University of Southampton

Faculty of Social Science

School of Law

A Legal and Technical Assessment of the Rights to Information and an Explanation

by

Alejandra Bringas Colmenarejo

ORCID ID 0000-0002-7968-9853

Thesis for the degree of Doctor of Philosophy

October 2025

University of Southampton

Abstract

Faculty of Social Science

School of Law

Doctor of Philosophy

A Legal and Technical Assessment of the Rights to Information and an Explanation

by

Alejandra Bringas Colmenarejo

This thesis explores why automated individual decision-making attracts the obligations of information and explanation, what these rights entail and how users of these automated processes may implement them, as referred to in the European General Data Protection Regulation. This thesis also explores why these data processing practices deserve special legal consideration when equivalent non-automated decisions are free from these onerous obligations. The problem at the heart of this thesis is the automation of everyday and high-consequence decision-making processes and the challenges and risks such technological transformation poses to the rights, freedoms, and legitimate interests of individuals. Particular relevance is given to the complexity and lack of neutrality that are introduced in decision-making through the automation of the process. This thesis provides an overview of the legal cases and disputes involving automated decisions reaching a court or a national Data Protection Authority (DPA) in a European member state. Likewise, this thesis provides both a doctrinal and a normative framework of the rights to information and an explanation that examine the legal foundations, rationale and intention of these obligations. This thesis also reflects on the distinction between explainability requirements for automated decision-making systems from a legal and a technical perspective and presents the most desirable properties technical and legal information and explanations about automated decision-making processes should attain according to both notions. Finally, this thesis examines the suitability of three concrete types of technical explainability methods to comply with the rights to information and an explanation. The expectations, reasoning, and rules exposed by a group of legal experts and practitioners regarding these explanations are complemented with an assessment of those same explanations using the doctrinal, normative frameworks and desirable properties proposed beforehand in the thesis.

Table of Content

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Alejandra Bringas Colmenarejo

Title of thesis: A legal and technical assessment of the rights to information and an explanation.

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   Bringas Colmenarejo A, State L and Comandé G,. How Should an Explanation Be? A Mapping of Technical and Legal Desiderata of Explanations for Machine Learning Models. *International Review Law, Computers and Technology* (2025). https://doi-org.soton.idm.oclc.org/10.1080/13600869.2025.2497633

   State, L., Bringas Colmenarejo, A., Beretta, A. *et al.* The Explanation Dialogues: An Expert Focus Study to Understand Requirements towards Explanations within the GDPR. *Artif Intell Law* (2025). https://doi-org.soton.idm.oclc.org/10.1007/s10506-024-09430-w

Signature:

Date: 2 October 2025

Acknowledgements

After four years researching, discussing, and getting paranoid about algorithms I could not help but ask an algorithm to write an acknowledgement paragraph about all this misfortune. Here is what it recommended me:

> This thesis was only made possible thanks to an automated decision-making system, whose efficiency and precision facilitated data analysis, modelling, and simulations throughout this work. I also acknowledge my own dedication and perseverance in turning this experience into a calamity.

Things are definitely much clearer now, thanks. Now let's get to work.

First and foremost, I want to thank my supervisors, Dr. Law and Prof. Uta Kolh for their insightful guidance, advice, and support during this journey. I would not have been able to untangle the intellectual conundrum that this thesis has presented without their invaluable expertise and mentorship. This thesis would not exist if it were not for NoBIAS and its members, who were good enough to offer me the opportunity to embark on this project. Special mention and recognition to Prof. Dr. Salvatore Ruggieri, Prof. Dr. Franco Turini, and Dr. Beretta who saw the value that my legal knowledge could have and decided to invest their time and advice in me. Many thanks to Prof. Giovanni Comandé, without whom I would not have seen it possible to achieve success and recognition in this area. Likewise, thanks to all my colleagues at NoBIAS who, despite possibly not understanding anything of what I say and causing them more than a headache, were there through thick and thin. You are all an amazing group of people I have the luck to find in my, for now, short research career.

And as none of this would have been possible without my people, here goes.

Gracias Mamá y Papá. Es imposible explicar lo increíble que es tener unos padres que te apoyan y te acompañan incondicionalmente aun cuando tomas las decisiones más raras y inesperadas. Muchas gracias, no sería quién soy sin vosotros.

Thanks to my life partner. These four years have been the most difficult, but also the most beautiful. I would never have done a PhD if it had not been for you and that link you sent me on a random day, so literally, this is for you, Pablo

Thank you girls.

Thank you, Andrea and Elena, for being with me even when we were thousand kilometres apart. Thank you, Mery, for (forcibly) picking me on the first day of Uni and not letting me go. Thanks for being the opposite side of the same coin, there is no Alex without Mery. Thank you, Andrea and Virag, for the cheese and wine throughout Europe, the headquarters in Brussels, the unexpected shared love for fantasy books and the incomprehensible intellectual drive we have. Thank you, Almudena and Arancha, for not acting surprised when I tell you about my next misadventure and for sticking with me no matter what happens or where I went.

Thank you all for conversations than grounded me to earth, the video calls that made me felt less alone and the meeting times every time I came home. We have seen each other grow and I could not be prouder of what we have achieved.

Definitions and Abbreviations

Artificial Intelligence (AI)

Automated decision-making processing (ADM)

DIverse Counterfactual Explanations (DiCE)

European General Data Protection Regulation (GDPR)

European Union (EU)

Explainable artificial intelligence (XAI)

LOcal Rule-based Explanations (LORE)

Machine Learning (ML)

Post-hoc explainability methods (XAI methods)

SHapley Additive exPlanations (SHAP)

Chapter 1: Introduction

1.1. Background

This thesis explores why automated individual decision-making, including profiling, attracts the obligations of information and explanation, what these rights entail and how users of these automated processes may implement them. This thesis also explores why these data processing practices deserve special legal consideration when equivalent non-automated decisions are free from these onerous obligations.

The scope of this thesis is precisely delimited to Automated Decision-Making (ADM) processing, irrespective of the specific technology employed. Given that the automation of decision-making predominantly utilizes algorithms—encompassing Artificial Intelligence (AI), Machine Learning (ML), and other computational methods—the resultant conceptual and normative frameworks herein will necessarily address the inherent challenges and risks introduced by algorithmic integration into the decision-making lifecycle. Crucially, this analysis is consistently framed through the lens of algorithmic deployment within ADM systems. In this context, ADM signifies the automation of the decision-making process itself, specifically the automated processing of an individual's data for the explicit purpose of rendering a decision.

A fundamental constraint of this investigation is its exclusive focus on the automation of decision-making processes. Consequently, the deployment of algorithms to assist human agents in decision-making is expressly excluded. The thesis's ambit is thus restricted to fully automated processes and decisions characterized by the absence of meaningful human involvement.

The central focus of this research pertains to the rights to information and an explanation concerning automated individual decision-making, as articulated within the European General Data Protection Regulation (hereafter GDPR)[1]. The inquiry seeks to critically examine the scope, relevance, and legal rationale of these rights, evaluating their adequacy in mitigating the complex challenges posed by ADM systems. To this

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection) 2016 (OJ L 1191/1).

end, the analytical depth of this thesis is purposefully limited to the GDPR's provisions, adopting a high-level approach to ADM and the associated information and explanation rights. The interpretation of the GDPR's provisions has garnered substantial attention in national and Europea courts, among Data Protection Authorities, and within the scholars and the society as a whole. In recent years, such attention has been reflected in the increased case-law and jurisprudence on the matter, both at the national and European level.

Special consideration must be accorded to the recently promulgated European Artificial Intelligence Act (AI Act). This new European regulatory instrument establishes the most extensive explainability and justificability requirements to date for AI systems. These requirements manifest as obligations concerning literacy, transparency, information provision, and human oversight, thereby functionally establishing a quasi-general obligation for interpretable and explainable AI. The AI Act mandates literacy requirements for all AI systems without exception and imposes specific explainability duties on both high-risk systems and General-Purpose AI Models (GPAI). Particularly pertinent to this research is the AI Act's inclusion of a right to an explanation for decisions predicated upon the output of specific classes of high-risk AI system. The statutory language and underlying principles of this provision strongly resemble the GDPR's stipulations regarding information and explanation in the context of algorithmic ADM.

It is important to acknowledge that the European Parliament initiated the legislative process for the AI Act subsequent to the commencement of this doctoral research. The initial draft proposal did not incorporate an independent right to an explanation; by the time this right was introduced in compromise amendments, the research was already in an advanced stage. Notwithstanding the academic resonance generated by this proposed new right for high-risk AI systems, the scope of this thesis was consciously delimited to the GDPR. This decision stemmed from concerns regarding the potential uncertainty of the new right's development in the immediate future and the consequent impact such an unknown variable might exert upon the trajectory and conclusions of the thesis.

Hence, the own relevancy of this thesis can be prejudged to be outdated after the enactment of the new European AI Act. To my understanding, however, the rights to information and an explanation as referred to in the GDPR remain a key driven force for the protection of individuals' rights, freedoms and liberties. Assessing the scope, relevance, and legal rationale of these rights as well as evaluating their adequacy in mitigating the complex challenges posed by ADM systems set a solid foundation for the future study and assessment of the AI Act's right to explanation development, along with the implementation of its interpretability and explainability requirement.

Without going into more detail than is necessary for a brief introduction to this thesis, I introduce, in the following section, the problem at the heart of this thesis; the automation of everyday and high-consequence decision-making processes, and put forward the major challenges it introduces in the process and the effects it provokes in individuals' lives, rights, and freedoms.

1.1.1. Algorithms in our everyday

We can think about multiple and diverse activities that we do in our daily lives without offering them much thinking. Examples include going to the bank, grocery shopping, checking the news, visiting the doctor, or requesting an ill absence from our jobs. Some of them could be considered common or ordinary practice, while others are accepted to have a direct bearing on our lives. We might have completely normalised how we browse for popular coffee shops near our hotel during our holidays. However, the recommendations obtained are not accidental; they impact our commercial habits, the coffee's clientele influx, and their reputation. Part of the relevance of these everyday decisions resides in the power relations and practices that are put in place and shape the society and the invisible practices that could occur behind the general public's view, leaving them unobjective[2]. Following the previous example, we might overlook that some establishments pay to be on the recommended list or that foreigners do not use the same applications we use, possibly biasing the results and review system. Again, our everyday decisions are not homogeneous but differ in the context and time they

[2] Michele Willson, 'Algorithms (and the) Everyday' (2017) 20 Information, Communication & Society 137 see quote n.2.

take place and in their decision-maker and recipient, logically affecting people in different and diverse ways[3]. We might consider the decision of where to have lunch to be superfluous. Still, other decisions of our everyday are usually made by professionals with several years of experience and normally require concrete and well-accepted reasons, like the concession of credit to buy a house or the order of urgency on the waiting list for a kidney transplant. Even if we as particular individuals do not ask for a loan every other day nor need to request medical coverage for a specific treatment every Monday, these are decisions affecting millions of individuals daily and that, up to a certain level, are considered mundane as they are necessary to achieve very common life goals. These are all decisions that are made on a day-to-day basis, even though their relevance may be completely different. In essence, we can think of innumerable decisions that make up our everyday life in the context of normalised, everyday practices

> They are the seemingly mundane or banal, recurrent and multiple activities and routines that we all engage with and that shape the form and flow of our individual and social lives in space and time. These activities and routines are replicated in countless ways by many people on a daily or regular basis. Through this process, practices become normalised or naturalised, usually enacted with minimal thought and often rendered invisible or in the background (or at the very least as largely unquestioned)[4].

As an attempt to justify the potential disparity that could arise from the decisions associated with everyday practices, there is a long-lasting trend to turn the decision-making process to empirical knowledge and reasoning while ensuring the respect of some fundamental principles such as accuracy, effectiveness, and impartiality[5].

---

[3] Jim Johnson', 'Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer'.
[4] Willson (n 2) p.138.
[5] Rabab Naqvi and others, 'The Nexus Between Big Data and Decision-Making: A Study of Big Data Techniques and Technologies' in Aboul Ella Hassanien and others (eds), *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)* (Springer International Publishing 2021); Nada Elgendy and Ahmed Elragal, 'Big Data Analytics in Support of the Decision Making Process' (2016) 100 Procedia Computer Science 1071; Milan Dordevic, 'Council Post: How Artificial Intelligence Can Improve Organizational Decision Making' (*Forbes*) <https://www.forbes.com/sites/forbestechcouncil/2022/08/23/how-artificial-intelligence-can-improve-organizational-decision-making/> accessed 10 December 2023; 'Decision Management Software &

Digitalising and automating private and public sector processes through algorithms and data-driven systems reflect in part the eternal desire to rid decisions of human partiality and prejudice and endow them with mathematical accuracy and neutrality[6]. Since algorithms are playing an increasing role in our everyday practices, the delegation of functions traditionally performed by individuals has been equally normalised, even though this delegation has brought with it a reconfiguration and reframing of social (everyday) dynamics and relationships[7]. When we ask the receptionist of our hotel for a good restaurant, we take the risk of being offered a recommendation entirely based on their personal experience or an agreement between the hotel and the restaurant around the corner; we possibly give it minimal thought or assume that it could happen. However, when we rely on recommendation algorithms, possible biases, be they caused by unintentional prejudices or business-aligned intentional practices, could create a feedback loop where previous recommendations would be repeated over time and possibly institutionalised, changing how consumers act and establishments function.

By automating the decision-making process of the everyday, we are also subjecting social relationships to the procedural logics of computational systems. More concretely, we are introducing algorithms in a myriad of spheres of our everyday life[8], enabling them to decide which information is relevant and required for our participation and inclusion in social dynamics[9],

Solutions | IBM' <https://www.ibm.com/decision-management> accessed 10 December 2023; Marc Schmitt, 'Automated Machine Learning: AI-Driven Decision Making in Business Analytics' (2023) 18 Intelligent Systems with Applications 200188; 'Decision Automation' (*DevX*) <https://www.devx.com/terms/decision-automation/> accessed 10 December 2023.

[6] Stefanie Hänold, 'Profiling and Automated Decision-Making: Legal Implications and Shortcomings', *Robotics, AI and the Future of Law* (Springer Singapore 2018) p.124 <http://link.springer.com/10.1007/978-981-13-2874-9>.

[7] Willson (n 2) p.146.

[8] Tarleton Gillespie, 'The Relevance of Algorithms' in Tarleton Gillespie, Pablo J Boczkowski and Kirsten A Foot (eds), *Media Technologies* (The MIT Press 2014) p.167 <https://academic.oup.com/mit-press-scholarship-online/book/14976/chapter/169333383> accessed 24 November 2023.

[9] ibid p.168.

> The algorithmic assessment of information, then, represents a particular *knowledge logic*, one built on specific presumptions about what knowledge is and how one should identify its most relevant components'[10].

As in my previous example, the recommendation system that we use to choose where to buy coffee is designed to offer us what is presumed to be the best coffee place in town. Nonetheless, this knowledge is based on pre-assumptions about the relevant and useful indicators for such information, factors such as individuals' review, establishment localisation, distance from our devices, or how crowded it is. However, the recommendation might also be based on presumptions about ourselves and our preferences, such as the possibility of going with our pets, our age and main interests, or our similarity with other presumed customers. Whether not intentionally, these presumed indicators of relevant knowledge have positive and negative impacts that might not even be considered during algorithm design[11]. If the algorithm fulfils its intended function, the way it does it and how it affects others might be overlooked unless someone raises awareness about unexpected results or unwanted effects. The problem escalates due to the technical difficulty in identifying the logic that leads to the final recommendation and the undesired assumptions behind it[12].

When algorithms are used to support us and others in making decisions that do not significantly impact our legitimate interests, rights, and freedoms, we might not be very concerned about the repercussions and effects they entail for our lives. However, when algorithms are used to automate decision-making without any meaningful human involvement to grant a loan, manage our work conditions, or grant us access to specific healthcare services, we might find ourselves more concerned and interested in the knowledge and logic behind the automated decision-making processing -hereafter ADM- and the particular decision affecting us. Particularly, we might want to be able to

[10] ibid.
[11] Lee Rainie and Janna Anderson, 'Theme 7: The Need Grows for Algorithmic Literacy, Transparency and Oversight' (*Pew Research Center: Internet, Science & Tech*, 8 February 2017) <https://www.pewresearch.org/internet/2017/02/08/theme-7-the-need-grows-for-algorithmic-literacy-transparency-and-oversight/> accessed 12 December 2023.
[12] ibid.

confirm that the decision was made fairly and lawfully and if we deemed it appropriate to challenge it, for which we will actually need to understand it.

Thinking about it, these expectations are not that far away from the expectations we have when a human decision-maker makes a decision. When the usual thing to do was to go to the nearest bank office and apply for a loan, we were aware that the reasons for granting or refusing it depended almost entirely on the bank's discretion. Even so, we were certain, or at least had the feeling, that we would be able to ask the reasons why this or that decision had been made and correct the information that could have been incorrect or incomplete. When the decision is made by an automated system, instead of a human, we no longer have the intuition but the right to obtain information and an explanation about the decision-making process and the particular decision. Something must have changed in decision-making for this change to take place, and that, together with what these rights mean for the individuals affected by an automated decision, is precisely what this thesis intends to examine.

1.1.2. From algorithms to profiling and automated decision-making

There is no common definition for an algorithm; however, Harold Stones offered in 1971 a still widely used definition: 'An algorithm is a set of rules that precisely define a sequence of operations'[13]. This definition is so generic and broad that it could describe a cooking recipe for roasted chicken as well as the deep neural network algorithms which mimic the human brain and are used, for example, to differentiate dialects of a language[14]. The definition provided by Stones, however, fundamentally represents algorithms as a list of instructions used to perform a task or solve a problem based on the understanding of available alternatives. In computer science and statistics,

[13] Harold S Stone, *Introduction to Computer Organization and Data Structures* (McGraw-Hill, Inc 1971); See also 'What Is an "Algorithm"? It Depends Whom You Ask' (*MIT Technology Review*) <https://www.technologyreview.com/2021/02/26/1020007/what-is-an-algorithm/> accessed 4 December 2023.

[14] Bernard Marr, '10 Amazing Examples Of How Deep Learning AI Is Used In Practice?' (*Forbes*) <https://www.forbes.com/sites/bernardmarr/2018/08/20/10-amazing-examples-of-how-deep-learning-ai-is-used-in-practice/> accessed 12 December 2023; "Neural networks are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way the biological neurons signal to one another" 'What Are Neural Networks? | IBM' <https://www.ibm.com/topics/neural-networks> accessed 5 December 2023.

algorithms are the instructions coded in different programming languages before being compiled into a machine-readable binary sequence that a computer executes to perform calculations, process data, automate reasoning or make decisions -among the many other actions algorithms can be used for-[15].

In other words, algorithms determine the concrete steps and processes a computer needs to follow or employ to complete a task or solve a problem[16]. Encoding these instructions is accomplished through algorithmic languages -usually part of a programming language[17]- specially designed 'to express mathematical or symbolic computations and thus to express algebraic operations in a notation, reminiscent of logic and related to mathematics'[18]. To understand an algorithm and assess its capabilities to perform the designated task, expert knowledge of both the algorithmic -mathematical- and programming -computational- language used to build the algorithm is thus required. This level of mathematical and computational knowledge is, nonetheless, rare to be in the possession of the general public or the average individual.

Some of the common tasks algorithms perform include prioritising, associating, classifying and filtering information[19]. Particularly, machine learning algorithms -hereafter ML- are designed to discover correlations and seek patterns through statistical inferences, measurements and analytics that would otherwise be difficult to identify[20]. Among the many tasks that algorithms can perform, they may also be employed to profile individuals for various purposes, such as personalised advertising[21],

---

[15] 'What Is an "Algorithm"? It Depends Whom You Ask' (n 13); Anton Vedder and Laurens Naudts, 'Accountability for the Use of Algorithms in a Big Data Environment' (2017) 31 International Review of Law, Computers & Technology 206.
[16] Willson (n 2).
[17] 'Algorithmic Language' (*Oxford Reference*) <https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095402323> accessed 5 December 2023.
[18] Vedder and Naudts (n 15) p.208.
[19] Nicholas Diakopoulos, 'Accountability in Algorithmic Decision Making' (2016) 59 Communications of the ACM 56.
[20] Usama Fayyad, Gregory Piatetsky-Shapiro and Padhraic Smyth, 'From Data Mining to Knowledge Discovery in Databases' (1996) 17 AI Magazine <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>; Mark MacCarthy, 'Fairness in Algorithmic Decision-Making' [2019] The Brooking Institution's Artificial Intelligence and Emerging Technologies (AIET) 14.
[21] Abid Haleem and others, 'Artificial Intelligence (AI) Applications for Marketing: A Literature-Based Study' (2022) 3 International Journal of Intelligent Networks 119.

precision medicine[22] or political microtargeting[23]. Algorithms are used in many contexts to decide on an individual's eligibility for a benefit, penalty, or service and are meant to improve the validity of the decision-making processes. Profiling, in this sense, involves the analysis of the aspects of an individual's personality, behaviour, interests and habits to make predictions or decisions about them by automated means without any human involvement[24][25]. Profiling is used to analyse or predict the conduct of a person, e.g., his or her performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements. It requires the processing and evaluation of their personal data.

An automated decision is a decision made only by technological means[26]. Quite often, if not always, automated decisions made by algorithmic systems include a prior construction and evaluation of profiles, where the final decision depends on the result of profiling. However, automated decisions can be made without prior profiling, just as profiling can be executed without the objective of making a decision, although both cases are rare in practice[27]. Think, for example, of monitoring an employee's productivity in her workplace, which results in a profile of her high work productivity. Initially, that profiling served no other purpose than to have control over her workflow and her ability to meet deadlines. The profile might seem innocuous and independent of any automated decision. Still, from a closer inspection, we can see that there was a decision not to act, as her high levels of productivity made it unnecessary to pass the case to the senior management or human resources. This example is rather simplistic, but it clearly shows how there are few, if any, examples in which someone will want to

[22] Jia Xu and others, 'Translating Cancer Genomics into Precision Medicine with Artificial Intelligence: Applications, Challenges and Future Perspectives' (2019) 138 Human Genetics 109.
[23] Janice Richardson, Normann Witzleb and Moira Paterson, 'Political Micro-Targeting in an Era of Big Data Analytics: An Overview of the Regulatory Issue', *Big Data, Political Campaigning and the Law* (Routledge 2019).
[24] Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008).
[25] ibid; Jean-Marc Dinant and others, 'Application of Convention 108 to the Profiling Mechanism' 35; Vedder and Naudts (n 15).
[26] Hänold (n 6).
[27] The possible combinations of algorithmic profiling and automated decisions will be discussed in more detail in Chapter 3, section 3.2.3.given that the mere possibility that these practices can be run independently and without the intention of feeding into each other has led to extensive legal debate. Just as preliminary note, there are few data processing techniques that use the personal data of individuals but do not use actual profiling, for example the Benford's Law.

analyse the aspects of an individual's personality, behaviour, interests and habits without then wanting to make a decision based on that knowledge, whether it is an active or passive decision.

The increasing presence of ADM -combined or not with profiling- in a wide number and diversity of social and economic situations in our daily lives[28] has been encouraged by the alleged increase in accuracy and effectiveness of the decision-making process. However, their inclusion has also prompted the discussion as to when these systems and processes are more or less acceptable on the grounds that their repercussions might be more or less significant and their effects more or less easily controlled.

1.1.3. The rights and wrongs of algorithmic automated decision-making

ADM relies on data-driven systems, usually ML algorithms, which are a building block of artificial intelligence -hereafter AI- 'that allows computer systems to learn directly from examples, data, and experience'[29]. AI, in turn, 'is a technology that enables computers and machines to simulate human learning, comprehension, problem-solving, decision-making, creativity and autonomy'[30]. Algorithms have the potential to accurately[31] perform prediction tasks regarding the likelihood of uncertain phenomena happening in the future, the level of risk of particular outcomes, and the weight of the determining

[28] Willson (n 2).

[29] -Doran Dominique Hogan, 'Computer Says "No": Automation, Algorithms and Artificial Intelligence in Government Decision-Making' 13 The Judicial Review: Selected Conference Papers: Journal of the Judicial Commission of New South Wales 345, p.23.

[30] Cole Stryker and Eda Kavlakoglu, 'What Is Artificial Intelligence (AI)? | IBM' (*IBM*, 9 August 2024) <https://www.ibm.com/think/topics/artificial-intelligence> accessed 17 January 2025.

[31] The future is uncertain. Hence, the capability of algorithms to "accurately" predict or detect a phenomenon to happen in the future will never be exact. Algorithms' accuracy varies and its highly contextual. For instance, it is not the same an algorithm that can predict the likelihood of an individual to repay a credit with an accuracy of 55 percent, than an algorithm that can predict the probability of an individual to suffer a rare disease with an accuracy of 55 percent. For both cases, the accuracy of the algorithm is not better than the flip of a coin, however, if the rare disease is currently almost impossible to predict for human doctors or the symptoms do not appear until the patient is gravely ill, that extra 5 percent could be worth the risk. The same will, likely, not be say for an algorithm used in the finance sector. Literature in algorithms' accuracy is vast, I point out here just a couple of examples Marina Sokolova, Nathalie Japkowicz and Stan Szpakowicz, 'Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation' in Abdul Sattar and Byeong-ho Kang (eds), *AI 2006: Advances in Artificial Intelligence*, vol 4304 (Springer Berlin Heidelberg 2006) <http://link.springer.com/10.1007/11941439_114> accessed 2 December 2024; Pierre Baldi and others, 'Assessing the Accuracy of Prediction Algorithms for Classification: An Overview' (2000) 16 Bioinformatics 412.

factors for those outcomes[32]. Relying on algorithmic predictive capabilities, ADM renders a decision based on correlations. Those correlations are, in turn, based on inferred rules and hidden patterns that, for the general human-cognitive analysis, remain unknown or inexplicable due to the complexity of the data sets[33].
Algorithms offer knowledge about correlations, that is, relationships, in the data sets that can be used to make decisions based on the predicted future. Algorithms can arguably select attributes to predict an unknown function more comprehensively and less subjectively than a human brain would do. Hence, proponents of predictive algorithms argue that by using this technique, companies and institutions can embrace evidence-based decision-making, which increases their accuracy, efficiency, and unbianess[34].

Detractors of ADM argue that algorithms are deterministic, opaque, and uncertain[35]. The theory of statistical discrimination exposes how certain groups can be over- or underrepresented due to insufficient or erroneous data[36], leading to discriminatory outcomes that might not be explicit at first sight. Likewise, the theory of computational injustice highlights how algorithms codify social inequalities and unfairness through hidden patterns and what can be seen as reasonable correlations[37]. For instance, since some minority groups and communities have historically suffered discriminatory practices, algorithms will consider -reasonable- correlations to the social biases behind those practices and, by using those correlations to provide an outcome, reinforce and restate discriminatory patterns and correlations. Without conscious preprocessing of the datasets or proper constraints in the algorithm logic, algorithms

[32] Irina Pencheva, Marc Esteve and Slava Jankin Mikhaylov, 'Big Data and AI – A Transformational Shift for Government: So, What next for Research?' (2020) 35 Public Policy and Administration 24.
[33] Joel Tito, 'Destination Unknown: Exploring the Impact of Artificial Intelligence on Government' [2017] Centre for Public Impact 7; Daniel B Neill, 'Using Artificial Intelligence to Improve Hospital Inpatient Care' (2013) 28 IEEE Intelligent Systems 92.
[34] Joel Nantais, 'Predictive Analytics in Government Decisions' (*Medium*, 31 July 2019) <https://towardsdatascience.com/predictive-analytics-in-government-decisions-8128ba019a77> accessed 6 December 2022 See also Chapter 1 section 2. .
[35] Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 2053951715622512.
[36] Bernhard Rieder, 'Big Data and the Paradox of Diversity' (2016) 2 Digital Culture & Society 39, p.50.
[37] Nick Thieme, 'We Are Hard-Coding Injustices for Generations to Come' (*Undark Magazine*, 20 February 2018) <https://undark.org/2018/02/20/ai-watchdog-computational-justice/> accessed 6 December 2022.

will, for example, perpetuate higher levels of police detection and jury conviction of black individuals compared to white individuals only on the basis of their race[38].

Additionally, algorithmic processing tends to be neither transparent nor interpretable as the logic between the model's input and output is obscured and difficult, if not impossible, for a human to understand[39]. To that effect, 'algorithm high dimensionality[40] of data, complex code and changeable decision-making logic'[41] creates an understandability problem, which usually makes it either impossible or difficult to explain the system's output, even where this information is critical for an individual's life choices.

For these reasons, there is an increasing awareness towards algorithmic tools whose performance is particularly unintelligible, untraceable, and, by extension, untrustworthy[42] as it will entail deciding upon unjustifiable, illegitimate or non-explainable reasons. Yet, the higher the system's accuracy, the better it is in its performance, and the more easily it solves the problem for which it has been deployed. This presents a dilemma concerning the trade-off between algorithms' performance and interpretability[43].

This thesis interrogates the law and legitimacy of ADM with no meaningful human involvement, including those decisions based on the profiling of an individual. Such

---

[38] Julia Angwin and others, 'Machine Bias - There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.' (*ProPublica*, 23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

[39] See the distinction between interpretable and non-interpretable algorithmic systems in Chapter 1 section 2.1.

[40] Algorithmic high-dimensionality refers to the large number of features or attributes the algorithm's dataset possess.

[41] Brent Daniel Mittelstadt and others, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3 Big Data & Society 205395171667967, p.6 referring to ; Burrell (n 35).

[42] Alejandro Barredo Arrieta and others, 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI' (2020) 58 Information Fusion 82; Luciano Floridi and others, 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 Minds and Machines 689; Davinder Kaur and others, 'Trustworthy Artificial Intelligence: A Review' (2023) 55 ACM Computing Surveys 1; Independent High-Level Expert Group On Artificial Intelligence Set Up By The European Commission, 'Assessment List for Trustworthy Artificial Intelligence (ALTAI)' (European Commission 2020) DOI:10.2759/002360.

[43] Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (arXiv, 21 September 2019) <http://arxiv.org/abs/1811.10154> accessed 20 September 2022.

ADM has become of increasing significance in many social, economic, and political domains, and algorithms have shaped the decision-making process they are part of according to its intrinsic characteristics of complexity, inscrutability, and lack of neutrality.

1.2. The Research Questions to Be Addressed

The main research question of this thesis is: Can the right to information and to explanation applicable to automated decisions affecting individuals adequately address the problems arising from their inscrutability and lack of neutrality?

From this umbrella question, these research sub-questions emerge:

(1) Why does society ask for explanations and information about automated individual decision-making processing? Moreover, what is the legal rationale behind the rights to information and an explanation about automated decisions?
(2) What objectives are to be achieved, and what is exactly to be provided through those explanations and information?
(3) What is the potential development of the exercise of the rights to information and an explanation taking into account the state-of-the-art of post-hoc explainability methods?

These are the three preliminary ideas that, although inevitably incomplete, guided the research and analysis carried out to answer these questions. First, the power imbalance between the actors using ADM and the individuals affected by them is key to understanding why data protection demands information and explanation and how complete and interpretable those explanations and information should be to allow individuals to assert their rights as the affected party. Secondly, information and explanations are prerequisites to address the challenges and potentials that arise from ADM and to understand and contest, if deemed appropriate, the legality of automated decisions, whether on the grounds of discrimination, arbitrariness, or other unlawfulness. Finally, as data controllers provide information and explanations about ADM, the transparency and interpretability of their algorithmic systems will be sought and prioritised in both design and development.

1.3. Statement Of Originality and Significance

The rights to information and an explanation for automated individual decision-making, in Articles 13(2)(h), 14(2)(g), 15(1)(h), and 22 (3) of the GDPR have been the object of an intensive debate since it adoption. The imprecise wording of the right to not be subject to automated individual decision-making, including profiling, established in Article 22 of the GDPR, has given rise to multiple and disparate interpretations about its meaning, enforceability and effectiveness[44]. Discussions dwell around the nature and content of the provision as a prohibition to be subjected to an automated decision or as a right to object to ADM[45], as well as on the derogations and safeguards associated with the right that affect its enforcement by individuals. The discussion expanded on the right to access and information about the existence of ADM and the right to an explanation about an automated decision as referred to in Articles 13(2)(h) and 14(2)(g), 15(1)(h) and Article 22 (3), respectively. Scholars and practitioners debated on the legal basis that buttresses the existence of such rights[46] and the importance of temporality in the effective enjoyment of the rights to information and an explanation about an automated

[44] Reuben Binns and Michael Veale, 'Is That Your Final Decision? Multi-Stage Profiling, Selective Effects, and Article 22 of the GDPR' (2021) 00 International Data Privacy Law; Diana Sancho, 'Automated Decision-Making under Article 22 GDPR: Towards a More Substantial Regime for Solely Automated Decision-Making' in Martin Ebers and Susana Navas (eds), *Algorithms and Law* (1st edn, Cambridge University Press 2020) <https://www.cambridge.org/core/product/identifier/9781108347846%23CN-bp-4/type/book_part> accessed 28 November 2024; Antoni Roig, 'Safeguards for the Right Not to Be Subject to a Decision Based Solely on Automated Processing (Article 22 GDPR)' (2017) 8; Luca Tosoni, 'The Right to Object to Automated Individual Decisions: Resolving the Ambiguity of Article 22(1) of the General Data Protection Regulation' (2021) 11 International Data Privacy Law 145; Maja Brkan, 'AI-Supported Decision-Making under the General Data Protection Regulation', *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law* (ACM 2017) <https://dl.acm.org/doi/10.1145/3086512.3086513> accessed 4 January 2025.

[45] Sandra Wachter, Brent Mittelstadt and Luciano Floridi, '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; Isak Mendoza and Lee A Bygrave, 'The Right Not to Be Subject to Automated Decisions Based on Profiling' in Tatiana-Eleni Synodinou and others (eds), *EU Internet Law* (Springer International Publishing 2017) <http://link.springer.com/10.1007/978-3-319-64955-9_4> accessed 28 November 2024; Lee A Bygrave, 'Article 22. Automated Individual Decision-Making, Including Profiling' in Christopher Kuner, Lee A Bygrave and Christopher Docksey (eds), *The EU General Data Protection Regulation (GDPR)* (Oxford University PressNew York 2020) <https://academic.oup.com/book/41324/chapter/352297995> accessed 4 January 2025; F Thouvenin, A Früh and S Henseler, 'Article 22 GDPR on Automated Individual Decision-Making: Prohibition or Data Subject Right?' (2022) 8 European Data Protection Law Review 183.

[46] Bryce Goodman and Seth Flaxman, 'European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"' (2017) 38 AI Magazine 50; Wachter, Mittelstadt and Floridi (n 45); Andrew D Selbst and Julia Powles, 'Meaningful Information and the Right to Explanation' (2017) 7 International Data Privacy Law 233.

decision[47] addressing the different alternatives in which these rights can be enjoyed or demanded and the legal scenarios and challenges that each one creates.

This thesis contributes to the debate by delimiting a framework of temporality, application, and type of information required by the right to information and by the right to an explanation and proposing a spectrum of compliance for both rights with a minimum and maximum threshold. This original contribution is offered in **Chapter 3** and used in **Chapter 6** to assess the compliance of three concrete types of post-hoc explainability methods, i.e., SHapley Additive exPlanations (SHAP), DIverse Counterfactual Explanations (DiCE), and LOcal Rule-based Explanations (LORE). The framework and spectrum of compliance, as well as its practical application to explainability methods used in real scenarios, is a significant novel contribution.

Besides the scholarly debate over the rights to information and an explanation, in recent years, the rights have also been subject to profound examination by EU member states' Data Protection Authorities and Courts. Disputes over the interpretation, extension and exercise of the rights have reached the European Court of Justice alike. This thesis provides, in **Chapter 2**, an analysis of this case law. The thesis solely focuses on cases whose facts involve ADM as referred to in Article 22 of the GDPR and whose legal disputes concern the impacts ADM has on individuals' rights and freedoms. Cases which does not fall within the scope of Article 22 either because the automated decision does not have legal or similarly significant effects on the rights, freedoms and legitimate interests of data subjects or because the profiling does not constitute an automated decision were excluded from the analysis.

In particular, this thesis contextualises and connects the black-box problem of ADM with the existent case law, highlighting the concrete normative challenges and potentials brought by algorithms to high-consequence decision-making processes and the concrete impacts they have on the lives of individuals. Although it is possible to find overviews of case law related to Article 22 of the GDPR, for instance, in Barros Vale &

[47] Wachter, Mittelstadt and Floridi (n 45); Mendoza and Bygrave (n 45); Selbst and Powles (n 46); Gianclaudio Malgieri and Giovanni Comandé, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 243.

Zanfir-Fortuna[48] or the GDPRhub[49], as well as an analysis of the concrete impacts ADM have for individuals in Allen Qc & Masters[50] or The Directorate General for Parliamentary Research Services[51], an up-to-date combination of both approaches is a novel contribution to this area of research.

The originality of this thesis is also brought forward through two articles that combine a legal and technical approach to the rights to information and an explanation. **Chapter 5** and **Chapter 6** are partially based on both articles, although the content included in the thesis has been revisited and adapted to the scope of this thesis.

In *How should an explanation be? A mapping of technical and legal desiderata of explanations for machine learning models* my colleagues and I propose a comparison between the notions of legal and technical explainability for ADM systems and identify the desirable properties explanations about them shall have. We also provide an overview of the current legal requirements for explainability and interpretability for ADM established in EU law. To the best of our knowledge, we were the first to do these things. In this thesis, I further these contributions in **Chapter 5Chapter 5: A Legal and Technical Approach to Explainability** by specifically relating them to Articles 13(2)(h) and 14(2)(g), 15(1)(h) and Article 22 (3) of the GDPR. Particularly, the connection of each desirable property with the provisions and wording of the GDPR in concrete and clear terms is an addiction to the content of the academic article that cannot be found in its published version.

Finally, in *The explanation dialogues: an expert focus study to understand requirements towards explanations within the GDPR*[52], we summarise the reflections, perspectives,

---

[48] Sebastiao Barros Vale and Gabriela Zanfir-Fortuna, 'Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities' (Future of Privacy Forum 2022) <https://fpf.org/blog/fpf-report-automated-decision-making-under-the-gdpr-a-comprehensive-case-law-analysis/>.

[49] 'Category: Article 22 GDPR' (*GDPRhub*) <https://gdprhub.eu/index.php?title=Category:Article_22_GDPR>.

[50] Robin Allen Allen Qc and Dee Masters, 'Technology Managing People – the Legal Implications' (AI Law Consultancy).

[51] European Parliament. Directorate General for Parliamentary Research Services., *Understanding Algorithmic Decision-Making: Opportunities and Challenges*. (Publications Office 2019) <https://data.europa.eu/doi/10.2861/536131> accessed 28 November 2024.

[52] Laura State and others, 'The Explanation Dialogues: An Expert Focus Study to Understand Requirements towards Explanations within the GDPR' [2025] Artificial Intelligence and Law.

and opinions of legal experts on the rights to information and an explanation for automated decisions towards four types of explainability methods. I expand the academic paper by offering my assessment of the four concrete types of explainability methods presented in the academic paper on the basis of the framework and spectrum of compliance and the technical and legal desirable properties for the rights to information and an explanation as referred to in the GDPR.

Therefore, the main significance of this thesis is the creation of a hybrid body of work that addresses the rights to information and an explanation from the point of view of the technological challenges of automated decision-making systems and how they can affect the effective exercise and enjoyment of the rights. A hybrid approach to the rights to information and an explanation is needed because the legal and technical notions of information and explanations about ADM are not alike. Hence, when the law seeks information and explanations, it may require completely different things than what the technical approach to these rights is aimed or designed to offer. A comprehensive legal assessment of the rights -conceptual, doctrinal, and normative- is necessary -and thus performed in the remainder of the thesis- to understand the rationale, scope, and motivation behind the rights and so assess the potential compliance of the technical approach. Likewise, examining the technical methods used to make algorithms understandable to humans and to provide information and explanations about their internal workings is also necessary to assess the shortcomings they might have when used to comply with legal requirements.

1.4. The Scope

This thesis is part of NoBIAS, a Marie Skłodowska-Curie Innovative Training Network[53] focused on developing 'novel methods for AI-based decision-making without bias by taking into account ethical and legal consideration in the design of technical solutions'[54]. Being part of this project, the scope of my thesis has been influenced by NoBIAS's research boundaries as well as by its intended objectives and desirables. I

[53] The Project was funded by European Union's Horizon 2000 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 860630.
[54] 'NoBIAS - About' (*NoBIAS*) <https://nobias-project.eu/about/> accessed 17 January 2025.

defined the research questions of this thesis and delimited the concrete parameters of my research and the scope of this thesis.

1.4.1. Automated decision-making systems – neither assisted decision-making nor artificial intelligence

In recent years, AI has become 'the topic' in almost every field of research and social, economic, political and cultural context. No aspect of our lives or our society has not been affected in one way or another by AI[55], and that is something that has had an impact on academia and research. However, when we look at the definition of AI, we find an umbrella term for an alleged technology able to simulate different facets of human intelligence[56]. Its very definition makes it a controversial term since many question whether it is really possible to understand the functioning of the human brain and, therefore, simulate its functioning and capabilities through technology[57]. Furthermore, AI is not a unique technology but a sum-up of different techniques used to simulate the human ability to discover, infer, or reason knowledge[58]. Thus, AI is compounded by techniques including natural language processing, vision, text and speech, motion, and prediction and decision.

ML is part of AI insofar as it is the technique -of AI- involved in predicting and making decisions based on data[59]. In turn, an algorithm is the set of rules that determine the concrete steps and processes a computer needs to employ to complete a task or solve a problem[60]. An ML algorithm is one whose task is prediction or decision-making[61].

[55] Sergio David Becerra, 'The Rise of Artificial Intelligence in the Legal Field: Where We Are and Where We Are Going' (2018) 11 Journal of Business, Entrepreneurship and the Law 27; Jasmin Praful Bharadiya, Reji Kurien Thomas and Farhan Ahmed, 'Rise of Artificial Intelligence in Business and Industry' (2023) 25 Journal of Engineering Research and Reports 85; Adam Bohr and Kaveh Memarzadeh, 'The Rise of Artificial Intelligence in Healthcare Applications' [2020] Artificial Intelligence in Healthcare 25; Jayden Khakurel and others, 'The Rise of Artificial Intelligence under the Lens of Sustainability' (2018) 6 Technologies 100; Henrik Palmer Olsen and others, 'What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration'.
[56] Stryker and Kavlakoglu (n 30).
[57] Simon Makin, 'The Four Biggest Challenges in Brain Simulation' (2019) 571 Nature S9.
[58] Stryker and Kavlakoglu (n 30).
[59] ibid.
[60] Stone (n 13); Willson (n 2).
[61] Zoubin Ghahramani, 'Probabilistic Machine Learning and Artificial Intelligence' (2015) 521 Nature 452.

