional Administrative Court in Lazio, case No 3769/2017, 22 March 2017, *Gilda v MUIR*. For a commentary, see Gianluca Campus, “Copyright protection of algorithms does not prevent the disclosure of their source code in the context of administrative proceedings” (*IPLens*, 30 March 2018) <[iplens.org/2018/03/30/copyright-protection-of-algorithms-does-not-prevent-the-disclosure-of-their-source-code-in-the-context-of-administrative-proceedings/](https://www.iplen.org/2018/03/30/copyright-protection-of-algorithms-does-not-prevent-the-disclosure-of-their-source-code-in-the-context-of-administrative-proceedings/)> accessed 7 February 2020.

¹⁹⁰ Henderson, *supra*, note 48, p 7.

¹⁹¹ C Ghezzi, “Software Qualities and Principles” in AB Tucker (eds), *Computer Science Handbook* (Chapman and Hall 2004) p 2377.

¹⁹² AE Bryson, “A Gradient Method for Optimizing Multi-Stage Allocation Processes” (1961) Harvard University Symposium on Digital Computers and their Applications (Vol 72).

¹⁹³ See note 175.

¹⁹⁴ Art 5 Directive on the legal protection of computer programs. See also Brkan (2017a), *supra*, note 133, p 5; Noto La Diega, *supra*, note 140, para 34.

determine the operation of the algorithm and, if its technical features allow for it, determine the factors that played a role in the algorithmic decision-making.¹⁹⁵

V. CONCLUSION: FEASIBILITY OF THE RIGHT TO EXPLANATION

Numerous technical and legal factors can shape the type and scope of explanation and information offered to the person affected by an algorithmic decision. These factors are usually intertwined and impact the final possibility of explanation of an algorithmic decision. In conclusion, therefore, we seek to bring the discussion on the interplay of those factors together and offer a final overview of different types of explanation in correlation with factors that impact the feasibility of the explanation. Based on the analysis in this article, we classify explanations into six different types summarised in Table 3: property checking (see Section IV.1.a.1), interpretable traces (see Section IV.1.b.1) local explanation (see Section IV.1.a.2), counter-factual faithfulness (see Section IV.1.a.2), decision decompilation (see Section IV.1.a.2), and providing arguments (see Section IV.1.b.2). Each of these types of explanations bears different characteristics and varies depending on the scope of explanation provided, the time it can be provided, the system it can be applied to and the way it is applied on the system.

As far as we are aware, from the computer science perspective, and more specifically XAI, there are no general impossibility results. An impossibility result is a formal proof showing that a given problem cannot be solved by a given computational model.¹⁹⁶ Therefore, XAI methods present scientific limitations for the type of explanation they can provide, depending on which information should be provided to the data subject and depending on the time when the explanation is provided. These technical constraints are depicted in the rows “Meaning of explanation” and “Time of issue”. The “Meaning of explanation” category refers to the information that can be provided when trying to explain an algorithmic decision, ranging from a simple verification of properties to more elaborate arguments that are used to support such a decision. The “Time of issue” category indicates when the explanations can be given: either in advance (*ex ante*) or after the algorithmic decision has been taken (*ex post*).¹⁹⁷ Furthermore, different types of systems (supervised/unsupervised) might be suitable to provide a certain type of information, as depicted in the column “Best fitted systems to explain”. The explanations further depend on different technical methods to provide such explanations (“Best fitted technical method”).

The “Feasibility to overcome legal obstacles” column depicts how difficult it is to effectively provide an explanation where legal obstacles are present, such as trade secrets or classified information over the algorithm or the source code. Generally speaking, external white-box methods necessarily need access to the source code or

¹⁹⁵ See Brkan (2017a), *supra*, note 133, p 6.

¹⁹⁶ For instance, M Minsky and S Papert, in *Perceptron: An Introduction to Computational Geometry* (first published MIT University Press 1969) p 188, showed that zero-hidden layer neural networks can only learn functions that are determined by linearly separable training sets of examples. Such a result paved the way to the design of current multi-layered neural networks.

¹⁹⁷ Wachter et al. *supra*, note 6, p 78, provide a classification that distinguishes between, on the one hand, the explanations of system functionality and specific decisions and, on the other hand, *ex ante* and *ex post* explanations.

Table 3: Feasibility of explanations of algorithmic decisions

Type of explanation	Meaning of explanation	Time of issue	Best fitted systems to explain	Best fitted technical method	Overall technical feasibility	Feasibility to overcome legal obstacles	Degree of correspondence with GDPR explanation
Property checking	Proving a given decision will always or never happen in a given situation	Ex ante	Unsupervised	External white-box methods	Moderate	Difficult	Low
Interpretable traces	Translating execution traces into natural language	Ex post	Supervised	Reflexive built on top methods	Difficult	Moderate	High
Local explanation	Identifying the main factors involved	Ex ante / Ex post	Unsupervised	External black-box methods	Easy	Easy	High
Counter-factual faithfulness	Evaluating the influence of different factors	Ex ante / Ex post	Unsupervised	External black-box methods	Easy	Easy	High
Decision decompilation	Approximating the logics involved	Ex ante / Ex post	Unsupervised	External black-box methods	Easy	Easy	High
Providing arguments	Providing pro and con arguments to support a decision	Ex post	Supervised / Unsupervised	Reflexive built within methods	Difficult	Moderate	Medium

specifications. More specifically, in the case of property checking, trade secrets and classified information can indeed constitute an obstacle to reach an explanation as this method requires access to the source code in order to allow for extraction of properties of both the algorithm and the program. Therefore, if the source code is subject to a trade secret, this might prevent the effective explanation of an algorithmic decision, unless one of the methods proposed earlier in this article is applied. As elaborated above in Section IV.2, observing, studying, disassembling or testing of a public product does not violate the trade secret.¹⁹⁸ Nevertheless, as also mentioned above, such reverse engineering does not always lead to satisfactory results.