ADM is the process of making a decision by technological means and without human involvement. ML is logically involved in the automation of many decision-making processes, but the automation of decision-making is not exclusively done through ML. Other simplest technologies and algorithms can also be used.

The distinction between these technologies -AI, ML, and other algorithms- and ADM is relevant to this thesis because, depending on the technology used, the effect on the ADM varies significantly. The scope of this thesis is delimited to ADM -automated decision-making processing-independently of the technology used to achieve such a goal. However, since the automation of decision-making is generally done through algorithms -independently of them being AI, ML or other algorithms- the conceptual and normative frameworks of this thesis will take into account the challenges and risks algorithms introduce in the decision-making process but always from the lenses of those algorithms been used in the decision making. In this sense, ADM refers to the automation of the decision-making process and, hence, the automated processing of the individual's data and information for the goal of making a decision[62].

Note that this thesis only focuses on the automation of decision-making processes, and it does not cover the use of algorithms to assist humans in the decision-making process. Thus, only automated processes and decisions without meaningful human involvement are included in its scope. There is indeed an extensive debate over the real meaningfulness of the role of humans in the final decision[63] and the problems that may arise when multiple stages of automated and human decision-making are combined[64]. So far, I not will elaborate on this debate, although the challenges of identifying what process is automated or not will be addressed in the doctrinal framework.

The thesis does not address, either, traditional human decision-making neither as an opposed concept to fully automated decision-making nor as a granted fundamental

[62] Note that in technical fields, one could refer to automated decision-making system, model and processing. The GDPR only refers to automated processing or automated decision making-processing, so that is the notion I generally used in this thesis. However, I will use the specific notion of ADM model or algorithmic model when to the concrete technology used to automate the decision-making model.
[63] Barros Vale and Zanfir-Fortuna (n 48).
[64] Binns and Veale (n 44).

right for individuals. In this regard, there is an open academic and social discussion[65] on whether there is or should be a fundamental right to human decision-making considering the risks and challenges ADM can pose to individuals' rights, freedoms, and liberties - I will dwell on these concerns in future sections-. However, for the purpose of this thesis I only address the right to a human decision as a safeguard towards ADM as referred to in the GDPR -please see section below-. The jump from being a safeguard in the European data protection legal framework to be considered a fundamental right is a normative analysis that fails out of the scope of this thesis.

1.4.2. The right to information and an explanation

The European General Data Protection Regulation[66] -hereafter, GDPR- has specifically regulated automated individual decision-making, granting every data subject a right to not be subject to automated decisions, including profiling, if it produces legal or similarly significant effects on him or her. However, the GDPR prescribes three exceptions for the prohibition: 1) the necessity to enter into or perform a contract, 2) the existence of a particular EU or national law that allows it, or 3) the explicit data subject consent.

When automated individual decision-making is allowed, the GDPR obliges data controllers to implement at least three suitable safeguards: the right to express his or her point of view, the right to obtain human intervention, and the right to contest the automated decision. The recitals of the GDPR extend those safeguards to the right to obtain specific information and the right to obtain an explanation of the decision reached. Moreover, the GDPR establishes the right to be provided with information where personal data is collected from the data subject or a third party, including information about the existence of automated processing, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

---

[65] Aziz Z Huq, 'A Right to a Human Decision' (2020) 106 Va. L. Rev. 611; Yuval Shany, 'From Digital Rights to International Human Rghts: The Emerging Right to a Human Decision Maker' (*Institute For Ethics in AI - University of Oxford*, 11 December 2023) <https://www.oxford-aiethics.ox.ac.uk/blog/digital-rights-international-human-rights-emerging-right-human-decision-maker>.
[66] General Data Protection Regulation.

In essence, the GDPR establishes transparency and accountability requirements for automated individual decision-making in the form of information and explanations. This thesis focuses on providing an in-depth analysis of why automated individual decision-making requires suitable safeguards and specific obligations of information and explanation in the context of personal data protection.

In essence, the scope of this thesis is limited to the rights to information and an explanation of automated individual decision-making, as referred to in the GDPR. The thesis enquires into the scope and relevance of the rights, their legal rationale, and their adequacy to solve the challenges posed by ADM.

1.4.3. European General Data Protection Regulation - and not other EU laws dealing with transparency[67]

In recent years, the European Union -hereafter EU- has enacted several laws[68] that lay accountability, transparency, and understandability at the centre of algorithmic and AI systems' governance. Within this legal framework, the use of assisted or fully ADM based on data-driven technologies requires, to some degree or another, the provision of explanations and justifications about the final decision and the decision-making process to different interesting parties. In this sense, the European General Data Protection Regulation -hereafter GDPR- is just one of the multiple European laws that establish specific rights and duties for the use of algorithms. Bibal, Lognoul, De Streel,

[67] This section is based on the academic article Alejandra Bringas Colmenarejo, Laura State and Giovanni Comandé, 'How Should an Explanation Be? A Mapping of Technical and Legal Desiderata of Explanations for Machine Learning Models' [2025] International Review of Law, Computers & Technology 1. I acknowledge the authorship of the article's section "Explainability, ML and Legal Desiderata" from which I extracted part of the content of this thesis section. The content of the academic article has been modified and adjusted to this thesis.

[68] Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 Amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as Regards the Better Enforcement and Modernisation of Union Consumer Protection Rules (Modernisation Directive) 2019 (OJ L 328/7); Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (Online Intermediary Service Regulation) 2019 (OJ L186/57); Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU)2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU)2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) 2024 (OJ L 2024/1689).

and Frenay[69]; Hacker and Passoth[70]; and Lognoul[71] have surveyed and synthesised the requirements on information and explanations for algorithmic and AI systems in the EU law, covering both public and private law. Other authors have analysed information and explanations requirements for algorithmic and AI systems in concrete sectors such as health[72], public administration[73], or law enforcement[74].

Special consideration needs to be given to the recently approved Artificial Intelligence Act[75]. The new European Regulation establishes the most extensive explainability and justificability requirements to date for AI in the form of literacy, transparency and information obligations, and human oversight, which, in consequence, establish a quasi-general obligation for interpretable and explainable AI. The Artificial Intelligence Act establishes literacy requirements for all AI systems with no exception, as well as specific explainability duties to high-risk[76] and to General-purpose models[77]. Quite relevant for this thesis, the Artificial Intelligence Act also includes a right to explanation for decisions made on the basis of an output from a specific class of high-risk AI system. The provision strongly resembles the wording and standpoint of the GDPR's information and explanation requirements concerning algorithmic ADM.

Furthermore, the GDPR has been implemented by national Member States' laws, covering the regulation of ADM. For a complete analysis of all the Member States's laws and the different approaches undertaken for the GDPR's implementation -i.e.,

[69] Adrien Bibal and others, 'Legal Requirements on Explainability in Machine Learning' (2021) 29 Artificial Intelligence and Law 149.
[70] Philipp Hacker and Jan-Hendrik Passoth, 'Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond', *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers* (Springer 2022).
[71] Michael Lognoul, 'Explainability of AI Tools in Private Sector: An Attempt for Systemization' [2020] SSRN Electronic Journal <https://www.ssrn.com/abstract=3685906> accessed 17 January 2022.
[72] Julia Amann and others, 'Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective' (2020) 20 BMC medical informatics and decision making 1.
[73] Olsen and others (n 55).
[74] Stephan Raaijmakers, 'Artificial Intelligence for Law Enforcement: Challenges and Opportunities' (2019) 17 IEEE security & privacy 74.
[75] Artificial Intelligence Act.
[76] For instance, Article 13 paragraph 3 established a transparency and information provision to deployers, which includes, among other things, the high-risk AI system's technical capabilities and characteristics that are relevant to explain its outputs -as per paragraph 3 (b)(iv).
[77] As referred to, for instances, in Annexes XI and XII of the Artificial Intelligence Act.

narrowing or widening the scope of the regulation- I refer to Malgieri[78]. Although this thesis would initially have covered these national laws and looked at their differences and similarities with respect to the GDPR, the scope of this thesis has been limited only to the GDPR and adopted a more high-level approach to ADM and the rights to information and an explanation. In recent years, the interpretation of the GDPR's brought considerable attention in national courts and data protection authorities. Therefore, the goal of this research is to contribute to the existing debate by offering an up-to-date framework of the rights to information and an explanation of the EU law level.

I also want to acknowledge that the European Parliament proposed the Artificial Intelligence Act once I had already started my PhD journey. The first draft proposal did not include an independent right to an explanation; by the time a possible right to an explanation was introduced in the draft compromise amendments, I was already quite advanced in my research. Despite the academic excitement this possible new right to an explanation for high-risk AI systems provoked, the scope of this thesis was consciously delimited to the GDPR. I was concerned with the uncertain development the new right of an explanation could suffer in the upcoming months and the impact such an unknown could have had on my thesis. That being said, I am convinced that most of the analysis offered in this thesis can be used in the interpretation of the Artificial Intelligence Act since, in the end, the final right to an explanation of the Artificial Intelligence Act is more limited than the right to an explanation as presented in the GDPR.

Moreover, despite the limitation of its scope to the GDPR, the analysis, arguments and conclusions provided in this thesis are of relevance and applicability in other specific contexts and national laws. This thesis is concerned with the use of algorithms to automate decision-making in high-consequence processes. Therefore, the conceptual and normative framework provided in this thesis can be used to understand the risks and challenges brought by algorithms to other types of contexts and processes and

[78] Gianclaudio Malgieri, 'Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations' (2019) 35 Computer Law & Security Review 105327.

even by other more complex technologies that use them. Likewise, the jurisprudence of the GDPR's rights to information and an explanation can inspire the interpretation of similar rights existent in other European and national laws, especially with regard to the future interpretation and exercise of the IA Act.

1.4.4. Data subjects – and not users, deployers or providers of the systems

Importantly, when considering information and explanation requirements, we should differentiate between those that are duties for the users, deployers, or providers of the system and those that are granted to the individuals directly affected by the system. The distinction is of high relevance due to the possible conflict of interest that may arise between all the involved parties and the different levels and categories of information that are granted to each pertinent party. This thesis is limited to the data subjects' rights of information and an explanations as referred to in the GDPR, and the consequent information obligations data controller are adhered to. Thus, this thesis does not dwell on the information, explanations or even instructions that shall be available and provided to all the actors involved in the development, deployment and use of the systems[79]. Whilst legal requirements for algorithmic and AI systems' interpretability, transparency, and accountability are indispensable parts of the governance framework for algorithmic and AI systems, this thesis will not cover that facet. Of course, requirements demanding more interpretable, transparent, fair and trustworthy systems will compel their developers, deployers and users to adopt a responsible approach to algorithms and AI and often have an impact on individuals' rights, given that fairer and more transparent and interpretable systems by design facilitate the provision of

[79] Of particular relevance in this regard is the recently approved Artificial Intelligence Act. For instance, Article 13 of the Artificial Intelligence Act specifies transparency and information duties to deployers of high-risk AI systems to "ensure that their operation is sufficiently transparent to enable deployers to interpret the system's output and use it appropriately". Among other information, high-risk AI systems would need to be accompanied by instructions indicating "[…] its intended purpose, the level of accuracy including its metrics, robustness and cybersecurity, its performance regarding specific persons or groups of persons on which the system is intended to be used, or specifications about the input data […]". Likewise, Article 14 impels high-risk AI systems to be designed and developed in a manner that ensures effective oversight by a natural person -also possible via appropriate human-machine interfaces-, the proper understanding of the relevant capabilities and limitations of the high-risk AI system, and the correct interpretation of its output.

information about their functioning and outcomes. This also makes it easier for data controllers to provide more relevant and useful information to data subjects.

1.4.5. Black-box systems and post-hoc explainable methods

In computer science, systems that use algorithms are considered interpretable when they are inherently designed and constructed in a way that allows the understanding of the logic behind the model performance and the reasons behind the decision reached[80]. Interpretable algorithms are commonly referred to as white-box models[81]. In contrast, black-box models refer to those algorithms that are very difficult to interpret, even for human experts in functional domains, and require the use of specific technical methods to provide some level of understanding of the functioning of the model - so-called post-hoc explainability methods-.

This thesis does not distinguish between technical interpretable and non-interpretable algorithms - white and black boxes, respectively- when referring to ADM as per Article 22 of the GDPR. However, it would not be feasible to offer an assessment of all the technical approaches and methods that can be used to provide information and explanations about both interpretable and non-interpretable algorithms that conform to the ADM. For this reason, the techno-legal assessment of the rights to information and an explanation is limited to the compliance of three concrete types of post-hoc explainability methods to the requirements of information and explanations established in the GDPR. The methods are SHapley Additive exPlanations (SHAP), DIverse Counterfactual Explanations (DiCE), and LOcal Rule-based Explanations (LORE).

The reason behind this decision resides in that the use of non-interpretable algorithms to automate the decision-making process adds another layer of complexity to the already challenging compliance with the rights to information and an explanation. This

[80] Francesco Bodria and others, 'Benchmarking and Survey of Explanation Methods for Black Box Models' (arXiv, 25 February 2021) <http://arxiv.org/abs/2102.13076> accessed 13 September 2022.

[81] A model is a mathematical construct that aims to simplify reality in order to understand an event. They range from econometrics models that aim to predict financial crisis, to predictions of crime rates or image recognition. Mathematical models are often encoded forming an algorithm. The algorithm is a set of rules that can be understood from a mathematical point of view, what might not be interpretable is the knowledge construct embedded in the model. Moving forward I would use model and system interchangeably as precise definitions do not affect the scope of this thesis.

is because data controllers need to understand the algorithmic system's logic and functioning using post-hoc explainability methods, filter and adapt the information about the model, and provide it to the data subjects. Thus, analysing some of the possible methods used by data controllers in this first stage offers insightful answers regarding the possibilities for effective implementation and exercise of the rights to information and explanation.

That said, post-hoc explainability methods will only be assessed in **Chapter 5** and **Chapter 6**. The analysis and frameworks provided in the rest of the chapters are applicable to technical black-and-white boxes irrespectively. The inner workings of an algorithm, irrespective of whether it is technically interpretable or not, require a level of expertise and knowledge that is certainly not possible for anyone and everyone. Thus, although the technical assessment is done with respect to non-interpretable systems, the arguments and conclusions provided in both chapters can guide the provision of information and explanation of interpretable systems.

1.5. The Structure Of The Thesis

In **Chapter 2: Automated Individual Decision-Making Processing - The Particular Risks and Challenges to Individuals' Rights and Freedoms**, I approach the use of ADM in our everyday and high-consequence decisions and expose the problems associated with their promised neutrality and inherent inscrutability as well as the so-called problem of black-box systems. Chapter 2 provides a review of cases and disputes involving automated decisions, including those based on profiling, that have reached a court or a national Data Protection Authority in a European Member State or at the Court of Justice of the European Union. The selected case law shows how even if the person's personal data was affected or impacted as a result of the automated processing, the impacts and effects on the individual's rights, freedoms and legitimate interests broaden to other areas of law besides the protection of their personal data. In fact, the selected case law reflects the possibilities in which utilising modern data-driven technologies can have an undesirable or unrequired effect on the individual's participation in society, e.g. their access to employment, credit or social activities as a football match. In turn, Chapter 2 also serves to expose the importance of this thesis as

it offers a contextualisation of automated individual decision-making processes, and so the ideal background to understand the relevance of the rights to information and an explanation.

In **Chapter 3: The Doctrinal Framework of the Right to Information and an Explanation,** I lay the legal foundations of this thesis. I present the rights to information and an explanation as well as of the challenges or problems that their wording in the GDPR has brought about. In concrete, Chapter 3 first outlines the key points of contention regarding the nature and content of the right not to be subject to automated decisions -as referred to in Article 22 of the GDPR- while it also sets out both derogations and safeguards associated with the right that may affect its enforcement by individuals. Subsequently, Chapter 3 approaches the right to access and information about the existence of ADM and the right to an explanation about an automated decision assessing Articles 13(2)(h) and 14(2)(g), 15(1)(h) and Article 22 (3), respectively. Finally, Chapter 3 proposes the minimum threshold of compliance and the desirable extensive compliance approach to the rights to information and an explanation. In essence, through Chapter 3, I offer a framework of conclusions from the doctrine rather than a unique interpretation of the rights.

When it comes to unravelling the significance and intention of having rights to information and an explanation for automated decisions in the European data protection law, we cannot forget the rationale of the GDPR. Even though the problematics and case law exposed in Chapter 2 can explain the rising concerns towards ADM, they do not offer an indefensible argument that explains why these two rights were deemed suitable safeguards for ADM. Neither do the identified problematics explain why the use of ADM was a concern addressed in data protection law instead of any other sectorial law. Hence, reasonable questions to pose are:Why would the rights to information and an explanation help data subjects when affected by an automated decision? and Why would individuals have a right to understand the decision-making processes affecting them in their daily lives from the perspective of personal data protection?. Answering those questions can offer, in consequence, more insights into

how the framework of compliance for the rights to information and an explanation can be concretised in real scenarios.

Chapter 4: The Normative Framework of Transparency and Explainability Requirements in the GDPR addresses and responds to those questions. I first explore the aggregated risks that arise from introducing algorithms in decision-making processes of governance both in the public and private sectors: 1) non-voluntariness and 2) arbitrariness. Then, I situate these aggregated risks within the context of personal data protection, identifying the specific risks posed by the increasing unbalance of power between the individual and the data controller. Subsequently, I focus on the normative basis of the legal solutions existing in data protection law to control or mitigate those risks. To do so, I examine the GDPR requirements on transparency and explainability for automated decisions through two approaches: 1) as resemblance to due process safeguards and 2) as risk mitigation and control mechanisms. Together, Chapter 4 provides a comprehensive understanding of the normative framework of the rights to information and an explanation for automated decisions.

Chapter 5: A Legal and Technical Approach to Explainability and Chapter 6: How Should It Be an Explanation about an Automated Decision but How Can It Really Be? offer a technical and legal assessment of the rights to information and an explanation for automated decisions. The former is based on the academic article *How should my explanation be? A mapping of legal and technical desiderata for Machine Learning models*[82], whereas the latter is based on the other academic article, *The Explanation Dialogues: An Expert Focus Study to understand requirements towards Explanations within the GDPR*[83].

On the one hand, **Chapter 5: A Legal and Technical Approach to Explainability** presents a techno-legal mapping covering the potentials and challenges of black-box systems to comply with the information and explanation requirements of the GDPR. In this chapter, I first provide an overview of technical notions of interpretability and explainability that explores the distinctive characteristics, implications, and roles of

[82] Bringas Colmenarejo, State and Comandé (n 67).
[83] State and others (n 52).

both notions as mechanisms to either ensure or achieve the understandability of algorithmic systems. Subsequently, I dwell on the interpretation of algorithmic explainability and interpretability from a legal perspective, critically assessing whether there are comparable notions in law and whether the law is actually preoccupied with the technical distinction or not. Chapter 5 concludes with an examination of the properties that technical and legal perspectives of explainability look to attain, which may or may not coincide specifically. To narrow this rather theoretical analysis on explainability, I connect each desideratum with the information and explanation requirements of the GDPR in concrete and clear terms.

On the other hand, **Chapter 6: How Should It Be an Explanation about an Automated Decision but How Can It Really Be?** assesses the level of compliance technical explainability methods can reach when used to fulfil the requirements of information and explanations about an automated decision. In this chapter, I offer a conceptual taxonomy of technical explainability methods and examine the different types of techniques available to make the functioning of algorithmic systems and the main reasons behind their outputs 'understandable'. This approach looks to identify the objectives behind the design, development, and application of some of the most commonly used and technically developed explainability methods in the field of XAI and appraise whether their rationale coincides with the rationale behind legal explainability. Chapter 6 also expounds on the perceptions, expectations, and reasoning offered by a group of legal experts and practitioners on legal explainability when questioned about four concrete explanations of an automated decision. Chapter 6 concludes with my assessment of the selected explainability methods that serve as a real case analysis of the *Framework of the rights to information and an explanation* and *The spectrum of compliance – minimum and maximum* thresholds presented in Chapter 3 and the *Technical Desiderata* and *Legal desiderata* analysed in Chapter 4.

1.6. Interdisciplinary Methodology

This thesis provides a legal and technical assessment of the rights to information and an explanation. Due to the interdisciplinary of the research carried out for this thesis, the methodology is not restricted to the doctrinal legal method.

Traditional desk-based/doctrinal legal research methodology was used to analyse the GDPR, related case law, and any other legal source of the EU and its Member states, as well as the existing literature that refers to rights to information and an explanation of automated individual decision-making. Whereas the examination of the law is focused on its wording, purpose, and rationale to avoid excess abstraction, commentaries and opinions about the particular interpretation and exercise of the rights provided both in case law and in legal literature are also included. In essence, this thesis aims (1) to critically analyse the existent legal sources to identify and describe the potentials and challenges of exercising the mentioned rights, (2) to seek the critical aspects of their enforcement, identifying possible ambiguities and legislative gaps of the law, and (3) provide new commentaries, solutions and opinions to those already offered in the literature.

To complement this traditional legal methodology, this thesis also engages with the rights to information and an explanation for automated decisions from the perspective of computer science and social science.

Therefore, I explore in Chapter 5 the technical notions of interpretable and understandable algorithmic and AI systems. In Chapter 6, I also investigate the different post-hoc explanation methods that can provide information and explanations about technical black-box systems. Further, I explore the various techniques that are used to make algorithms understandable to humans, the goals that these techniques are intended to achieve, and the particularities of the functioning and outcome of the system that they are designed to reveal.

The interdisciplinarity of this thesis is largely brought out in the incorporation of the two academic articles I co-authored. *How should an explanation be? A mapping of technical and legal desiderata of explanations for machine learning models* we combined the traditional desk-based/doctrinal legal method with a straightforward literature review methodology involving the research, reading, analysis, evaluation and summary of scholarly literature on technical and legal explainability. The proposition of desirable properties for technical and legal explanations of ML models -used in ADM- was made through a simple qualitative methodology.

By incorporating the content of this academic article in my thesis, I intend to address [the thesis'] research question of "What objectives are to be achieved, and what is exactly to be provided through those explanations and information?" Providing a qualitative analysis of the literature that assess both the legal and technical scholarly on the desirable properties of explanations and interpretations for ADMs systems aims to delimits the main characteristics one could and should expect and require from them.

In turn, *The explanation dialogues: an expert focus study to understand requirements towards explanations within the GDPR* uses the qualitative methodology of thematic analysis based on ground theory[84]. Qualitative research in technology can be understood as a methodological approach that emphasizes the contextual interpretation of various stakeholders in their interactions with technological systems. Its primary objective is to capture stakeholders' attitudes, behaviours, and perspectives in order to inform and enhance the design of future technologies. Common strategies for data collection, analysis, and organization within this paradigm include ethnography, narrative inquiry, thematic analysis, and grounded theory. Despite their methodological differences, these approaches share a commitment to generating a nuanced understanding of participants' lived experiences and viewpoints. Typically, such studies involve small, purposefully selected samples based on relevant criteria, for example, the participants' specific expertise, and data collection is characterized by sustained and direct engagement between researchers and participants.

To obtain such knowledge*, The explanation dialogue* was designed as a user study in which the behaviours, preferences, and opinions of the target audience were observed and analysed. This process typically entails gathering both qualitative and quantitative data through diverse methods like surveys, interviews, and usability testing. In our case, we provide an initial questionnaire and subsequently carried out individual interviews

[84] The part of this methodology section describing ground theory is heavily based on the content of the academic article *The explanation dialogues: an expert focus study to understand requirements towards explanations within the GDPR*. I acknowledge Dr. Andrea Beretta as the principal author of the article's section *Qualitative Research,* although the wording and content of such article's section has been edited and adapted to the scope of this thesis. Additional content and references have been included when considered appropriate for the purpose and scope of this chapter.

with the willingly participants. The study was approved by the University of Southampton Faculty Ethics Committee, ERGO ID: 80482.

In accordance with the methodology of a user study, we contacted a small purposely defined sample of legal experts, to whom we asked to perform as the bank's responsible of compliance individuals, and later on, interviewed in accord to their own role and expertise of legal experts. Hence, we used qualitative research in technology to collect the potential attitudes and insights of two types of stakeholders (i.e. responsible of compliance and legal experts) towards explanations and interpretations of ADMs systems.

While the general methodology of user studies are a crucial to evaluate technology, *The Explanations Dialogue'* user studies attempts to concretise such evaluation on clarifying how various XAI methods are interpreted, accepted, and employed, as well as how these techniques are subjectively experienced and perceived by different stakeholder groups. Conducting a user study involves examining users' perspectives, expectations, levels of trust, and other qualitative (as well as quantitative) dimensions. It also requires linking these insights to professional practices, as well as to legal and political considerations. To link the data collected in *The Explanations Dialogue* with the legal considerations that conform the background of our article, we use three methodologies: grounded theory, quantitative evaluation through aggregation, and a qualitative comparison of the interviews questions.

Ground theory is a qualitative research method designed to generate new theories that are rooted in the qualitative data collected during the research process. We apply grounded theory as an epistemological framework. We opted for this approach given the exploratory nature of our study as grounded theory furnishes a structured framework for organizing the knowledge derived from the process of data collection. However, I refrain from presenting in this thesis the complete analysis of both the questionnaire and the interviews that was performed using grounded theory methodology due to space and relevance constraints in regard to this thesis, for which I do not expand on this type of methodology it any further.

The thesis include, nonetheless, 1) the summary of the legal experts' reflections on each type of explanation, both in their roles of responsible of compliance and legal experts, and 2) the summary of the whole project results, provided in the form of concrete answers to the research questions of this thesis. The latter unavoidably feeds from the ground theory methodology followed in The Explanation Dialogue academic article[85], hence the short reference to this type of methodology above.

It is worth noting that the following are the research questions that underlie our study:

> RQ1 How do legal experts reason about explanations for ADM systems, and how do they judge the legal compliance of existing methods? Some aspects to consider for a presented explanation:
>
> a) Is the explanation complete or incomplete with regard to the expectations of the legal scholars, and is some information given by the XAI more relevant than others?
> b) Is the explanation compliant with the GDPR, and is there a preference towards a specific method or presentation type?
> c) Does the legal reasoning change when presented with the explanation of a true positive/false positive?
>
> RQ2 Do legal experts understand and trust explanations for ADM systems, and what are the steps identified to go forward? Some aspects to consider are:
>
> a) How well are the presented explanations understood?
> b) Which gaps in presented explanations are identified? How can the presented explanations be improved?

The Explanations Dialogues aims to close the gap between XAI and legal definitions of explainability and interpretability by inquiring about the legal compliance of the former with respect to the information and explainability requirements established in the GDPR. Furthermore, the project attempts to assess the capability of XAI to make automated decisions understandable and scrutinized in regard to their lawfulness and fairness. Incorporating part of the results obtained and conclusions reached in the

[85] State and others (n 52) p.10.

academic paper to this thesis aims to add to the answers to its second and third questions. It is important to highlight that the questionnaire's and interviews' questions were designed with these premises and the aforementioned research questions in mind, which not casually resemble the own research questions and premises that guide this thesis.

Finally, although they are not directly used to analyse the core data of this thesis nor to answer any of its research questions, it is necessary to mention that three technical explainability methods are, certainly, used. On the one hand, we use these three methods to develop the explanations presented to the legal experts in *The Explanations Dialogues*. On the other hand, those same methods and the visualisation of their explanations were the main object of analysis of Chapter 6. The methods in question are SHapley Additive exPlanations (SHAP), DIverse Counterfactual Explanations (DiCE), and LOcal Rule-based Explanations (LORE). Due to the particularities of their use in this thesis, I will present and describe them in Chapter 6 rather than in this section.

In essence, the presented interdisciplinary methodology approach pursues to critically examine how the rights to information and an explanation can be effectively exercised to strengthen individuals' rights and empower them in the unbalanced relations resulting from the use of ADM.

Chapter 2: Automated Individual Decision-Making Processing - The Particular Risks and Challenges to Individuals' Rights and Freedoms

2.1. Introduction

Just looking at the cases of ADM that have reached courts in Europe, there are, as will be shown below, multiple and diverse practices of automating the decision-making processes that pose myriad challenges and threats to society. The issues arising from these cases also highlight the difficulty involved in finding effective solutions to these situations, as they can create both individual and collective harm alike. To facilitate the necessary discussion and conversations towards finding appropriate solutions and mitigation measures, it is a prerequisite to identify and define the potential threats and challenges arising from ADM. For this reason, this chapter considers important harms that may arise when ADM is introduced into the decision-making processes of our everyday life. To begin, this section elaborates on the nature and particular elements that generally characterise algorithms, which, nonetheless, also bring several potential challenges both for society and its individuals. This first section does not seek to demonise algorithms or deny their benefits to society but rather to provide some clarity on why their use and implementation have generated so much hostility and distrust. Particularly, it addresses the claims referring to the algorithms' neutrality and objectivity as well as the so-called *black-box problem*.

To continue, section 2.3 offers an exploration of particular cases of algorithmic decision-making that explore the ways these practices could be, and have been, used. The aim is to show that they are not isolated technologies but part of a broader picture, simultaneously impacting different areas of commercial practice and law and regulation.

The discussion draws on cases and disputes involving automated decisions, including profiling, reaching a court or a national Data Protection Authority (DPA) in a European member state. The situations presented below are not intended to be an exhaustive

representation of all the scenarios where ADM occur but illustrate the harms and threats to individuals and society. Despite their multiple discrepancies, the cases covered in section 2.3 have the common denominator of involving ADM, including profiling, as referred to in Article 22 of the GDPR. Although Chapter 3 will dwell in great detail on the content, development, and interpretation of Article 22, it is worth mentioning at this point of the thesis that the provision is only applicable to decisions based solely on automated processing of personal data, including profiling, which produces legal effects concerning an individual or a similarly significantly affects that individual[86]. Consequently, the situations presented in this chapter will expose some of the reasons why automated decisions -as referred to in Article 22 of the GDPR- could go wrong and the different legal regimes and rights that may be implicated in the controversies. The cases provide an overview of both illegal and unfair practices, the former referring to harms that are illegal under national or European laws and the latter covering actions that may typically be legal but trigger notions of unfairness[87].

The facts and circumstances that brought these cases to the courts and DPAs in Europe might not exactly coincide with the potential harms that we identified in each of them but still provide a sense of the circumstances and changes that automated decisions are bringing to our legal, social, and economic world. Consequently, our analysis will not be exclusively centred nor elaborate on all the disputed facts and legal reasoning behind the cases' judgments, but on the underlying problematics that could be identified in the state of the facts and the legal grounds of the courts' and DPAs' decisions.

It is important to clarify that although the contexts where automated decisions take place in our daily lives encompass a myriad of interactions -including our rather naïve example of using a recommendation system to find a place to buy our daily dose of caffeine- the cases selected for this chapter are limited to situations that produce a legal or similarly significantly effect on individuals, hereafter called high-consequence decisions. The aforementioned common denominator triggers some specific legal

[86] General Data Protection Regulation.
[87] Future of Privacy Forum, 'Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making' (2017) p.5.

protections under Article 22 of the GDPR, particularly the right not to be subject to decisions of that nature and the requirement of implementing concrete safeguards[88] on the limited cases in which the use of automated decisions are allowed. A logical conclusion to draw from these circumstances is that not all automation of everyday decisions requires the same legal attention. However, this deduction is certainly simplistic and requires further analysis. Chapter 4 will present and examine the normative framework that could support the legal distinction between automated decisions and why transparency and explanation requirements have been put forward as an appropriate measure to ensure the protection of individuals legally or similarly affected by them. Rather than develop further on this matter, this chapter presents real scenarios of automated decisions as showcases of the importance of this thesis, which I also find notably important to keep in mind throughout the rest of the thesis.

The cases presented below will address three contexts where ADM, including based on profiling, has spurred a social debate and legal discussion about the impact of their use on individuals and the transformation they have brought to their respective environments. The contexts in question are 1) the workplace, concretely concerning algorithmic management of employees; 2) finance services, particularly regarding credit scoring; and 3) private protection of the public interest, crowd control and automated facial recognition.

Sections 2.2 and 2.3 provide an overview of the reasons why this PhD has relevance and importance and contributes to the current state of the art.

2.2. Automated Individual Decision-Making – The Problem

2.2.1. Promised objectivity and neutrality

Despite being correct in the sense of 'statistical validity'[89] to the extent to which the conclusions drawn from the statistical inferences are mathematically accurate and

[88] Among those safeguards, -and due to the object of this PhD- we emphasize the information and explanations requirements established in Article 13(2)(h), Article 14(2)(g), Article 15(1)(h), and Article 22(3) of the GDPR.
[89] Lilian Edwards and Michael 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.

reliable, algorithms will never be neutral as they are always built on the choices made by their developers and owners, e.g., which data is included in the dataset, which fairness metrics are introduced in the code, which constrictions are set, or under which labels are the data classified[90]. The conceptualisation of algorithms as neutral and objective poses a threat to society, given that they are neither neutral nor objective. Algorithms are designed and created by individuals and are unavoidably predicated on certain values that they incorporate into their internal workings and processes[91]. Furthermore, even when they are ethically charged or value-laden, assessing the harmlessness of the algorithm and whether its performance met those established thresholds cannot often be assessed through the evaluation of those design choices alone[92]. In other words, even if the developers have established some restrictions to the correlations and presumptions the algorithm can produce, introduced some constraints to the knowledge logic, or set certain standards to be respected, the mere use of these techniques does not always ensure a fair or desirable outcome as the algorithm could still work in a way that was not predicted nor expected. Algorithms' final desirability -in terms of potential social and legal repercussions- might need to be evaluated within the final context where they are employed; assessing the correlations, assumptions, and outputs developed by the algorithms as well as the effects that those have on the individuals.

However, what makes an algorithm biased and its outcomes unfair is the subject of a contested debate[93]. Fairness is essentially a contested concept[94] as it is context-

---

[90] ibid

[91] Vedder and Naudts (n 15) P.208; Willson (n 2) p.139; Marc Steen, 'Upon Opening the Black Box and Finding It Full: Exploring the Ethics in Design Practices' (2015) 40 Science, Technology, & Human Values 389.

[92] Vedder and Naudts (n 15).

[93] Michael Rovatsos, Brent Mittelstadt and Ansgar Koene, 'Landscape Summary: Bias in Algorithmic Decision-Making: What Is Bias in Algorithmic Decision-Making, How Can We Identify It, and How Can We Mitigate It?' <https://www.research.ed.ac.uk/en/publications/landscape-summary-bias-in-algorithmic-decision-making-what-is-bia> accessed 20 May 2022; Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' [2016] SSRN Electronic Journal <https://www.ssrn.com/abstract=2477899> accessed 20 May 2022; Abigail Z Jacobs and Hanna Wallach, 'Measurement and Fairness', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) <https://dl.acm.org/doi/10.1145/3442188.3445901> accessed 20 May 2022.

[94] Shira Mitchell and others, 'Algorithmic Fairness: Choices, Assumptions, and Definitions' (2021) 8 Annual Review of Statistics and Its Application 141; Jon Kleinberg and others, 'Algorithmic Fairness', *Aea papers and proceedings* (American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2018); Deborah Hellman, 'Measuring Algorithmic Fairness' (2020) 106 Virginia Law Review 811.

dependent and highly conflicts with different ethical, political, and cultural understandings. Still, fairness needs to be statistically defined to model fair data-driven systems, leaving the question of which values need to be operationalised into variables unsolved. For this reason, the literature on fair algorithms mainly derives its fairness constructs from a legal context where a process or decision is considered fair if it does not discriminate against people based on their membership to a protected group[95]. Indeed, fairness in law entails that everybody is equal, but what equal treatment means is still the object of a strong and context-dependent debate. Fairness may entail equality in the sense that everyone must be treated the same. Still, fairness may also mean that everyone must be offered equal opportunities and be treated depending on their needs[96]. Fairness can be understood as equality or equity, so the instruments and ways to achieve and ensure the goals of each one highly differ[97]. In essence, fairness can be understood differently depending on its nature, formal or substantive; the context it applies to, legal or technical, or the actor it refers to, public or private.

From the point of view of computer science, fairness is the absence of legal discrimination. It could be achieved through two measurement approaches: 'individual fairness' and 'group or statistical fairness'[98]. The former requires that a measurement of individual fairness must compare an algorithmic behaviour across similar individuals. Simultaneously, the latter seeks that to achieve fairness, a statistical measurement must compare the algorithmic behaviour across different demographic groups and then seek the approximate parity of some desirable measure across the groups[99]. Although

[95] Songül Tolan, 'Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges' 25.
[96] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law' [2021] SSRN Electronic Journal <https://www.ssrn.com/abstract=3792772> accessed 9 February 2022.
[97] In truth, fairness ca be understood in many ways, representing different political, social, or economic values. The COMPASS case reveals how the notions of fairness and so the algorithmics metrics used to reach those notions can conflicted and exclusive Angwin and others (n 38); Marcello Di Bello, 'Algorithmic Fairness – ProPublica v. Northpointe' (*Marcello Di Bello*, Fall 2021) <https://www.marcellodibello.com/algorithmicfairness/handout/ProPublica-Northpointe.html> accessed 28 November 2024..
[98] Wachter, Mittelstadt and Russell (n 96).
[99] Lori Bowen Ayre and Jim Craner, 'The Baked-in Bias of Algorithms' (2018) 10 Collaborative Librarianship <https://digitalcommons.du.edu/collaborativelibrarianship/vol10/iss2/3>; Dino Pedreschi, Salvatore Ruggieri and Franco Turini, 'Discrimination-Aware Data Mining' 9; Atoosa Kasirzadeh and Andrew Smart, 'The Use and Misuse of Counterfactuals in Ethical Machine Learning', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021).

both approaches attempt to ensure unbiased and fair outcomes, each industry, sector, and service requires specific fairness constructs. Consequently, in practice, there are currently more than twenty different definitions of algorithmic fairness[100].

In truth, algorithmic fairness may confront different political, ethical, and cultural understandings. Therefore, a widely acknowledged problem in algorithmic fairness is the need to choose between competing values and measures, which, in the end, results in a contest between different political values and conceptions of fairness[101].

Besides their fairness, algorithms invariably 'create moral consequences, reinforce or undercut ethical principles, and enable or diminish stakeholder rights and dignity'[102]. Algorithmic inferences and conclusions may be politically and socially unacceptable because they may discriminate and discern in an undesirable way, resulting in the misrepresentation and invisibility of certain groups and the consolidation of discriminatory access to goods and services, which may reinforce both distributive and symbolic inequalities[103]. Indeed, finding evidence of discriminatory conclusions and unjust disparities in data-driven systems used in criminal justice, healthcare, education, or employment contexts is not difficult[104].

Specifically, algorithms can lead to discrimination through two different types of inequality (re)producing mechanisms: (1) those that stereotype and prejudice the different groups in the society, affecting its equal and accurate representation, and (2) those which reflect structural forms of inequality by relying on past and hierarchical discriminatory data[105]. The algorithm's reliability is highly dependent on the reliability of the data they are trained on and its similarities and differences with the data they will use when deployed. Therefore, there is an unavoidable risk that training and post-

---

[100] Arvind Narayanan, 'Fairness Definitions and Their Politics' [21] Youtube: Arvind Naranayan, Available online: https://www. youtube. com/watch.
[101] Pak-Hang Wong, 'Democratizing Algorithmic Fairness' (2020) 33 Philosophy & Technology 225.
[M]Kirsten Martin, 'Ethical Implications and Accountability of Algorithms' (2019) 160 Journal of Business Ethics 835, p.835.
[103] Xavier Ferrer and others, 'Bias and Discrimination in AI: A Cross-Disciplinary Perspective' (2021) 40 IEEE Technology and Society Magazine 72, p.2.
[104] Angwin and others (n 38); Nicolas Kayser-Bril, 'Austria's Employment Agency Rolls out Discriminatory Algorithm, Sees No Problem' (*Algorithm Watch*) <https://algorithmwatch.org/en/austrias-employment-agency-ams-rolls-out-discriminatory-algorithm/>.
[105] Barocas and Selbst (n 93).

training data reflect patterns of biases and discrimination or even offer statistically distorted pictures of reality. Interpreting this data in an undesirable manner would allow for the codification, replication, and even amplification of social injustices based on statistics and generalisations[106].

For this reason,

> To assess whether an algorithm is free from biases, there would be necessary to analyse the entirely of the algorithmic process. It would entail confirming that (1) the algorithm's underlying assumption and its modelling are not biased, (2) its training data does not include biases and prejudices, and (3) that it is adequate for making decisions in that specific context and task[107].

We cannot expect, however, for this fairness assessment to be available or accessible to every person interacting with the algorithm. Intellectual property or trade secret rights will strongly delimit the amount of information an individual is allowed to access, equally delimiting their options to check the fairness of the algorithm. Even if the algorithmic profile or decision were to be contested and its neutrality and objectivity verified, the assessment of the pertinent decision-making processing would be difficult as algorithms need to be understood as the intersections of two elements: their mathematical arithmetic or computational logic, and the particular context where they were used[108]. Particularly, assessing algorithmic decision-making would entail some difficulties as these two sides of an algorithm are embedded in varying political, technical, cultural, and social interactions that impact and influence the whole algorithmic process.

Furthermore, in the majority of the cases, the algorithm interacts with their environments. Algorithms construct meaning by bringing out particular ways of seeing the world that are later communicated to or read by other systems, entities, or

---

[106] Slava Polonski PhD, 'Mitigating Algorithmic Bias in Predictive Justice: 4 Design Principles for AI Fairness' (*Medium*, 24 November 2018) <https://towardsdatascience.com/mitigating-algorithmic-bias-in-predictive-justice-ux-design-principles-for-ai-fairness-machine-learning-d2227ce28099> accessed 13 December 2023.
[107] Xavier Ferrer and others, 'Bias and Discrimination in AI: a cross-disciplinary perspective' 40 IEEE Technology and Society Magazine 72-80 p.2
[108] Willson (n 2).

interested parties. Equally, the real world presents algorithms with disparities, stereotypes, reify practices, and world views that are assumed –typically through the datasets but also incorporated in the encoding of the algorithm and the goals to be achieved- and reproduced by the algorithms, restricting choices and reiterating discriminatory practices. Algorithms, thus, exhibit[109], as they construct, meaning but also behave on the basis of past or misconstrued patterns rather than on actual realities. Therefore, to assess an algorithm, understand its role in a decision and the effect it provokes, the algorithm cannot be conceived as a stand-alone or isolated process but as a contextual and relational entity[110].

Therefore, presenting algorithmic processing and profiling - as the neutral and objective[111] alternative to human discretion is problematic given the significant consequences on the users' access to products and services, job opportunities, insurance or housing. Assumptions about its neutrality and objectivity allow the decisions to go unquestioned and unchallenged.

2.2.3 Inherent inscrutability and complexity

Beyond their dubious neutrality and objectivity, algorithms' obscure inner workings would make us ponder over whether we would accept their judgement and performance without major reticence, or rather, we would require and expect basic knowledge and understanding about how the algorithm actually works –including relevant design and development choices[112].

In science, computing, and engineering, the notion of interpretability refers to the level of closeness of the internal essence of an algorithmic system[113]. Algorithms are considered interpretable when they are inherently designed and constructed in a way that allows the understanding of the logic behind the model performance and the

[109] ibid p.141.
[110] Willson (n 2).
[111] Vedder and Naudts (n 15).
[112] Bartosz Brożek and others, 'The Black Box Problem Revisited. Real and Imaginary Challenges for Automated Legal Decision Making' [2023] Artificial Intelligence and Law p.2 <https://link.springer.com/10.1007/s10506-023-09356-9> accessed 30 November 2023.
[113] Amina Adadi and Mohammed Berrada, 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)' (2018) 6 IEEE Access 52138.

reasons behind the decision reached[114]. Interpretable algorithms are commonly referred to as *white-box models,* but thus far, only four types of algorithmic models are considered inherently interpretable or understandable for a human, i.e. linear models, decision trees, rule lists, and decision sets[115]. Aside from these algorithms, the rest of the models are not inherently interpretable or understandable, leaving their internal design, structure, and working completely obscure for humans. By being non-interpretable by design, the logic and functioning of these algorithms need to be explained through external means[116]. In essence, *black-box* is a term used for labelling all those algorithmic models that are very difficult to interpret and explain, even for human experts in functional domains, and require specific methods to offer some level of explainability to offer some level of explainability[117]. This difficulty in providing a suitable explanation about how the system arrived at a concrete answer is called 'the black-box problem', and so a great challenge when using algorithmic decision-making designed with a black-box approach relies on developing and employing the appropriate post-hoc methods to facilitate or enable the *opening of the black-box*. The field of eXplainabe AI -hereafter XAI- investigates the possible methods to 'make [AI systems'] behaviour more intelligible to humans by providing explanations'[118], aiming to develop techniques and methods that increase the transparency and explainability of the model while maintaining high-performance levels[119]. The concrete name of the methods used to open-up the black-box is post-hoc explanation methods -hereafter XAI methods-.

Chapter 5 and Chapter 6 will offer more clarity regarding XAI and the different approaches and XAI methods existent to make algorithms more interpretable and less obscure. For the purpose of this PhD, I pointed out the difference between white and black boxes, as referred to in the computer science literature, primarily because the inherent technical inscrutability of the latter poses additional challenges when used for

[114] Bodria and others (n 80).
[115] Riccardo Guidotti and others, 'A Survey Of Methods For Explaining Black Box Models'; Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2019).
[116] Bodria and others (n 80) p.4.
[117] Rudin (n 43); Octavio Loyola-González, 'Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View' (2019) 7 IEEE Access 154096.
[118] David Gunning and others, 'XAI—Explainable Artificial Intelligence' (2019) 4 Science Robotics eaay7120, p.1.
[119] Adadi and Berrada (n 113).

algorithmic decision-making. Nonetheless, it should be noted that white boxes are still mathematical algorithms with a high level of complexity. While its inner workings can be interpreted and its knowledge logic identified, such understanding is certainly not possible for anyone and everyone. Likewise, the problems referred to algorithms' apparent neutrality and objectivity -presented in the previous section- affects both black and white systems. The interpretable nature of the latter may facilitate the evaluation and examination of its neutrality and objectivity, but the challenges identified above would still need to be considered and balanced.

The inscrutability of algorithms is not restricted to the mathematical sequence but to the technical context in which algorithms are ultimately deployed. As explained above, algorithms can be described as computational recipes to perform a concrete task or make a particular decision. Besides the complexity of the algorithmic *set of instructions* -algorithmic codes-, their inscrutability also resides in the inner workings and the usually unidentifiable patterns and correlations identified and used by algorithms to carry out such tasks on the basis of the offered instructions. Furthermore, the complexity of algorithms is structural in such a way that they are not technically isolated but work as part of a larger structure[120]. Algorithms are inert and without meaning unless they are paired with databases[121]. The information fed to the algorithm needs to be previously collected, prepared, and formalised in a way in which the algorithm can understand and use automatically. Available data can have a diverse nature, including structure -quantitative- data in the form of numbers and values, and unstructured -qualitative- data in the form of text, audio, image, or video films[122]. Before its use in training or deployment, this data needs to be organised in a unique data source. In many cases, this process comes with removal, amendment, and adjustments in the original data, which can have an impact in the latter stages of the algorithmic inner working as it could -intentionally or unintentionally- limit its capacity to identify some patterns or correlations.

[120] Vedder and Naudts (n 15) p.209.
[121] Gillespie (n 8) p.169; Vedder and Naudts (n 15).
[122] Kevin Normandeau, 'Beyond Volume, Variety and Velocity Is the Issue of Big Data Veracity' [2013] Inside big data.

Additionally, the datasets that are fed to the algorithm for training, as well as those that are used during its deployment, might contain unnecessary, duplicated, or extraneous data. Pre-processing techniques intend to process the data to remove discrimination before the algorithm could learn unrequired patterns, correlations, or assumptions[123]. The techniques, among which we can find the cleaning and structuring of the dataset, are complex and require qualified expertise since the quality and representativeness of the training data can determine the fairness of the algorithm outcomes.

Algorithms can also be fed with incomplete or partially incorrect information due to, for example, inadequate or insufficient datasets[124]. The correlations reached by the algorithm and the outputs offered on the basis of them would be equally incomplete or incorrect. Whether intentional or unintentional or caused by technical limitations, these events can also result in spontaneous or unexpected results.

Considering algorithms obscure is also based on the idea that the rationale behind their working may not be apparent, or if possible, it could be highly difficult to explain simply and straightforwardly.

In this context, the opaqueness of algorithms becomes more problematic if considering the imbalance that can emerge from it. Although the obscurity of algorithms affects both the developers and users and the individuals affected by them equally, the former will tend to know their products and systems in a better way than the latter would ever possibly do[125]. Furthermore, due to the need for datasets and personal information for the algorithms to work, developers and users may even know the individuals more than people know themselves, at least in the areas that concern that particular algorithmic process, i.e., their consumer habits, work performance, or ability to repay a loan. Individuals may not even know which concrete personal information is collected, pre-

[123] Suad A Alasadi and Wesam S Bhaya, 'Review of Data Preprocessing Techniques in Data Mining' (2017) 12 Journal of Engineering and Applied Sciences 4102; Faisal Kamiran and Toon Calders, 'Data Preprocessing Techniques for Classification without Discrimination' (2012) 33 Knowledge and information systems 1.
[124] Thieme (n 37).
[125] Mateusz Grochowski and others, 'Algorithmic Transparency and Explainability for EU Consumer Protection: Unwrapping the Regulatory Premises' (2021) 8 Critical Analysis of Law 43, p.47.

processed, and processed or how it is done[126]. This informational imbalance is prevalent among algorithmic systems used in everyday practices. The potential economic benefit that companies might gain when using algorithmic automated decisions while keeping them intentionally opaque for their customers may also be one of the causes that perpetuate their inscrutability[127]. Likewise, the cost of developing methodologies and techniques that make algorithms less inscrutable and more explainable can also limit the resources allocated to provide more transparency[128].

Ultimately, due to this obscurity and complexity, individuals might not know why they were offered an opportunity or excluded from a certain service. They might also be unable to understand how the social (everyday) dynamics and relationships they are part of are configured and framed, limiting and restricting their participation and involvement. It is also very likely that this will lead to a decrease in their trust[129] in the system and a potential decline in the benefits and gains they could have enjoyed or received. The other side of the coin, however, is that algorithms' inscrutability would also allow for -generally unintentional- unfairness and illegal practices of users of algorithmic systems, where they would be difficult to identify, let alone challenge.

The lack of interpretability of algorithmic systems can be considered to present a threat to human dignity. The use of systems that do not reveal the rationale behind their decision undermines the capacity of humans to understand the systems, reducing their possibilities of observing and, if necessary, exercising control over them[130]. Algorithmic decision-making may threaten human dignity as they force humans to adapt their conduct and lives to -and to be at the service of - inflexible and inscrutable technologies. Specifically, algorithmic systems should 'be at the service of human self-

---

[126] Christoph Schmon, 'AUTOMATED DECISION MAKING AND ARTIFICIAL INTELLIGENCE - A CONSUMER PERSPECTIVE' p.9.
[127] Alexander Buhmann, Johannes Paßmann and Christian Fieseler, 'Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse' (2020) 163 Journal of Business Ethics 265.
[128] Grochowski and others (n 125) p.48; Jean Dessain, Nora Bentaleb and Fabien Vinas, 'Cost of Explainability in AI: An Example with Credit Scoring Models' in Luca Longo (ed), *Explainable Artificial Intelligence*, vol 1901 (Springer Nature Switzerland 2023) <https://link.springer.com/10.1007/978-3-031-44064-9_26> accessed 13 December 2023.
[129] Kaur and others (n 42).
[130] Luciano Floridi and others, 'AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations' 28 Minds and Machines 689-707

determination and foster societal cohesion, not undermining human dignity or human flourishing'[131]. In essence, black-box systems appear to threaten human dignity because, due to their opaque and unintelligible nature, it is difficult to ensure that their design, assessment, and deployment are approached from a 'tolerant care and fostering respect for people (both as individuals and as a group), their cultures and their environment'[132]. Otherwise, humans would lose self-worth as well as the systems their legitimacy.

2.2.4. The black-box problem

In computer science, the so-called *black-box problem* refers to the -frequent- problem of not knowing how and why an algorithm offers a particular output or makes a concrete decision. The problem is commonly associated with three dimensions of black-box algorithms:

(1) the high number of features or variables that characterises algorithmic dataset,
(2) the high complexity of algorithmic codes, and
(3) the almost -humanly- unpredictable way in which the internal decision-making logic of the algorithm may work -and vary[133].

Based on the assumption that the problem arises from the inner inscrutability of *black-box* algorithms, technical scholarship addresses the issue from the perspective of making algorithms understandable through external means, i.e. XAI and XAI methods. In essence, XAI aims to achieve two objectives; (1) 'produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)'[134] and (2) 'enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners'[135]. Both goals are not easy, so different concepts shape the landscape and contribute to the field of XAI. I do not dwell

[131] ibid p.694
[132] Corinne Cath and others, 'Artificial intelligence and the 'good society': the US, EU, and UK approach' 24 Science and engineering ethics 505-528 p. 21
[133] Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' 3 Big Data & Society 2053951715622512
[134] Matt Turek, 'Explainable Artificial Intelligence' (*DARPA Defense Advanced Research Projects Agency*) <https://www.darpa.mil/program/explainable-artificial-intelligence> accessed 22 September 2022.
[135] ibid.

on XAI in this chapter[136], however; I point out here that XAI focuses on 'opening the black-box' of inscrutable algorithms, leaving white-boxes out of the field's consideration and concern.

The intentional exclusion of some types of algorithms from this black-box problem, whereas it might be logical from a technical perspective, may also be problematic from a legal and social point of view. As we presented in the two previous sections, algorithms -whether black or white boxes- have a series of characteristics intrinsic to their nature, which will undoubtedly impact the processes to which they are implemented. When considering algorithmic decision-making, the alleged neutrality of the algorithm, as well as its high mathematical complexity, will affect the process and the outcome obtained from it. Even if the starting point of white-boxes may be considered more advantageous in terms of intelligibility, and without denying that the use of external techniques to approximate the behaviour of the black-boxes -using XAI methods- creates another set of challenges to add to the already controvert qualities of algorithms, this thesis argues that the black-box problem is more than an issue only concerning inscrutable algorithms. In my understanding, the black-box problem arises when an algorithm is included in a decision-making process; as a result, the -potentially problematic- normative features of the algorithm are equally introduced in the process and the resulting decision. The difficulties of straightforwardly understanding the workings of the black-box undoubtedly aggravate this problem but do not justify the exclusion of white-boxes from our reflection and debate. Furthermore, understanding the black-box problem through this extensive lens forces us to deem algorithms as 'an intentional product that serves a particular goal, or multiple goals in a given domain of applicability'[137]. A solution for this understanding of the black-box problem would entail the justification of the algorithm's use and design and each individual decision resulting from it. However, I will not advance on my conclusion as Chapter 3 and Chapter 4 address the black-box problem from a doctrinal and normative perspective.

[136] I present a more extended overview of the different existing methods and approaches in *Chapter 6*.
[137] Michele Loi, Andrea Ferrario and Eleonora Viganò, 'Transparency as Design Publicity: Explaining and Justifying Inscrutable Algorithms' (2021) 23 Ethics and Information Technology 253.

Nonetheless, from this point onwards, when referring to a concept such as system, process, or decision, with the particularity of being a black-box, I will be denoting this extensive understanding of the black-box problem that encompasses any type of algorithm and its normative features. To avoid possible misunderstandings, I will explicitly specify it in the concrete cases where I need to refer to a technical black-box or distinguish between types of technical inscrutability.

As explained beforehand, the subsequent section presents several real cases of ADM which have occurred in European countries. From the facts of the cases, it is not possible to know or deduce whether the algorithmic systems involved were technically white or black boxes. Without jumping to a hasty conclusion, the apparent lack of interest for the courts and DPAs in the technical distinction between white and black boxes can be seen as an early indicator in favour of my extensive interpretation of the black-box problem.

2.3. The Algorithmic Controversies in Courts

The following discussion shows the type of legal disputes that have been triggered by algorithmic decision-making processes, the values challenged by them, and the legal regimes implicated in their resolution. The cases highlight that algorithmic decision-making processes do not just fall within the realm of data protection law, but also challenge employment law, anti-discrimination law, surveillance and privacy law.

Furthermore, these cases highlight that the problem of algorithmic system does not merely rely in their inner technical inscrutability, but in the opaqueness and lack of neutrality that they bring to the whole context where they are deployed. The cases below will also offer an overview of the effects ADM can have on individuals, and the imbalances they might create between the users of the algorithmic systems and the people impacted by them.

2.3.1. Workplace and algorithmic management systems

Algorithmic management refers to a the use of algorithmic techniques and systems to remotely manage workforces, relying on data collection and surveillance of workers to

enable automated or semi-automated decision-making[138]. More specifically, algorithmic management refers to systems encompassing;

1) The collection or creation of any information (whether [personally] identifiable or not) with a view to organising, monitoring, supervising, or evaluating work performance or behaviour; and/or
2) the use of that information to support, augment, or fully automate decisions that affect working conditions, including access to work, earnings, occupational safety and health, working time, promotion and contractual status, and disciplinary as well as termination procedures[139].

The options offered by this technology are multiple and diverse, from the optimization of logistics routes and deliveries times to the remote tracking and management of employees, or the organisation of their schedules and the evaluation of their promotion through consumer-sourced rating systems. The purpose of algorithmic management is to improve the production and service models of companies alongside different types of industries and businesses. For instance, by incorporating techniques and tools in the work schedule it becomes possible to structure, on a micro-level, the conditions of works remotely.[140] Algorithmic management include systems and tools of different degrees of complexity, but whether combining algorithmic management with existing practices or replacing them, most of the features have the peculiarity of expanding the scope, scale, and purpose of workers surveillance. Incidentally, they also alter the power dynamics between the employee and the employer, as well as the relationship between customers and providers, and companies and regulatory agencies[141]. Algorithmic management brings benefits to the employers, and often alter, for better or worse, business models and economic and legal structures.

[138] Min Kyung Lee and others, 'Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers', *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM 2015) p.1 <https://dl.acm.org/doi/10.1145/2702123.2702548> accessed 11 November 2023.
[139] Jeremias Adams-Prassl and others, 'Regulating Algorithmic Management: A Blueprint' [2023] 14 European Labour Law Journal (forthcoming) 126 <https://ssrn.com/abstract=4373355>.
[140] Lee and others (n 138).
[141] ibid; Adams-Prassl and others (n 139).

Even though algorithmic management systems are becoming omnipresent, many of these tools were designed by companies of the so-called *sharing* or *gig* economy, which 'involves the exchange of labour for money between individuals or companies via digital platforms that actively facilitate matching between providers and customers, on a short-term and payment by task basis'[142]. Gig companies are also well-known for their strong reliance on temporary and part-time worker positions, usually filled by independent contractors and freelances, rather than full-time permanent employees[143]. The tendency of gig companies to classify their workers as independent contractors even when they exert control over such *independent* workforces consistent with an employment relationship has given rise to legal controversies.

The issue gained importance after several Member states started to legislate on the matter, coinciding with the disputes about algorithmic management and gig companies before national courts and Data Protection Authorities (DPA) in Europe. In January 2020, the Supreme Court of Italy confirmed a Turin appeal decision[144] whereby riders working for the food delivery app service Foodora would fall within the scope of employee-like protection and workers' rights under Italian legislation[145], despite being ostensibly self-employed. The Supreme Court asserted that the workers were organised 'from outside' and so suffered the disciplinary power of the employer, which justify the employee-like protection even in a self-employment context. In the same year, the Palermo Tribunal reinstated a Glovoo rider and reclassified him as a full-time, permanent employee, to be remunerated according to the collective bargaining agreement for the service sector, on the grounds that his autonomy was merely notional, since the platform could organise the execution of work and discipline noncompliance with rigorous instructions issued through the internal booking system[146].

[142] Department for Business, Energy & Industrial Strategy, 'The Characteristics of Those in the Gig Economy' (2018).
[143] Lee and others (n 138).
[144] *Sentenza Foodora - n 778* (Tribunal di Torino).
[145] *Cass sez lav n 1663/2020 (Foodora)* (Corte Suprema Di Cassazione, Sezione Lavoro).
[146] *Sentenza n 3570/2020, causa civile n 7283/2020* (Tribunal di Palermo Sezione Lavoro).

Similarly, in 2021 the Spanish Riders' Law[147] established a presumption of employment to any delivery provider whose working conditions are determined using a digital platform and algorithmic management system. The Spanish law also compels for certain transparency requirements concerning the algorithms used to exert control over the workers. Under the presumption established in the Spanish Riders' Law, any worker for a gig company, unless proved otherwise, would benefit from the Spain's Workers Statute Law and enjoy the same protection as any other employee. The Spanish Riders' Law consolidated the view of the Spain's Supreme Court[148] which had already ruled in 2019 that Glovo's riders need to be considered employees, not self-employed workers or economic dependent self-employed workers, a third category existing under the Spanish law. In its ruling, the Supreme Court highlighted the importance of adapting the workers classification in accordance with the new economic and societal dynamics. The Court explained that Glovo established and controlled all aspects related to the form of the provision of service while the riders are subject to comprehensive micro-management through Glovo's platform. According to the judgment, Glovo is not a mere electronic intermediary in the contracting of services between the business and the riders, but it carries out the coordination and organization of its services, exercising effective control and management over the riders through its digital platform.

As can be observed in the Spanish and Italy legislative modifications, the inclusion of algorithmic management systems was one of the causes of change in the new legal treatment of gig workers and the consolidation of modern employee authorities.

The collection and processing of personal data for the purpose of algorithmic management made it inevitable that ADM in the workplace has become a prominent issue for DPAs and courts.

After the Italian Supreme Court ruling of 2020 conceding delivery riders workers' rights, the Italian and Spanish DPA started several investigations concerning the handling of

[147] Ley 12/2021, de 28 de septiembre, por la que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el Real Decreto Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales (Ley Rider) 2021 (BOE-A-2021-15767).
[148] *STS 2924/2020* [2020] Tribunal Supremo, Sala de lo Social 805/2020 ECLI: ES:TS:2020:2924.

employees' data by the food delivery company Foodinho which operates in both Members labour markets and is owned by the Spanish-based holding company GlovoApp23. The DPAs concluded that Foodinho used algorithms to opaquely micromanage platform workers' labour, and in particular (a) the assignment of slots to its riders; (b) their possible automated exclusion from the platform by rating their performance on the basis of information such as the riders' communication with Foodinho's customers; (c) their real-time geolocalization; (d) estimated and actual delivery times; (e) details of the management of previous and ongoing orders; (f) feedback from customers and partners; (g) the percentage of orders each rider accepted; and (h) how long it took them to accept each order. The Italian DPA ascertained that, although the digital platform used algorithmic profiling to automate assigned slots to their riders, it failed to communicate any information regarding the collection and use of the data to their riders[149]. Foodinho was, therefore, managing the workflow of their riders to the extent that they could be excluded from the platform, analogous to a disciplinary action and even to dismissal of an employee, but as it occurred through automated decisions-making traditional protective rules were side-lined.

A similar situation was identified by the Italian DPA in another case which involved Deliveroo, and its algorithm called *Frank*, which was indispensable to manage its contractual obligations with its riders. The DPA observed that 'the company carries out the processing of personal data of the riders in the context of an employment relationship concerning the transport of food or other goods from restaurants or other partner merchants, though the use of a digital platform'[150]. The DPA highlighted the seriousness of the effects produced by Frank on the riders as it could result in the exclusion of Deliveroo workers from the digital platform or the reduction of their job opportunities. Frank's algorithm automatically ranked and assigned riders to certain delivery slots based on the riders' previously manifested availability in critical time slots and their reliability regarding that manifest availability, i.e., whether they accepted or

[149] *Ordinanza ingiunzione nei confronti di Foodinho s.r.l n 9675440* (Il Garante per la Protezione dei Dati Personali).
[150] *Ordinanza ingiunzione nei confronti di Deliveroo Italy s.r.l 9685994* (Il Garante per la Protezione dei Dati Personali).

refused offered services and actually participated or not in their booked slots[151]. In other words, Deliveroo used Frank as a tool to manage its employees and grant or deny them access to different job opportunities.

Following the finding of the DPA on Frank processing of Deliveroo's data riders and subsequent employees' management, the Labour division of the Italian Bologna Court found that Frank's profiling system did not take into consideration that the absence of Deliveroo riders during their manifest available hours could have reasonable and legitimate reasons. The Tribunal asserted that Frank penalised riders when they were not available or cancelled a given service, irrespective of whether they have a trivial or legitimate reason to do so, such as sickness or participating in a strike action[152]. Regardless of the intentionality behind the wrongdoing, Frank decided the allocation of rides on the grounds of factually incorrect or incomplete information about the riders, making the decision based on that assessment unreasonable and unfair. Furthermore, Frank's profiling of Deliveroo riders and the automated decision upon the allocation of rides resulted in a discriminatory loss of access to job opportunities and less favourable positions for riders who were merely exercising their labour rights[153].

In September 2021 the Amsterdam District Court dealt with three similar cases involving the ride-hailing service companies Uber[154] and Ola[155]. In one of the rulings, the First Instance Court characterised the drivers who offered their services on Uber's platform as the company's employees under the relevant labour law[156]. Similar to the employer-employee relationship identified in the above cases, Uber was recognised to exercise the power and prerogatives of an employer through their algorithms, i.e. assigning rides to drivers and determining the payment obtain for each ride. Likewise, the Amsterdam District Court determined that drivers would be automatically excluded for future rides and logged off the platform if they cancelled previous rides, in a manner that resembled the subordination of an employee to their employer and the disciplining

[151] ibid.
[152] *Cass sez lav n 2949/2019 (Deliveroo Italia SRL)* (Tribunale Ordinario di Bologna Sezione Lavoro).
[153] ibid.
[154] *C/13/692003/ HA/RK 20-302* [2021] Rechtbank Amsterdam ECLI:NL:RBAMS:2021:1018; *Case 8937120 CV EXPL 20-22882* [2021] Rechtbank Amsterdam ECLI:NL:RBAMS:2021:5029.
[155] *Case C/13/689705/HA RK 20-258* [2021] Rechtbank Amsterdam ECLI:NL:RBAMS:2021:1019.
[156] *Uber - Automated termination contract* (n 154); *Uber - employment relationship drivers* (n 154).

and instructive effects that arise from a modern employer authority[157]. Importantly, the algorithmic micro-management of an employee in a workplace context may impact on the employee's rights and freedoms.

In a second ruling, the Amsterdam District Courts reasserted that ADM can have severe effect on the individual's (here the drivers) access to job opportunities and their remuneration. The Court restated that algorithmic management systems determined the match between a rider and a particular client, allocation that initially would not hold any significant legal effect on the riders. The Court asserted, however, that on a long-term basis and under certain circumstances, as could be the reasons behind why specific rides are adjudicate to particular riders, the match between rider and client could become similarly legally relevant[158]. In this regard, Uber and Ola algorithmic micro-management are typical scenarios where ADM can be used in a workplace in a way which may appear harmless at first sight and only upon further scrutiny reveals exploitative practices that attract, or should attract, protective legal regimes. This view was reinforced in a third case only involving the drivers of the ride-hailing service company Ola where the Court upheld that

> [A]utomated decisions to impose fare deductions and/or fines on its drivers on the basis of the performance data it collects about them have effects that are important enough to deserve attention and that significantly affect the conduct or choices of the person concerned. [...] Such a decision leads to a penalty which affects the rights of [the applicants] under the agreement with Ola[159].

In this case, algorithmic automated decisions were not merely used to assign rides to suitable riders but involved a disciplinary aspect that potentially resulted in a decrease in their remuneration and economic penalties directly linked to their performance. The Court asserted the capacity of algorithmic systems to influence people's behaviours on the assumption that certain actions would have some foreseen consequences. Individuals' autonomy and free choice (here the drivers) would have been restricted

[157] *Uber - Automated termination contract* (n 154); *Uber - employment relationship drivers* (n 154).
[158] *Uber - Automated termination contract* (n 154).
[159] *Ola - automated detection fraud* (n 155) unofficial translation, para 4.51.

upon the inclusion and use of these systems as individuals adjust their behaviour to these systems that ensure the most advantageous outcome or to allow the negative consequences to be reduced as possible with those at their disposal. These uses of algorithmic decision-making are not innocuous or indifferent to the rider but could severely affect their livelihood.

2.3.2. Finance services – credit score

Banks and financial entities are well acquainted with the evaluation of risk, management of economic uncertainty and the maximisation of revenue. The processing of large quantities of data to minimise that uncertainty and risk -particularly in the case of the handling of cash, credit, and other financial transactions for individual consumers and businesses- is, therefore, part and parcel of their daily commercial practice. The use of ADM, including profiling, became the logical and expected step in the financial setting to generate significant incremental value for customers, partners, as well as the banks and other financial institutions[160]. Algorithms could be used to improve the customers experience with personalised messages, increase the lifetime value of customers, lower their operating costs, or lowering the credit risk by more accurately detection behaviours that signal higher risks or potential for fraud[161]. These latter use have, in recent years, been subject to scrutiny in several EU member states, particularly in regard to credit scoring companies that offer their services to banks and other customers alongside the territory of Europe.

In a case before the Supreme Court of Cassation, several individuals had voluntarily uploaded documents containing personal data to an online platform managed by the Associazione Mevaluate Onlus, a company that specialised in developing reputational profiles concerning natural and legal persons. The service provided by the Associazione

[160] Layla Abdel-Rahman Aziz and Yuli Andriansyah, 'The Role Artificial Intelligence in Modern Banking: An Exploration of AI-Driven Approaches for Enhanced Fraud Prevention, Risk Management, and Regulatory Compliance' (2023) 6 Reviews of Contemporary Business Analytics 110.
[161] Akshat Agarwal, Charu Singhal and Renny Thomas, 'AI-Powered Decision Making for the Bank of the Future' [2021] McKinsey & Company.–2021.–March.–URL: https://www. mckinsey. com/~/media/mckinsey/industries/financial% 20services/our% 20insights/ai% 20powered% 20decision% 20making% 20for% 20the% 20bank% 20of% 20the% 20future/ai-powered-decision-making-forthe-bank-of-the-future. pdf (дата обращения 15.04. 2021).

aimed primarily at the creation of 'impartial, reliable, and objective' alphanumeric indicators capable of measuring the reliability of individuals in the economic and professional fields for the benefit of the Associazione customers, i.e. counterparties such as contractors and subcontractors, suppliers, distributors, business partners, aspiring employees, employees in force and customers. The uploaded documents would be evaluated, and the system would calculate an overall score to be assigned and sent to the interested parties. This score would vary over time and in accordance with the elements transmitted resulting in five sub-ratings or categories: criminal, tax, civil, work and civil commitment, and studies and training.

The case in hand revolved around the validity of the consent given by the individuals in respect of their uploaded documents; or more specifically whether or not adhering to the platform's terms and voluntarily uploading documents manifests an individual's consent to the ADM, and the profiling, used to arrive at their reputational rating. In 2021 the Supreme Court of Cassation concluded that this consent to join the platform and upload document did not directly manifest consent to the subsequent automated processing, including profiling, given that the individuals were not aware of the algorithmic executive scheme, and neither its underlying logic nor its constitutive elements[162]. The creation of reputational scores for legal and natural persons has the aim of making socio-economic relationships more efficient, accurate, and secure. However, this case evidenced that these same systems also presupposed the collection of large and varied types of personal data which would likely significantly affect in different degrees and relevance the economic and social representation of a large audience of subjects (customers; employees; candidates; entrepreneurs; freelancers; suppliers; citizens; etc.). The reputational rating, in fact, could have a serious impact on the life of the individuals surveyed, influencing choices and prospects and conditioning their own admission to, or exclusion from, specific opportunities, services or benefits. Give an example the Supreme Court asserted that extreme caution was necessary when dealing with such delicate matters. For the Court,

[162] *Civile Ord Sez 1 Num 14381* (Corte Suprema di Cassazione).

reputation, as purported to be measured in these circumstances, is closely related to how individuals see themselves and thus their own social projection and dignity[163].

In September 2020, the Icelandic DPA asserted that the use of an individual' financial information by Creditinfo Credit Ltd. to evaluate or predict their economic situation and attribute a certain creditworthiness rate must be considered automated profiling. Thus, the Court upheld that profiled individuals are entitled to specific transparency rights such as information about how the credit score is calculated or about the factors that downgrade or upgrade their credit rating[164]. Similarly in March 2022, the Swedish DPA imposed on Klarna Bank AB a fine for not providing information about the ADM used for the purposes of deciding on customers' credit applications in the period between March and June 2020[165]. The DPA asserted that Klarna Bank had not explained to their customers which circumstances may be decisive for a negative credit concession decision. The same lack of information regarding the automated credit decision was also recognised in the use of ADM, including profiling, for detecting potential cases of fraud and money laundering[166].

In 2019 the Finnish DPA ordered the financial credit company Svea Ekonomi to provide credit applicants with information about how its ADM worked, the role it played in the final credit decision, and the consequences that it could have for its customers. Besides its relevance in terms of customers information rights for automated decisions, the Finnish DPA defined a concrete standard for correct data processing in the context of credit applications when it ordered the financial credit company to correct its processing practices related to creditworthiness assessments. In its decision, the DPA stated that the use of an upper age limit - as an automatically excluding factor from having a credit application further analysed - was not acceptable under the definition information set out in the Credit Information Act, as 'the mere age of the credit applicant does not describe their solvency, willingness to pay or ability to deal with their

[163] ibid.
[164] *Vinnsla Creditinfo Lánstrausts - 2020010592* (Persónuvernd).
[165] *Beslut efter tillsyn enligt dataskyddsförordningen - Klarna Bank AB* [2022] Integritetsskyddsmyndigheten DI-2019-4062.
[166] ibid.

commitments'[167]. Based on the account submitted by the credit company, the Finnish DPA concluded that the customer's financial position was not considered in the ADM, nor consequently, in the customer's refusal of credit. This case illustrated how the inaccuracy, or lack of relevance, of the data used in the application processing could exclude individuals from accessing financial services or result in a differential or discriminatory access to them by members of certain groups.

In 2023, the Berlin Commissioner for Data Protection and Freedom of Information (BInBDI) imposed a fine on a bank for lack of transparency in the ADM used to decide the concession of credit cards. In this case, an algorithm analysed the information and automatically rejected the individual's request without any specific justification, despite the fact that the individual had both a good credit score and a high income. The bank refused to provide the individual with information about the reasons for the automated rejection of the credit card application arguing that the algorithm used for the automated individual decision is based on criteria and rules previously defined by the bank. In other words, the bank initially argued that the definition by a human staff member of such criteria and rules would invalidate the consideration of the algorithm as an ADM. When asked by the rejected individual, the bank only provided information about the scoring process. However, the bank did not provide information about the automated rejection, nor the database and factors on which the rejection was based or the criteria on which the credit card application was accordingly rejected. In essence, since none of that information was provided to the data subject, they could neither way understand them. Interestingly, the BInBDI called all these specific information as the 'individual justification' needed by a data subject 'to challenge the automated individual decision in a meaningful way'. In the words of Meike Kamp, Berlin Commissioner for Data Protection and Freedom of Information,

> When companies make automated decisions, they are obliged to justify them in a sound and comprehensible manner. Those affected must be able to understand the automated decision. [...] A bank is obliged to inform customers

[167] Office of the Data Protection Ombudsman, 'The Data Protection Ombudsman Ordered Svea Ekonomi to Correct Its Practices in the Processing of Personal Data' (1 April 2018) <https://tietosuoja.fi/-/tietosuojavaltuutettu-maarasi-svea-ekonomin-korjaamaan-kaytantojaan-henkilotietojen-kasittelyssa>.

of the main reasons for a rejection when making an automated decision about a credit card application. This includes specific information on the database and the decision-making factors as well as the criteria for rejection in individual cases[168].

In 2022, the nature of credit scoring became the subject of two independent requests for a preliminary ruling by the Wiesbaden Court[169] and the Vienna Court[170]. The Court of Wiesbaden sought clarity regarding whether or not the automated creation of a probability value concerning the ability of an individual to service a loan in the future already constitute an automated decision with a particularly significant impact on the individual's rights and freedoms. The importance of this preliminary request lies in the existence of two independent actors benefiting, or making use, of the individual's credit score. In the case at hand the financial credit company SCHUFA Holding S.A, processed the individual's data and provided their credit rating, and an interested bank to whom the customer had applied for a credit would use the score provided by SCHUFA to grant or deny the application. What this and analogous disputes highlight is the strong reliance that some actors place on the assessment and evaluation of personal data from an individual's different spheres of life for their decision-making processes, regardless of whether they or their customers understand the determinant rating or score. Yet such rating can lead to differential access to credit, goods or services, such as insurance, or housing, potentially based on inaccurate or discriminatory information, hidden or unacknowledged by the interested parties.

The Court of Vienna focused its request on the amount and type of information that an individual should be provided with when they are subjected to an automated decision, including profiling. The Court expressed its uncertainty concerning the extent of information offered to the individual about the profiling process and the automated decision. It provided, nonetheless, some examples that could guide the response of the European Court of Justice, like 1) the input data used for profiling, 2) the particular input

[168] *Computer say no - BInBDI* ("Berliner Beauftragte für Datenschutz und Informationsfreiheit).
[169] *C-634/21 SCHUFA Holding (Scoring) - Request for Preliminary Ruling from the Verwaltungsgericht Wiesbaden* [2023] ECJ OJ C/2024/913.
[170] *C-203/22 Dun & Bradstreet Austria GmbH - Request for preliminary ruling from the Verwaltungsgericht Wien* [2022] ECJ OJ C 222, 7.6.2022.

variables used in the particular rating, 3) its origin and influence, 4) the reasons why an individual was assigned that particular rating and the implications of such rating, 5) the possible profile categories an individual could fall under, and 6) the implication associated with each profile categories. Nevertheless, the Court of Vienna seemed to be uncertain about some of the challenges posed by credit scoring in terms of data inaccuracy and relevance. Thus, the Court asked if the information provided to the profiled person should be the information on which the final decision was actually based and if that information should respect some level of accuracy[171].

This case draws attention to the reluctance of some providers to reveal the intricacies of their ADM in order to protect their intellectual property rights and commercial secrets, without taking into account the possible detriment to their customers. Likewise, systems' transparency is not a straightforward matter, as some of the mentioned hesitation to provide accurate and broad information about automated systems relies on the threat that this disclosure of information could pose to the privacy and data protection rights of other individuals - in relation for example to the pseudo anonymization of personal information- and the challenges that explaining a black-box entails.

The determination and evaluation of individuals' creditworthiness could be considered, along with gig-workers' micro-management, the context where most controversies and uncertainties have, so far, arisen due to the use of ADM.

2.3.3. Private protection of the public interest – crowd control and automated facial recognition

Law enforcement bodies and security agencies commonly use different forms of biometrics such as fingerprints, iris, voice, DNA, particularly in investigative and criminal justice settings. The use of these biometrics has grown drastically in the last decades often as in response to terrorism, and violent and serious crimes, or to control borders and frontiers. However, the use of biometrics has increased in many different areas of our lives, from unlocking smart devices, to accessing online banking services

[171] ibid.

or reducing the human touchpoint at the airports' security. Facial recognition technologies are one of the increasingly common form of biometrics, used indistinctly both in our day-to-day activities, as well as in law enforcement and security settings.

Automated facial recognition (AFR) technology is based on 'algorithms that perform a series of functions, including detecting a face, creating a digital representation -or 'temple'- of the face, and comparing this representation against other images to determine the degree of similarly between them'[172]. AFR can be considered an automated decision in so far as the technology performs two main functions -verification and identification- with the purpose of taking a decision either to carry out an action or to remain inactive. The verification function aims to confirm the identity of the individual by comparing it with a single stored image in a one-to-one basis[173]. Identification, on the other hand, compares an image to many images in a dataset to find a match to the target image[174]. Verifying whether your face matches the image on your passport in an automated border control gate is an example of the former, whereas, confirming that neither your face nor your passport photograph appears in any criminal or terrorist list are examples of the latter. Logically, AFR opens a wide range of possibilities in more quotidian settings which has prompted a public debate on their adequacy and proportionality.

In 2019 the Danish DPA received a request for approval from Brøndbyernes I.F. Football A/S -hereafter Brøndby- for the processing of biometric data with the intention of control access at the Brøndby Stadium by means of automated facial recognition. This request for the processing of sensitive data was justified by Brøndby on the basis of that facial recognition was necessary to ensure an essential social interest in the security of the Brøndby[175] stadium spectators. Brøndby's AFR wanted to verify whether any of the individuals aiming to enter the Stadium were or were not in quarantine due to a violation

[172] Kay L Ritchie and others, 'Public Attitudes towards the Use of Automatic Facial Recognition Technology in Criminal Justice Systems around the World' (2021) 16 PLOS ONE e0258241.
[173] Erin Sullivan, 'Facial Recognition Technology: Verification vs. Identification' (Montana State Legislature, Economic Affairs Interim Committee 2021) <https://leg.mt.gov/content/Committees/Interim/2021-2022/Economic%20Affairs/Meetings/November%202021/Facial-verification-vs-facial-identification.pdf>.
[174] ibid.
[175] *Tilladelse til behandling af biometriske data ved brug af automatisk ansigtsgenkendelse ved indgange på Brøndby Stadion* (Datatilsynet meddeler).

of the rules of procedure for Brøndby IF and the Super League. The use of the requested AFR extended to both football and training matches with the participation of teams from the super league, the 1st and 2nd divisions as well as at football matches under the auspices of The Union of European Football Associations (UEFA) at Brøndby Stadium.

This case revolved around the necessary and proportionate legal grounds that could justify, and ensure the lawfulness of, the processing of biometric data for facial recognition purposes. The Danish DPA took the view that the processing of match attendants' sensitive data through AFR would be necessary and proportionate to attain the objective of substantial public interest even though all the stadium ´s spectators would be subject to a AFR process without their explicit consent. The permission to install and use AFR at the stadium gates would irrevocably lead to increased surveillance but leave in doubt the emotional cost and the impact on people's dignity and social repercussion it would entail.

Notwithstanding these unresolved concerns, the Danish DPA specified a series of concrete measures that would need to be observed. Firstly, all sensitive information must be removed, not stored, after the verification that there was not any match within the dataset of banned spectators. Secondly, Brøndby shall comply with the pertinent disclosure obligations when collecting personal data. Additionally, Brøndby must provide signage or other clear information that access checks were carried out, including the use of AFR systems. Thirdly, Brøndby shall transfer and store on a server with up-to-date and widely recognized encryption algorithms any personal information processed as part of the AFR system. In essence, the Danish DPA authorised the use of AFR asserting that its lawfulness was based on an existent substantial public interest but required Brøndby to ensure specific safeguards to be in place to consider the processing lawful and guarantee the protection of the spectators' rights.

In total contrast with this decision, the Spanish chain of supermarkets, Mercadona was denied the permission to install and use an AFR in their establishment with the intention of preventing the entry of two judicially banned individuals[176]. Mercadona ´s request of

[176] *Auto 72/2021* [2021] Audiencia Provincial Penal de Barcelona Seccion 9 Rec 840/2021 ECLI: ES:APB:2021:1448A.

permission for the use of an AFR system needs to be contextualise under the light of a previous sentence from the Criminal Court of Barcelona that condemned two persons as perpetrators of an attempted crime of robbery with violence against people and prohibited their entry to one particular Mercadona supermarket for the period of two years. In 2019, Mercadona claimed the factual impossibility to ensure compliance with said sentence due to the incapacity of its supermarket workers to be constantly aware of the people who enter the supermarket chain premises, much less if they had been convicted and banned from the particular establishment. Under these circumstances, Mercadona requested permission to use electronic means, consisting of a closed circuit of video recording, with the objective of detecting the entry of the two mentioned convicted individuals to its establishments. In the permission request, Mercadona proclaimed the alleged proportionality and necessity of such system under the premise of ensuring the enforcement of the criminal sentence and sustained that the legal grounds for the processing were based on the principle of public interest expressed in the Spanish Private Security Law and the legitimate interest of the commercial entity.

After the Criminal Court rejected the petition, Mercadona contested this decision to the Court of Appeal without success. The Court of Appeal held that Mercadona had no legal ground to process personal data for the purpose of automated facial recognition. The Court dismissed the petition upholding that Mercadona intended to use a AFR system to identify natural persons, a purpose that is in principle prohibited in the General Data Protection Regulation (GDPR). Additionally, the Court of Appeal asserted that any public interest justifying the processing of personal categories of data would have to be grounded on a specific law, which for the case in hand (here AFR) was currently not present in any Spanish law. In the absence of such public interest, Mercadona would be required to ask for the consent of the individuals intended to be identify, which in the given case would never be freely offered as it will be a precondition to enter the supermarket establishment.