Differently, in case of interpretable traces, the danger of infringing a trade secret or classified information is minor. Because a module is designed to provide a simple textual explanation of the execution traces left by the algorithm taking a decision, no direct access to the latter algorithm or the source code is necessary. However, this algorithm or this source code must be especially designed to produce those explanations, which is difficult to be legally enforced. On the contrary, explanations using external black-box methods (local explanation and counter-factual faithfulness¹⁹⁹) do not need a special design of the program. There is also no need to analyse the code of the program, since the program itself is probed or tested.

In cases of decision decompilation, which is a kind of reverse engineering of the decision-making system, access to the code or the algorithm is equally not necessary. Decision decompilation functions in a comparable fashion to local explanation, since both approaches rely on probing the system by giving it inputs and observing the outputs. However, differently from local explanation, decision decompilation does not focus on one main factor impacting a decision, but rather builds a simple model (such as a decision tree) that approximates the logic of the decision.

Finally, the explanation that provides arguments in favour and against a decision would not really face legal obstacles in terms of trade secrets or classified information. Since the program generates arguments that can be used as explanations, the question of these types of legal obstacles would not really arise. However, as in the case of interpretable traces, the system that is able to provide arguments needs a special design of the program, which is difficult to legally enforce.

The column “Degree of correspondence with GDPR explanation” aims to explain which type of explanation is best fitted to correspond to the scope of explanation required by the GDPR. As discussed in this article, the explanation given to the data subject should enable her to understand the reasons behind the decision, in order for her to be able to contest such decision.²⁰⁰ In practice, this means that the explanation should be able to be translated into natural language and be understandable to the average data subject without expertise in computer science.²⁰¹

When it comes to property checking, it is quite apparent that this approach does not correspond to the abovementioned requirements. Not only does this type of explanation

¹⁹⁸ Art 3(1)(b) Trade Secrets Directive.

¹⁹⁹ See more specifically Wachter et al, *supra*, note 6, p 871.

²⁰⁰ See *supra* note 6.

²⁰¹ The challenge of laymen understanding explanations is discussed also by Wachter et al, *supra*, note 6, pp 851, 861.

focus on providing the logic of the program rather than concrete reasons underpinning a decision, its result is also rather complex and understandable mostly to experts. Differently, interpretable traces can be easily understood by non-experts because the explanation provided by this method is rather simple. This method is suitable for decisions with clear pre-determined criteria on the basis of which a decision is made. For example, this method could explain a decision of a non-complex university admission system that uses the average high-school grade as a benchmark for admission. However, this approach cannot be used for more complex machine-learning systems. For example, it cannot explain how a credit rating system based on a neural network makes decisions because such network can merge in a single value both postal codes and birthdates, which is semantically difficult to interpret.

Furthermore, local explanations are suitable in terms of GDPR requirements. Local explanations probe the system to determine the correlations between input and output as well as to extract the main factors of the decision.²⁰² While this approach individualises the explanations, it can be quite time consuming if the system is probed for every decision of every individual. Counter-factual faithfulness functions in a similar fashion, with the difference being that it evaluates how a change in a particular factor influences the decision. It may be used, for instance, to assess the fairness of a decision, based on how each factor within a given specific input influences the decision. Therefore, this type of explanation is useful in terms of the GDPR.²⁰³

The method of decision compilation may be also useful because it approximates the whole logic of the decision, instead of focusing only on the main factors of the decision independently. Finally, providing arguments in favour and against a decision could, at least theoretically, be a helpful method to explain more complex algorithmic decisions, for example decisions made in public administration based on previous similar precedents. This method could also be used as a decision support in legal decision-making more generally, as it allows for the weighing of various arguments. However, the models providing for such explanations are rather difficult to build, which hinders their wider use in practice.²⁰⁴

The analysis in this article demonstrates that algorithmic decision-making, in particular the question of the right to explanation, necessitates a broader and clearer policy approach. Specifically, it must be clarified which type of explanation is required by the legal rules (GDPR), whether explanations need to be provided also for decisions based on non-personal data and whether those explanations can allow for flexibility depending on the model used.

Moreover, XAI methods also present different needs in terms of access to the systems to be explained. While the attention of the recent literature has been very much on “opening” black boxes, an alternative option, as Rudin suggests, could lie in building

²⁰² See *supra*, notes 105–108 and 116, for LIME, SPLIME, DLIME, LEAFAGE and LORE.

²⁰³ Wachter et al, *supra*, note 6, p 883.

²⁰⁴ See for example DM Gabbay et al, “Present and Future of Formal Argumentation” (2018) 7 Dagstuhl Manifest 65, 89–90. The authors highlight that many argumentation formalisms are abstract and still need important implementation work; moreover, there are no automated methods to design arguments that fit all kinds of application.

and using more inherently interpretable models.²⁰⁵ This idea follows the same line of reasoning as using reflexive explanation methods, which may produce arguments but require a special design of artificial agents. However, this idea depends not only on the general possibility of using certain models for certain types of decisions, but also on the willingness of industry and public administration that deploy algorithmic decision-making to choose one model rather than another.²⁰⁶ Indeed, it might seem disproportionate to (legally) oblige those stakeholders to use a specific model in order to increase the explainability of their decisions. Hopefully, further research in both law and computer science will help dissipate the current *fata morgana*s around the right to explanation, and allow for clearer insights into the reasons and causes behind algorithmic decision-making.

²⁰⁵ See *supra*, note 10.

²⁰⁶ Rudin, *supra*, note 10, pp 5–6, for example advances an argument that corporations benefit economically from using decision-making models that are protected as trade secrets.