Despite this negative resolution for the petition of use of AFR, in 2021 the DPO opened an investigation against the supermarket chain for the alleged use of these systems in their premises throughout Spain. In views of the news published in the Spanish media, Mercadona would have acted against Barcelona Criminal Court's ruling and deployed

and tested AFR systems with the purpose of detecting persons with sentences and restraining orders in force against Mercadona or any of its employees. The investigation ended up in a multimillionaire fine against the supermarket chain company as the allegations were verified. In its decision, the Spanish DPO reasserted that the private interests of the company cannot be considered a lawful justification and ground for the processing of biometric data with AFR purposes. The Court equally stressed that AFR systems could not be used to enforce any previous judgement, unless it explicitly identifies technological means as suitable measures to achieve such purpose. In the absence of lawful ground for the AFR processing, Mercadona would need to have requested explicit consent to the individuals affected by the AFR. The Court confirmed, nonetheless, that Mercadona has not proceeded as such, thus failing in its duties towards them and effectively preventing them from exercising their rights.

Although the circumstances surrounding Mercadona's cases differ substantially from Brøndby's, the cases refer to the same lawful ground for the processing of biometric data, being this the protection of the public interest. The differing resolution of the cases, however, suggests that the processing of such sensitive data with the consequent potential loss of liberties for the individual and the increased surveillance requires a more detailed assessment of interests than it may appear at first glance.

I will include in this section an Amsterdam District Court case involving Twitter International Company -hereafter Twitter- and the practice to restrict users' visibility (i.e. shadow banning) in the social media platform. In this specific case, Twitter temporally restricted the account of a data subject -shadow banned- for posting a message that included the word 'child pornography'. The message reads

> The chats of hundreds of millions of people will soon be scanned to detect a relatively small number of criminals, no matter how bad. Strong criticism of European plans against child pornography: 'Not proportionate'' [link to a newspaper article][177].

[177] *C/13/742407 / HA RK 23-366* [2024] Rechtbank Amsterdam ECLI:NL:RBAMS:2024:4019.

Twitter automatically detected the post as a potentially violation of their policy to combat child abuse and sexual exploitation and imposed a restriction in the user's account meaning for which the account and its posted messages temporarily did not appear in searches. Although the user was not notified of this restriction nor the motives behind it by Twitter, he learnt of the situation through third parties who could not find him in the platform. The user, then, proceeded to request a general information access regarding the information processed about himself by Twitter with relation to the search and search suggestion ban. Later on, the request of data access was supplemented, requesting specific information in the context of the restrictions on the functionality and/or reach and/or visibility of the account and posts, that the user had on the platform. That request of information included

- T]he origin and source of his personal data.

- A list of the recipients to whom his personal data have been or will be discreet, including the safeguards put in place for this purpose.

- The identity of all (joint) controllers of his personal data.

- A full copy of the personal data that are or have been processed about him.

- Whether automated decisions (including profiling) are or have been made, and if so:

    - which automated decisions have been made;

    - the logic behind these decisions;

    - the importance and expected (duration of the) consequences of these decisions to me;

    - what measures have been taken to prevent errors, bias and discrimination; and

    - explanation of how I can explain my position and challenge these decisions.

- Any other information necessary to ensure proper processing, as required under recital 60 GDPR[178].

After the request for information access was made, Twitter lifted the restriction to the user's account but did not notified him about the change. Correspondence between Twitter and the user was shared with no agreements between the parties on the amount and type of information Twitter shall provide the user about his former account restriction. The user brought the case to Amsterdam First Instance Court in regard to his request for access within the meaning of Article 15 of the GDPR and his request to obtain information on the ADM within the meaning of Article 22 of the GDPR.

I include the case in this section because the user argued that when Twitter's account shadow banning is due to possible child abuse, the platform automatically reports it to the National Advocacy Group for Consumer Protection and Corporate Fair play (NCMEC). This sharing of data would entail relevant and serious consequences for the individual in regard to their right and freedoms, and a clear attempt of private enforcement for the protection of social interests. The allegations were made on the basis of one email from the former CEO of Twitter, Elon Musk. However, Twitter has repeatedly contested such email and the alleged NCMEC's automated report. In this particular case, the mentioned Elon Musk's mail was subsequent to the user's account restriction and, to the Court Opinion, insufficient to assumed that such report was made. Notwithstanding this decision, the user's based part of his Court request on the assumption that such report was made and that his interest and freedoms were significantly impacted without notice nor information from Twitter.

After assessing the facts, the First Instance Court of Amsterdam sentenced Twitter to comply with the user's general request of access as referred to in Article 15 (1) of the GDPR. Particularly, the Court highlighted that Twitter shall have informed the user about the restriction, the existence of an automated decision, its underlying logic and its importance and expected consequences for the user. By not doing so, the Court asserted that Twitter failed to offer the pertinent information to the users in such a way

[178] ibid para. 2.3.1.

that 'the context of the restriction was unclear and verification of correctness and legality was not possible for[applicant]'[179]. Furthermore, the Court also sentences Twitter to address the ADM as referred to in Article 22 of the GDPR. In doing so,

> Twitter must provide information that is useful for [the applicant] to challenge the decision, such as information about the factors considered in the decision-making process. The information must be complete enough for [applicant] to understand the reasons for the decision.[180]

2.4. Discussion

ADM, including profiling, has become a practice of our everyday life. It is used to decide upon our interests, rights, and freedoms in context that have high-consequences for our participation in society and the achievement of our life goals. More and more actors not only use these systems, but are totally dependent on them, either because they base their business model on the use of these systems as it is the case of Glovo or Uber, or because they consider that no other means could perform the assigned tasks under the current circumstances, as alleged by Mercadona or Brøndby. The reality is that automated individual decision-making has shaped, and still shapes, many spheres of our lives. The gig economy cannot be understood nor developed without the use of algorithmic management systems, as the monitoring of all the content existent in the digital platforms could be a very tedious and difficult job to do only by humans. Banks and financial institutions have being using simple algorithms for more time that we might be aware of, but their current dependency of them is almost palpable. ADM have enhanced the accuracy and effectiveness of decision-making processes along multiple sectors, created new business models, and improved old ones. They should not be perceived as intrinsically bad.

However, ADM come with some pitfalls. The intrinsic lack of neutrality of algorithms is brought to the decision-making process in the same manner that their complexity and inscrutability. If the position of the individual in a high-consequence decision is usually

[179] ibid para. 4.25.
[180] ibid para 4.27.

disadvantageous, introducing those factors in the equation does little to improve the situation.

When reconsidering the real cases of ADM, including profiling, presented in this Chapter, we can agree that most, if not all, of the cases have an element in common; individuals could not avoid being subject to certain type of private institutions if they want to participate in our current society. Without undermining the role of consent and lawfulness for data processing practices as they alone could entail another PhD thesis, it is undeniable that our current society requires -to some degree or another- the collection, processing, and use of one's personal data to access a wide number of services and products, for example, credit or an employment.

In this Chapter I provided an analysis of the use of ADM in our everyday and high-consequence decisions and the problematics associated with their promised neutrality and inherent inscrutability as well as the so-called problem of black-box systems. The conclusion I reached after this analysis is that real cases of ADM, including profiling, as referred to in Article 22 of the GDPR -with legal or significantly effects on individuals' rights and freedoms- do not merely impact on their right to personal data protection, quite the contrary. The selected case-law show how even if the person's personal data was affected or impacted as a result from the automated processing, the concerns and threats arisen from such processing distant in a great manner from privacy or personal data processing concerns alone. In fact, the selected case-law reflects the possibilities in which utilizing modern data-driven technologies can have an undesirable or unrequired effect on the individual's participation in society, e.g. their access to employment, credit or social activities as a football match. Still, the right for the protection of personal data remain at the core of the discussion around algorithmic systems and AI systems in general. The own development, training and use of these systems depend on extensive datasets and the use of pertinent or necessary information of the concrete individual affected or impacted by it. Hence, assuming a regulatory path that tackles the most basic and necessary element of the functioning and use of algorithmic and AI systems -i.e. data- seems the most logical decision. After all, the use of algorithms for ADM, is no more than an specific methodology or practice involving the processing of personal data.

Chapter 3: The Doctrinal Framework of the Right to Information and an Explanation

3.1. Introduction

Chapter 3 introduces the legal doctrinal framework of the rights of information and an explanation. To do so, it first presents the right not to be subject to automated decisions, analysing both the Data Protection Directive and the current General Regulation on Data Protection. Chapter 3 seeks to lay the legal foundations for the other chapters of this thesis. It presents the rights to information and an explanation as well as of the challenges or problematics that their wording in the GDPR has brought about. Chapter 3 introduces concepts and ideas that will be evaluated in detail in later chapters, such as the raison d'être of these rights and their connection to the principles of fairness, legality and transparency, the technical challenges that these rights imply in themselves, and the necessity to ensure their effectiveness and feasibility.

In concrete, Section 3.2 outlines the key points of contention regarding the nature and content of the right not to be subject to automated decisions while it also sets out both derogations and safeguards associated with the right that may affect its enforcement by individuals.

Section 3.3 approaches the right to access and information about the existence of an ADM and the right to an explanation about and automated decision assessing Articles 13(2)(h) and 14(2)(g), 15(1)(h) and Article 22 (3), respectively. This section analyses the legal basis that buttress the existence of a right to an explanation, i.e. the safeguard right to contest an automated decision as referred in paragraph 3 of Article 22, and the information and access requirements established in Article 13,14, and 15 which grant individuals a right to information. Likewise, section two also examines the importance of temporality in the effective enjoyment of the rights to information and an explanation about an automated decision addressing the different alternatives in which these rights can be enjoyed or demanded and the legal scenarios and challenges that each one creates. This second section, ultimately, explores the debate regarding the existence,

enforceability and effectiveness of the right to information and an explanation about automated decisions. Section 3.4 reflects on the previous two sections proposing the framework and the spectrum of compliance -minimum and maximum thresholds- to the rights to information and an explanation. Consequently, this closing section will , offer a framework of conclusions for the doctrine, rather than a unique interpretation of the rights, with the aim of facilitating the connection of ideas between the legal doctrinal framework and subsequent more technically grounded chapters.

3.2. The Right To Not Be Subject To Automated Decision-Making, including profiling

3.2.1. The Origin: Article 15 of the Data Protection Directive

The EU firstly transposed its concerns about ADM into the 1995 Data Protection Directive[181] (DPD) which harmonised data protection law across European countries and envisaged for its protective regime to extend across manual and automated decision-making, albeit already foreseeing the rise of ADM, as per Recital 27,

> Whereas the protection of individuals must apply as much to automated processing of data as to manual processing; whereas the scope of this protection must not in effect depend on the techniques used, otherwise this would create a serious risk of circumvention; whereas, nonetheless, as regards manual processing, this Directive covers only filing systems, not unstructured files.[182]

The Directive covered purely machine-based in Article 15:

> 1. Member States shall grant the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate

[181] European Parliament and Council Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data 1995 (Data Protection Directive) 1995 (OJ L281/31) OJ L281/31.
[182] ibid recital 27.

certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc.

2. Subject to the other Articles of this Directive, Member States shall provide that a person may be subjected to a decision of the kind referred to in paragraph 1 if that decision

(a) is taken in the course of the entering into or performance of a contract, provided the request for the entering into or the performance of the contract, lodged by the data subject, has been satisfied or that there are suitable measures to safeguard his legitimate interests, such as arrangements allowing him to put his point of view; or

(b) is authorised by a law which also lays down measures to safeguard the data subject's legitimate interests[183].

Article 15 was the first attempt to regulate purely automated decisions at the European level. Until then, provisions addressing purely machine-based decisions were rare amongst data protection instruments at both national and international level[184]. However, Recital 8, 9 and 10[185] reflected the intention of the Directive to upgrade the protection of personal data across the EU by offering a harmonised framework with the same level of protection for individuals with regard to the processing of their personal data. Whilst Member States were left with a margin for manoeuvre for implementing the Directive, it established the general conditions for lawful data processing[186]. Likewise,

[183] ibid.
[184] Provisions along the lines of Article 15 (1) of the Data Protection Directive could be found in the French Act Section 2 and 3, (*Loi no. 78-17 du 6. janvier 1978 relative à l'informatique, aux fichiers et aux libertés),* Article 12 of the first Spanish data protection law (Ley orgánica 5/1992 de 29 de octubre 1992 de la Regulación del Tratamiento Automatizado de los Datos de Carácter Personal) and Article 16 of the first Portuguese data protection law (Lei no. 10/91 de 12 de Abril 1991 da Proteccão de Dados Pessoais face à Informática); see also Lee A Bygrave, 'Automated Profiling' (2001) 17 Computer Law & Security Review 17.
[185] Recital 8 of the Data Protection Directive stated that "in order to remove the obstacles to flows of personal data, the level of protection of the rights and freedoms of individuals with regard to the processing of such data must be equivalent in all Member States" whereas Recital 9 asserted that EU Member States "shall strive to improve the protection currently provided by their legislation" and Recital 10 which addressed that the "approximation" of Member States' data protection laws pursuant to the Directive "must not result in any lessening of the protection they afford but must seek to ensure a high level of protection in the Community".
[186] Bygrave (n 184) p.17.

the problematic of automating decision-making processes was expressed by the drafters of the Directive, as per the EC Commission,

> [T]his provision [Article 15] is designed to protect the interest of the data subject in participating in the making of decisions which are of importance to him. The use of extensive data profiles of individuals by powerful public and private institutions deprives the individual of the capacity to influence decision-making processes within those institutions, should decisions be taken on the sole basis of his 'data shadow'[187].

Article 15 referred to the right for a person not to be subject to automated decisions[188]. Whereas the DPD did not explicitly use the term profiling, it referred to the 'automated processing of data intended to evaluate certain personal aspects relating to him'[189] which closely resembles to the definitions of profiling provided by several authors, and afterwards included in the General Data Protection Regulation (GDPR)[190], which explicitly mentions the notion of profiling[191]. Hilderbrandt and Gutwirth argued that for a particular data processing to be considered profiling it shall

> Denote the process of (1) inferring a set of characteristics about an individual person or group of persons (i.e., the process of creating a profile), and/or (2) treating that person or group (or other persons/groups) in light of these characteristics (i.e., the process of applying a profile).[192]

The authors, thus, drew a distinction between the process of evaluating certain personal aspects of an individual, and the decision made upon such profiling. In this regard, Article 15 did not prohibit the process itself but an automated decision based on

---

[187] Council Directive Proposal 19990/0287/COD Concerning the Protection of Individuals in Relation to the Processing of Personal Data 1990 (COM(90) 314 final – SYN 287 90/C 277/03, 29).
[188] Data Protection Directive.
[189] ibid.
[190] General Data Protection Regulation.
[191] Profiling as referred in General Data Protection Regulation Article 4(4) means "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements".
[192] Hildebrandt and Gutwirth (n 24).

such processing. Dinant clarified this distinction by explaining that profiling, as referred in Article 15, is a method that allows to 'place individuals, with a certain degree of probability, and hence with a certain induced error rate, in a particular category in order to take individuals decisions relating to them'[193]. In other words, profiling entails a process by which personal data about an individual is used to create a picture of his or her behaviour, tastes, or personality and then used for different purposes, such as decision-making, personalised marketing, or service provision.

Whether Article 15 entailed a default strict prohibition of automated decisions or an individual right to challenge an automated decision and ask for a re-examination of the decision or a human in the loop was left to the interpretation of each EU member state. On the one hand, the wording of the provision 'Member States shall grant the right to every person not to be subject to a decision [...]' could be understood either as requesting for a default prohibition of being subject to automated decisions or for an enforceable data subject right to not be subject to automated decisions. On the other hand, neither the preamble nor the preparatory works made any reference to the nature of Article 15. As European Directives require their implementation into the national legislation[194], the ambiguous and open nature of Article 15 led some discretionary room in its transposition by Member States regarding its form and means as long as it achieves its objectives. However, Directive Recitals 41 and 42 emphasised that in accordance with Article 15, national legislations needed to allow any person to exercise their right to access their processed data in order to verify its accuracy and the lawfulness of the processing. In this sense, the Recitals clarified that any possible limitations to the individual's rights shall have been based on the interest of the data subjects or the protection of others' rights and freedoms. In no case, none of the possible limitations to the rights of the data subjects established in the national legislations could imply a total denial of the right of access.

Moreover, Article 15 itself had several characteristics that made it special compared with other data protection norms. Firstly, it addressed a specific type of decision rather

[193] Dinant and others (n 25).
[194] Consolidated Version of The Treaty on the Functioning of the European Union (TFUE) 2016 (OJ C202/1) art 288.

than the general concept of data processing. Secondly, it embedded a new data subject right that demanded that inferred characteristics or induced categorization of an individual should not be the only basis for a decision[195]. Thirdly, it directly tackled aspects of automated profiling for the first time in European legislation[196].

Besides the right for a person not to be subject to automated decision, the DPD also conceived a right to an information about the logic involved, as per Recital 41,

> Whereas any person must be able to exercise the right of access to data relating to him which are being processed, in order to verify in particular the accuracy of the data and the lawfulness of the processing; whereas, for the same reasons, every data subject must also have the right to know the logic involved in the automatic processing of data concerning him, at least in the case of the automated decisions referred to in Article 15 (1) [...].[197]

Article 12 of the DPD complemented this last aspect by providing a right to obtain information about whether and how their particular personal data was processed, giving individuals the specific right to obtain 'knowledge of the logic involved in any automatic processing'[198] of their data.

These two rights – right to information and right to not be subject to automated decisions- were not specially used by citizens or by lawyers, nor were they litigated before the Court of Justice of the European Union or any national court. However, they regained importance with the latest European data protection law modification, the General Data Protection Regulation (GDPR), when the essence of Article 15 and 12 of the DPD was transposed in Article 22 and Articles 13, 14 and 15 of the GDPR, respectively.

[195] Mendoza and Bygrave (n 45).
[196] ibid.
[197] Data Protection Directive recital 41.
[198] ibid.

3.2.2. The subsequent development in Article 22 of the General Data Protection Regulation

Article 22 of the GDPR. formerly Article 15 of the DPD, states that

> 1.The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
>
> 2.Paragraph 1 shall not apply if the decision:
>
> (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
>
> (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
>
> (c) is based on the data subject's explicit consent.
>
> 3.In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
>
> 4.Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place[199].

As Article 15 did previously, Article 22 now grants the data subject a right to not be subject to ADM. However, Article 22 introduces some new features compared to its

[199] General Data Protection Regulation.

predecessor[200]. Firstly, Article 22 (2) refers to two new exceptions: the necessity to enter in a contract[201] and the consent of the data subject[202]. Secondly, it increases the mandatory safeguards for the exceptions where ADM is accepted explicitly addressing the right to obtain human intervention, to express own's point of view, and to contest the decision[203]. Thirdly, paragraph (4) establishes an additional limitation to automated decisions referred to in paragraph (2), whereby they can only be based on special categories of personal data if the conditions set forth in Article 9 (a) and (g) are met. That is, if data subjects has given explicit consent for the processing of their special categories of personal data or if there is a substantial national interest proportionate to the aim pursued through such processing. If Article 9 (a) or (g) applies, automated decisions can be based on protected attributes but suitable measures to safeguard the data subject's rights and freedoms and legitimate interests need to be put in place, same condition as the one introduced in Article 22(3). Finally, profiling[204][205] is explicitly referred as a type of processing included in the notion of automated processing.

In other words, Article 22 obliges the data controller to not fully automate decision-making with legal or similar effects for the data subject, establishing an individual right to not be subject to such decisions unless one of the exceptions stated in Article 22 paragraph 2 applies. Article 22 again leaves the question open, as did its predecessor, as to whether the right needs to be exercised by the data subject a posteriori of an automated decision took place or whether it is a default prohibition to carry out ADM unless under the conditions established in Article 22 (2). However, given that the GDPR is a Regulation and not a Directive, it is directly applicable in the Member States per Article 288, leaving no residual room for interpretation in the national implementation. What interpretation should be given to this article remains, nonetheless, an important open question.

[200] Isak Mendoza, 'The Right Not to Be Subject to Automated Decisions Based on Profiling - Applied to Examples of Online Scoring Technology, Weblining, and Behavioral Advertising' (PhD Thesis, University of Oslo, Faculty of Law UiO 2016).
[201] General Data Protection Regulation art 22 (2) (a).
[202] ibid art 22 (2) (c).
[203] ibid art 22 (3).
[204] ibid art 22 (1).
[205] ibid article 22 (1).

If Article 22 were to be read simply as a right to object to an ADM, it would make its application conditional on the proactive role of data subjects[206] and their willingness to make use of the right when none of the exceptions are met. Following this interpretation, the right to object would not apply if one of the exceptions encompassed in Article 22(2) apply, but these exceptions would only come to light once data subjects exercise their right to object. Automated decisions that do not fall within one of the exceptions would continue to take place until the data subject exercise his or her right to object, when they would have to inevitable stop, not triggering as well any of the safeguards[207] presented in Article 22(3)[208]. This interpretation of Article 22(1) put the burden of the right on the data subjects, as the provision would require an active engagement of data subjects in monitoring and objecting to automated decisions[209].

By contrast, treating Article 22(1) as a *prohibition* would prohibit automated decisions unless they fall under the exceptions included in Article 22(2)[210]. According to this interpretation, data controllers would need to assess in which exception their use of automated decision fall (necessary to enter or perform a contract, authorised by European or national law, or explicit consent) and implement the appropriate safeguards[211]. If none of the exceptions listed in Article 22(2) cover such automated decisions, then the decisions are prohibited and data subjects are protected by default against them, and an active objection is not necessary. The burden of complying with Article 22 of the GDPR, then, would fall on the data controllers and the Supervisory Authorities.

---

[206] Wachter, Mittelstadt and Floridi (n 45).
[207] According to Recital 71 of the General Data Protection Regulation such suitable safeguards should include "specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision".
[208] Margot E Kaminski, 'The Right to Explanation, Explained' <https://lawcat.berkeley.edu/record/1128984> accessed 28 November 2024; Marco Almada, 'Automated Decision-Making as a Data Protection Issue' [2021] SSRN Electronic Journal <https://www.ssrn.com/abstract=3817472> accessed 28 November 2024; Ugo Pagallo, 'Algo-Rhythms and the Beat of the Legal Drum' (2018) 31 Philosophy & Technology 507; Wachter, Mittelstadt and Floridi (n 45); Sancho (n 44).
[209] Wachter, Mittelstadt and Floridi (n 45).
[210] Mendoza (n 200); Mendoza and Bygrave (n 45); Wachter, Mittelstadt and Floridi (n 45); Kaminski (n 208).
[211] Wachter, Mittelstadt and Floridi (n 45).

At first sight it is noticeable that the first interpretation could entrench the power imbalance between data controllers and individuals as the latter could be subject to unlawful automated decisions, as referred in GDPR, but only be aware of it once they inquire as to whether they comply with the requirements of Article 22. The second interpretation seems to be more reasonable as it more closely reflects the values of GDPR. An analysis of the wording of the GDPR could offer some insights regarding this debate.

The GDPR includes several other data subject rights, such as the right to determine whether personal data can be collected and transmitted to others[212], the right to access[213], update, and delete[214] personal data; the right to rectification[215], and the right to refuse the processing of such data[216]. The Regulation offers data subjects the possibility to exercise, or not, such rights when their personal data is processed. The wording of these Articles differs in the wording of Article 22 in that the former read as a *right to* while the latter provides a *right not to be*. Furthermore, Recital 71 referred to Article 22 makes a double distinction regarding the right not to be subject to a decision. On the one hand, it uses the same words as its referred Article 22 by expressing that 'the data subject should have the right not to be subject to a decision'. On the other hand, it makes a clarification by denoting that:

> However, decision-making based on such processing, including profiling, should be allowed where expressly authorised by Union or Member State law to which the controller is subject, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent.[217]

Given that the Recital itself makes it clear that in certain circumstances automated decisions *should be allowed*, it can be interpreted by extension that automated

[212] General Data Protection Regulation's right to information in arts. 13 and 14, and right to data portability in art.20.
[213] ibid Regulation's right to information in arts 13 and 14 and right to access in art 15.
[214] ibid's right to rectification in art 16 and right to erasure in art 17.
[215] ibid's right to rectification in art 16.
[216] ibid's right to withdraw consent in art 7(3).
[217] ibid recital 71.

decisions are in principle prohibited without the need for the data subjects to engage and exercise their right to not be subject to them. This interpretation can also be inferred from the *travaux préparatoires* of the GDPR. The initial proposal of the European Commission stated in its Article 20(1) that:

> [E]very natural person shall have the right not to be subject to a measure which produces legal effects concerning this natural person or significantly affects this natural person, and which is based solely on automated processing intended to evaluate certain personal aspects relating to this natural person or to analyse or predict in particular the natural person's performance at work, economic situation, location, health, personal preferences, reliability or behaviour.

This provision was amended by the European Parliament in such a manner that it read as follows; 'without prejudice to the provisions in Article 6, every natural person shall have the right to object to profiling in accordance with Article 19. The data subject shall be informed about the right to object to profiling in a highly visible manner'. Furthermore, the preliminary Article 20(2) specified that:

> Subject to the other provisions of this Regulation, a person may be subjected to profiling which leads to measures producing legal effects concerning the data subject or does similarly significantly affect the interests, rights or freedoms of the concerned data subject only if the processing (then follow to state the exceptions) [...].

And finally, preliminary Article 20(4) read as follow:

> Profiling which leads to measures producing legal effects concerning the data subject or does similarly significantly affect the interests, rights or freedoms of the concerned data subject shall not be based solely or predominantly on automated processing and shall include human assessment, including an explanation of the decision reached after such an assessment. The suitable measures to safeguard the data subject's legitimate interests referred to in paragraph 2 shall include the right to obtain human assessment and an explanation of the decision reached after such assessment.

In other words, the European Parliament's proposal included a right to object to profiling, and this right had to be clearly communicated to the individual. The premise for such profiling, however, was that it should fall within one of the exceptions listed in the second paragraph of Article 20 and that solely or predominantly profiling should in no case have legal repercussions or similar results. The relationship between these provisions was that controllers could only carry out profiling on specific grounds and in all cases had to inform the individuals of their ability to exercise their right to object. The final version of the GDPR makes no distinction as far as profiling is concerned and does not include in its Article 22, any reference to the right to object. In turn, it is Article 21 of the GDPR that is entirely devoted to the individuals' right to object to the processing of their data, referring to profiling but not to automated decisions including profiling. This distinction suggests that data subjects do not have to exercise their right to object in order not to be subject to ADM, but that it is a right per se that they enjoy without the need to actively exercise it.

Analysing the *travaux préparatoires* of the GDPR, it can be concluded that the burden of the right to not be subjected to ADM was, therefore, debated as to whether it had to fall on the active engagement and exercise of the data subject or as an obligation to the data controller. The final wording of the GDPR implies that the regulator does not seek to offer data subjects an alternative in the face of automated decisions that may affect them, but directly a context in which, with limited exceptions, the data subjects do not have to worry about being affected by such decisions.

Furthermore, Article 22 (3) presents some of the suitable safeguards that data controllers should implement to protect data subjects' rights and interest when one of the exceptions referred in Article 22 (2) applies and so individuals can be subject to ADM[218]. It would not make sense if, in order to exercise these rights, the individuals were first required to enforce their right not to be subject to ADM, on the assumption that the controller would waive the exception allowing such processing, and that only after such an exchange would data subjects be able to contest or challenge the decision.

[218] Wachter, Mittelstadt and Floridi (n 45); Mendoza (n 200).

Following the same logic Mendoza and Bygrave concluded that Article 22 most likely 'functions as a (qualified) prohibition with which the decision-maker has to comply regardless of whether the right holder invokes it or not'[219]. Indeed, this position was supported by the Article 29 Working Party Guidelines on Automated Individual Decision-Making, when arguing that:

> The term *right* in the provision does not mean that Article 22(1) applies only when actively invoked by the data subject. Article 22(1) establishes a general prohibition for decision-making based solely on automated processing. This prohibition applies whether or not the data subject takes an action regarding the processing of their personal data[220].

Moreover, the notion of prohibition was repeatedly used in Article 29 Working Party Guidelines to address the right to not be subject to ADM defined in Article 22(1)[221]. Therefore, Article 29 Working Party Guidelines aligned with the view that Article 22 (1) established a prohibition on ADM, not a mere right to object to it. Following this argument, Article 22 would compel companies not to use solely ADM unless their use falls under one of the exceptions included in Article 22, i.e., contract, explicit consent, or Member state law.

Overall, whether Article 22 provides the data subject with a right to object or establishes a general prohibition on ADM, are still uncertain. However, considering that the GDPR requires for some of the rights encompassed in its Chapter III an active role of the data subject or an obligation to the controller to ensure its effectiveness, it would make sense that the wording of Article 22 involves a passive role for the data subject and therefore a general prohibition towards automated decisions. Furthermore, if Article 22 grants a right to object, any automated decision would be legally unchallenged, unless the data subjects exercise their right and by extension, only after this initial exercise of the right to object would the data subject be entitled to avail himself of the safeguards afforded by the exceptions set out in paragraph 3 (i.e., right to obtain human

[219] Mendoza and Bygrave (n 45) 87.
[220] Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (A29WP Guidelines on ADMs and Profiling) 2017 (WP251rev01) 19.
[221] see A29WP Guidelines on ADMs and Profiling9, 12, 19, 20, 23, 34, 35.

intervention, to express one's view and to contest the decision). Interestingly, one of these safeguards already entail a right to contest the decision, what would be redundant if, in order to exercise this safeguard, the subject had to object to the decision beforehand. Additionally, this logic of Article 22 as a particular right to object may make little sense, as Article 21 is directly dedicated to the right to object, and it does not refer to ADM, but simply to profiling. In conclusion, although the nature of Article 22 remains a matter of debate, it is possible to conclude, after the above analysis, that the provision highly likely establishes a general prohibition of ADM with three concrete exceptions and its subsequent safeguards.

The importance of this debate with regard to the object of study of this thesis, the right to information and an explanation, lies in the ease of its exercise, as well as of the subject who has to actively exercise them in the first place. If Article 22 were to be considered a right to object, data subjects would have to exercise their right to an explanation after exercising their right not to be subject to an automated decision. If considered a prohibition, Article 22 would be up to the data controller to ensure that the suitable safeguards are in place at the time an automated decision is made and therefore provide an explanation along with that decision. The Court of Justice of the European Union (CJEU) still has the final authority in this debate and will presumably offer clarification on the basis of the two preliminary rulings to clarify the content of Article 22, the *SCHUFA Holding (Scoring)* and the *Dun & Bradstreet Austria GmbH* cases. Indeed, this uncertainty regarding the nature of Article 22 has lead scholars as Wachter et al. to argued that the more suitable safeguard towards automated decision is a right to be informed, rather than a right to an explanation[222]. This discussion as well as the legal reasoning introduces in the mentioned cases, whereas introduced here, will be analysed in deep detail in the subsequent sections.

[222] Wachter, Mittelstadt and Floridi (n 45) 21.

3.2.3. Content of Article 22: profiling, automated processing, and automated decisions

Beyond the debate regarding the nature of Article 22 of the GDPR, the content of the provision is also under an intense debate. The official tittle of the article, 'Automated individual decision-making, including profiling', is often considered misleading and confused[223] as it connects the concepts of automated individual decision-making and profiling in a manner that seems to conflict with the definition (of profiling) provided by Article 4 of the GDPR, which reads as follows,

> [Profiling is] any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements[224].

Article 4(4) of the GDPR defines profiling as a form of automated processing but makes no reference to any possible decision resulted from such profiling or whether the profiling itself shall be considered a decision itself.

However, the tittle of Article 22 might suggest that profiling could be considered a form of automated individual decision-making and therefore, trigger the application of Article 22. While this argument will be discussed in more detail below, the logic behind such an assertion lies, for the most part, in the existence of scenarios in which the acts of profiling and decision-making are carried out by different actors, but with a strong influence of the profiling result on the decision made. In these cases, it could well be argued that the final decision was not based purely on profiling and therefore would not trigger the application of Article 22, but the concern arises from those occasions where the profiling result itself already entirely determines the decision. The debate lies in determining whether such profiling could be considered an advance decision and therefore fall under Article 22. The consequences of such extensive interpretation are

[223] Mendoza and Bygrave (n 45).
[224] General Data Protection Regulation.

ample. Article 22(1) seems, nonetheless, to limit this initial interpretation by stating that 'the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling,'. In this sense, Article 22 would not be applicable to profiling in a broader sense, rather only to decisions based on profiling, defined, as seen above, as a form of automated processing by Article 4(4).

The intention of regulators when titled Article 22 would be reasonably clearer if the GDPR has included a definition of automated individual decision-making, what was not the case. However, the *travaux préparatoires* of the GDPR could be of help to delimit the content of Article 22. The first draft of European Commission encompasses in its Article 20 'measures based on profiling', as referred,

> Every natural person shall have the right not to be subject to a measure which produces legal effects concerning this natural person or significantly affects this natural person, and which is based solely on automated processing intended to evaluate certain personal aspects relating to this natural person or to analyse or predict in particular the natural person's performance at work, economic situation, location, health, personal preferences, reliability or behaviour[225].

The Draft of the European Commission, then, identified automated processing to what later on in the GDPR was defined as profiling, excluding from the content of the prohibition encompassed in its Article 20(1) any other automated processing that might not involve profiling. In the subsequent draft from the European Parliament and the Council, the whole wording of Article 20(1) was replaced by a simple statement 'without prejudice to the provision in Article 6, every natural person shall have the right to object profiling in accordance with Article 19 [...]'[226]. As presented in the above section, the Parliament and the Council shifted the nature of Article 20 from a prohibition to a right to object but maintained the limits of the article's applicability to only profiling instead

[225] European Commission Proposal for a Regulation 2012/0011 of 25 January 2012 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (European Commission Proposal GDPR) 2021 (2012/0011 (COD)).
[226] European Parliament and Council Legislative Resolution 2012/0011 of 12 March 2014 on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (European Parliament and Council Proposal GDPR) 2012 (COM(2012)0011).

of to any automated processing that leads to measures or decisions with legal effects or similarly significantly as the final GDPR does. Both drafts lead to consider that, initially, the regulators only contemplated the impact that decisions and measures based on profiling could have on the interests, rights, and freedoms of data subjects. However, the final wording on the GDPR suggests a broader evaluation of automated processing where profiling is just an option among many.

The acceptability of this interpretation lies in finding any example of processing of personal data that does not entail the classification or profiling of the data subject which still gives rise to legal consequences or of similar significance. Arguably, most of ADM used today relies on some sort of profiling with few examples of automated decisions not based on some evaluation and categorization of the individual that could result in these required legal or similar effects. An unlikely but possible scenario can be found in the processing of personal data to ensure the randomness of a decision. In other words, the processing of personal data whose aim is not to create a profile of the individual, but to remove all informative value from these data by obtaining only a random variable upon which a decision is made. For instance, the so-called Benford's Law holds that in the great variety of data and numbers that exist in the real world, the first digit is 1 much more frequently than the rest of the numbers[227]. Based on this law, different methods have been developed to perform data processing leading to fraud detection processes[228] and forensic audits[229] in sectors such as international trade, banking, civil registry, social security, health, services allocation. Therefore, a broader interpretation of automated processing as referred in the title of Article 22 will ensure that practices where there is data processing but not proper profiling -as could be those using Bedford's Law- will clearly fell within Article 22 if they give rise to legal or similar consequences.

---

[227] EW Wesstein, 'Benford's Law' (*Wolfram MathWorld*) <https://mathworld.wolfram.com> accessed 23 September 2022.
[228] J Carlton Collings, 'Using Excel and Benford's Law to Detect Fraud' (*Journal of Accountability*, 1 April 2017) <https://www.journalofaccountancy.com/issues/2017/apr/excel-and-benfords-law-to-detect-fraud.html> accessed 19 January 2025; Andrea Cerioli and others, 'Newcomb–Benford Law and the Detection of Frauds in International Trade' (2019) 116 Proceedings of the National Academy of Sciences 106.
[229] Mark J Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection* (John Wiley & Sons 2012).

Without any further guidance as to which interpretation -restrictive or extensive- of the scope of Article 22, scholars and practitioners still wonder whether the title of the article 'including profiling' refers to automated processing or an automated decisions with legal or similarly significant effects[230]. Either profiling 1) presents an example of a type of automated processing that can lead to an automated decision, 2) it denotes the only automated processing that triggers the application of Article 22, or 3) it refers to an alternative to automated decisions that leads, itself, to legal or similarly significant effects for the data subject.

The first interpretation establishes the existence of an automated processing or profiling which results in an automated decision as a pre-requisite for Article 22. This interpretation would immensely expand the content of the provision as any automated processing that led to an automated decision, with legal or similarly significant effects, whether it is based on the profile or the evaluation of an individual or not, would also fall under the umbrella of Article 22[231]. On the contrary, the second interpretation will limit the scope of Article 22 only to situations where profiling takes place and leads to a decision. Finally, the third approach diverges in that it would allow the application of Article 22 to mere profiling that has legal or similar consequences without the need for a subsequent decision, noting the trouble to demonstrate that an individual has been affected by legal or similar consequences without a *posteriori* decision having been taken that gave rise to such effects. The debate, therefore, revolves around what needs to be consider a decision or to what extent profiling can be consider a decision itself. As mentioned about, this interpretation tries to offer coverage to situations where the automated processing and the decision can be distinguish in two independent acts but where the profiling is considered, by many, the actual real decision.

As the preparatory work of the GDPR was strongly focused on profiling rather than on automated processing it deems reasonable to assert that at least one of the focal points of the final Article 22 of the GDPR is the potential of profiling to harm the interest, rights and freedoms of data subjects. However, the *travaux préparatoires* always

[230] Mendoza and Bygrave (n 45).
[231] ibid.

referred to profiling that leads either to a decision or a measure that affects the data subject legally or similarly significantly. The trigger of the provisions of both the Commission, and the Parliament and Council drafts seemed to be linked to the pre-existence of a profiling which brings an action affecting the individual. Suffice to say none of wording of the drafts were directly incorporated in the final version of the GDPR for which we could presume that the regulators finally decided to expand the scope of Article 22 not merely to profiling but to other types of processing leading to automated decisions that could end up harming the interests, rights, and freedoms of individuals.

Whereas some scholars, as Mendoza, suggest that the right interpretation of the provision is omitting the expression 'automated processing' from the title of the article and therefore from its application concluding that the automated decision needs to be based on profiling to trigger the exercise of Article 22, Article 29 Working Party Guidelines repeatedly refers to profiling and ADM as independent practices going so far as to mention that 'the GDPR introduces provisions to ensure that profiling and automated individual decision-making (whether or not this includes profiling) are not used in ways that have an unjustified impact on individuals' rights'[232] and that 'the GDPR does not just focus on the decisions made as a result of automated processing or profiling. It applies to the collection of data for the creation of profiles, as well as the application of those profiles to individuals'[233]. From these statements and the subsequent independent sections in the Guidelines related to profiling and ADM, it seem that they are treated as different practices, that can both provide the base for a decision as referred in Article 22. The Guidelines, however, confirmed that ADM may partially overlap with profiling, despite having a different scope.

> Automated decisions can be made with or without profiling; profiling can take place without making automated decisions. However, profiling and automated decision-making are not necessarily separate activities. Something that starts off as a simple automated decision-making process could become one based on profiling, depending upon how the data is used[234].

[232] A29WP Guidelines on ADMs and Profiling 6.
[233] A29WP Guidelines on ADMs and Profiling.
[234] ibid 8.

Despite this, apparently, clear distinction of practices, the content of Article 22 is still confusing. However, it is important to consider that both the final version of the GDPR and The Guidelines emphasize the difference between profiling and processing, so this difference must be taken into account when interpreting Article 22. Likewise, it is worth recalling that the interpretation of the article does not need to be according to the strict literal wording of the article but it is advisable to take into account the reality in which the article is to be applied as well as the scenarios in which granting the protection offered would respond to the interests and values of the GDPR.

In fact, the Administrative Court of Wiesbaden[235] has expressed its concerns regarding the legal lagoons that following a restrictive interpretation of the title of Article 22 can create in the Case SCHUFA Holding (Scoring). The referring Court assumes that the 'establishment of a score by a credit information agency is not merely profiling that serves to prepare the decision of the third-party controller but constitutes an independent 'decision' within the meaning of Article 22(1) of the GDPR'[236]. Contrary to the restrictive interpretation of Article 22 of the GDPR which would imply that the act of collecting, compiling, and processing personal data for the purpose of a credit score is independent from the final decision of a third-party controller such as a bank or a lender, the referring Court argues that the 'automated establishment of a score by credit information agencies for the prognostic evaluation of a data subject's financial capacity is an independent decision based on automated processing within the meaning of Article 22(1) of the GDPR'[237]. Furthermore, the referring Court highlights the importance of such credit score on the later decision made by the data controller and expresses its concerns regarding the lack of individual assessment and evaluation by a human being that a restrictive interpretation of Article 22 would have towards automated profiling such as credit scoring. According to the referring Court, the ultimate aim of Article 22 is to offer data subjects with the necessary understanding of the underlying assumptions and evaluation standards of automated decisions affecting them. Furthermore, the 'regulatory aim is to create transparency and fairness in decision-making processes' as

[235] *C-634/21 SCHUFA Holding (Scoring) - Request for Preliminary Ruling from the Verwaltungsgericht Wiesbaden* (n 169).
[236] ibid para 24.
[237] ibid para 24.

well as 'make the human corrective for automated data processing mandatory in principle and to allow derogations only in the limited exceptional cases that is thwarted'[238]. I concentrate on ADM transparency and fairness in subsequent sections, but it can brought now to focus that from the referring Court's opinion, a restrictive approach to Article 22 would collide with these legislative aims in multiple real cases. For example, the referring Court affirmed that a situation like the one presented in this case would prevent data subjects' from enjoying the right of access under Article 15 of the GDPR as credit score agencies would not have to disclose the logic and composition of the parameters that are essential to determine the score, while third-party controllers would be also unaware of such logic and parameters or even unallowed to offer them to the data subject. As per the words of the referring Court, credit scoring :

> [This] gives rise to a lacuna in the legal protection: the party from whom the information required for the data subject could be obtained is not obliged to provide access to information under Article 15(1)(h) of the GDPR because it allegedly does not engage in its own 'automated decision-making' within the meaning of Article 15(1)(h) of the GDPR, and the party that bases its decision-making on the score established by means of automation and is obliged to provide access to information under Article 15(1)(h) of the GDPR cannot provide the required information because it does not have it[239].

The only way of filling such lacuna would be to confirm that activities that initially appear to be automated profiling, such as credit scoring, actually fall under the scope of Article 22 of the GDPR.

In its recent judgment, the First Chamber Court has offered some clarity concerning the correct interpretation of Article 22, by asserting that the interpretation of EU law provisions requires not only the literal wording of the article, but also its context, objectives and purpose. Accordingly, Article 22 shall be interpretated to established three cumulative conditions, (1) the existence of a decision, (2) the basis of that

[238] ibid para 26.
[239] ibid para 31.

decision being solely based on automated processing, including profiling, and (3) the production of legal effects or similarly significantly effects on the interested party[240]. In the considerations of the questions referred by the Administrative Court of Wiesbaden, the First Court supported Advocate General's Opinion in a number of points concerning these three conditions. In this section, however, we would only addressed the First Court points regarding the concepts of *decisions* and *profiling*. The argumentation around the two other conditions would be addressed in section 3.2.4., as this section provides clearance and guidance regarding the interpretation of the concepts of *solely automated decision-making* and *legal or similarly significant effect*, the other aspects - and conditions- of Article 22 beyond its applicability to profiling or data processing.

Thus, the Fist Chamber asserted that the concept of *decision* shall be interpreted in a broad scope as implied in recital 71 of the GDPR according to which a decision evaluating personal aspects relating to a person, to which that person should have the right not to be subject, 'may include a measure' which either produces legal effects concerning him or her, or, similarly significantly affects him or her. Agreeing with point 38 of the Advocate General's Opinion, the First Chamber asserted that the concept of *decisions* under the meaning of GDPR is capable of including a number of acts which may, as well, affect the data subject in many ways[241]. This first conditions drew importance not on the act of data processing or profiling, but on the decision resulting from either of those acts. As such, although the First Chamber does not clarify whether the scope of Article 22 encompasses only data processing or only profiling, it could be deduce that this distinction has limited practical relevance for the intention and purpose of the provision.

In the concrete case at hand, the First Chamber agreed with Advocate General's Opinion in that the establishment of a probability value based on personal data relating to a person and concerning that person's ability to repay a loan in the future -the common activity carried out by SCHUFA- constitutes profiling as appears in Article 4(4) of the GDPR. Precisely, the probability value obtained from calculating a person's

[240] *C-634/21 SCHUFA Holding (Scoring) - Judgement of the Court* [2023] ECJ OJ C, C/2024/913 para 43.
[241] ibid para 46.

creditworthiness shall be encompassed under the broader definition of *decision* as referred in Article 22 'in the event where a loan application is sent by a consumer to a bank, an insufficient probability value leads, in almost all cases, to the refusal of that bank to grant the loan applied for'[242]. Shortly advancing over the interpretation of the second and thirds conditions of Article 22, it is worth mentioning that SCHUFA's aforementioned probability value of repayment is considered, by the First Court, a fully automated decision producing legal or similarly significant effects on the individual as the bank to whom the value is transmitted draws strongly on it in granting credit.

Thus, whether profiling or data processing leads to a decision is not sufficient condition to trigger the applicability of Article 22. The SCHUFA Holding (Scoring) ruling draws special attention to the second and third conditions for which the individuals' rights and freedoms need to be significantly affected by a solely ADM. Mere profiling without a subsequent effect or with a relevant human involvement will likely fell out of the scope of the protection granted to data subjects in Article 22 of the GDPR. The critical concept in the delimitation of Article 22 would rather be the legal or similarly significant effects arising. As stated above, whether or not data subjects are impacting by legal or similarly significant consequences will be key to assess if automated processing, including profiling, shall be considered to trigger the protection offered by Article 22. As will be exposed in the subsequent section such condition is controvert as well

3.2.4. Decisions based solely on automated processing with legal or similarly significant effects

The analysis of the requirements set out in Article 22(1) for a decision to be considered based solely on ADM is relevant for this thesis since Articles 13(2)(h), 14(2)(g), 15(1)(h), and Article 22(3) of the GDPR only apply to the automated decisions that fall under the exceptions of Article 22(2), namely automated decisions based on solely data processing, including profiling, with legal and significant effects concerning the particular data subject, necessitated by a contract, the explicit data subject's consent, or a European or national law. In essence, the internal limits established in paragraphs 1 and 2 of Article 22 limit the application of the rights to information and an explanation to

[242] ibid para 48.

those automated decisions that fulfil the mentioned requirements. Thus, algorithmic recommendations or decisions with a human-in-the-loop would not be considered automated decisions.

According to the wording of Article 22, paragraphs 1 and 2, two very specific limits concerning the right to not be subject to ADM and the implementation of its suitable safeguards can be noting: (1) the nature of the decision and (2) its effects and consequences.

With regard to the first internal limit, it is worth recalling that decisions based solely on automated processing are those which are taken by technological means without any meaningful human involvement, understanding that, otherwise, the decision will be taken by someone who has the authority and competence to change the decision and who will consider all the relevant data before deciding - the so-called human-in-the-loop-. So for a decision to fall outside of Article 22 the human involvement must be meaningful in the sense that the person must have the authority and competence to change the decision and must, in addition, have access to additional information beyond the algorithm's output. Three types of automated decisions can result from processing and profiling:

a) decisions where the automated output applies straightforwardly;
b) automated decisions with human nominal involvement, where a human actor intervenes in the application of the automated output without revising or assessing it; and
c) human-based decisions, where a human analyst revises the automated output and makes a decision[243].

Based on Article 29 Working Party Guidelines, Article 22 application will cover cases a) and b), whereas cases c) will fall out of its scope. Analysing the current European case law regarding ADM, Barros Vale & Zanfir-Fortuna argued that to not be considered automated, the decision should be held in either of these contexts;

[243] Sancho (n 44) 143.

> When organizational measures are put in place to ensure structured and substantial human involvement, such as when multiple persons analyse automated individual potential fraud flags and have to unanimously agree on whether fraud was committed taking into account additional elements and correlating facts; or when
>
> internal procedure requires a written assessment made by case officers on the basis of an automated assessment, which then needs to be vetted by the head of the organization; or when
>
> employees are specifically trained and provided with detailed guidelines on additional elements to take into account in order to make decisions on the basis of automated assessments and recommendations[244].

However, as shown below, academics have long discussed the feasibility of proving the meaningful participation of a human in a decision based fully or partially on processing or profiling.

With regard to the second internal limit, *decisions with legal or similarly significant effect* refers to decisions that affect 'someone's legal rights, legal status or their rights under a contract'[245] as well as decisions with a 'significantly great or important [effect] to be worthy of attention'[246]. Examples of both effects are decisions that lead to a refused admission to a country, the denial of a particular benefit, or the refusal of an online credit application or refusal of employment applications. Inevitably these provisions pose several challenges to protecting individuals in the context of algorithmic systems as it is left to the court and Member State discretion to determine which is their exact extent and meaning in real-world cases. The lack of guidance or standards to determine what might be considered *legally or similarly significant effects* lead its interpretation to contextual and subjective considerations. Article 29 Working Party has clarified that the effect of processing must be 'sufficiently great or important

[244] Barros Vale and Zanfir-Fortuna (n 48) 29.
[245] A29WP Guidelines on ADMs and Profiling.
[246] ibid.

to be worthy of attention'[247], these are decisions that 'significantly affect the circumstances, behaviour or choices of the individuals concerned, have a prolonged or permanent impact on the data subject, or, at its most extreme, lead to the exclusion or discrimination of individuals'[248]. To bring light to this claim, Barros, Vale & Zanfir-Fortuna compiled some of the elements that were taken into account by European courts when addressing the effects of the automated decisions. The authors highlighted the following criteria;

i. the categories of personal data on the basis of which the automated decisions are produced and whether they include data points and/or inferences about the behaviour of data subjects;
ii. the immediate consequence the decisions have on data subjects;
iii. the temporary or definitive effect of the decisions;
iv. whether the decisions affect conduct or choices of the data subjects;
v. whether the decisions limit opportunity for income or are followed by a quantifiable financial loss for data subjects;
vi. whether the data subjects are able to demonstrate the impact of decisions on them are not trivial where enforcers do not find the facts of the case sufficient to show a legal or similarly significant effect[249].

Barros, Vale & Zanfir-Fortuna's work highlights how identifying whether a decision is purely automated and whether it gives rise to legal or similar consequences pose both certain challenges. On the one hand, it requires an inquiry into the internal organization of a company as well as its decision-making processes. Moreover, it makes it necessary to establish appropriate instruments and safeguard to ensure that the decision-maker is not merely accepting the recommendation of the support decision-making system and assuming the result as the final decision. On the other hand, the effects that a decision may have on the individual may depend greatly on the context in which the decision is made. Whilst certain standards can be identified regarding the impact on the individual's life, such relevance may also arise from a presumably innocuous decision

[247] ibid 21.
[248] ibid.
[249] Barros Vale and Zanfir-Fortuna (n 48) 35.

that results in a not foreseeable discrimination or social isolation. Hence, an initial decision without legal or similarly significant effects may become a misconduct with severe impact on the individual. Certainly, the concepts of (1) *solely automated processing* and (2) *legal [...] or similarly significantly affects* have given rise to an ongoing academic debate as the wording of the article is tricky itself and the clarification offered by Article 29 Working Party did not resolve the problem as a whole until the CJEU will offer some insight as to their meaning.

As addressed above, the Court of Justice of the European Union was asked to interpret Article 22 on the Preliminary Ruling *SCHUFA Holding (Scoring)* in which the Administrative Court of Wiesbaden, Germany, requested an interpretation regarding:

> The existence of a solely automated decision based on profiling with legal or significant effects when a value, determined by means of personal data of the data subjects, is transmitted by the controller to a third-party controller and the latter draws strongly on that value for its decision on the establishment, implementation or termination of a contractual relationship with the data subject[250].

Regarding the second condition of Article 22 -solely automated processing- the First Chamber has clarified that a decision based on credit score offered by a controller -credit information agency- to a third party -bank- needs to be considered an automated decision without nominal human involvement, if a human actor intervenes in the application of the automated output without revising or assessing it. On words of the referral Court, SCHUFA Holding case presents an scenario where:

> it is ultimately the score established by the credit information agency on the basis of automated processing that actually decides whether and how the third-party controller enters into a contract with the data subject. Although the third-party controller does not have to make his or her decision dependent solely on the score, he or she usually does so to a significant extent[251].

[250] *C-634/21 SCHUFA Holding (Scoring) - Request for Preliminary Ruling from the Verwaltungsgericht Wiesbaden* (n 169) para 1.
[251] ibid para 25.

Agreeing with the referral Court argument, the First Court clarified that as the rationale of Article 22(3) is to protect the subject from automated decisions with no human involvement, assessment or evaluation, the data subject should not be left defenceless against an exclusively technical and non-transparent process carried out by a data controller for the benefit of the final decision-maker. Hence, both the Referral and the First Court highlighted how the data subjects should understand the underlying assumptions and evaluation standards and intervene, if necessary, by exercising their rights. Upon the question of the Administrative Court regarding whether:

> The establishment of a score by a credit information agency is not merely profiling that serves to prepare the decision of the third-party controller but constitutes an independent 'decision' within the meaning of Article 22(1) of the GDPR[252].

The First Chamber confirmed that SHUFA's probability value concerning the individual ability to meet payment commitment in the future needs to be interpret as an automated decision as referred in Article 22 when a third party, to which that probability value is transmitted, draws strongly on that probability value to establish, implement or terminate a contractual relationship with that person[253].

As argued above, through this request of preliminary ruling, the European Court of Justice has had the opportunity of clarifying some of the main questions regarding the content and scope of Article 22, mainly in the context of credit scoring. It is worth pointing out that the actual use of profiling to make automated decisions involves practices that go beyond both a clear distinction between the data processing and the decision, and a clear causal relationship between the profile and the final decision. Although this ruling offers some clear guidance regarding how to interpret this condition in other situations, social and economic factors surrounding the decision and impacting the individual would need to be taken into account to determine whether the data processing and profiling in other fields should be considered an automated

[252] *C-634/21 SCHUFA Holding (Scoring) - Request for Preliminary Ruling from the Verwaltungsgericht Wiesbaden* (n 169) para 21.
[253] *C-634/21 SCHUFA Holding (Scoring) - Judgement of the Court* (n 240) para 75.

decision with legal or similarly effects. To protect individuals' rights, freedoms, and interests from the threats that automated decisions can pose it would be highly desirable to adopt a similar broad interpretation -as the one followed by the First Court in the *SCHUFA Holding (Scoring)* case- or at least a contextual interpretation to the extent to which the social and economic factors surrounding the decision and impacting the individual are taken into account to determine whether the data processing and profiling should be considered an automated decision with legal or similarly effects.

To this regard, Bygrave highlighted the difficulty of determining the involvement as to whether a person can or fails to 'actively exercise any real influence on the outcome of a particular decision-making process'[254]. Likewise, Hildebrandt has criticised that Article 22(3) would not apply to those decisions with a routine human involvement, even when the algorithmic recommendations and the final decision will always remain the same[255]. Following this line of thought, Wachter et al.[256] argued that if both conditions are interpreted in a narrow sense, the applicability of the safeguards presented in Article 22(3) will be extremely limited. Moreover, they suggested that even a trivial human involvement might prevent the application of the safeguards defined in Article 22(3). The same concerns were presented by Selbs and Powles[257], thus, with an optimistic perspective towards the future interpretations of the application of the article. Edwards and Veale have strongly emphasised how a restricted interpretation of these two requirements would leave behind decisions whose effect on people's lives are significant but whose decision-making processes 'are not usually fully automated - instead used as decision support- since their full automation seem inappropriate or far off'[258]. Indeed, Mendoza and Bygrave[259] further criticised that the ambiguity of these provisions has been exacerbated by a lack of final guidance about how they should be interpreted and exercised.

[254] Bygrave (n 184) 22.
[255] Hildebrandt and Gutwirth (n 24).
[256] Wachter, Mittelstadt and Floridi (n 45).
[257] Selbst and Powles (n 46).
[258] Edwards and Veale (n 89) 45.
[259] Mendoza and Bygrave (n 45).

I share the same concerns expressed by these academics insofar as determining the level of involvement of a human in the final decision-making is a challenge. However, it should be noted that given the prohibition referred to in Article 22, businesses will be obliged to establish the appropriate processes and mechanisms to ensure an active and meaningful human role when not seeking to make an automated decision. It does not therefore seem appropriate to argue that the existence of profiling or data processing already poses a direct threat to the interests and rights of individuals, although it does in fact comes with its risks and challenges -Chapter 2.2 and Chapter 4.2.-. If such practices support the decision-making process, then the decision maker will have to respect and guarantee the corresponding requirements and safeguards and offer justifications and explanations when necessary. Ultimately, the aim of Article 22 is to offer guarantees and protection to individuals in the face of advances in automation, ensuring the fairness, lawfulness and justice of the decision-making process and the final decision, but in no case does it seem to seek to disregard it altogether. However, given the dangers that an unidentified automated decision may bring to individuals, the extensive interpretation of Article 22 conditions seems to be the best option to guarantee the protection to individuals so as to take into account the circumstances of each use and the impact it may have.

Moreover, Castets-Renard[260] emphasised how easy it is to pretend that other processes were used to make a decision, although it would be only an automated decision what took place. Likewise, the scholar highlightes the difficulty to determine the level of influence of the algorithm score in the decision-maker and so confirm the independence of the final decision from the processing. Lastly, Binns' & Veale's work[261] pay attention to the decision-making process with multiple stages, potentially both manual and automated, which together might difficult the application of Article 22(3) safeguards.

The scholars argue that national and European courts would need to extensively redefine some of the provisions of the GDPR and 'transform stubborn *ex-ante* concepts

[260] Celine Castets-Renard, 'Accountability of Algorithms in the GDPR and Beyond: A European Legal Framework on Automated Decision- Making' [2019] 30 Fordham Intell. Prop. Media & Ent. L 91.
[261] Binns and Veale (n 44).

like lawful bases into *ex-post* oversight'[262] if they want to ensure the application of these provisions to complex cases of automated decisions. Despite acknowledging the legal uncertainties and complexities that might result from courts' reinterpretations, Binns and Veale defend the need for a wider approach to the protection of personal data as the scholars present five possible complications concerning Article 22:

a) The potential for selective automation on subsets of data subjects despite generally adequate human input;
b) The ambiguity around whether to locate the decision itself,
c) Whether 'significance' should be interpreted in terms of any 'potential' effect or only selectively in terms of realised effects;
d) The potential for upstream automation processes to foreclose downstream outcomes despite human input; and
e) A focus on the final step that may distract from the status and importance of upstream processes[263].

While the possibility of developing ex-post oversight seems useful to identify covert cases of automated decisions, it also has certain disadvantages because it will create legal uncertainty and complexity both to the data controller and the data subject. Moreover, while this monitoring can be done after invoking Article 15 of the GDPR, which asks to obtain knowledge of the possible existence of an automated decision, the decision itself would already have been taken and in most cases the safeguards of Article 22(3) would be redundant. To ensure the effective protection of individuals, it would be more convenient to develop standards, principles or concrete guidelines to help identifying a-priori the automated processing practices and decisions which fall under Article 22 and which do not.

Additionally, some authors have analysed how the literal wording of Article 22(3) may exclude some automated decisions from the safeguards offered by the GDPR. Wachter and Mittelstadt note how Article 22(3) might exclude from its protection decisions based on inferred data, or information obtained anonymously or from third parties,

[262] ibid.
[263] ibid.

although this type of information can be used 'to infer our preferences, weakness, sensitive attributes, and opinion'[264]. They highlight, that decisions that would, otherwise, be considered as decisions based solely on automated processing and with legal or similarly significant effect will be excluded from the protection of the GDPR since it cannot be considered personal data - of the particular individual - as such. Likewise, Mendoza and Bygrave[265] suggest that automated processing, as referred to in Article 22(3), only applies to processing personal data or data related to the person, excluding automated decisions based only on generalisation of personal data from the particular individual.

I am under the impression that the wording of the provision is rather clear, the problem arises when one wants to fit the real applications and daily-day uses of ADM into these precepts. Generally, the use of such systems is not as clear or identifiable in plots as the regulator may have assumed. Therefore, proving that a concrete decision shall fall under the protection of Article 22 can be a matter of finding the right evidence or presenting the right argument. For example, the human-in-the-loop can be affected by automation bias, which denotes the increased unlikelihood that humans will challenge the output or decision suggested by the system[266] reducing the final decision to an automated acceptance of the systems recommendation. To what extent the granting of a mortgage could be, unintentionally, based solely and exclusively on a person's credit score would largely depend on the good practice of the decision-maker. Equally, a decision that for one individual may certainly be innocuous, such as closing a profile on a social media platform, can have enormous effects on the business activity of another person if that profile was the fundamental means of carrying out his or her professional activity, as was the situation in the case involving Twitter's Shadow Banning. For this reason, the judgment as to whether we are faced with a decision that triggers Article 22 or not must be made on the basis of the interrelation of its two conditions but with the central focus on the protection of the individual. It cannot be forgotten that the central

[264] Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' [2019] Colum. Bus. L. Rev. 494, 610.
[265] Mendoza and Bygrave (n 45).
[266] Sarah Chander, 'Recommendations for a Fundamental Rights-Based Artificial Intelligence Regulation - Addressing Collective Harms, Democratic Oversight and Impermissable Use.' (European Digital Rights EDRI 2020) <https://edri.org/wp-content/uploads/2020/06/AI_EDRiRecommendations.pdf>.

rationale of this article revolves around protecting the dignity of the data subject in such a way that the automation of processes and the processing of personal data does not lead to the dehumanization of data subjects. I perceived necessary, therefore, an extensive interpretation of the conditions of Article 22 that allows and encourages a context-dependant definition of automated decision with legal or similarly significant effects.

Due to these reasons, considerable importance will lie in the burden of proof as to who and how it is to be proved that the decision in question was taken automatically or have legal or similar effects Still, a certain degree of legal agreement and social understanding will need to be established as to what procedures or protocols ensure an active role of the decision-maker, and what consequences are understood to be of sufficient relevance to trigger the requirement of legal or similar effects. The role of the European Court of Justice will again be essential in determining the legal basis and conditions of these requirements.

3.2.5. Safeguards: Right to obtain human intervention, expressing one's views, and contesting the decision as referred in Article 22 (2)

As previously stated, Article 22 provides a right of not be subject to a decision based solely on automated processing, including profiling, with legal or significant effects. However, the right does not rise if one of the bases of Article 22(2) exists, i.e. if the decision:

(1) Is necessary for entering into or performance of, a contract,

(2) Is based on the data subject's explicit consent, or

(3) Is authorised by the Union or Member State law.

Regardless of these exceptions, the GDPR remains concerned with the gravity of the outcomes that ADM can have regarding situations with legal and significant effects. For this reason, Article 22(3) presents as suitable safeguards:

(1) The right to obtain human intervention on the part of the controller,

(2) To express own's point of view and

(3) To contest the decision.

These measures aim to protect the data subjects' rights by introducing processes through which individuals could systematically verify the accuracy and correctness of automated decisions as well as the relevance of the process[267].

As argued by Bayamlıoğlu, through these safeguards the GDPR adopts a transparency scheme for solely automated decisions that would allow data subjects to contest the decision through transparency mechanisms intended to make the decision-making process interpretable[268]. However, although the literal wording of the Article seems to presume the individuality and cumulative nature of these safeguards rights, how they are chronologically enforced can be extremely relevant for data subjects as it can result in the impracticality of the transparency mechanisms[269]. In fact, as Watcher and others highlighted, 'whether these rights are interpreted as a unit that must be invoked together, or as individual rights that can be invoked separately, or in any possible combination, would determine how a decision could be contested'[270].

Initially, neither the right to express own's point of view nor the right to request human intervention creates any special duty for the data controller to respond to the data subject with further information. In other words, the data controller is not compelled to reply to data subjects once received their opinion nor it is obliged to provide any specific information or explanation regarding the decision-making process followed by the human. For the former, the GDPR does not specify any follow-up procedure or the need of an official respond from the data controller. Not only that, but the content of the right is not specified in Article 22 so that the data subject may well express his conformity or disagreement with the decision. For the latter, the human-in-the-loop needs to have the authority to change the decision but is not obliged to do so nor to take

[267] *Ordinanza ingiunzione nei confronti di Deliveroo Italy s.r.l 9685994* (n 150).
[268] Emre Bayamlıoğlu, 'The Right to Contest Automated Decisions under the General Data Protection Regulation: Beyond the So-called "Right to Explanation' (2022) 16 Regulation & Governance 1058, 1060.
[269] Claudio Sarra, 'Defenceless? An Analytical Inquiry into The Right to Contest Fully Automated Decisions In the GDPR' [2020] An Anthology of Law, Claudio Sarra, 'Defenceless? An Analytical Inquiry into The Right to Contest Fully Automated Decisions In the GDPR' 235..
[270] Wachter, Mittelstadt and Floridi (n 45).

the arguments of the data subject into account whether it is to his or her benefit or detriment. Despite it may be logical to think that the human in the loop could inform his or her decision with the point of view put forward by the data subject, at no time does the GDPR establish such an obligation. Therefore, the right to express own's point of view ends exactly after the data subject express his or her opinion while the right to request for human intervention ends once there is a human in the loop who can make a new decision.

The right to contest, on the contrary, entails a defensive act which 'implies a specific kind of argumentative effort directed to specifically challenge the decision'[271]. Such a contestation would pose a duty to the controller, either to respond the arguments made by the data subject or to clarify whether they have had any impact on the original decision to render a new one. Otherwise, the right to contest an automated decision would be reduced to either the right to express one's own opinion or the right to object to the use of automated processing, both of which are already explicitly included in the GDPR, Articles 22 paragraph 3 and Article 21 respectively. Having noted this however, a distinction must be made in the enforcement between of the right to contest and the right to request human intervention, likely resulting in a scenario where the data controller is compelled to offer a reflexive response or action as a respond to the challenge of the data subject.

It can be argued, therefore, that data subjects can make use of the three safeguards encompassed in Article 22 paragraphs 3 cumulatively and as chronologically they deem most appropriate, even though the controllers may respond to such exercise in different ways, thus influencing the data subjects' plan of action. However, the rights apply exclusively to automated decisions, so that the use of one prior to the other may exhaust the exercise of the rest, as will be discussed below.

A cumulative approach to the rights suggests that data subjects can use the three safeguards one after the another. However, this is unlikely to be the case since the rights only applied to decisions based solely in ADM and the right to request human

[271] Sarra (n 269) 6.

intervention by its very nature invalidates the use of any other safeguards after its exercise. In other words, by requesting the intervention of a human (a human-in-the-loop), the original automated decision becomes human-made, excluding it from the protection of Article 22. Strictly speaking, the safeguards referred in Article 22(3) will only be exhausted if the human-in-the-loop is a person with sufficient competence and authority to weigh the reasons that gave rise to the automatic decision and decide whether to confirm or modify it[272]. However, whether or not these requirements are met will involve an independent judgment as to whether or not the decision was made with sufficient human intervention. In any case, data subjects will see their rights exhausted with respect to the initial automated decision, without prejudice to the exercise of the same rights against the refuted (not so human-made) decision. In addition to this scenario in which the right exercised is that of human intervention, the right to contest an automated decision would surely involve some kind of reaction on the part of the data controller, which, if it is some kind of competent human intervention, would take us again to the aforementioned argument. The new decision, whether it confirms or modifies the original automated one, would no longer be considered solely based on ADM and would exhaust the exercise of any of the other safeguards referred in Article 22 paragraph 3.

Therefore, it can be concluded that the three rights towards automated decisions encompassed in Article 22(3) work only cumulative in pairs if data subjects exercise their right to express their points of view and lately either their right to request human intervention or their right to contest the decision. Any other type of chronological use of the rights would exclude the possibility of exercising the other two. Likewise, the unit approach does not seem to offer a logical understanding of the rights since the wording of Article 22 paragraph 3 appears to offer the data subject the option to use the right at his free choice and independently.

Having examined the three safeguards, it is necessary to emphasize that the right to contest a decision is of high relevance in the debate regarding the existence and usefulness of the right to an explanation. Despite the GDPR does not clearly specify

[272] A29WP Guidelines on ADMs and Profiling 27.

what the right to contest is or entails for both the data subject and the data controller, from the rights established in the GDPR it can be interpreted what the right to contest is not. It should not be a right to rectify the information on which the ADM was based, as Article 16 already states a right to correct inaccurate data. Likewise, it cannot merely entail a right to object, as the same is also included in Article 21 as well as it cannot be limited to a right to express one's point of view, as it would collide with other of the safeguard referred in Article 22 paragraph 3. In other words, the right to contest needs to differ in content and form from these other rights or it would be redundant and unnecessary.

Additionally, the right to contest cannot be configured as a right to effective judicial remedy. Article 22 paragraph 3 applies to legitimate uses of ADM; therefore, it would be certainly problematic to interpret the right to contest as a right to effective judicial remedy. This approach would imply that automated but still legitimate decisions are so problematic that they may almost certainly violate the rights and freedoms of the data subjects who would need access to judicial remedy, directly after being the object of such a decision. While both the GDPR and its *travaux préparatoires* are particularly cautious about the threat of automated decisions to data subjects, this interpretation is perhaps exorbitant. Even more so if one takes into account that arguing so would presume that data controllers have at their disposal data subjects' right to an effective remedy. Article 22 paragraph 3 obliges data controllers to implement these safeguards but does not clarify what standards, procedures or mechanism to follow. For this reason, interpreting the right to contest as a right to judicial remedy would be presuming that automated decisions are so threatening to the rights and freedoms of data subjects that the GDPR prescribes direct access to judicial remedy, but at the same time is making access to such remedy dependant to the discretion of the data controller, without making any further specification. The interpretation is contradictory to say the least, all the more so as the right to effective remedy is protected at the highest level in the Charter of Fundamental Rights of the European Union and is always guaranteed[273].

[273] Charter of Fundamental Rights of the European Union 2012 art 47.

This argument could be refuted by indicating that the right to be heard is also a fundamental right as referred in Article 47 of the Charter of Fundamental Rights of the European Union and therefore, Article 22 paragraph 3 would only reiterate the most relevant rights for the data subject in a case of ADM. However, the wording of the Article does not merely reaffirm these fundamental rights but establishes a duty for data controllers who 'shall implement suitable measures' presumably in their organizations. In other words, although the safeguards of Article 22 paragraph 3 reflect fundamental and procedural rights of individuals, the responsibility of their effective exercise falls under the data controller who needs to create suitable platforms for the exercise of these rights by the data subject. As Sarra explained, Article 22 paragraph 3 'imposes to create places and structures within the organization with the end to let the data subject make use of a specific version of her/his fundamental rights, that the data controller in primis should take care of'[274]. In other words, despite the safeguards referred in Article 22 paragraph 3 resemble to individuals' fundamental rights, they are independent and additional safeguards for data subjects when they are affected by ADM whose suitable and practical exercise falls under the duty of data controllers. I will delve into this argument in Chapter 4.3.3..

3.3. Right To Information and an Explanation

3.3.1. Right to explanation pursuant to Article 22(3) in combination with Recital 71

The right to contest serves, according to Kaminski, 'to perfect more substantive rights of fairness and justice and to preserve rule of law values, by correcting, preventing or changing unjust outcomes, and enhancing predictability and consistency of decision'[275]. Translated into the GDPR, the right to contest an automated decision aims to allow data subjects to examine if the particular decision that affected them are fair, just and lawful in accordance with Article 5 of the GDPR as well as other substantive laws, e.g. non-discrimination law, employment law, or consumer law. In particular, the

[274] Sarra (n 269) 12.
[275] Margot E Kaminski and Jennifer M Urban, 'The Right To Contest AI' (2022) 121 Columbia Law Review 1975.

right to contest obliges data controllers to create a mechanism within the ADM that permits data subjects to challenge automated decisions and to receive an adequate respond upon that challenge. Despite being unclear regarding its content, the right to contest of the GDPR 'obliges the data controller either to render automated decisions contestable or to cease [automated decision-making]'[276]. Making a decision contestable, therefore, implies some level of transparency and interpretability regarding the particular decision and the ADM around it. In essence, the right to contest presumes the contestability of the particular automated decision and the implementation of individual transparency and process rights allowing the data subject to inspect the adequacy of the decision in light of the GDPR and the pertinent sectorial laws affecting it, i.e. contract law, employment law, anti-discrimination law.

For these reasons, the right to contest of Article 22 paragraph 3 demands a right to an explanation about the particular decision and the normative grounds of it. ADM are designed and built to achieve certain objectives or serve certain ends and so their outputs - automated decision- are the result of certain inputs. In other words, the automated decision is based on some facts, rules, or norms followed by the decision-making system that give rise to specific legal or similarly significant consequences for the data subject. To effectively contest an automated decision, data subjects would not only need to understand these facts and rules but also the values and principles that surround and hold those norms, hence the normativity of the decision-making system[277]. So the right to explanation that precede the right to contest must address the consequences of the particular automated decision as well as how and why the facts and norms led to that particular outcome or decision. Bayamhoglu clarified that; 'as the initial step of contestation, we need the knowledge of what the system learns about persons, places or events, and how people are represented as inputs to the algorithm'[278]. In other words, to contest a decision, data subjects initially need to obtain an explanation about the relevant features and the inferences relied upon then. They also need to know the normative basis of the decision, what were the decision rules

[276] Bayamlıoğlu (n 268) 1063.
[277] Bayamlıoğlu (n 268).
[278] ibid 1064.

encompassed by hypothesis and assumptions followed to translate those features and inferences in the particular automated decision. Additionally, data subjects need to know the private interests that can determine a decision, e.g. the economic risk of a decision granting a loan or the commercial benefit of a decision denying the rental of a premises.

In essence, the right to an explanation becomes a pre-condition for the right to contestation. Neither the GDPR nor the Article 29 Working Party Guidelines elaborate on the content or form of the safeguard, currently limiting it to a standard rather than a set of specific procedural rules to which controllers must adhere[279].

Interestingly, the safeguards and protections required by Article 22(3) are supported by Recital 71 as well as extended to explicitly incorporate the right to an explanation. As per Recital 71, these safeguards are:

a) Specific information to the data subject,
b) The right to obtain human intervention,
c) The right to express his or her point of view,
d) The right to obtain an explanation of the decision reached,
e) The right to challenge the decision[280],

Despite the non-binding nature of Recitals, the CJEU has already applied them to determine the intent of the valid law, interpreting and establishing its meaning[281]. Thus, it is safe to assume that Article 22 paragraph 3 must be interpreted along the lines of Recital 71, thus accommodating the right to an explanation.

3.3.2. Right to information and access concerning Article 13(2)(h), Article 14(2)(g), and Article 15(1)(h)

Since the beginning of the debate related to new digital practices, the automation of decision-making processes which used personal information to render a decision has

[279] Kaminski and Urban (n 275) 1957.
[280] General Data Protection Regulation recital 71.
[281] *C-503/08 TNT Express Nederland BV v AXA Versicherung AG Reference for a preliminary ruling: Hoge Raad der Nederlanden* [2010] ECJ ECR I 4137- ECLI:EU:C:2010:243 paras 47, 49, 50, 53, and 54.

risen concerns about the ability of data controllers to offer explanations regarding those automated decision and the capacity of data subjects to understand the automated processes and react to them. For this reason, the principle of transparency has always played a key role for the legal approach to ADM as an underlying limit to the risks and harms created by the opaqueness of the systems. In this context, the DPD already compelled Member States to ensure that data controllers provide information to the data subject regarding the purposes of the collection and usage of personal information, as refer in Section IV Articles 11 and 12 of the DPD. Likewise, the Directive granted data subjects with the right to access regarding the use of their personal data, the undergoing processing, and the knowledge of the logic involved in any automated processing of data concerning them[282]. In essence, the DPD contemplated the adverse effects that automated processing of personal data can have for data subjects and create the original framework of transparency which has been consolidated in the GDPR. The Regulation did not only maintain the transparency individual rights firstly outlined by the DPD but included different mechanisms and tools for a better and more concrete exercise of such rights. The safeguards established Article 22 paragraph 3 for automated decisions are reinforced by the information and access requirements established in Articles 13, 14, and 15 of the GDPR. On the one hand, Articles 13(2)(h) and 14(2)(g) respectively compel the data controller to provide the data subject with particular information, establishing the rights to be provided with information where personal data is collected from the data subject or a third party, including information regarding:

> The existence of automated-decision making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject[283].

[282] Data Protection Directive.
[283] General Data Protection Regulation art 13(2)(h) and 14(2)(g).

On the other hand, Article 15(1) includes, a right to confirm as to whether or not personal data concerning him or her are being processed and access to the personal data**,** and again:

> The existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject[284].

So, Articles 13-15 of the GDPR establish a duty to information about the data processing that reaches beyond automated decisions according to which controller shall provide the data subject with information regarding, for example, 'the purpose of the processing or the categories of personal data concerned'. However, knowing the existence of ADM, Article 15 allows data subjects to scrutinise the lawfulness of automated decisions, whether it results from the individual's explicit consent, entails cases of contractual necessity or Member state law. Consequently, the right to information is based on the duty of information encompassed in Articles 13, 14 and 15 of the GDPR.

In particular, the wording of Articles 13, 14 and 15 of the GDPR -i.e., *meaningful information about the logic involved* and *the significance and the envisaged consequences of such processing for the data subject*- directly correspond with the information required by data subjects to contest a particular decision, the decision rule and the consequence of such decision. The difference between the information requirements embedded in these provisions and the respective explainability requirements of the right to contest in Article 22(3) lies in the scope of the information to be provided. Whether Article 22(3) compelled data controllers to provide information about the particular automated decision affecting the data subject, Articles 13, 14 and 15 of the GDPR refers to the ADM as a whole which process the personal information of the data subject. The applicability of all these provisions is restricted, nonetheless, to decisions based solely on automated processing and then will not apply to decisions with a human-in-the-loop. For this reason, the requirements of information as stated in

[284] ibid art 15(1)(h).

13(2)(h), 14(2)(g), and 15(1)(h) will only apply to those ADM which fall under the scope of the exceptions listed in Article 22(2). From this it is stressed, again, the importance as to the interpretation of the title and first paragraph of Article 22, since the application of a more extensive or more restrictive interpretation of its conditions and scope will lead to a more or less exhaustive protection of individuals' rights, freedoms and interests.

Although, as seen in the previous section, Recital 71 referred to Article 22 and makes a direct reference to a right to an explanation and Articles 13, 14, and 15 establish, respectively, rights of information and access in respect of the automated decisions referred to in Article 22(2), the wording of the Articles themselves and the non-binding nature of recitals have given rise to an intense debate on the existence, limits and characteristics of such rights. The scope and exercise of the rights to information and an explanation are, therefore, uncertain and require an in-depth analysis which will provided in the next section.

3.3.3. Access, information and contestability requirements as foundations for the rights to information and an explanation

The existence of safeguards against automated decisions in the form of a right to an explanation has been the subject of debate in academia, particularly given that it is presumed to be a precondition for the right to contested as referred in Article 22(3) and its mention in Recital 71 instead of in the main text and provisions of the GDPR

Goodman and Flaxman[285] first discussed the existence of the right and claimed that the protection was constructed relatively narrow in the GDPR. In a more technical than legal fashion, the authors argued that the right to explanation could be satisfied reasonably easily if algorithms were not designed to be merely efficient but transparent and fair. Goodman and Flaxman drew attention to the challenges of explaining an algorithm's decision when machine-learning systems lack interpretability, which 'refers to the degree of human comprehensibility of a given black-box model or decision'[286].

[285] Goodman and Flaxman (n 46).
[286] PJG Lisboa, 'Interpretability in Machine Learning – Principles and Practice' in Francesco Masulli, Gabriella Pasi and Ronald Yager (eds), *Fuzzy Logic and Applications* (Springer International Publishing 2013); Brent Mittelstadt, Chris Russell and Sandra Wachter, 'Explaining Explanations in AI', *Proceedings*

This first academic impression regarding the right to an explanation highlighted an already traditional technical trade-off between accuracy and interpretability that exists around the algorithmic systems. The duty to provide explanations and information on the operation of algorithmic systems brought this trade-off to the fore as well as reactivated the voices clamouring for the development and use of more interpretable systems or systems whose functioning could be at some level figure out post-hoc. I enhance in this discussion in Chapter 5 and 6.

The response to this first contribution came from Wachter, Mittelstadt and Floridi[287], who disputed the existence of a right to an explanation given the non-binding nature of recitals in the European regulation and doubted the technical feasibility of the provision. The scholars made a contribution to the debate by presenting a framework of possible algorithmic explanations along with chronicle and functional dimensions (i.e., ex-ante and ex-post explanations). Nevertheless, the scholars defended the existence of a right to be informed which provides the data subject with, at minimum, 'a right to an explanation of system functionality [...] subject to restrictions by the interest of data controllers and future interpretations'[288]. Wachter et al. raised their concerns regarding the ambiguity and limited scope of the safeguards and protections offered in Article 22 and claimed that a right to an explanation was not provided for in the GDPR. Still, they claimed that the right to access in Article 15, even if ambiguous, provides a right to be informed of the general system functionality, rather than a right to explanation of specific decisions.

Selbs and Powles sought to rebut both Goodman's and Flaxman's claims and Wachter et al.'s analysis and framework. They stated that the right to explanation 'should be interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise his or her rights under the GDPR and human rights law'[289], for example, the right to not be discriminated according to Article 21 of the European Charter of Fundamental

of the Conference on Fairness, Accountability, and Transparency (ACM 2019) <https://dl.acm.org/doi/10.1145/3287560.3287574> accessed 28 November 2024.

[287] Wachter, Mittelstadt and Floridi (n 45).

[288] ibid 87.

[289] Selbst and Powles (n 46) 233.

Rights[290]. They argued that Article 13-15 of the GDPR provide such a right, even if not directly addressed or named.

To my mind, the relevance of Wachter et. al.'s and Selbs & Powles' claims relies in the importance all of this authors gave to Article 15 and the temporality of its exercise in relation to the possible existence of a right to an explanation. In essence they all argue that a right to know about the logic of the algorithm's performance arises from the exercise of Article 15 when a decision has already been made and the individual wants to ensure its fairness or legality. In short, when the individual may be concerned that he or she has been the subject of an automated, rather than a human, decision and that the decision may not respect the principles and provisions of the GDPR or any other law applicable to the particular case, be it non-discrimination, employment, or health. What is worth recalling about these arguments is that the authors seem to forget the existence of a right to contest such a decision, as referred in Article 22(3), which would directly open the possibility of challenging the lawfulness and fairness of the potential automated decision. While it is true that the subject must be aware of the existence of an automated decision either ex ante on the basis of articles 13 and 14, or ex post on the basis of article 15, the right of explanation would not arise from these precepts, but from the safeguard right contained in paragraph 22 to contest such decisions. Arguably, the authors seek to offer an interpretation of Articles 22 and 15 too convoluted when Article 22, paragraph 2 offers a more straightforward way to claim that the right to an explanation is guaranteed in the GDPR.

Besides these three initial and prominent claims, other scholars contributed to the debate, broadening the academic interpretations of the existence of the right to explanation in the GDPR.

Edwards and Veale[291], in turn, did not deny the existence of the right but claimed that if it did exist, such a right would not be an appropriate safeguard towards ML algorithms due to the difficulties in its enforcement in the context of black-box systems. Indeed, they argued that the legal provisions which embedded the right are 'restrictive, unclear,

[290] ibid 242.
[291] Edwards and Veale (n 89).

or even paradoxical concerning when an explanation-related right can be triggered', which would also create difficulties for computer scientists to provide the information required by the legal provisions (i.e., explanation regarding meaningful information about the logic of processing).

More comprehensive interpretations of the right to explanation were proposed in three other contributions by Malgieri and Comandé[292], Pagallo[293], and Mendoza and Bygrave[294].

Firstly, Malgieri and Comandé defended the necessity of a systematic interpretation of Articles 13-15 and Article 22 and incorporated an original concept to the ongoing academic debate, algorithm legibility. In essence, the scholars argued that the

> Legibility of data and analytics algorithms is a concept able to combine comprehensibility of the functioning of the algorithm (for which we will use the term 'architecture') with transparency about the commercial use of that algorithm (for which we will use the term 'implementation') in an effective way[295].

Hence, Malgieri and Comandé asserted that algorithm legibility 'offers the most appropriate interpretation of the right to know the meaningful information about the logic involved in a decision making, Article 15 (1)(h) combined with Article 22 of GDPR'[296], as it integrates algorithm's transparency and comprehensibility. This proposal brings together both technical and legal attempts to encourage the use and development of ADM systems that together allow for a higher level of transparency and therefore accountability. The authors argue that both information and explanation rights must be interpreted from a comprehensive point of view that is not limited to providing information about the decision and the logic of the processing. I strongly agree with the academics' claim that the information and explanation requirements of Articles 13, 14, 15 and 22 of the GDPR call for a transparent and accountable use of ADM, therefore;

[292] Malgieri and Comandé (n 47).
[293] Pagallo (n 208).
[294] Mendoza and Bygrave (n 45).
[295] Malgieri and Comandé (n 47) 245.
[296] ibid.

requiring the interpretability and explainability of the model in order to ensure respect for individuals' rights, freedoms and interest.

Secondly, Pagallo presented the right to an explanation as a conjunction of Article 22 and its recital 71, along with the rights to information and access encompassed in Articles 13-15. The scholar drew attention to the extent of such right, arguing that the 'legal problem around the right does not revolve around whether a right to an ex-post explanation exists in EU law. Rather, the issue concerns the extent of such right, e.g., whether the right to an ex-post explanation includes the explanation of how algorithms work'[297]. Whilst it is true that providing the data subject with information about how the algorithm works may conflict with the interests and rights of the data controller, if one accepts the existence of the right to an explanation it seems logical to affirm that one accepts the right to know the reasons for that explanation and the manner or logic through which it has led to that decision. Therefore, the scope of the explanation provided should balance the interest of both parties while respect the objectives of the GDPR.

Mendoza and Bygrave resumed the debate regarding the chronicle and the functioning dimension of the right to explanation. The authors explored whether the GDPR supports the ex-post explanation of a particular decision besides the ex-ante explanations of systems functionality provided by Articles 13-15. According to the scholars, the wording of Article 15 does not necessarily exclude the possibility that it embraces a right to an ex-post explanation of an automated decision as referred to in Article 22. On the contrary, the scholars' claimed that

> A right of ex-post explanation of automated decisions is implicit in the right 'to contest' a decision pursuant to art. 22(3). The term 'contest' connotes more than 'object to' or 'oppose'; simply put, a right of contest is not simply a matter of being able to say 'stop' but is akin to a right of appeal. If such a right is to be meaningful, it must set in train certain obligations for the decision maker, including (at the very least) an obligation to hear and consider the merits of the appeal. If the appeal

[297] Pagallo (n 208) 518.

> process is to be truly fair, it must additionally carry a qualified obligation to provide the appellant with reasons for the decision. The need to give reasons is buttressed by the general principle of 'lawfulness, fairness and transparency' in art. 5(1)(a) which animates most of the basic norms of the Regulation, including the provisions of art. 22[298].

Mendoza and Bygrave assertion agrees with the arguments presented in Section 3.2.5. which assert that the right to an explanation is born as a prerequisite to the right to contest an automated decision. Arguably, the extend of the information provided based on such right needs to be as extensive as to guarantee an effective exercise of the individuals' guarantee rights referred in Article 22(2). In essence, the right to an explanation would not be respected merely providing as little information as possible about the decision, but the data controller must provide sufficient content for the individual to understand the reasons, normative basis, and logic that led to the final decision. The opposite scenario would contravene the principle enshrined in Article 5 of the GDPR as it would attempt to elude the data controller obligations towards fairness and transparency during the processing of personal data.

Finally, other scholars[299] suggested that the scope and usefulness of the rights to information and explanation could be further limited by the transparency requirement encompassed in Article 12, which compels data controller to make an effort to communicate information in a way understandable to individuals.

> The controller shall take appropriate measures to provide any information referred to in Articles 13 and 14 and any communication under Articles 15 to 22 and 34 relating to processing to the data subject in a concise, transparent, intelligible and easily accessible form, using clear and plain language[300].

[298] Mendoza and Bygrave (n 45) 94.
[299] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' [2017] SSRN Electronic Journal <https://www.ssrn.com/abstract=3063289> accessed 10 January 2022.
[300] Dino Pedreschi and others, 'Meaningful Explanations of Black Box AI Decision Systems' (2019) 33 Proceedings of the AAAI Conference on Artificial Intelligence 9780, 9780.

Ananny & Crawford[301] included a clarification of the above claim defending that Article 12 also prevents data controller from flooding individuals with irrelevant and useless information or abusing notice and consent notification to create some sort of obscurity through information overfloods. Kaminski[302] merge these two conceptions of Article 12 by asserting that the provision would compel companies to provide individuals with comprehensible, intelligible and actionable information, rather than reducing the value of the rights to meaninglessly high-level or simplistic information. According to Kaminski, 'the rights to information and an explanation shall provide enough information that an individual can act on it -to contest a decision, or to correct inaccuracies, or to request erasure'[303]. This same argument was followed as well by Selbs and Powles as they stated that the data subjects' safeguard to contest any automated decision is reinforced by the emphasis on meaningful transparency assessed in Article 12[304].

It is conceivable that given the complexity and inaccessibility of ADM, data controllers could take advantage of this circumstance and, while complying with the information and explanation requirements of the GDPR, offer such an obscure and technically complex information that the data subject would be hardly able to understand nor decipher it. As defended by Ananny & Crawford and Kaminski, this scenario would contravene the objectives of the GDPR and infringe the principle of transparency referred in Article 12 of the GDPR. To my impression, this principle does not only act as a limitation to the potential misconduct of data controller, but reinforce the safeguard rights established in Article 22(2) by ensuring that the information and explanations they are provided with are not only potentially useful for the exercise of their rights but undeniably concise, intelligent, and easily accessible.

Despite the unsettled academic debate, the Article 29 Working Party Guidelines seem to confirm the existence of a right to explanation independent of the data subjects' right to information by mentioning on three occasions the 'safeguard to obtain an

[301] Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973.
[302] Kaminski (n 208).
[303] ibid 213.
[304] Selbst and Powles (n 46).

explanation of the decision reached after such assessment and to challenge the decision'[305]. The Guidelines address the need for this type of transparency safeguard because individuals would only be able to contest a decision or express their particular view if they 'actually understand how it has been made and on what basis'[306]. Furthermore, the Guidelines proposed that

> Instead of providing a complex mathematical explanation about how algorithms or machine-learning work, the controller should consider using clear and comprehensive ways to deliver the information to the data subject, for example: the categories of data that have been or will be used in the profiling or decision-making process; why these categories are considered pertinent; how any profile used in the automated decision-making process is built, including any statistics used in the analysis; why this profile is relevant to the automated decision-making process; and how it is used for a decision concerning the data subject[307].

In other words, the Article 29 Working Party seemed to present the right to an explanation in Article 12 as a precondition to the exercise of other data subjects' rights in the GDPR, i.e., right to contest a decision or to be heard and express their view.

Again, despite the lack of clarity in the wording of the GDPR, from the analysis of the text itself, as well as the academic debate and the guidance offered by the Article 29 Working Party, several conclusions can be drawn.

First, paragraph 3 of Article 22 makes direct reference to three suitable safeguards towards ADM: (1) right to obtain human intervention on the part of the controller, (2) to express his or her point of view and (3) to contest the decision. It is important to acknowledge that for the proper exercise of two of these three protections data subjects should understand the decision affecting them and have sufficient information about it to develop a personal opinion in order to express their point of view and, if they deem it

[305] A29WP Guidelines on ADMs and Profiling 19, 27, 35.
[306] ibid 27.
[307] ibid 31.

appropriate, to contest it. It is arguable that the explicit inclusion in the GDPR of the notion right to an explanation is not necessary to confirm its existence.

Second, Articles 13(2)(h), 14(2)(g), and 15(1)(h) grant the data subject a right to information about the existence and logic of the automated decisions to which he or she may or will be exposed. The existence of such a right is not in doubt, although its scope is.

Third, one of the major problems associated with the existence of the rights to information and an explanation lies in their feasibility. The information that these rights deliver can be differentiated between information related to the legitimate interest of the data controller in making use of an ADM (i.e., significance and the envisaged consequences of such processing for the data subject) and the information about the functioning of the algorithm and the intricacies of the decision (i.e., meaningful information about the logic involved). While the first category may certainly be affordable for data controllers as it would require the simplification of an internal and business decision that has already been made and reflected upon, the second category will undoubtedly entail technical and non-technical challenges. Such challenges should not, however, override or overshadow the importance of both rights, nor justify their denial or dismissal as part of the compendium of data subject rights included in the GDPR.

Lastly, the rationale behind these two rights can be presumed from the need to ensure that ADM respects the principles of fairness, legality, and transparency encompassed in Article 5 and Article 12 of the GDPR. For all these reasons, although there are undeniably problems and uncertainties with regard to the exercise and implementation of these rights, their existence seems assured.

This thesis, therefore, follows the argument that there is a logic between the data subject's rights stated in the GDPR -i.e., right to contestation, correction, and erasure- and the individualised transparency equally stated on the GDPR. In general terms, individuals have a right to information – accessible and comprehensive- to ensure fair and transparent data processing. For example, individuals need to know the possible

errors to correct the information about them, and they need to know the factors used in a decision to contest it. Otherwise, the rationale behind these data subject rights would remain useless, and the fairness of processing would be questioned.

The academic debate surrounding the rights to information and explanation based on the wording of the GDPR and the Article 29 Working Party guidance highlights the importance of the scope of their application in practice and the extension of the safeguards offered to the data subjects in different contexts and cases. The different interpretations introduced in academia regarding the rights to information and an explanation demonstrate that the nature of ADM and their application create challenges for the enforcement and exercise of the legal requirements and provisions created to protect the individuals' rights.

3.3.4. The relevance of temporality: ex-ante and ex-post information and explanations

The initial main difference between Articles 13(2)(h) and 14(2)(g), and Article 15(1)(h) in the context of ADM resides in their scope and moment of exercise concerning automated decisions, as referred to in Article 22. Articles 13(2)(h) and 14(2)(g) create an ex-ante notification duty about the existence of an automated decision that does not require any previous awareness by the data subject about the processing of their data nor grant more information than a general disclosure about the system functionality. Article 15(1)(h), in conjunction with Article 22, establishes both an ex-ante and an ex-post notification duty. The former is about the existence of an automated decision, the logic involved and the significance and the envisaged consequences of such processing. The latter deals with specific information about the logic behind the decision taken towards the data subject that allows the individual to understand and, if appropriate, contest the decision in accordance with Article 22(3) safeguards. Additionally, Article 22(3), establishes an ex-post explanation duty for the data controller regarding the particular decision with the objective of allowing the data subject to contest the decision if decided so.

The distinction between ex-ante and ex-post information and explanations has sparked academic debate possibly on the same level as the very existence of the rights, giving rise in many cases to arguments from both sides of the debate that refute and support each other. Scholars have discussed whether the rights to information and an explanation would distinguish between ex-post and ex-ante explanations. Thus, the discussion delved into the extent of this differentiation in the GDPR provisions and the impact it can have in the exercise and enforcement of the rights. The debate elaborates on how ex-ante and ex-post explanations may strengthen or weaken the rights and legitimate interests of affected data subjects.

In the context of the GDPR, ex-ante generic explanations are usually associated with the right to meaningful information about the logic involved and the significance and the envisaged consequences of such data processing for the data subject as referred in Article 13(2)(h) and 14(2)(g). In other words, this type of ex-ante explanations would offer information about the system functionality and so recall to the traditionally well-accepted right to informed consent. Specifically, ex-ante information would offer enough information to ensure that individuals are well informed before deciding to consent or entering into a contract with an automated decision[308]. Indeed, the Italy's Corte Suprema de Cassazione appears to follow this argument as referred in the case Civile Ord. Sez. 1 Num. 14381 where it rules that the consent to an ADM requires the awareness towards the executive scheme (i.e. logic involved) and the constitutive elements of the algorithm[309].

Ex-post explanations, by contrast, refers to the specific decision and are associated with a 'right to a remedial explanation as a precondition for placing trust intelligently'[310] as well as a way to ensure fair, lawful and transparent processing of data. Hence, ex-post explanations ensure that the decision-maker respond fairly and responsibly to possible wrongful processing of personal data, offering the affected individual a way to

[308] Kristina Astromskė, Eimantas Peičius and Paulius Astromskis, 'Ethical and Legal Challenges of Informed Consent Applying Artificial Intelligence in Medical Diagnostic Consultations' (2021) 36 AI & SOCIETY 509.
[309] *Civile Ord Sez 1 Num 14381* (n 162); as referred in Barros Vale and Zanfir-Fortuna (n 48).
[310] Tae Wan Kim and Bryan R Routledge, 'Informational Privacy, A Right to Explanation, and Interpretable AI', *2018 IEEE Symposium on Privacy-Aware Computing (PAC)* (IEEE 2018) 64 <https://ieeexplore.ieee.org/document/8511831/> accessed 28 November 2024.

redress the grievances and defend their interests. In the context of the GDPR, ex-post explanations will entail 'the logic or rationale, reasons, and individual circumstances of a specific automated decision'[311], as established in Article 22(3) referring to the right to contest a particular decision. Ex-post explanations require a prior commitment from the decision-maker but ultimately provide information about a specific decision once it has been made. Therefore, ex-post explanations would have to offer and explain the values and metrics used to obtain the outcome that result in the particular final decision.

The *SCHUFA Holding (Scoring)* case shows, however, that the distinction between ex-ante and ex-post explanations can be also affected by the existence and interaction of different third parties for the collection and use of personal data through automated processing. Indeed, the interest of the referral Court falls on the effective enforcement of rights by data subjects when the data controller and the decision-maker are two different entities. In turn, the importance of *SCHUFA Holding (Scoring)* case regarding the temporality of the rights to information and an explanation relies on the determination of the specific party entitled to provide information to the data subject.

Specifically, by estimating credit scoring as an independent solely automated decision as referred to in Article 22, the data subject would have the right to request information and explanations from the data controller following Article 13 and 14. By contrast, if such a decision shall not be recognised by the CJEU as an automated decision as stated in Article 22, the data subject would find the options to exercise their rights of information and explanation quite restricted. In other words, data controllers -such as credit information agencies- would not be obliged to disclose the logic and composition of the parameters that are decisive for the establishment of a score, nor would the decision-making third parties be able to offer the same such information as the logic involved is not disclosed to them.

As stated by the Referral Court, the latter situation would give rise to a potentially -serious- lacuna in the legal protection:

[311] Edwards and Veale (n 89) 52.

> the party from whom the information required for the data subject could be obtained is not obliged to provide access to information under Article 15(1)(h) of the GDPR because it allegedly does not engage in its own 'automated decision-making' within the meaning of Article 15(1)(h) of the GDPR, and the party that bases its decision-making on the score established by means of automation and is obliged to provide access to information under Article 15(1)(h) of the GDPR cannot provide the required information because it does not have it[312].

The First Chamber ruling whereby credit scoring -as established in the case circumstances- shall be considered an automated decision under Article 22 prevents the circumvention of Article 22. Furthermore, the First Court ruling compels controllers to lay down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, including the right to challenge the decision taken in his or her regard. The Ruling does not clarify who in practice shall provide the information and explanations about the credit score to the individual, e.g. whether it should directly be the credit scoring agency or the bank after being properly informed by it. However, what becomes clear from the SCHUFA Holding (Scoring) ruling is that data subjects cannot remain defencelessness under the pretext that the credit scoring agency is not compelled to provide information and explanations to them nor that the bank is unaware of the such knowledge and therefore unable to provide it by itself.

Notwithstanding, it has become clear in the discussion amongst scholars that the right to information and the right to an explanation entails three different types of explanations regarding the algorithmic systems*:*

- Ex-ante about the general logic,
- Ex-post about the general logic, and
- Ex-post about the specific decision.

In the context of the GDPR, the first seems to apply to Articles 13(2)(h), 14(2)(g), and Article 15(1)(h) rights to ex-ante notification about the processing of personal data and

[312] *C-634/21 SCHUFA Holding (Scoring) - Request for Preliminary Ruling from the Verwaltungsgericht Wiesbaden* (n 169) 25.

ex-ante general information about the algorithm functionality. The second would apply to Article 15(1)(h) when exercised after an automated decision has been made affecting the data subject. Finally, the third would relate to Article 15(1)(h) when exercised after an automated decision took place and therefore actives the safeguard rights referred in Article 22(3), or when Article 22(3) applies as an automated decision is taking place and the data controller is laying down the required suitable safeguards, hence the right to an explanation about the 'specific details on a concrete decision process which has concerned the data subject'[313] as a precondition for the exercise of the right to contest. The academic debate regarding ex-ante and ex-post explanation is, nonetheless, worthy of presenting.

Wachter et al. argued in favour of a limited right to be informed (instead of a right to an explanation) that includes a duty to provide ex-ante information about the general functionality of an ADM, rather than ex-post information of how specific automated decision on an individual was made[314]. Especially, they argued that providing meaningful information about the logic involved does not essentially entail providing information on the rationale and circumstances of a particular decision. Certainly, the authors favour an interpretation of the GDPR limited to ex-ante information. Contrary to this claim, Mendoza and Bygrave defended that 'the possibility of a right of ex-post explanation of automated decision is implicit in the right to contest a decision'[315]. Moreover, Selbst & Powles[316] argued that access to a system-level explanation would and should provide the necessary information to understand the specific decision. In this regard, meaningful information would not necessarily require information on a specific decision but an explanation regarding the whole system. Through the legibility test, Malgieri and Comandé[317] argued, in turn, that data controllers should provide meaningful information about the architecture and the implementation of the decision-making algorithm. Their proposal, then, encompasses an approach to information duties more focussed on ex-ante information, but including algorithm's accountability

[313] Malgieri and Comandé (n 47) 247.
[314] Wachter, Mittelstadt and Floridi (n 45).
[315] Mendoza and Bygrave (n 45) 16.
[316] Selbst and Powles (n 46).
[317] Malgieri and Comandé (n 47).

and audit as a key aspect of the GDPR framework presented in Articles 13, 14 and 15. This framework of information rights that presents a combined approach to accountability would force developers to make their systems understandable and transparent in their general functionality and their potential impacts on the data subjects. Likewise, it would foster users' interest in understanding the systems' functioning to provide explanations to end-subjects and increase trust in the system on their own behalf and behalf of individuals.

Additionally, Article 29 Working Party Guidelines asserts that the information provided to the data subject as result of Articles 13(2)(h) and 14(2)(g) 'has to be sufficiently comprehensible for the data subject to understand the reasons for the decision'[318]. Although Article 29 Working Party does not specifically mention whether it refers to ex-ante or ex-post explanation, it is reasonable to presume that these articles refer to ex-ante explanations as it is the data controller who must provide the information to the data subject when obtaining personal data about him/her. Indeed, 'the GDPR requires the controller to provide meaningful information about the logic involved, not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm'[319]. The information provided to the data subjects is intended to explain the possible effects of the system on the individual case as well as its rationale and the main characteristics used to reach the decision, the source of the information and its relevance.

Article 29 Working Party Guidelines acknowledge that Article 15 (1)(h) obliges the controller to provide information 'about the envisaged consequences of the processing, rather than an explanation of a particular decision'[320]. However, the Guidelines also recognised that 'data controllers should provide the data subject with general information -notably on factors taken into account for the decision-making process, and on their respective 'weight' on an aggregate level-'[321]. Importantly, the Guidelines clarify that the information provided to the data subject 'has to be useful for the data

[318] A29WP Guidelines on ADMs and Profiling.
[319] ibid 25.
[320] ibid 27.
[321] ibid.

subject to challenge the decision'[322]. In other words, the Guidelines seems to refer to the normativity and decision rule of the whole decision-making process as necessary information for data subjects to contest a decision.

Although the wording of the Article 29 Working Party Guidelines does not clarify whether the right to information and an explanation involves ex-ante or ex-post explanations, it emphasised the priority to protect the subject data rights and interests, which, in turn, would demand extensive provision of information to facilitate the compliance with the suitable safeguards presented in Article 22(3) and, therefore, the required comprehension of both the ADM and the particular decision under question. Whether some of these information requirements would demand ex-ante information about the logic involved and the envisaged consequences, others would ask for more particular details, and therefore ex-post information, about the decision-making process and the particular automated decision. Finally, under the Guidelines 'a complex mathematical explanation about how algorithms or machine-learning work will generally not be relevant, it should also be provided if this is necessary to allow experts to verify further how the decision-making process works'[323]. Unfortunately, how or when this extensive disclosure of information becomes necessary has not yet been clarified under the Article 29 Working Party Guidelines. Notwithstanding, the right to contest an automated decision as referred in Article 22(3) can offer an appropriate mechanism for data subjects to request further information and to engage with the data controller in a discussion regarding the automated decision affecting them and its compliance with the principles and provisions of the GDPR.

3.4. Discussion

3.4.1. Framework of the rights to information and an explanation

Based on the analysis presented in the previous sections, the following frameworks outlines the stance of this thesis concerning the temporality, application of the Articles 13(2)(h), 14(2)(g), 15 (1)(h) and 22(3) of the GDPR, and the type of the information to be provided upon them. This framework aims to provide a clear and individualised view on

[322] ibid.
[323] ibid 29.

the position of this thesis regarding the precepts of the GDPR that address both the right to information and the right to an explanation.

Article 13(2)(h) & Article 14(2)(g)	
Temporality	-Ex-Ante Duty of the Data Controller to Provide Information
Application	- Active Compliance by the Data Controller
Type of information	- Existence of automated decision-making, including profiling - Meaningful information about the logic involved - The significance and the envisaged consequences

Figure 1: Framework Article 13(2)(h) & Article 14(2)(g)

Article 15(1)(h)	
Temporality	- Ex-Ante Duty of the Data Controller to Provide Information - Ex-Post Right of the Data Subject to Access and Require Information
Application	- Active Exercise by the Data Subject
Type of information	- Existence of automated decision-making, including profiling - Meaningful information about the logic involved - The significance and the envisaged consequences

Figure 2: Framework Article 15(1)(h)

Article 22(3)	
Temporality	- Ex-Ante Duty of the Data Controller to Provide Information - Ex-Post Duty of the Data Controller to Provide Explanation
Application	- Active Compliance by the Data Controller
Type of information	- Enough information of the automated decision-making to consent or enter in a contract (ex-ante) - Enough explanation of the decision reached after such assessment to contest the decision (ex-post) - Enough explanation of the decision reached after such assessment to express his or her point of view (ex-post)

Figure 3: Framework Article 22(3)

This thesis holds the view that the rights to information and an explanation arise from different Articles of the GDPR. Firstly, Articles 13(2)(h) and 14(2)(g) of the GDPR establish a duty for the data controller to provide information to the data subject

regarding the existence of ADM, meaningful information about the logic involved, and the significance and the envisaged consequences. As a result of this duty, the data subject has an *ex-ante right to* information which does not need to be actively exercised. Additionally, Article 15(1)(h) stipulates the data subject's right to access information, including information regarding the existence of ADM, meaningful information about the logic involved, and the significance and the envisaged consequences. As the exercise of this right resides in the active exercise of it by the data subject, it is arguable that it can entail *ex-ante* or *ex-post information* about the decision. Given that it is the subject who has to exercise that right, it is apparent that Article 15(1)(h) covers situations where either no automated decision affecting the subject has yet been taken, or a decision has been taken and the subject sought to corroborate whether it is an automated or not. In both cases, the information received by the data subject shall be the same as for Articles 13(2)(h) and 14(2)(g), yet, if an automated decision has already took place, the information shall be complemented as per Article 22 (3) and its Recital 71. Thus, this ex-post information about the particular decision shall also explain the decision reached after such assessment to contest the decision and express the point of view of the data subject. Indeed, the first case will involve a right to information, while the second will entail a right to explanation. Finally, Article 22 establishes both a right to information and a right to an explanation. The first will arise from the ex-ante necessity of the data subject to consent to and enter into a contract with ADM as referred to in the exceptions of Article 22(2). The second will arise from the ex-post effective exercise of the safeguards set out in Article 22(3), namely, to express his or her point of view and to contest the decision.

This chapter concludes that the GDPR establishes a framework of transparency and explainability around ADM by incorporating different mechanisms and tools for data subjects to assert their rights. Specifically, these mechanisms are based on the transparency and explainability requirements referred to in Articles 13(2)(h) and 14(2)(g), Article 15(1)(h) and Article 22. However, whether these requirements have been starting to be addressed at different levels, there are still no practical specifications or requisites to implement them. The GDPR only assesses the aims and motives for the existence of the rights to information and an explanation but fails to

concretise the exact information and explanations that would comply with the transparency and explainability obligations. It is worth noting that real case scenarios of ADM pose different threats to individuals' rights, freedoms, and interests. Therefore, the expectations and goals to be achieved by data subjects when enforcing their rights to information and explanation may differ in scope and significance beyond the common aim of verifying the lawfulness, accuracy and fairness of the processes they are subject to.

3.4.2. The spectrum of compliance – minimum and maximum thresholds

It would be counterproductive for this thesis to offer a single argument as to what exact information and explanations have to be provided in order to comply with the obligations set out in Articles 13(2)(h) and 14(2)(g), Article 15(1)(h) and Article 22 of the GDPR. On the contrary, this thesis argues that the uniqueness of each case, and the interest of data controllers in respecting, to a greater or lesser extent, the framework of transparency and explainability encompassed in the GDPR, will be essential in the fulfilment of the rights of information and explanation. As expressed by the Austrian Administrative Court in the Request of preliminary ruling C-203/22[324], the content requirements that must be met by the information provided in order to be classified as sufficiently *meaningful* with the meaning of Article 15(1)(h) -shall be understood as extending to Articles 13(2)(h) and 14(2)(g)- is still uncertain. Furthermore, this uncertainty applies to the other concepts referred to in the GDPR, such as *the logic involved* and the *significance and meaningful consequences*. As the referral Court put it:

> The information provided under the GDPR is only sufficiently meaningful if the person requesting the information is enabled to actually, profoundly and promisingly exercise the rights guaranteed to him/her by Article 22(3) GDPR to express his/her own point of view and to contest the automated decision within the meaning of Article 22 concerning him/her?[325].

[324] *C-203/22 Dun & Bradstreet Austria GmbH - Request for preliminary ruling from the Verwaltungsgericht Wien* (n 170).
[325] ibid.

For the referral Court, all the data subject's rights related to ADM -as referred in Article 22- are founded on the protection of individuals against the particular risks to their rights and freedoms represented by the automated processing of personal data. Hence, the minimum content of Articles 13(2)(h) and 14(2)(g), Article 15(1)(h) cannot be determined without taking into account the purpose of Article 22 and so the way in which the data subject may effectively use the rights -safeguards- conferred on him or her by the same Article 22. This argument highly resonates with the arguments presented in the previous sections by which the right to information and an explanations are prerequisites for the effective exercise of the right to contest an automated decision.

In the views of Advocate General Pikamäe and Advocate General De la Tour, this argument seems appropriate,

> The obligation to provide 'meaningful information about the logic involved' must be understood to include sufficiently detailed explanations of the method used to calculate the score and the reasons for a certain result. In general, the controller should provide the data subject with general information, notably on factors taken into account for the decision-making process and on their respective weight on an aggregate level, which is also useful for him or her to challenge any 'decision' within the meaning of Article 22(1) of the GDPR[326]

Advocate General De la Tour asserts that the purpose of Article 15(1)(h) -and by extension Articles 13(2)(h) and 14(2)(g)- is to achieve the objectives of the GDPR, particularly a consistent and high level protection for natural person within the EU and an strengthening of data subjects' rights. The Advocate General clarifies that, in general, these rights 'must enable the data subject to ensure that the personal data relating to him or her are correct and that they are processes in a lawful manner.'[327]

---

[326] *C-634/21 SCHUFA Holding and Others (Scoring) Opinion of Advocate General Pikamäe* [2023] ECJ ECLI:EU:C:2023:220 point 58.

[327] *C-203-22 Dun & Bradstreet Austria GmbH - Request for a preliminary ruling from the Verwaltungsgericht Wien - Opinion of Advocate General Richard de la Tour* [2024] ECJ ECLI:EU:C:2024:745 referring to; *C-487/21 Österreichische Datenschutzbehörde and CRIF - Request for a preliminary ruling from the Bundesverwaltungsgericht* ECJ EU:C:2023:369 para 34.

In essence, this thesis defends that the minimum threshold of compliance of the rights to information and an explanation is delimit by the relationship between these with the safeguards set out in Article 22(3) and the rights of data subjects' rights established in the GDPR. This is to say, that the right to information and an explanation shall also allow the data subject to check the compliance of the ADM and the particular decision with the principles related to the processing of personal data; i.e., lawfulness, fairness, accuracy and transparency. If the data subject is not provided enough information and explanations about the decision-making process and the particular decision to exercise their rights if deemed necessary and to confirm the normativity of the process and decision, it could not be said that the data controller has not complied with his or her duties of transparency and explainability.

Having established this, the right to information – as per Articles 13(2)(h) and 14(2)(g), Article 15(1)(h)- and an explanation -as per Article 22(3) and Recital 71 -require the provision of meaningful information about to the rationale behind the logic followed by the decision-making process and the criteria used to reach the final decision as well as the significance and the envisaged consequences of such processing for the data subject. The information and explanation needs to be contextualise to the particular individual case, so the data subject can verify that the particular ADM affecting him or her, and it pertinent decision- do respect the processing principles of the GDPR. Therefore, this thesis follows Advocate General De la Tour Opinion by which information about the logic involved in ADM shall be in the first place 'concise, easily accessible and easy to understand, and formulated in clear and plain language [and about] the method and criteria used for that decision'[328] and in second place

> Sufficiently complete and contextualised to enable that person to verify its accuracy and whether there is an objectively verifiable consistency and causal link between, on the one hand, the method and criteria used and, on the other, the result arrived at by the automated decision.[329]

[328] *C-203-22 Dun & Bradstreet Austria GmbH - Request for a preliminary ruling from the Verwaltungsgericht Wien - Opinion of Advocate General Richard de la Tour* (n 327) para 71.
[329] ibid para 71.

To my understanding, this minimum threshold obliges data controllers to make their systems sufficiently transparent and intelligible, so data subject can understand the normative reasons of the decision affecting them and the main elements of the system affecting them, without necessarily offering complex explanations about the system's functioning. Additionally, the minimum threshold requires providing information regarding the consequences an automated decision could have for the data subject, addressing how the rights and freedoms of data subjects could be affected.

In essence, the minimum threshold of compliance still requires a comprehensive approach to the transparency and explainability of ADM by encompassing the basic decision-making process's rules and consequences, as well as the characteristics of the data subject that were used as the main criteria to reach the decision. A necessary condition to this minimum threshold of compliance is set out in Article 12 (1) by which the information and explanations need to be concise, easily accessible, and easy to understand, and that clear and plain language is used. This condition prevent the provision of technical or mathematical information and explanations about the decision-making system and the automated decision, since it will contravene the content of Article 12 (1).

Therefore, the minimum threshold of compliance not only requires data controllers to provide information to data subjects about the logic of the system and the main criteria used in some checklists but would oblige the former to consciously design and use systems that respect the final aims of the transparency and explainability requirements of the GDPR. In other words, this minimum level of compliance would impact the ADM beyond the final decision affecting the data subject requiring that the systems are as intelligible and transparent as possible. By requiring this, this level of compliance would provide data subjects with sufficient information to understand the ADM affecting them -e.g. the decisions made during the design of the system, the importance of the criteria for the logic of the system, or their classification and determinant criteria for the final decision-. In essence, the minimum threshold allow data subjects to verify the lawfulness, accuracy, and fairness of the decision-making process and the decision and to effectively exercise the safeguards encompassed in Article 22(3). The level of compliance assumes that the GDPR's requirements concerning transparency and

explainability cannot be understood as independent clusters or checklists but must be adopted as a comprehensive framework that aims to protect and strengthen individuals' rights, freedoms, and interests. For this reason, the information and explanation provided to data subjects following the rights to information and an explanation would need to offer sufficient *meaningful* information to ensure that data subjects can verify the lawfulness and fairness of the processing and contest the decision if needed.

The level of compliance with the rights to information and an explanation does not depend on mere legal analysis or interpretations. However, it is largely a response to technical issues and constraints. Building on the arguments and conclusions offered in the doctrinal analysis presented in this chapter, Chapter 6 will set out the main methods of explanation and the advantages and disadvantages they present in the GDPR framework of transparency and explainability.

This thesis argues that even the minimal threshold of compliance with transparency and explainability rights requires ex-ante and ex-post information and explanations about the decision. On the one hand, the general logic of the system and the consequences of ADM related to ex-ante general information as referred to in Articles 13(2)(h) and 14(2)(g) and Article 22(2) and ex-post information as established by Article 15(1)(h). On the other hand, ex-post information and explanation about the concrete decision affecting the data subject related to Article 22(3).

Beyond this minimum threshold of compliance, the framework of transparency and explainability established in the GDPR can be confronted so that a further extensive approach to the rights to information and explanation is embraced in an spectrum of compliance. To my understanding, beyond this minimum threshold, data controllers can decide the amount of information they provide to data subjects when complying with the rights to information and an explanation. By doing so, they can decide to provide more technical information about the decision-making system and accompanied it with non-technical information that allow the data subject to understand and comprehend it.

Likewise, this spectrum of compliance can vary depending on the needs of the data subjects in so far as they can request more information that the one originally provided for not considering it sufficient to exercise their rights and verify the lawfulness, accuracy, and fairness of the processing. In this regard, it could be said that rather than an spectrum of compliance, the minimum threshold can vary depending on the data subject and the context the decision has took place. As pointed out in this Chapter, the right to contest a decision can be considered to entail some level of confrontation between parties, so for the purpose of simplification, I assume that the possible back and forth between the data subject and the data controller is part of the spectrum of compliance rather than the minimum threshold.

The spectrum of compliance of the rights to information and an explanation, and it minimum threshold, will compel data controllers to make their systems as transparent and explainable as possible, understanding the decision-making process and the consequences that the decisions reached could have to the extent that they can provide this information to data subjects clearly and thoroughly.

That said, the spectrum of compliance is not infinite. Article 12 (1) already makes the compliance to the rights conditional to the easy, accessible, and understandable nature of the information and explanation provided. Although additional information can be offered to ensure the comprehensibility of technical information about the ADM, there is undoubtedly a limit to what an average individual can properly understand. Article 12 (1) may be more relevant to determine the minimum threshold of compliance than the maximum threshold, but still it is relevant to have it in mind. However, the maximum threshold of compliance of the rights to information and an explanation is certainly delimited by the legitimate interests of third parties. The compliance with the rights to information and an explanation shall not infringe legitimate trade secrets or property rights of the data controller, nor shall entail a violation of third parties data protection rights.

The conclusion to be drawn from this chapter is, therefore, that information rights would enable users to enforce their rights and interests. Otherwise, it would be hard for citizens to enjoy the rights granted to them in the GDPR and other applicable laws.

Nonetheless, the extent of the information required by individuals may vary depending on the rights they intend to exercise and the decision they want to contest. Consequently, this thesis argues that to contest an automated decision, individuals need to know how that decision was made and, therefore, the general functionality of the system and the specific factors that lead to the final decision.

The doctrinal analysis presented in this chapter demonstrates that the rights to information and explanation are not merely rights to ensure fair processing of personal data but are intended to facilitate the protection of other individuals' rights, for example, by proving a contractual relationship between data subjects and data controllers and so protect and exercise the rights and duties associated with a such contractual relationship. For this reason, the information and explanations provided due to the exercise of these rights can ensure the protection of even higher rights and freedoms, such as the right not to be subject to discrimination or the right to a fair trial and effective remedy. In other words, the rights to information and explanation are not to be understood as isolated elements limited to the data protection law but rather as having their raison d'être in the larger interest of society to protect individuals from the increasingly use of ADM. This argument can well be deduced from the general information requirements about the logic involved and the envisaged consequences associated with the exceptions in Article 22(2) of the GDPR -i.e., consent and contract necessity-, as it can be extracted from the right of data subjects to request comprehensive information about particular automated decisions to verify the correctness and lawfulness of the data processing and act upon that.

Chapter 4: The Normative Framework of Transparency and Explainability Requirements in the GDPR

4.1. Introduction

Data protection rules intend to 'make it possible to use personal data in a manner acceptable to society. In this way, it should sustain the possibilities of utilising modern information technologies'[330]. The protection offered by the GDPR does not exclude from other sectorial or concrete regulations that also tackles the impacts and challenges created by algorithmic and AI systems, such as the recently approved European Regulation on Artificial Intelligence, the European Digital Markets Act or the European Digital Service Act. However, as explained in Chapter 1, the scope of this thesis is specifically limited to the rights to information and an explanation as referred to in Articles 13, 14, 15 and 22 of the GDPR. When it comes to unravel the significance and intention of having such rights to transparency and explainability for automated decisions in the European data protection law, we cannot forget the own rationale of the GDPR. Even though the problematics mentioned in Chapter 2 can explain the rising concerns towards ADM, they are not sufficient to justify the inclusion of the aforementioned rights within the data protection framework nor why they were deemed appropriate solutions for the discussed challenges. Hence, reasonable questions to pose are: Why would contestation and information rights help data subjects when affected by an automated decision? and Why would individuals have a right to understand the decision-making processes affecting them in their daily lives from the perspective of personal data protection?. Answering those questions can offer, in consequence, more insights on how the framework and spectrum of compliance for the

[330] Peter Blume, 'Data Protection and Privacy – Basic Concepts in a Changing World' Scandinavian Studies In Law 154.

rights to information and an explanation can be concretised in real scenarios – see Chapter 3.4. - *Discussion*.

Chapter 4 provides the normative framework affecting the rights to information and an explanation under the GDPR. To do so, the chapter is divided into two main sections: *Section 4.2. The Aggregated risks of Automated Decision-Making* and *Section 4.3. A Tool To Rectify Power and Information Imbalances.* The former explores the aggregated risks that arise from introducing algorithms in decision-making processes of governance both in the public and private sectors. Section 4.2.1 first addresses the technological and institutional factors through which algorithms can built and reinforce autocratic and non-voluntary structures of power. Meanwhile, Section 4.2.2 dwells on the arbitrariness introduces by algorithms in the decision-making processes of those same structures of power, exploring how arbitrary decisions can prevent individuals from accessing services and products and fulfil their life goals and aspirations. Hence, *Section 4.2* situates the aggregated risks created by algorithms withing the context of personal data protection, identifying the specific risks posed by the increasing unbalance of power between the individual and the data controller.

The latter *Section 4.3* shifts focus to the potential of the rights to information and an explanation to be tools to rectify power ad information imbalances. This section examines the GDPR requirements on transparency and explainability for automated decisions through two approaches: as resemblance to due process safeguards and as risk mitigation and control mechanisms. *Section 4.3.3* explores the traditional resemble of the fair data protection principles to traditional due process safeguards, and analyses the reasons behind such closeness under the umbrella of the aggregated risks proposed in *Section 4.2.*. This section also critically examines the limits to such similarity and whether this approach aligns with principles and fundamental basis of data protection law, particularly on regard to its pragmatism toward personal data processing practices. On the other hand, *Section 4.3.4* examines the differences between a right-based and a risk-based approach to data protection law. This section considers whether data protection law might have followed an hybrid right-risk approach to regulate the risks associated with data processing practices -specially automated individual decision-making-. Hence, *Section 4.3.2.* studies how a partial

risk-based approach to algorithms could have strengthen the regulatory data protection framework by introducing innovative mechanisms of risk mitigation and control in the form of requirements to transparency and explainability.

Together, Chapter 4 aims to provide a comprehensive understanding of the normative framework of the rights to information and an explanation for automated decisions.

4.2. The Aggregated risks of Automated Decision-Making

4.2.1. Algorithms autocracy and non-voluntariness

I have already described in Chapter 2 that public and private institutions alike have turned to automated individual decision-making to aid or substitute human decision-makers with the promise of consistency and effectiveness. Furthermore, algorithms arguably prevent human biases to influence and affect decision-making processes, although their neutrality is highly controversial[331]. The increasing use of algorithmic systems comes with the repercussion that not all the public and private institutions turning to algorithms have the financial and technical capabilities to develop their own algorithms, or even if they do, they decide to resort to the same private providers[332]. Hence, a small number of algorithms with certain modifications and adjustments in respond to the specific purpose or domain where they are deployed, are replacing and assisting human decision-makers. Even if the algorithm is designed for an specific purpose or by the own institution which will finally use it, an unique algorithm will likely supplant or assist multiple human decision-makers, replacing or impacting in the original human set of decision-making criteria.

---

[331] Bibin Xavier, 'Biases within AI: Challenging the Illusion of Neutrality' [2024] AI & SOCIETY 1; Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press` 2018); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press 2018).

[332] Hirevue, an interview technology company, is alleged used by more than seven thousand companies including a third of Fortune 500 companies, which sum-up to over 10 million job interviews. Moder Hire, another interview technology platform saw a forty percent increase in its users in 2019, which entailed a support of over twenty million job candidates assessments and interviews according to Peter Rubinsteain, 'Asynchronous Video Interviews: The Tools You Need to Succeed' *BBC* (Remote Control, 6 November 2020) <https://www.bbc.com/worklife/article/20201102-asynchronous-video-interviews-the-tools-you-need-to-succeed>.

Regardless of the ownership of the algorithm, the turn to algorithms tends towards an overall standardization of the decision-making logic and criteria. When humans were the decision makers, they had to stick to a set of criteria pre-defined by their institutions, although room for manoeuvre was possible within some reasonable limits. Hence, the reach and impact of each individual in the overall scheme of decision-making was limited. The good or harm a human decision-maker could have made was constrained to the number of decision they were handed. On the contrary, the standardization brought by algorithms make the impact and reach of algorithms' criteria an issue at large scale.

Kathleen Creel and Deborah Hellman argued that two reasons intensify the warnings concerning the systematicity of algorithms' influence. On the one hand, 'a limited number of algorithms produced by the same companies are uniformly applied across wide swathes of a single domain'[333]. On the other hand, a single algorithm can hold in its hands the decisions of entire institutions and therefore the allocation of their products and services[334]. Since no decision-maker is perfect -human or algorithmic- some people will be misclassified or unreasonable denied access to the desired product or service[335]. However, the standardised and uniform use of algorithms can lead to a systematic exclusion and misclassification of people, preventing them for accessing a significant number of important opportunities. Alarmingly, even algorithms that are considered fair and accurate on technical standards metrics can fail to correctly and justly classify an individual[336]. Creel and Hellman emphasise how the systematic limitation of people's

[333] Kathleen Creel and Deborah Hellman, 'The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems' (2022) 52 Canadian Journal of Philosophy 26, 2.
[334] See for example the cases of Glovo, Uber, or SCHUFA presented in Chapter 2 *Finance services – credit score*.
[335] For instance, as presented in *Chapter 2 subsection Workplace and algorithmic management systems*, the allocation of Deliveroo riders through Frank's profiling system was originally designed in such a manner that it penalised riders when they were not available or cancelled a given service, irrespective of whether they had a trivial or legitimate reason to do so, such as sickness or participating in a strike action.
[336]One of the most well-known examples of the algorithm's fairness criteria debate involved the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a software developed by and property of Northpointe, Inc., which calculates the likelihood of a subject being re-arrested. In 2016 Propublica singled out the racial disparities of COMPAS in classification errors. Northpointe responded asserting the less significant racial disparities that COMPAS showed in accordance with prediction errors. The debate arose since COMPAS was proved to satisfy predictive parity, not classification parity fairness. The former takes into account the equal classification rate across group, the later the equal prediction error rate across groups. In other words, 23.5% of whites who did not re-offend were misclassified as "high-risk" versus 44.9% of blacks (low levels of classification parity), whereas it was also true that among labelled

opportunities can become a moral concern when unjust algorithms are use at large scale in high-consequence decisions[337].

Either because the transfer of our data provides us with a service or a product that we believe offers us a certain easiness or improvement in our daily lives, or because our voluntariness is rather limited due to unavoidable social structures, our data-driven society demand the collection and processing of individuals' personal data to assure the accomplishment of institutions' goals and individuals' live prospects. The most delicate scenario in this dynamic comprehends those cases where individuals have almost no say in whether they want to be subject to certain type of private institutions, which in turn puts them under the power of hierarchical institutions and within a rule-governed social structure. Alluding to the parties involved in the case-law presented in **Chapter 2**, SCHUFA Holding AG acts as a central place where information about consumers is sourced from banks and saving banks, companies in stationary or online retail, telecommunication companies, and energy suppliers. The reach of the company is so broad that it receive data from around ten thousand cooperating partners in every industry and provide a credit rating score for all German residents, which follows them everywhere in their everyday life[338]. The use of food delivery platforms as Glovoo or Deliveroo can be considered a more trivial service which access and use is strongly based on the easiness and eagerness of the consumer, but their trivially can be question under specific circumstances, i.e. imagine an individual with mobility problems or the not so distant scenario in which the majority of the world's population was confined to their own homes. Irrespectively of whether we decide, for example, to use a food delivery service or not for our personal enjoyment, adhering to the control of an algorithmic management system is not usually a decision that one can make independently of our employee as was the case for Glovoo and Deliveroo riders. Same lack or limit voluntariness can be found in the processing of our personal information by

"high-risk", 41% of whites and 37% of blacks did not re-offend (acceptable levels of predictive parity). Whether the system should be considered racially biased is open to debate, particularly assuming that both fairness criteria cannot be satisfied by an algorithm, forcing a decision on which fairness criteria -and by consequence definition- should be implemented in each particular context. Di Bello (n 97).

[337] Creel and Hellman (n 333) 2.

[338] 'How SCHUFA Works - Our Principle: Reciprocity: SCHUFA and Its Contractual Partners' (*SCHUFA*) <https://www.schufa.de/en/ueber-uns/schufa/schufa-works/> accessed 21 January 2025.

hiring or health insurance services, for which the possibility to reject the processing comes with the high stake of renouncing to a job position or a particular health insurance provider. As exposed by Kate Vredenburg, the problem generated by this limited or complete lack of voluntariness focused on the possible abusive, coercive or manipulative power that private hierarchical institutions, offering and allocating relevant products and services, can develop and enforce over the individuals who only wish to participate in the society and who has not much saying in how our world works[339].

Both with regard to the uniform and systematic use of algorithmic systems over a wide range of sectors and with regard to the imposition of an autocratic algorithmic system to get access to certain product and services, individuals can decide they want to regain some control and power and so 'represent one's interests and values to decision-makers and to further those interests and values within an institutions'[340], in what Vredenburg defined as *informed self-advocacy*[341]. As part of the -arguably unavoidable- algorithmic society, individuals will benefit from knowing the rules of the institutions they are required to be part of and decide whether they want to intentionally adjust their behaviour to comply with those set of rules or risk being completely excluded from them. Vredenburgh argued that the interest of individuals towards informed self-advocacy stands up for both the uniform and systematic use of algorithmic systems as well as for the -arguably- abusive and coercive power of private hierarchical institutions. At the end, in both scenarios an individual would be interested in knowing the rules determining their access to services and products, conforming or not their behaviour to those rules, and contesting the mistaken or unfair decision[342]. In essence, Vredenburg presents informed self-advocacy as a plausible solution to the problems of algorithmic systematisation and uniformity in decision-making identified by Creel and Hellman.

[339] Kate Vredenburgh, 'The Right to Explanation' (2022) 30 The Journal of Political Philosophy 209.
[340] ibid p.5.
[341] Vredenburgh (n 339).
[342] ibid p.5.

Public institutions are expected to be guided by democratic values and its processes lead by the principle of equal opportunity[343]. The same expectation does not generally apply to private institutions, but even so, people will expect that the rules followed by these institutions will not be abusive, coercive, or manipulative. Individuals will also seek for their own interests to be taken into account to a foreseeable limit and that their persona is effectively and accurately represented. The challenges associated with algorithms -inscrutability and lack of neutrality- makes us question these expectations. Barocas, Hardt and Narayanan argued that if we are now driven to accept and participate in the algorithmic society – as conceptualised by Creel & Hellman- we shall understand the system of rules determining the conditions under which we live. The authors assert that we shall aspire for ADM to be predictable, fair, and accountable for their wrongdoings and mistakes, to the same extent we will expect a human decision-maker to be so[344]. To my understanding, the authors' arguments complements Vredenburgh's *informed self-advocacy* insofar that to represent one's interests and values to decision-makers we need predictable, fair and accountable ADM. Furthermore, the interest towards self-advocacy and understandable ADM is exacerbated when the access to basic and primary services and products is at hand or when the effects of the decision making process can significantly affect our rights and freedoms.

At the end of the **Chapter 3** of this thesis -*Access, information and contestability requirements as foundations for the rights to information and an explanation*-, I have argued that the right to information and an explanation to an automated decision as referred to in Article 22 can be presumed from the need to ensure that the decision respects the principles of fairness, legality, and transparency as referred to in Article 5 of the GDPR. In concrete, I stated that individuals are granted with these rights along with the rights to contestation, correction, and erasure to ensure fair and transparent data processing. This argument highly resonates with the interest towards informed self-advocacy. The GDPR explicitly prohibits the use of automated individual decision-

[343] Chapter III establishing equality rights in Charter of Fundamental Rights of the European Union.
[344] Solon Barocas, Moritz Hardt and Arvind Narayanan, 'When Is Automated Decision Making Legitimate?', *Fairness and Machine Learning - Limitations and Opportunities* (MIT Press 2023) <https://fairmlbook.org>.

making 'which produces legal effects [concerning the data subject] or similarly significantly affects him or her'[345], unless it falls under one of the three exceptions: 1) explicit data subject consent, 2) the necessity to enter or carry out a contract, and 3) EU or national law. By doing so, Article 22 of the GDPR stresses the scenarios where ADM -as an automated individual high-consequence decision- deserves informed self-advocacy in the form of the rights to challenge the decision or to be heard, as well as the right to get access to the pertinent information about the systems overall logic and relevant metrics.

4.2.2. Algorithmic arbitrariness

Despite the arguments presented above, we could still argue that ADM does not deserve a special treatment nor specific mechanisms of individual self-advocacy to the one designed for human decision-making[346]. Human brains are also black-boxes and yet, individuals do not generally enjoy rights to information and an explanation for decisions made by a human in the private sector, nor a general right to contest such decision. Still, there is a strong difference between the critical examination that ADM suffered compared to human decision-making, although this is a point of contention[347]. As argued above, the use of ADM can be describe as systematic and homogeneous. However, Barocas, Hardt and Narayanan take a step further and argue that an additional circumstance should be taken into account, i.e., algorithms are replacing humans in tasks traditionally perform under bureaucratic structures, either public or private. The authors allege that the same motives that forward ADM, i.e., humans' subjectivity, arbitrariness, inefficiency and inconsistency, do also explain the rise of bureaucracies and the creation of institutionalised rules and procedures which principal goal was to control and minimise the negative influence of humans in decision-making. According to the authors, this situation is tricky because even though in principle bureaucratic and ADM are both bound and guided by formal rules and procedures, usually several actors are involved in the former -each with different

[345] Article 22 General Data Protection Regulation.
[346] Barocas, Hardt and Narayanan (n 344).
[347] John Zerilli and others, 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?' (2019) 32 Philosophy & Technology 661; Barocas, Hardt and Narayanan (n 344).

responsibilities and roles-, whereas the latter is known by its uniformity and homogeneity[348].

Furthermore, the authors claim that the overly formalistic nature of bureaucracies can also be a disadvantage for the individuals who are push into a constricted and hard to navigate system of rules and procedures, which can feel defenceless, inscrutable and dehumanizing. They further note that the bureaucratic formalistic nature that attempt to reduce human frailties can also stands in the way of accountability and discipline for the decision-makers. Hence, to avoid the adverse circumstances that bureaucracies can create, bureaucracies include; 'mechanisms that ensure that decisions are made transparently, on the basis of the right and relevant information, and with the opportunity for challenge and correction'[349]. Following this reasoning, Barocas, Hardt and Narayanan expound that if human bureaucratic processes are been replaced by algorithmic bureaucratic processes, then, similar -or alike- requirements shall be expected and required for the latter.

Recounting the arguments above, ADM is systematic and homogeneous whereas bureaucracies were designed to protect individuals from arbitrary and inconsistent decision-making. It is not difficult to see the concerns that may arise if ADM goes awry and no proper protections prevent the internalisation of arbitrary and inconsistent decision-making in the bureaucratic process.

According to Creel and Hellman, arbitrariness can have two aspects. On the one hand, arbitrariness can be defined as 'unpredictable, unconstrained or unreasonable decision-making'[350]. On the other hand, arbitrariness can be describes as 'decision-making for no reason at all or as lacking means/ends rationality'[351]. Either way, arbitrariness is problematic because it affects the fairness, lawfulness, and correctness of the decision-making process. Furthermore, arbitrary decisions are problematic

[348] Barocas, Hardt and Narayanan (n 344).
[349] ibid.
[350] Creel and Hellman (n 333) 4.
[351] ibid 8.

because they can unreasonable and unjustifiably prevent individuals from pursuing their life goals and aspirations, negatively limiting their personal autonomy[352].

As I have reiterated, public institutions often have a legal requirement for rational decision-making that will prevent or minimise the risk of arbitrariness. For private institutions such requirement does not exist. It can even be say that private institutions are entitled to make poor decisions, even arbitrary, if deemed so. However, Barocas, Hardt and Narayanan argue that there is a general expectation that a private company will not go against its own interests or that at least it will try to make the decision that more effectively push for its benefit and profit. Whether that goal and the reasoning followed to achieve it align with the interests of individuals is another matter. That said, the authors reason that when the decision affects important matters of our lives and have major consequences, we often expect for a good reasoning[353]. This expectation responds to the own gravity of the decision and the impact it can have to our rights and freedoms. Individuals will not likely accept a treatment that is completely undignified, arbitrary or discriminatory.

Algorithms are rule-based, therefore, by definition there are not supposed to be arbitrary neither for being unreasonable or unpredictable, nor for a lack of goal or ends. Algorithmics are defined as 'a process or set or rules to be followed in calculations or other problem-solving; (in later use spec.) a precisely defined set of mathematical or logical operations for the performance of a particular task'[354]. Thus, algorithms by nature need to fulfil a predefined goal and keep a high level of consistency on reaching that end. Unconstrained, unpredictable and lacking a means is not supposed to be part of an algorithm. That does not preclude them from failure or error – for example due to lapses in following the preconceived established logic or as result of underfitting[355],

---

[352] Solon Barocas, Moritz Hardt and Arvind Narayanan, 'Fairness and Machine Learning. Limitations and Opportunities. When Is Automated Decision Making Legitimate?' (*Fairmlbook*, 13 December 2023) <https://fairmlbook.org/legitimacy.html>.

[353] ibid.

[354] 'Algorithm, n. Meanings, Etymology and More | Oxford English Dictionary' (*Oxford English Dictionary*) <https://www.oed.com/dictionary/algorithm_n> accessed 12 January 2025.

[355] Underfitting is an undesirable behaviour that occurs when the model "performs poorly both in training and new (validating) data. It occurs when the model is too simplistic to capture or learn the underlying patterns in the training data" 'Overfitting vs. Underfitting: What's the Difference?' (*Coursera*, 11 April 2024) <https://www.coursera.org/articles/overfitting-vs-underfitting> accessed 20 January 2025.

overfitting[356] or hidden bias-. Even if any of these undesired scenarios do not occur, the arbitrariness of a (rule-base) algorithm can emerge from its unintelligibility or unrepeatability. As explained by Creel and Hellman 'an algorithmic decision might be rule-governed but complex in a way that make it difficult or impossible for the person affected to understand just what that rule is'[357].

Recalling the black-box problem -see Chapter 2 *section 2.2.4. The black-box problem.*- I argued that algorithms regardless of whether they are considered interpretable or noninterpretable in technical terms -*white-box* or *black-box* respectively- introduce in the processing of personal data a series of inherent characteristics that impact and influence the decision-making processes where they are implemented. Hence, if the own inscrutability of algorithms is not tackle through transparency or explainability measures that facilitate their understanding for the individual affected by them, they could be perceived to be completely arbitrary and undermining the perceived legitimacy of the decision-making process they are part of. Either if the arbitrariness of algorithms results from not meeting the desirable goal or from a lack of understanding of their functioning, the decision-making process could be potentially affected by a perception of arbitrariness, and discriminatory and unlawful treatment.

The Advocate General Richard de la Tour in the Dun & Bradstreet Austria GmbH case argues on that regard that the notion of *meaningful information* as referred to in Article 15 of the GDPR has, among other purpose, the goal to ensure the data subject can check the correct processing of her personal data insofar as the data subject is granted with a right to understand the decision affecting her, and the features used in it as well as the connection between those features and the final decision. The Advocate General does not particularly mention a right to check whether the decision was arbitrary or not, but by arguing in favour of a right to know why and how the characteristic of the individual were relevant for the final decision, it seems to advocate for a right to not be affected by an arbitrary decision, or being it, a right to know why it was it. Furthermore,

[356] Overfitting is an undesirable behaviour that occurs when the model "offers ideal predictions when tested against training data but fails against new, unidentified (validating) data." It occurs when the model is too complex or convoluted or when the training data is not applicable information. ibid..
[357] Creel and Hellman (n 333) 6.

the Advocate General mentions how a decision could be also made on the basis of a lack of information, emphasising that that -lack thereof- information should also be made known to the data subject.

4.3. A Tool To Rectify Power and Information Imbalances

4.3.1. A metaphor for automated decision-making processes - Orwell's' *Nineteen Eighty-Four* or Kafka's *The Trials*

Orwell's *Nineteen Eighty-Four (1949)* presents a totalitarian society in the future ruled by an omnipotent dictator called Big Brother[358]. People of this society -called Oceania- are continuously monitored in their thoughts and actions. Orwell's work settled the foundations for notions and conceptions about data protection where the relationship between individuals' own personal information and the actors collecting and processing it was set on the individual's protection towards *oversurveillance*.

Close to the Big Brother's metaphor, the notion of *dataveillance* was firstly proposed by Roger Clark to describe how database stores of personal information facilitate and intensify surveillance practices, particularly in regard to the state power[359].

However, as I asserted above, the power to collect and process personal data is currently not a concern only posed by traditional state actors. Dan Solove suggests that *dataveillance*, as we experience nowadays, is better represented in Kafka's *The Trial* metaphor, than in Orwell's *Big Brother* metaphor[360]. In Kafka's work, an individual is arrested and forced to attend a series of hearings at a mysterious Court without being explained the reasons for the arrest nor the changes against them[361]. *The Trial* delves on ideas of labyrinthine bureaucracy and legal systems, guilt and innocence, alienation and isolation, and the search for meaning. In Solove words, *The Trials* offers a more appropriate metaphor for current *dataveillance* than Kafka's traditional concern on secrecy or surveillance since *The Trials'* presents a ' thoughtless process of

[358] George Orwell, *Nineteen Eighty-Four (1984)*.
[359] Roger Clarke, 'Information Technology and Dataveillance' (1988) 31 Communications of the ACM 498.
[360] Daniel J Solove, 'Privacy and Power: Computer Databases and Metaphors for Information Privacy' (2000) 53 Stanford Law Review 1393.
[361] Franz Kafka, *The Trial: A New Translation Based on The Restored Text*.

bureaucratic indifference, arbitrary error, and dehumanization, a world where people feel powerless and vulnerable, without meaningful form of participation in the collection and use of their information'[362] . Authors such as De Gucht[363] or Kim Taipale[364] endorse Solove's argument in respect of how Kafka's *The Trial* offers the more suitable metaphor to the sense of loss of control that individuals experience when they are affected by ADM in their everyday. In truth, the aggregated risks associated to automated individual decision-making resonates with the concerns highlighted by Solove in Kafka's *The Trial*. Automated decision making brings forward systematization, uniformity, arbitrariness and an increase in the unbalance of power between data subjects and data controllers.

Although *The Trial*'s meaning is difficult to decipher in its entirety, Kafka's work accentuates the ordinary person's struggle against an unreasoning and unreasonable authority. Without knowing the reasons for being arrested nor the charges for which being judged, the protagonist in Kafka's work is left defenceless to the actions of the authority. Extrapolating this to the aggregated risks pose by ADM, I argue that we can find ourselves in a situation where we do not know the reasons nor motives behind the automated decisions affecting us, quite similarly as in Kafka's *The Trial*. This situation is aggravated due to the inscrutability -and arbitrariness[365]- of ADM since it can even preclude the user to understand the logic and reasons of the process. Hence, the black-box problem bolster the loss of control of individuals as -possibly- not even the institutions that are forcing them into a 'process of bureaucratic indifference, arbitrary error, and dehumanization' are aware of the principles governing the ADM.

In the current context of -arguably- unavoidable ADM that resembles Kafka's *The Trials*, the rights to information and an explanation offer individuals the possibility to understand the decision-making process logic and the reasons behind the particular

---

[362] Solove (n 360) p.1398.
[363] Paul De Hert and Serge Gutwirth, 'Data Protection in the Case Law of Strasbourg and Luxemburg: Constitutionalisation in Action' in Serge Gutwirth and others (eds), *Reinventing Data Protection?* (Springer Netherlands 2009) <http://link.springer.com/10.1007/978-1-4020-9498-9> accessed 28 November 2024.
[364] KA Taipale, 'Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy and the Lessons of King Ludd' (2004) 7 Yale Journal of Law & Technology and International Journal of Communications Law & Policy 123.
[365] As defined by Creel and Hellman (n 333).

decision affecting us. Through the right to contest we are also able to challenge the assumptions about ourselves that lead to that decision and to adjust our behaviour if we seem it appropriate. In the aforementioned normative framework, the rights to information and an explanation help individual to fight against ADM of 'bureaucratic indifference, arbitrary error, and dehumanization' since the rights allow them to understand the reasons for the decision and the information in their favour or against them, as opposed to the scenario presented in *The Trials*.

4.3.2. Perceptions of privacy and data protection

The contrast between Orwell's *Big Brother* and Kafka's *The Trials* resembles the contrast between the transatlantic and continental conceptions of privacy presented by Whitman[366]. Whitman argues that sensibilities about privacy differ in the United States and the countries of Western Europe on the basis that the most intrinsic value to protect through the right to privacy is different primarily 'due to old differences in social and political traditions'[367]. For the former is the value of human liberty, whereas for the latter is the value of human dignity which is ought to be protected. Whitman maintains that there is some resemblance and similarities between the systems, but also many differences. On the one hand, 'continental privacy protections are, at their core, a form of protection of a right to respect and personal dignity'[368]. On the other hand, American privacy 'is much more oriented toward values of liberty, and especially liberty against the state'[369].

To my understanding, these two disparate conceptions of privacy mirror the threats posed to the individual in Orwell's and Kafka's works. Orwell's *Big Brother* features an omniscient power and the concerns around oversurveillance, whereas Kafka's *The Trials* points out the dehumanization and loss of control that come from an unreasonable and inscrutable power. Orwell's and Kafka's works presented scenarios that primarily threat human liberty and dignity respectively, however; as pointed out by Whitman, the contraposition is not absolute. Both conceptions of privacy coexist in the

[366] James Q Whitman, 'The Two Western Cultures of Privacy: Dignity versus Liberty' The Yale Law Journal.
[367] ibid 1160.
[368] ibid p.1161.
[369] ibid.

American and the Continental traditions of privacy, but a higher prevalence on one over the other is possible to decipher[370].

In the continental perception of privacy as respect presented by Whitman, privacy is 'a set of rights over the control of one's image, name, reputation, and over the public disclosure of information about oneself'[371]. By contrast, American perception of privacy is better represented in the words of Aland F. Westin; privacy is the 'claim of individuals, groups, or institutions to determine for themselves when, how, and to what extend information about them is communicated to others' [372]. This understanding of privacy is closer to Orwell's work that it is to Kafka's as it grants individuals the power to decide the way information about themselves is offered to external actors as an opposed scenario of an omniscient surveillance power. However, Westin's notion of privacy also reconceptualises the concept in terms of control over personal information, which gets closer to continental notions of privacy. Thus, this facet of privacy requires a set of requirements and rules to ensure and arrange the effective and reasonable individuals' control over their personal information. Arthur R Miller construction of privacy as informational privacy follows this last perception of privacy as control and stresses the need for requirements and prerequisites when data related to the individual is processed[373]. Informational privacy delimits the practices that must be done and under which conditions when personal information about individuals is processed. *The Trials* critics the loss of control that a person suffers when subject to a process of bureaucratic indifference, arbitrary error, and dehumanization. Informational privacy, in turn, does not prescribe any necessity to refrain from the processing -or in *The Trials'* metaphor the Court hearing- but demands compliance with a set of rules to ensure the individual control over the processing of their data.

There is undoubtedly a jump from privacy to data protection perceptions. However, Miller's notions of informational privacy and fair information practices facilitates that jump at the same time that approaches current notions of data protection. Differences

[370] ibid p.1162-1163.
[371] ibid p.1167.
[372] Alan F Westin, 'Privacy and Freedom' [1967] New York: Atheneum p.7.
[373] Arthur R Miller, *The Assault on Privacy: Computers, Data Banks, and Dossiers* (University of Michigan Press 1971).

theories and conceptions exist around the intrinsic connection between both rights including arguments in favour of the interchangeability of both rights, the dependency of data protection on privacy, or their distinct nature[374]. Due to how extensive this discussion is, I refrain from entering in it beside acknowledging that data protection was originally related to privacy and so was bult-in on the premise of protecting own private life when one's personal information is processed. Thus, continental notions of data protection feeds from the continental perspective of privacy as respect of one's dignity, and all the aforementioned debate can be extended to data protection. That said, the ideas and goals that determine data protection has diverged from the privacy discourse, becoming something itself and developing its own legal legislations and institutions.

According to De Hert & Gutwirth, the concept of data protection is a wildcard word for a series of notions related to the processing of personal data. In their view, data protection serves to reconcile fundamental values that tend to be in conflict when personal data is at stake, such as individuals privacy, the free flow of information, the occasional need of government surveillance, or the allocation of services and goods[375]. In this context, the data protection regime is built upon the assumption that personal information will be process by both public and private powers because this information processing has almost become necessary for our society to properly function. De Hert & Gutwirth argue that data protection is forced to be pragmatic and to accept this reality by offering 'the protection of individual citizens against unjustified collection, storage, use and dissemination of their personal details'[376]. Data protection, therefore,

[374] The discrepancies in the conceptions of the rights have their contraposition in the distinct recognition they gained at the European and international level. The right to privacy is recognised in Article 12 of The Universal declaration of human rights, whereas at the European level, the right is enshrined in Article 8 of The Council of Europe's (CoE) European Convention for Human Rights (ECHR), which protects "the right to respect for private and family life", and Article 7 of the European Union Charter for Fundamental Rights (EUCFR), which recognises the "right to private and family life". In contrast, The Universal Declaration of Human Right does not explicitly envisages a right to personal data protection not does the European Convention for Human Rights. Only the European Charter does, in fact, sanction in its Article 8 that "everyone has the right to the protection of personal data concerning him or her". The right to data protection is also recognised at a (quasi) constitutional level in Article 16 (1) of the Treaty of the Functioning of the European Union (TFUE) and gained considerable importance after the Data Protection Directive (DPD) that introduced data protection principles within EU law and set main benchmarks for the protection of personal data in the EU.

[375] Serge Gutwirth and others (eds), *Reinventing Data Protection?* (Springer Netherlands 2009) p.3 <http://link.springer.com/10.1007/978-1-4020-9498-9> accessed 28 June 2024.

[376] ibid p.4.

does not protect individuals from data processing itself, but for data processing techniques that are unlawful and/or disproportionate[377]. Again, confronting this notion of data protection to Orwell's and Kafka's metaphors, the latter offers a better understanding of the concerns that can arise when individuals are forced to participate in data processing processes without concrete requirements and safeguards.

De la Corte explores data protection by describing it as instrumental, even a procedural right or a 'right to a rule'[378]. As consequence, the author claims that data protection shall provide necessary procedural guarantees or objectives requirements involving both requirements for the lawful, fair, and transparency processing of personal data, as mechanism to allow data subjects to participate in the information process, i.e. rights to access or correction. In this perception of data protection, personal data processing - particularly if it involves decision-making and impacts on individuals' interests, freedoms and rights- resembles *The Trials* events of being arrested and subjected to a hearing insofar as the concerns do not arise from the events perse, but for the possible lack of procedural guarantees when participating in such events.

Blume also elaborates on this notion of privacy arguing that 'data protection is specifically related to the legal rules that regulate to which extend and under which conditions information related to individual physical persons may be used'[379]. Data protection regulation formulates the conditions and requisites under which the processing of personal data is legitimate and defines the techniques and practices which shall be prohibited.

In essence, what I gather form this normative framework is that ADM and the aggregated risks associated with it can quite easily fit in The Trials' metaphor of a 'process of bureaucratic indifference, arbitrary error, and dehumanization'. The solution to that process is not merely a right to protect our private live, but a right to control the conditions under which our personal data is collected, processed and used and hence

[377] ibid p.3.
[378] Lorenzo Dalla Corte, 'A Right to a Rule: On the Substance and Essence of the Fundamental Right to Personal Data Protection' in Dara Hallinan and others (eds), *Data protection and privacy* (Hart Publishing 2020).
[379] Blume (n 330) 153.

a right to ensure that those conditions respect specific legal rules and values. Following this argumentation, I hold that data protection aims to ensure that the unavoidable use of personal data is done in a socially acceptable manner, which in the context of the GDPR, meaning a respect of EU principles and values of liberal and democratic societies. There is undoubtedly a discordance between the applicability of some of those principles to the private actors that process personal data in our everyday and high-consequence decisions, but I expand on this argument in the following section.

4.3.3. Resembles to due process safeguards in data protection to achieve the procedural and pragmatic objective of data protection

The rule of law is traditionally described as the 'restrain of power, in particular arbitrary power'[380]. The term is a contested concept with a history of more than two millennia that exceeds the normative framework of this thesis. However, for an analysis of the historical and political development of the rule of law, I refer to Paul Burgess[381] who offers an illustration of the evolution of the concept over time and of its pertinent distinctive elements under the views of Aristotle[382], Dicey[383], Hayek[384] and Locke[385]. It can be said, nonetheless, that the traditional approaches to the rule of law have in common a series of elements; their focus on the state power and their aim to restrain its potential arbitrariness. The rule of law problem does not traditionally deals or worries about the private relationships between individuals and entities. Although private relationships were conceivable to these traditions of the rule of law, the problems that they sought to address were not reflected or encompassed in non-state based exercises of power. In other words, a non-state-centric approach to the rule of law was beyond

[380] Ioannis Kampourakis, Sanne Taekema and Alessandra Arcuri, 'Reappropriating the Rule of Law: Between Constituting and Limiting Private Power' (2023) 14 Jurisprudence 76.
[381] Paul Burgess, 'Googling the Equivalence of Private Arbitrary Power and State Arbitrary Power: Why the Rule of Law Does Not Relate to Private Relationships' (2021) 17 International Journal of Law in Context 154.
[382] Aristotle, *Aristotle's Art of Rhetoric* (Robert C Barlett ed, University of Chicago Press 2019); Aristotle, *The Politics* (Sinclair T and Saunders TJ (trans), Penguin UK 1981).
[383] AV Dicey, *Introduction to the Study of the Law of the Constitution* (10th edn, Palgrave Macmillan 1885).
[384] FA Hayek, *The Road to Serfdom. Text and Documents*. (Bruce Caldwell ed, 2007).
[385] John Locke, *Two Treatises of Government* (Peter Laslett ed, 3rd edn, Cambridge University Press).

the intended meaning of these theories and would potentially collide with their original aim of limiting the state's interference in transactions between individuals.

From the traditional conceptions, the restrain of power that defines the rule of law entails a framework for which all individuals are subject to the law and must obey it and the state acts in accordance with the law, rather than arbitrarily or capriciously.

Due process or procedural fairness is based on the idea that once the rule of law provides a framework to operate within, procedural fairness set the rules and processes to ensure the fair and just treatment of individuals by the state and its institutions. Particularly, procedural fairness or due process intends to set the constrains and requirements that state powers need to respect when there is a adjudication or deprivation of an individual's liberty or property[386]. Due process also sought to ensure the proper separation of state powers in that the legislator who enact the law differs and act independently from those who enforce it in specific circumstances or are called to judge its possible infringement. Particularly, Crawford conceptualises due process as a 'form of management technique' which sought to create 'schemes and incentives to normatively circumscribe government actions within the bounds of law'[387] and 'discover error, identify causes, and implement corrective actions'[388].

Similarly to the rule of law, traditional conceptions of due process are state-centric insofar as due process is triggered when a significant right is meaningful affected by the exercise of state powers, generally through adjudication or depravation processes. In the word of Citron, 'procedural due process protects the important interest of individuals while constraints on rulemaking served as legitimate substitutes for individual adjudications'[389].

The rule of law and due process find that the main form of problematic power is the power of the state, however; arbitrariness is by no means exclusive to the exercise of

[386] Kate Crawford, 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms' 55 p.111.
[387] Kate Crawford, 'BIG DATA AND DUE PROCESS: TOWARD A FRAMEWORK TO REDRESS PREDICTIVE PRIVACY HARMS' 55 p.121.
[388] ibid referring to Richard H.Fallon, Jr.
[389] Danielle Keats Citron, 'Technological Due Process' 85 p.1251.

power by the states. Kampourakis, Taekema and Arcuri claim that the global economy has facilitated the rise of private powers which can equally impact citizens' interests as they have assumed characteristics of functional sovereignty, meaning that 'private actors exercising power in ways that are comparable to and often indistinguishable from state power'[390]. In this regard, Peter Rott wonder if the power gained by some digital players and their relevance for people's lives and their impact on their freedoms and rights have led these players to become somehow of a state-alike actor that shall observe the rule of law and the procedural due process principles[391].

Issues arise, however, when trying to assimilate private to state powers as the traditional core characteristics of states are not widely applicable to all the private parties involved in the global digital economy. As defined by Weber, 'a state is a human community that (successfully) claims the monopoly of the legitimate use of physical force within a given territory'[392]. Hence, four essential elements characterised the modern state: 1) population, 2) territory, 3) government, and 4) sovereignty. The definition and elements of the state can be loosely extended to the relationships that a concrete number of private (digital) players maintain with their users, e.g., digital players such as Google, Meta or Amazon which coincide to be some of the biggest personal data processors of our times. However, most actors that process individuals' personal data would not easily fall into the category of state, according to Weber's definition, and the big actors, although closer to the definition, do not fit exactly either. For example, we can claim that these massive digital players exercise a monopolistic control and coercive force over its users -population- within its virtual -territorial- sphere. Claims of monopolistic exercise of power have being rise worldwide against these players[393] illustrating how individuals have to -arguably forcefully- accept imposed

[390] Kampourakis, Taekema and Arcuri (n 380) p.77.
[391] Peter Rott, 'Powerful Private Players in the Digital Economy: Between Private Law Freedoms and the Constitutional Principle of Equality' (2020) 18 Baltic Yearbook of International Law Online 32.
[392] Max Weber, *Politics as a Vocation* (1921).
[393] Federal Trade Commission, 'FTC Sues Amazon for Illegally Maintaining Monopoly Power' (*Federal Trade Commission*, 26 September 2023) <https://www.ftc.gov/news-events/news/press-releases/2023/09/ftc-sues-amazon-illegally-maintaining-monopoly-power>; European Commission, 'Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service - Factsheet' (*European Commission - Press Corner*, 27 June 2017) <https://ec.europa.eu/commission/presscorner/detail/es/memo_17_1785>; McCabe, '"Google Is a Monopolist", Judge Rules in Landmark Antritrust Case' (*The New York Times*, 5

terms and conditions, what in turn can be considered a form of coercion in the access and use of the services provided by these players. Still, these big digitals players lack sovereignty over their user and the monopoly to exercise physical force. Hence, their assimilation with traditional notions of state powers requires more than a theoretical leap. If instead of these big digital player, we consider other actors using ADM in everyday and high-consequence decisions -such as credit scoring agencies, banks, food delivery platforms, or transportation companies-, the leap seems an insurmountable obstacle.

Although algorithmic non-voluntariness and arbitrariness are close to the problems sought to be solved by the rule of law and due process, the aggregated risks of ADM -as well as most of the data-driven technologies involved in the digital global economy- could not be easily solved through ideas that did not had them into account nor could have possibly imagine them. In this regard, Rott calls for a reconceptualization of the rule of law and procedural due process ideas to encompass the new problems related to private relationships in the data-driven economy. At the end, the author claims that the law is dynamic and there might be reasons to revisit traditional legal concepts[394], particularly if power has shift from the state to private players and the problem of arbitrariness now encompasses relations carry out by private players but which looks very alike to -traditional- state adjudications and deprivation actions.

In this last detail of Rott's argument is where I find one of the main elements of this normative framework, and which takes me back to the problems posed by ADM presented in **Chapter 2**. ADM requires specific requirements of information and explanation because they are used in high-consequence decisions that significantly impact individuals' freedoms, rights, and interests. Who use this systems is important, but more relevant are the aggregated risks and problems that algorithmic automation

---

August 2024) <https://www.nytimes.com/2024/08/05/technology/google-antitrust-ruling.html>; European Commission, 'Commission Fines Meta €797.72 Million over Abusive Practices Benefitting Facebook Marketplace' (*European Commission - Press Release*, 14 November 2024) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_5801>; Jody Godoy, 'Meta Will Face Antitrust Trial over Instagram, WhatsApp Acquisitions' (*Reuters*, 13 November 2024) <https://www.reuters.com/legal/meta-will-face-antitrust-trial-over-instagram-whatsapp-acquisitions-2024-11-13/>.

[394] Rott (n 391).

introduces in the decision-making process. Since those problems and risks resemble those posed by power states when allocating and adjudicating goods and services, it is not ludicrous to look for an answer in the rule of law and due process.

The coexistence of state and private powers is not new, nor are the power asymmetries created with respect to individuals. However, algorithms have widened and intensified power asymmetries to the point where authors as Aneesh or Richards & King drew attention to a new stage of 'governance by algorithms or algocrazy'[395] and the 'power paradox of Big Data and AI'[396] and denounced how the datafication of our society drives institutions and companies using data processing techniques to ultimately benefit at the cost of the individuals whose data is mined, analysed, sort out and used. O'neil goes as far as denoting algorithms as 'weapons of math destruction'[397]. In turn, Ivanova claim that this 'algorithmic society' 'enables novel forms of exclusion, subordination and discrimination that should find appropriate legal response and remedies'[398].

However, these new notions of algorithmic power -or algocracies- are not by far limited to the relationships between big digital players and their users. In this case, the concerns does not exactly arise from the actors in power, but from the means and instruments used by those actors and the threat they can pose to the interests, freedoms, and right of individuals. The swift is slime, but it allows from a change in the normative framework from traditional conceptions of rule of law and due process, to resemblances of it in data protection. For instance, the main aim of data protection could resemble one of the functions of traditional procedural due process, particularly in regard to administrative law. The rationale behind data protection is the knowledge that external powers can easily infringe fundamental rights through the processing of one's data, therefore, personal data processing shall be controlled and structured under specific rules. Data protection makes the core assumption that data will be

[395] A Aneesh, 'Technologically Coded Authority: The Post-Industrial Decline in Bureaucratic Hierarchies', *7th International Summer Academy on Technology Studies, Deutschlandsberg, Austria* (2002) <https://web.stanford.edu/class/sts175/NewFiles/Algocratic%20Governance.pdf>.
[396] Neil M Richards and Jonathan H King, 'Three Paradoxes of Big Data' (2013) 66 Stanford Law Review Online 41, p.45.
[397] Cathy O'neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2017).
[398] Yordanka Ivanova, 'The Role of the EU Fundamental Right to Data Protection in an Algorithmic and Big Data World' (2021) 13 Data Protection and Privacy, Volume 13: Data Protection and Artificial Intelligence 145, p.19.

processed, acknowledging that power imbalances will -unluckily- result from these practices. When defining the rules and requirements to proceed with such processing, data protection is not only transferring control to data subjects but it is also demanding transparency in regard to such processing. To offer data subjects some control on what data is processed as well as on the why, how, and when, data protection demands some level of openness and participation in the processing process. This idea also highly resonates with the possible responses to the loss of control and dehumanization posed in Kafka's *The Trials* as well as with the notion of informed self-advocacy.

The distinction between public and private sector is not one that data protection seems to be concerned with. Again, the stream of personal data primarily flows from the weak actor -the data subject- to the strong -the data controller-. The case-law presented in **Chapter 2** exemplifies how individuals do not only need to provide personal information to public authorities to achieve their goals in life, but also to private data controllers. Likewise, the systematicity and homogeneity brough by ADM as well as the potential arbitrariness and perceived lack of legitimacy of the processes has create a stronger sense of unbalance between the parties involved in the personal data processing. Thus, it is possible to say that the informational power furthered and sustained by algorithms does not distinguish between public or private. Although far from the traditional conception of state power, the use of data processing techniques has provide public and private actors with some or other type of coercive power over individuals when they attempt to access or use the product and services provided by the data processors.

That can explain why legal tools under the form of data protection rules and requirements resemble traditional procedural due process rights. If the intention is to address power imbalances, arbitrariness, and legitimacy, it may be a good idea to resort to the mechanisms which have long addressed such issues and whose foundations are deeply rooted in western legal tradition. In that regard, we can argued that current and previous data protection frameworks came with the concrete intention to provide various specific procedural safeguards to protect individuals' fundamental rights and promote accountability by the actors involved in the data processing. In essence, "the

citizen gets procedural guarantees as a compensation for the lack of testing of the reasonableness of the intrusion"[399].

The Fair Information Practice Principles

Perhaps the most straightforward example of resemblance to procedural due process mechanisms developed to address ADM, including profiling, can be found in one of the first attempts to regulate the processing of personal data; the Fair Information Practice Principles. In the early days of 1970, concerns about the technological development and the increasing use of automated processing of personal data lead to the adoption of the Fair Credit Reporting Act by the United States Congress. Contemporary to the Act, an influential report was published by the United States' Department of Health, Education, and Welfare (HEW) analysing the consequences of using computers to keep records about people. The Secretary's Advisory Committee on Automated Personal Data Systems assessed the impact of computer records keeping on private and public matters and recommended safeguards against the potentially adverse effects[400]. The Committee's Report reviewed historical development of records and records keeping to focused on the increasing challenges created by the application of computers to these practices. Hence, the Committee's Report brought attention to the enlarged data processing capabilities offered by computers as well as the easier access to data they offered and the new class of remote records keepers they create. All in all, the Committee's Report explored some of the consequences of these changes and assesses their potential for adverse effect on individuals, organizations, and the society as a whole[401]. The Committee concluded by highlighting that

> The net effect of computerization is that it is becoming much easier for record-keeping systems to affect people than for people to affect record-keeping systems. Even in non-governmental settings, an individual's control over the use

[399] Serge Gutwirth, 'Privacy, Data Protection and Law Enforcement. Opacity of the Individual and Transparency of Power' (2021) 18 p.33.
[400] U.S. Department of Health, Education & Welfare., 'Record Computers and the Rights of Citizens' (Secretary's Advisory Committee on Automated Personal Data Systems 1973) (OS)73-94 p.10.
[401] ibid p.xxi.

> that is made of personal data he gives to an organization, or that an organization obtains about him, is lessening.

Furthermore, the Committee highlighted that, although there is nothing inherently unfair in permitting the use of personal data to access the services a -public or private- organisation provides, "both parties to the exchange should participate in setting the terms"[402]. However, the Committee considered that under the law at that time, "a person's privacy was poorly protected against arbitrary or abusive record-keeping practices"[403]. Although referred as privacy, the arguments' concerns, and solutions offered in the Committee's Report highly resemble the same discourse that enhanced the development of data protection regimes. Hence, elements such as the inevitability of the data processing to participate in society, the arbitrariness or abusive data controllers processing practices, the tendency to secrecy and opaqueness of those same practices, and the limited individual control over them are not foreign elements in the data protection discourse.

The Committee's Report put forward the necessity to reconsider the meaning of personal privacy in relation to records and record-keeping practices based on the argument that the concept traditionally equated with secrecy or seclusion, a quality that is not inherent in most record-keeping systems as they tend to be public and available to anyone to see and use. The expectations derived from personal privacy collides with the same essence of records and becomes fruitless if the individuals who voluntary provide their personal information expect it to be kept for limited purposes or intentions. Notwithstanding the conflict between public records and personal privacy, it is still reasonable that individuals who allow the process of own's data for specific purposes would keep some privacy expectations. Hence, the Committee's Report highlighted that that conventional, and ill-fitting, conception of personal privacy needed to be constrained at least in regard to record-keeping practices and reformulated in a

[402] ibid p.xx.
[403] ibid p.xxi.

manner that it acknowledges some disclosure of data but affords affected individuals at least some agency in deciding the nature and extend of such disclosure[404].

To deal with all these issues, the Committee stressed the necessity to enact a Federal Code of Fair Information Practices (FIPPs) for all automated personal data systems, which rest in five basic principles that would be given legal effect as safeguards requirements for these type of practices. The principles reads as follows;

- There must be no personal data record-keeping systems whose very existence is secret.
- There must be a way for an individual to find out what information about him is in a record and how it is used.
- There must be a way for an individual to prevent information about him that was obtained for one purpose from being used or made available for other purposes without his consent.
- There must be a way for an individual to correct or amend a record of identifiable information about him.
- Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take precautions to prevent misuse of the data.

As could be perceived in the aforementioned principles, the Committee's Report embraced a highly procedural approach to records processing with mechanism resembling due process principles at its core. Furthermore, the Committee's Report undertook a broad scope of application considering public and private actors alike when they engage in the processing of individuals' information through automated systems. The recommended FIPPs mentioned concerns regarding the automated processing of personal data, particularly regarding its accuracy, fairness, individualised flexibility, dignity, and dehumanization. Particularly, the Committee's Report observed that automated data processing could sacrifice flexibility and accuracy in the name of efficiency while constraining the freedom of data subjects to provide explanatory

[404] ibid p.40.

details in responding to questions, contributing to the so-called dehumanizing image of computerization[405]. Likewise the Committee's Report identified the problem of *statistical stereotyping* concerning the prediction of individuals' future behaviour based on their association with statistical defined groups[406]. Looking at the concerns identified by the Committee, it is not difficult to find strong resemblances with the challenges presented in **Chapter 2** and this **Chapter 4** regarding the use of ADM in high-consequence decisions.

To these alike challenges, the Committee's Report FIPPs play a double function; they restrict the process by which data is process -to ensure its fairness- and yet do so by requiring the processing to conform with specific substantive values such as accuracy, lawfulness, and security. Hence, the FIPPs assumed a procedural approach insofar as the processing of data needed to be done under specific conditions and through concrete procedures, but those procedural requirements needed to be judged under the lenses of substantive vales such as lawfulness and legitimacy. Therefore, the FIPPs did not merely require compliance with a checklist of procedural requirements, but the adherence and respect to some substantive values through the application of that same procedural requirements[407]. The key element of the Committee's Report revolve around the redistribution of power between the organisations that process the records and the individuals affected by them, shaping the traditional understanding of personal privacy. To do so, the Committee's Report recommended, apart from the FIPPS, a number of safeguards conceived to offer the individual procedural protections against the decisions affecting them, and which have traditionally received little input from their part. The Committee's Report noted that these safeguards did not

> Provide the basis for determining a priori which data should or may be recorded and used, or why, and when. [They do], however, provide a basis for establishing procedures that assure the individual a right to participate in a meaningful way in

[405] ibid p.14.
[406] ibid p.26.
[407] Karen Yeung and Lee A. Bygrave, 'Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship' 16 Regulation & Governance 137.

> decisions about what goes into records about him and how that in formation shall be used[408].

Resembling due process or procedural safeguards[409], the Committee's Report proposed mechanism alike to the right to notice and to be heard established in our GDPR. Relevant to our thesis's right to information and an explanation, the Committee's Report also introduced as a suitable safeguard the individual's right to contest information, which would granted individuals with the power of disputing the 'accuracy and completeness of information maintained about him [or her] by a consumer-reporting agency'[410]. The right comes with the agency duty to 'reinvestigate and record the current status of that information, or delete the information if it is found inaccurate or cannot be verified'[411]. The right to contest information was repetitively mentioned as a suitable safeguard to protect individuals, possible as a way to counterbalance the power dynamics between organisation and individuals. The Committee's Report went as far as to affirm that ' theoretically, the adverse consequences of 'statistical stereotyping' can be avoided by permitting an individual to know that he has been labelled a risk and to contest the label as applied to him'[412]. The Report, nonetheless, highlighted the difficulties that a lone individual could face when attempting to contest a statistical stereotype.

The FIPPs and safeguards proposed in the Committee's Report could be seen, nonetheless, a strange regulatory hybrid as they aimed to restrict the processes by which personal data is processed, yet did so through substantive values such as requirement of accuracy, lawfulness, security, or fairness. This hybrid, nonetheless, is also followed in the GDPR as can be seen in Articles 5 and 6 in regard to the lawful grounds and principles for data processing. Looking at the previous section, the Committee's Report could be perceived as the pioneer attempt to develop a data protection legal regime in so far as the Report assumed a procedural approach to data processing and compelled for the adherence to a set of routine procedural checklist,

---

[408] U.S. Department of Health, Education & Welfare. (n 400) p.40.
[409] Citron (n 389); Kaminski and Urban (n 275).
[410] U.S. Department of Health, Education & Welfare. (n 400) p.xxvi.
[411] ibid p.70.
[412] ibid p.26.

while assuming that that pragmatic approach might not be suffice. Thus, the Committee's Report also incorporated explicit reference to norms that require judgements about compliance with substantive values, such as lawfulness and legitimacy[413].

A central point of the current European data protection regime dwells around the principles laid down in Article 5 of the GDPR, which require the processing of personal data to occur under the specific conditions of lawfulness, fairness and transparency, purpose limitation, data minimization and integrity and confidentiality. The principles lay the foundations for the data protection rules and requirements established in the GDPR, and, interestingly, resemble quite closely the FIPPs. Although the GDPR included accountability as an additional principle to those already covered in the Committee's Report, the close resemblance between the principles contained in both documents evinces the relevance and resilience the Report has maintained over time. By incorporating accountability to the original FIPPs, lawmakers could have, perhaps, attempted to reinforce the Report's principles and adapted them to the new needs and technological developments[414]. Particularly, Articles 5 and 6 of the GDPR preclude that the processing of personal data shall occur lawfully, fairly, and in a transparent manner, for specific legitimate purposes and not for subsequent incompatible purposes, that the data shall be adequate, relevant, and limited to what is necessary in relation to the purpose for which they are processed, and that the data shall be subject to appropriate security measures. It could be say, then, that the GDPR reinforced and updated the FIPPs to the current -and future- necessities of data subjects, yet failing again to offer clear reassurance on how precisely they shall be understood and implemented.

4.3.4. A combination of a right and risk approach to data processing practices

Taking as an example ***The Algorithmic Controversies in Courts*** of **Chapter 2** and **The Aggregated risks of Automated Decision-Making** presented in this chapter, it can be seen that ADM can easily and quickly turn into datafication, profiling and adversarial

[413] Yeung and A. Bygrave (n 407) p.140.
[414] Yeung and A. Bygrave (n 407).

practices. Although the first to be impacted by harmful data processing practices are particular individuals, the aggregation of individual cases and the generalization of these type of practices despite their adverse impacts can often cause more severe social, economic, cultural, and political consequences[415]. Harmful data processing practices can have the power of shaping our reality in an undesirable and unintended way. Therefore, data protection cannot only assume a pragmatic approach to data processing practices, but also an instrumental role in the protection and achievement of other fundamental rights.

In the normative framework presented above, data protection is perceived as individual control over the practices that can become 'processes of bureaucratic indifference, arbitrary error, and dehumanization', but such control might not only be transposed in a regime of permissive check and balances. Data protection as control also possesses a prohibitive nature towards those data processing practices that are not risky but threatening to the collective values, rights and freedoms of individuals and society. In this sense, data protection assumes an ex-ante prohibitive instance for certain data processing practices that can creates a severe and adverse harm. Following *The Trials'* metaphor, data protection as control does not only grant data subject with rules, rights and requirements to navigate through the charges and trials they have to endure, but data protection would also prohibit certain acts and practices throughout the process as it deems them too threatening to the individuals' rights, freedoms, and interests.

Risk regulations emerged to control and mitigate 'the risks that emerged from new technologies and industries and to address related market failures such as information asymmetry or unwanted side-effects of the progressive advancements'[416], in context such as health and safety or environment development and sustainability. Macenaite defines risk-based regulations as those which 'involve the development of decision-making frameworks and procedures to prioritise regulatory activities and the deployment of resources, principally inspection and enforcement activities, organised around the assessment of the risks that regulated firms pose to the regulators

[415] Creel and Hellman (n 333).
[416] Milda Macenaite, 'The "Riskification" of European Data Protection Law through a Two-Fold Shift' (2017) 8 European Journal of Risk Regulation 506, p.509.

objectives'[417]. In essence, risk-based regulations are designed as a regulatory strategy by governments and regulatory agencies to deal with societal and institutional risks. As clarified by Spina, it is a type of regulation traditionally associated with the 'uncertain negative outcomes connected with the products of industrialized manufacturing'[418]. As could be appreciated in the case-law presented in Chapter 2, it is becoming self-apparent in the increasing and often uncontrollable processing of personal data that there is a relationship between data processing and risk. Hence, there is a reasonable argument -followed among others by Spina- proposing risk-regulation as a solution to the challenges emerged from the use of algorithmic and AI techniques[419]. For instance, Spina asserts that a risk-regulation approach can be the logical contribution both to the governance framework developed around the data-driven digital economy and the own advancement in data protection law[420].

For Ivanova, the impact that data processing can have to individuals' rights and freedoms is located at the centre of the GDPR regime, thereby the GDPR assumes a risk-based approach where safeguards are imposed to certain high risky processing activities with the main goal of minimizing those risks and the power of information imbalances associated with them[421]. According to this approach, there is (some level of) dominance in the relationship between the data controller and the data subjects that is determined by two premises: 1) the data subject is in a situation of vulnerability with respect to the data controller due to the high risk data processing scenarios in which they need to get involved to engaged in their daily lives, and 2) those -risky- data processing scenarios have the potential to strongly and negatively impact and affect the rights and freedoms of the individuals.

From Ivanova's risk-based approach to data protection, threats or violations to individuals' data protection rights shall be evaluated taken into account these premises. In this regard, Ivanova argues that data protection as

[417] ibid p.510.
[418] Alessandro Spina, 'A Regulatory *Mariage de Figaro* : Risk Regulation, Data Protection, and Data Ethics' (2017) 8 European Journal of Risk Regulation 88, p.88.
[419] ibid.
[420] ibid p.89.
[421] Ivanova (n 398) p.20.

> Control should not be perceived, therefore, literally only in terms of data subject's individuals choices and available individual remedies, but it should be also conceived as a system of precautionary safeguards over the processing and an oversight architecture of control at different levels.[422]

Looking at the GDPR we distinguish between the individuals remedies transposed into, for instance, the right to access, to rectification, or to object, and the precautionary safeguards of a general prohibition to be subject to automated processing or the duty to carry out a data protection impact assessment[423]. The two facets of data protection as control are, therefore, complementary rather than exclusive.

The GDPR has a distinctive element as it mainly relies on individual actors rather than governments or regulatory agencies. Data controllers are, to a large extend, entrusted and delegated to assess, manage and communicate the risks associated with data processing techniques[424]. The GDPR brought a regulative transformation insofar as it present a model where the data controller not only shall comply with the law, but shall demonstrate such compliance through the principle of accountability[425]. In this regulatory framework, data controllers -irrespectively of their public or private nature- are compelled to assess and control the risk that their data processing techniques can pose to the rights and freedoms of data subjects in an structured and formal manner[426].

Such 'riskification' of the Union data protection regulation is, according to Ivanova, limited. Regulators have assessed concrete data processing practices and deemed them risky enough to require specific requirements and safeguards, such as in regard to automated decision making, including profiling[427] -see *Chapter 3.1. The right to not be*

---

[422] ibid p.21.
[423] Articles 15, 16, 21, 22, and 35 of the GDPR, respectively.
[424] See, for instance, Article 35 of the GDPR which requires data controller to "carry out an assessment of the impact of the envisaged processing operations on the protection of personal data", when those data processing operations " is likely to result in a high risk to the rights and freedoms of natural persons".
[425] See Article 5 paragraph 2 of the GDPR which establishes that "the controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 [of the aforementioned Article] ("accountability"). Such accountability refers to the principles relating to processing of personal data, e.g. lawfulness, fairness, transparency, data minimisation, or legitimate purpose. Hence, the data controller is compelled to demonstrate its own compliance to those principles and how those principles are respected by its data processing practices alike.
[426] Spina (n 418) p.89.
[427] Ivanova (n 398) 165.

subject to automated decision-making-. For Spina, this awareness towards the risks and challenges associated with personal data processing practices has made the GDPR to adopt an hybrid right-risk based approach that considers both the technical aspects of the techniques and methodologies used in personal data processing as well as the ethical aspect of the real impacts of those practices in the freedoms and right of individuals[428].

Looking at Recital 75 of the GDPR, it is easy to perceive that the concrete risks potentially resulting from personal data processing are extend and diverse:

> The risks to individuals' right and freedoms] may result from personal data processing which could lead to physical, material or non-material damage, in particular: where the processing may give rise to discrimination, identity theft or fraud, financial loss, damage to the reputation, loss of confidentiality of personal data protected by professional secrecy, unauthorised reversal of pseudonymisation, or any other significant economic or social disadvantage; where data subjects might be deprived of their rights and freedoms or prevented from exercising control over their personal data; where personal data are processed which reveal racial or ethnic origin, political opinions, religion or philosophical beliefs, trade union membership, and the processing of genetic data, data concerning health or data concerning sex life or criminal convictions and offences or related security measures; where personal aspects are evaluated, in particular analysing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, in order to create or use personal profiles; where personal data of vulnerable natural persons, in particular of children, are processed; or where processing involves a large amount of personal data and affects a large number of data subjects[429].

Despite the argument towards the riskification of the GDPR, it is important to reiterate that the GDPR respect a right-based approach insofar as, in words of Gellert, 'it applies

[428] Spina (n 418) p.89.
[429] General Data Protection Regulation recital 75.

irrespective of the level of risk [of the processing practice], and therefore provides an even level of protection to all [processing practices]'[430]. Article 29 working Party insights that 'the EU data protection legal framework provides for a minimum and non-negotiable level of protection for all individuals'[431]. This minimum and non-negotiable protection is reinforced by the EU Charter of Fundamental Rights which enhances the right to personal data protection to the level of fundamental right and therefore grants it the highest level of legal protection. The right-based approach to data protection is completely contrary to its risk-based approach, which is based on the predictable level of risk put at stake by the processing practice upon calculations in terms of harms and benefits[432]. For Gellert, the GDPR can be defined as hybrid, as far as its right-based approach provides an even -and primary- level of protection to the data subject that applies to every single processing operation irrespective of whether harm has been created and the risk-based approach is applicable as a complement to certain types of data processing[433].

The hybrid right-risk regulatory mode of the GDPR pose two consequences. On the one hand, any data processing that would, if implemented, clearly violate fundamental rights would, likewise, violate the data protection's principles of lawfulness, fairness and transparency. On the other hand, any data processing that would potentially but uncertainly, violate fundamental rights when implemented, would require a risk-based approach insofar as the data processing operation would need to be assess in its severity and probability to be a threated risk to fundamental rights. For Gellert, when assessed as high, these data processing operations would demand 'greater level of scrutiny, more demanding safeguards and appropriately greater caution' before such processing could be allowed. Yeung and Bygrave specify that the level of the safeguards stringency will vary on the level of severity and probability of the threatened risk[434].

[430] R Gellert, 'We Have Always Managed Risks in Data Protection Law': (2016) 2 European Data Protection Law Review 481, p.483.
[431] Opinion 1/98 Platform for Privacy Preferences (P3P) and the Open Profiling Standard (OPS) 1998 p.2.
[432] Gellert (n 430) p.483.
[433] ibid.
[434] Yeung and A. Bygrave (n 407) p.146.

The right-based approach to data protection entails that processing is either legal or illegal. The risk-based approach follows a granular, scalable logic that determines the level of risk a processing pose to the individual. It is not a matter of whether the processing is risky and therefore should be illegal or not, but on whether the level of risk could be handle through specific safeguards. Both approaches follow different logics and modus operandi.

The benefits and possibilities created by data processing technologies are undeniable, yet, data processing technologies have also the potential to erode the foundations of our society, therein lies the link between the data protection framework and fundamental rights. However, the threat of data processing practices are not only at the collective -human rights- level, as data processing technologies undoubtedly affect society in a highly individualised level.

The idea of acknowledging and accepting that data processing techniques may have a negative impact or effect on the rights and freedoms of individuals is not easy to accommodate within the fundamental rights discourse. Hence, the cost-benefit calculation of a risk-based approach is conflicted with the protection of fundamental rights and it is unclear how the benefits of the processing techniques should be weight against the risks to individuals[435]. However, data protection regulation was designed in such a manner that the protection of fundamental rights is envisioned through securing the compliance with specific principles and provisions, rather than through controlling data processing-related risks[436]. The aim of data protection is greater than protecting individuals from harm through risk control and mitigation, as it is forth and most a fundamental right that shall be safeguarded regardless of harms or possible adverse effects on individuals that aligned with risk regulation. In other words, data protection is pragmatic insofar as it assumes that the processing of personal data is necessary for the participation of individuals in the current society, and so, it present a legal framework in which individuals' fundamental right to data protection is safeguarded when specific principles and safeguards are respected and putting into place. The pilar

[435] Macenaite (n 416) p.521.
[436] ibid p.533.

of data protection is based on a right based approach to data processing practices, they are either legal or illegal depending.

The core principles of the GDPR referred to in Articles 5 and 6 sustain the right-based approach of the European data protection regime. Yet, the risk-based approach is included in a series of limited situations referring to compliance obligations such as accountability duties or data protection impact assessments (DPIAs). In other words, the core principles of data processing are still right-based and offer an even protection to data subjects, still a risk-based approach is compelled to specific processing techniques such as ADM or those which use sensitive information. The even protection of the GDPR is granted in a general level, but for those techniques whose risks to society and individuals are particularly and especially acute and yet uncertain -see discussion on Chapter 2 about inscrutability and lack of objectivity of algorithms as well as the aggregated risks of algorithmic systemisation and arbitrariness in 4.2-, the GDPR offers a second layer of protection. For example, ADM is generally prohibited as it is considered to violate fundamental rights and then the core foundations of our society. However, risk-assessments are forced to be taken if specific conditions under paragraph 2 and 3 of Article 22 applied and the processing is allowed. Permitting ADM followed the double logic of right and risk approaches, the second demanding specific safeguards to be put in place to prevent the potential uncertain risks linked to automated data processing.

In essence, by addressing that the GDPR follows an hybrid right-risk based approach, I am agreeing with Macenaite's argument that

> Risk has become a new boundary in the data protection field and a key indicator in deciding whether additional legal and procedural safeguards are required in a particular context in order to shield data subjects from potential negative impacts stemming from specific data processing activities.[437]

The potential risk of data processing techniques need to be assessed, however; that assessment do not determine the final safeguard put in place. Data protection

[437] ibid p.507.

safeguards apply regardless of the level of harms or benefits created by the processing techniques[438].

This claim is not contradictory with the assertion that the rights of data subjects towards ADM resembles procedural rights, but complementary. ADM come with some risks that the regulator deemed risky enough to deserve specific legal safeguards. Those legal safeguards have took the form of mechanism that are traditionally used to control the unbalance of power and the possible arbitrariness of the state, i.e. procedural rights. For automated processing the data subject shall have the right to, at least, be heard, to have a human intervention, and to contest the automated decision.

4.4. Discussion

I presented two questions at the beginning of this chapter 1) Why would contestation and information rights help data subjects when affected by an automated decision? and 2) Why would individuals have a right to understand the decision-making processes affecting them in their daily lives from the perspective of personal data protection? There are not two straightforward and irrefutable answers to these questions. However, in this chapter I provided a normative framework that can help to understand the rationale behind the rights to information and an explanation, and the rights to contest, and by extension, understand an automated decision.

Automated decision making introduces in the decision making process of private and public data processors alike a set of aggregated risks that threat individuals' rights, freedoms and legitimate interests. On the one hand, ADM can potentially built and reinforce autocratic and non-voluntary structures of power that can turn abusive, coercive or manipulative towards the individual that needs them to participate in society and achieve her life goals' and aspirations. On the other hand, ADM can introduce arbitrariness in the decision-making process, which unreasonably and unjustifiably exclude and limit individuals from accessing a wide range of opportunities. The arbitrariness can be either cased from a malfunctioning – or undesirable

functioning- of the ADM model or by its own complexity, which makes the reasons and motives behind a decision non-understandable to the individuals affecting by them.

These aggregated risks situate ADM very close to Kafka's metaphor *The Trials* where individuals face a 'thoughtless process of bureaucratic indifference, arbitrary error, and dehumanization, a world where people feel powerless and vulnerable, without meaningful form of participation in the collection and use of their information'[439]. In this context, the right to information and an explanation help individuals to fight against the potentially dehumanizing, arbitrary and bureaucratic nature of ADM by allowing them to understand the reasons for the decision and the information in their favour or against them. These two rights, together with the right to contest a decision, offer data subjects a possibility to exercise informed self-advocacy, i.e., to know the rules determining their access to services and products, to conform or not their behaviour to those rules, and to contest the mistaken or unfair decision.

ADM requires specific requirements of information and explanation because they are used in high-consequence decisions that significantly impact individuals' freedoms, rights, and interests. Who use this systems is important, but more relevant are the aggregated risks and problems that algorithmic automation introduces in the decision-making process. Since those problems and risks resemble those posed by power states when allocating and adjudicating goods and services, it is not ludicrous to look for an answer in the rule of law and due process.

When considering why these rights were granted to individuals in the data protection context, it is necessary to understand that data protection does not protect individuals from data processing itself, but for data processing techniques that are unlawful and/or disproportionate. In the end, the right to human dignity is considered the ultimate philosophical and legal foundation of data protection rights. Therefore, when considering the aggregated risk of ADM and the concept of informed self-advocacy, it is not unreasonable to borrow the words of O'Neil's and wonder "How do you justify evaluating people by a measure for which you are unable to provide an explanation?"[440].

[439] De Hert and Gutwirth (n 363) 6 agreeing with ; Solove (n 360).
[440] O'neil (n 397) p.8.

O'Neil's question makes us wonder how an automatic decision neither explained not informed can fail to jeopardize people's very dignity. Without information and explanation duties put in place, as in the rights to information and an explanation of the GDPR, it would be if not impossible, rather difficult, to verify that the ADM is not acting in an unfair, unlawful or arbitrary manner and, hence, directly threatening the fundamental right to dignity of individuals. Article 1 of the Charter of Fundamental Rights of the European Union[441] asserts that a person must never be treated merely as a means to an end, but always at an end themselves. The metaphor of *The Trials* illustrates how easily and hastily an opaque and arbitrary decision-making process can hinder the fundamental dignity of individuals.

The normative framework presented in this chapter frame the rights to information and an explanation around procedural fairness, power asymmetry and contestability. Those three concepts could not be understood without the ultimate goal of protecting human dignity. To be treated with dignity, it means to be treated as an autonomous being capable of making rational choices about one's own life, that idea is directly linked to the concept of informed self-advocacy. Likewise, human dignity requires a person to have a certain degree of control over their own life, image, and identity, information and explanation duties understood from the lenses of procedural rights, are the tools that restore and protect individuals' control over their personal information, hence dignity.

I conclude in this chapter that ADM and the aggregated risks associated with it can quite easily fit in The Trials' metaphor of a 'process of bureaucratic indifference, arbitrary error, and dehumanization'. The solution to that process is not merely a right to protect our private live, but a right to control the conditions under which our personal data is collected, processed and used and hence a right to ensure that those conditions respect specific legal rules and values. Following this argumentation, I hold that data protection aims to ensure that the unavoidable use of personal data is done in a socially acceptable manner, which in the context of the GDPR, meaning a respect of EU principles and values of liberal and democratic societies. In essence, data protection regulation formulates the conditions and requisites under which the processing of

[441] Charter of Fundamental Rights of the European Union.

personal data is legitimate and defines the techniques and practices which shall be prohibited.

Chapter 5: A Legal and Technical Approach to Explainability

5.1. Introduction

This chapter is based on the academic article *How should my explanation be? A mapping of legal and technical desiderata for Machine Learning models*[442], -hereafter *How should my explanation be?*- which I co-authorised with Dr. Laura State, and Prof Giovanni Comandé. Specific declarations of authorship will be included in footnotes throughout the chapter.

In the previous chapters I have analysed the conceptual, doctrinal and normative frameworks of the rights to information and an explanation for ADM as referred to in the GDPR. I have concluded that the GDPR establishes a framework of transparency and explainability around ADM by incorporating different mechanisms and tools for data subjects to assert their rights. Specifically, these mechanisms are based on the transparency and explainability requirements referred to in Articles 13(2)(h) and 14(2)(g), Article 15(1)(h) and Article 22. I have claimed that this framework of transparency and explainability arise as a solution to the challenges automated decision making can pose to the rights and freedoms of individuals, particularly in regard to their potential arbitrariness, systematicity, inscrutability and lack of neutrality. I have also asserted that the black-box problem arises when an algorithm is included in a decision-making process; as a result, the -potentially problematic- normative features of the algorithm are equally introduced in the process and the resulting decision. It is indifferent whether the algorithm would be considered a white-box or a black-box in terms of its technical interpretability. However, technical black-boxes add another layer of concern when providing information and explanations about an ADM -as referred to in the GDPR-.

[442] Bringas Colmenarejo, State and Comandé (n 67).

I have concluded that the rights to information and an explanations function as a resemble of procedural due process rights to adjust the unbalance power between data users and data controllers when personal data is processed through automated means.

This *Chapter 5* presents a techno-legal mapping covering the potentials and challenges of technical black-boxes to comply with the information and explanation requirements of the GDPR. To this regard, the term of "techno-legal mapping" is used to describe the process of establishing a correspondence between the technical and legal notions of explainability. The objective of such mapping, as with the one included in *How Should my Explanation Be?,* is to provide an analysis and correlation of technical concepts and methodologies with the laws and regulations that govern them. In essence, the term "techno-legal mapping" is used to understand how technology and law interact, particularly where one constrains or enables the other. The process of correlating the technical and legal notions of explainability aims to identify the specific points where technical approaches and processes intersect with legal requirements and expectations.

Although I have already clarified that this thesis does not distinguish between technical white and black boxes for the purpose of ADM, it will not be feasible to analyse all the approaches and methods used in technical fields to ensure the compliance of both type of algorithms -when used in ADM- to the GDPR's transparency and explainability framework. For this reason, the technical assessment of the rights to information and an explanation for ADM will be limited to post-hoc explainability methods for technical black-boxes -see *Chapter 2 Section 2.2.3. Inherent inscrutability and complexity* and *Chapter 6 Section 6.2. A Review Of Technical Explainability Methods*-. The reason behind this decision resides in that the inherent obscurity of non-interpretability of black-box add another layer of complexity to the already complicated compliance with the rights to information and an explanation. This is because data controllers first need to understand themselves the black-box's logic and functioning using post-hoc explainability methods and then, filter and adapt the information about the model and provide it to the data subjects.

Then, analysing the explainability techniques used in this first stage by the data controller can offer insightful answers regarding the possibilities for an effective implementation and exercise of the rights to information and explanation for ADM based on black-boxes. S*ection 5.2.* 'Understandable' Automated Decision-Making Systems provides an overview of the technical notions of interpretability and explainability. *Section 5.2.1.* A technical perspective: the notions of interpretability and explainability explores the distinctive characteristics, implications, and roles of both notions as mechanisms to either ensure or achieve the understandability of algorithmic systems. In Section 5.2.2. How does technical explainability match with requirements on transparency and explainability? I dwell on the interpretation of algorithmic explainability and interpretability from a legal perspective, critically assessing whether there are comparable notions in law and whether the law is actually preoccupied with the technical distinction or not. A problem that I observe when confronting technical and legal explainability is that, although both attain to make the ADM systems 'understandable' to humans, they do not actually mean the same.

Having stated this discord in the interpretation of explainability, I deepen this argument in *Section 5.3.* Desired Properties (Desiderata) For Explanations Of Automated Individual Decision-Making Systems. In *Sections 5.3.1.* Technical Desiderata and 5.3.2. Legal desiderata I present the properties that explanations from both technical and legal perspectives of explainability look to specifically attain, which may or may not coincide. To narrow this rather theoretical analysis on explainability, I connect each desideratum with the information and explanation requirements of the GDPR in concrete and clear terms. Hence, *Section 5.3.* aims to land the techno-legal mapping by drawing the dynamics among and between the two sets and linking the to actual requirements. In this mapping, which based on the one provided in *How Should my Explanation Be?,* I draw the interdependencies and the intersections between the technical and legal desiderata, creating an image that visualises the assessment of the technical and legal driving forces (desiderata matching requirements) in the design and provision of explanations in accordance with the GPDR.

In sum up, Chapter 5 aims to provide the first insights to the potentials and challenges of black-box systems to comply with the rights to information and explanation for ADM.

5.2. ‘Understandable’ Automated Decision-Making Systems

5.2.1. A technical perspective: the notions of interpretability and explainability

The notion of interpretability from a technical perspective refers to the ability of an algorithmic model to ‘describe [explain or present] the internals of a system in a way that is understandable to humans’[443]. For an algorithmic model to be considered interpretable – or transparent by design-, the cause-effect of its performance must be determinable and describable in a way that humans easily understand. Interpretable models intend to create ‘the situation where the user [of the model] is able to comprehend and generate explanations of how the model works (its way of functioning), without being offered any description of the process within the learning model’[444]. Hence, an algorithmic model is interpretable if it is possible to routinely offer the same cause-effect logic for certain inputs and outputs. In other words, interpretability refers to the algorithm’s property to display its supposed real meaning/working to the extent to which a human is able to predict what is going to happen and why changing a parameter of the input will lead to another output. This display includes the system’s mathematical formulas, operation and parameters. Logically, to ensure the interpretability of the model, it needs to be designed in such a way that it allows humans to understand it by just looking at its functioning without the need to be offered any other information about their learning process. A drawback of interpretable models is that only small models remain understandable for humans, i.e. once the rules or parameters of the model increase, its mathematical formulas, operations, and parameters become non-interpretable to humans[445]. Therefore, originally interpretable or transparent by design models can become non-interpretable

[443] Leilani H Gilpin and others, ‘Explaining Explanations: An Overview of Interpretability of Machine Learning’, *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE 2018) p.81 <https://ieeexplore.ieee.org/document/8631448/> accessed 15 September 2022; Finale Doshi-Velez and Been Kim, ‘Towards A Rigorous Science of Interpretable Machine Learning’.

[444] Mateusz Szczepański and others, ‘The Methods and Approaches of Explainable Artificial Intelligence’ in Maciej Paszynski and others (eds), *Computational Science – ICCS 2021*, vol 12745 (Springer International Publishing 2021) 4 <https://link.springer.com/10.1007/978-3-030-77970-2_1> accessed 15 September 2022.

[445] Timo Speith, ‘A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods’, *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (2022).

if they are making bigger and more complex[446]. It is worth mentioning that considering a model transparent by design -or interpretable- does not entirely correlates with agreeing that it is actually understandable. Although by technical means they are considered interpretable, outside of the field that denomination can be highly debated since being interpretable does not necessarily mean that the algorithm model is understandable for an average individual.

By contrast, explainability refers to the action of answering why questions about the model functioning in order to justify why the model acts as it acts[447], primarily 'Why does this particular input lead to that particular output?'[448]. In other words, explainability refers to the actions taken to make the inner workings of non-interpretable or opaque systems clear to humans in a manner that allows them to 'comprehend and literally explain the mechanisms that drive the learning of the systems'[449]. Explainability entails an active mechanism indicating the actions taken to offer clarity or details regarding the model's internal learning process, literally explaining what is happening. Thus, the explainability of non-transparent models by design requires post-hoc explainability methods[450]. Contrary to interpretability, the user [of the model] does not look at it and understands the model in such a way that they can explain the model, but it is through third means -explainability methods- that the user obtained such explanations.

What is, then, the difference between the interpretability or explainability of an ADM model? Doshi-Velez and Kim defined the former as 'the ability to explain or to present in understandable terms to a human'[451] the inner workings of the model. By contrast, the latter 'is about an interaction or an exchange of information'[452] by which the logic behind the model's outcome is intended to be determined. Interpretability is an internal characteristic of the model and, in strict terms, determines whether it should be

---

[446] Aniek F Markus, Jan A Kors and Peter R Rijnbeek, 'The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies' (2021) 113 Journal of Biomedical Informatics 103655.
[447] Bringas Colmenarejo, State and Comandé (n 67).
[448] Gilpin and others (n 443).
[449] Szczepański and others (n 444) p.4.
[450] Speith (n 445).
[451] Doshi-Velez and Kim (n 443).
[452] Laura State, 'Logic Programming for XAI: A Technical Perspective', *ICLP Workshops* (CEUR-WS.org 2021) p.2.

considered interpretable or non-interpretable. By contrast, explainability is, independent of the interpretability of the model, an external action to enlighten why a model has acted as it has[453]. Interpretability is a characteristic of the model that indicates whether a human can understand how the model determine the given output. Explainability is an action by which the inner working of the model is made understandable in human terms. The former is said to be internal because the model is either interpretable or non-interpretable, whether the latter is said to be external because the explainability of the models is achieved through third means -i.e. ad-hoc and post-hoc explainability methods see *Chapter 6 section 6.2. A Review Of Technical Explainability Methods*-.

Even though interpretability and explainability differ in nature and approach, they revolve around making algorithmic models comprehensive to humans, either as an inner quality of the algorithmic model or as a purpose to be achieved through external means. Extrapolating these notions to the information and explanation requirements of the GDPR, it can be argued that making more interpretable models generally facilitates the provision of information about 1) how the ADM worked, 2) what general logic was followed in the decision-making process, or 3) what were the normative basis, and the relevant features and inference in the AMD. Likewise, designing interpretable algorithms eases how a concrete decision affecting an individual is explained; interpretability can help to clarify the individuals' characteristics and features that determine the decision, their weight in the final decision, or the particular correlations applied to the case. On the other hand, working to achieve a higher level of explainability for inherent inscrutable models attempts to reduce the initial imbalance between white and black box systems. Through different methodologies and techniques, explainability experts aim to interpret how the model works, which knowledge logic could have been followed, or which factors were more relevant. However, the model will remain -inherently- inscrutable regardless of the level of explainability reached.

[453] Up to this point, this paragraph was extracted from *How should my explanation be?* Bringas Colmenarejo, State and Comandé (n 67).

When considering explainability and interpretability from a technical perspective is not possible to avoid the question of: *Interpretable and explainable to whom?* Individuals with the required technical expertise will be, probably, the only ones to whom the model is interpretable or who understand the model's explanations obtained through XAI methods. Hence, a model could still be considered interpretable (or explained) but its degree of understandability for a normal human would be under question. When considering the use of interpretability or explainability to ensure the compliance of ADM to the GDPR requirements on transparency and explainability I cannot fail to wonder how exactly would these two approaches help. Indeed, interpretable models are understandable to humans and so the provision of information in accordance with the pertinent GDPR provision is just a matter of selecting the appropriate information. Still, the user of the model -i.e. the data controller- needs to adjust the technical information to the receiving audience -i.e. the data subject-. In the case of explainability, such adjustments are greater as the user cannot select the information from its own understanding, but from the information obtained from third means. These third means or methods, then, need to be calibrated in a manner that provide the information and explanations required by law.

I had the opportunity to work with other PhD students and academics from the field of computer science and there was a requiring question in our conversations: How can we develop [technical] explanations about ADM that are compliant with the law?

The question itself lead to multiple sub-questions;

1. What law?
2. What do you consider an explanation about the ADM?
3. Would you [user] intend for just the [technical] explanation to provide all the information required under the law?

The scope of this thesis is limited to the GDPR, so the first question can be quite straightforwardly answered in this concrete scenario. In real live, it can be very possible that several laws apply at the same time, e.g., the GDRP along with the Modernisation Directive. The second question addresses one of the main line of thoughts of this thesis.

Still, in this Chapter 5 I analyse the contrast between the way technical disciplines approach an explanation of an ADM and how the legal discipline -particularly referred to the GDPR- does it. In concrete, I will examine in the following section the differences between the technical and the legal approach to explainability for ADM and the properties that both disciplines considered desirable when measure the quality of explanations. This analysis is relevant if we want to gather a full view of how technical explainability can -or cannot- help in the provision of information and explanations as referred to in the GDPR. Thus, I will tackle the last question in **Chapter 6** through a experts' group study involving three concrete examples of XAI methods used to make automated decision making systems *understandable* in real live.

5.2.2. How does technical explainability match with requirements on transparency and explainability?[454]

Having analysing the differences between technical interpretability and explainability of a model, and by extension an ADM model, I find necessary to assert that technical interpretability and explainability do not have direct counterparts in the transparency and explainability requirements of the GDPR. In my view, interpretability is an intrinsic technical term that is sometimes used interchangeably with that of understandability and even transparency in legal and social science contexts. Moreover, when the understandability and transparency of a ADM are expected or demanded, the law does not seem to be concerned with whether it is achieved through technical interpretability or explainability[455]. For this reason, I consider that the appropriate comparison should be made between the technical interpretability and explainability of a (ADM) model and

[454] Section 5.2.2. *How does technical explainability match with requirements on transparency and explainability* is heavily based on the academic article *How should my explanation be? A mapping technical and legal desiderata of explanations for machine leaning models*. I acknowledge the major authorship of the article's section *Explainability, ML and Legal Desiderata,* although the wording and content of such article's section has been edited and adapted to the scope of this thesis.

[455] Article 13 (3) (d) of the Artificial Intelligence Act refers to the "technical measures put in place to facilitate the interpretation of the outputs of the high-risk AI systems by the deployer". The new European regulation demands the interpretability of the system for a human, but does not specifies whether the ability to interpret the output of a system would come from the inherent interpretability of the model or due to the use of XAI methods. The GDPR does not seem to be concerned about whether the information about for example the logic involved in the automated decision-making process or the metrics that influence the final decision need to be obtained through the interpretability or the explainability of the system.

its required legal explainability, englobing the latter notions of explainability and justificability. I dwell on the differentiation of these last two concepts hereunder.

On the one hand, a normative reason 'is a consideration that counts in favour of someone's actions'[456]. In other words, normative reasons justify or make it right for someone to act in a certain way[457]. Therefore, legally justifying a decision requires proving its correctness, fairness and lawfulness as referred to in the appropriate laws, norms, and principles. As maintained by Malgieri[458], a legal *justification* of a [automated] decision: 'means not merely explaining the logic and the reasoning behind it, but also explaining why it is a legally acceptable (correct, lawful, and fair) decision'.

Motivating reasons, on the contrary, are reasons that either count in favour of the agents' actions or explain their behaviour[459]. We can distinguish between reasons that motivate and reasons that explain. The former addressed the motives of the decision-maker and their beliefs regarding the reality at hand, while the latter exposes the connection between the knowledge that is available prior to the decision and the following effect[460]. In other words, motivating reasons refer to the subjective knowledge or belief that the decision-maker have about some concrete facts at the moment of making a certain decision, while explanatory reasons allude to the actual facts and the relationship of cause and effect between those facts and the final result or action[461]. In legal terms, the notion of *explanation* contains both meanings, i.e., the provision of information that 'attempts to render a state of affairs, an event or a process understandable'[462] under a motivating reasons perspective. Therefore, if a decision results from an algorithmic decision-making process, its explanation shall disclose the

[456] Thomas M Scanlon, *What We Owe to Each Other* (Harvard University Press 2000) p.18.
[457] Arturs Logins, *Normative Reasons: Between Reasoning and Explanation* (Cambridge University Press 2022).
[458] Gianclaudio Malgieri, '"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation' (2021) 1 Law and Business 16, p.19.
[459] Maria Alvarez and Jonatha Way, 'Reasons for Action: Justification, Motivation, Explanation.', *The Stanford Encyclopedia of Philosophy* (edited by Edward N Zalta&Uri Nodelman, Metaphysics Research Lab, Stanford University 2024) <https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/.>.
[460] Malgieri (n 458).
[461] Alvarez and Way (n 459).
[462] Aulis Aarnio, *The Rational as Reasonable: A Treatise on Legal Justification*, vol 4 (Springer Science & Business Media 1986) Chapter 4 p.22.

connection between the input data and the final decision or the intentions and objectives that motivated such a decision[463].

In consequence, explanations are descriptive and intrinsically grounded on the ADM system with the goal of allowing individuals to understand a single decision or the whole system. Meanwhile, justifications are extrinsic, intended to assess the legality and validity of the decision. Justifications can demonstrate that the decision is grounded on the pertinent rule of law, against which the legality and validity of the decision will be assessed[464]. Using the Twitter Shadow Banning[465] case, we can see the difference between explanations and justifications. On the one hand, the Court states that Twitter must provide information that is useful for the applicant to challenge the decision, such as information about the factors considered in the decision-making process. The information must be complete enough for the data subject to understand the reasons for the decision. On the other hand, the Court sentences that the disclosure offered by Twitter to the applicant was provided in a way that made the context of the restriction unclear, and so the verification of correctness and legality was not possible for the applicant. The Court makes specific reference to the lack of information regarding the existence of ADM, its underlying logic, and its importance and the expected consequence for the applicant. We can see from the case that explanations and justifications about an automated decision are not fully isolated as both help the affected persons exercise their rights fully. Still, we can infer the differences in the goals to be attained.

I conceive legal explainability as a set of legal information requirements that specify the rationale and motivation of ADM. Likewise, I envision legal justificability as the set of information requirements directed to demonstrate the normativity, lawfulness, and legitimacy of ADM as whole[466]. In other words, explainability -in legal terms- is about

---

[463] Malgieri (n 458).

[464] Clément Henin and Daniel Le Métayer, 'A Framework to Contest and Justify Algorithmic Decisions' (2021) 1 AI and Ethics 463.

[465] *Rb Amsterdam - C / 13 / 742407 / HA RK 23-366* [2024] Rechtbank Amsterdam (Amsterdam Court) ECLI:NL:RBAMS:2024:4019.

[466] Our definition of *justificability* foster from the notion of *just* algorithms introduced by Malgieri (n 458) 16. on the basis of which "society want a sustainable environment of desirable AI systems, we should aim not only at transparent explainable, fair, lawful, and accountable algorithms, but we also should seek for "just" algorithms, that is, automated decision-making systems that include all the above-mentioned qualities

explaining how an ADM reached the decision but does not clarify whether that decision was made in a legally compliant way. On the contrary, justificability shows that the applicable legal requirements have been satisfied, both regarding whether the decision was made in a certain way and whether it fulfils the legal reason or conditions for that type of decision.

Looking at the **Chapter 3**, justificability will require to show, for instance, that the data used for its training was lawfully collected and used according to other applicable provisions of the GDPR or that the ADM were made on the basis established in Article 22 of the GDPR with regard to the particular exceptions and safeguards included in it. Explainability, on the other hand, will require to provide the logic involved in the ADM, as well as the features of the individuals used and their relevance in the final decision. Explanations about the ADM would also encompass the motivation behind the use of ADM as well as the consequences its use would have on the individual.

Hence, the main problem I observed when confronting technical and legal explainability is that, although both attain to make the ADM *understandable* to humans, they do not actually mean the same. The technical approach to explainability is concerned with the algorithmic model and its functioning whether legal explainability is concerned with the data processing and process that involve the automated decision-making. In the next section of this chapter, I deepen this argument by presenting and comparing the desirable properties of explanations of ADM systems from both the technical and legal perspective.

(transparency, explainability, fairness, lawfulness, and accountability)". We understand that legal justificability requirements for automated decision-making systems seek to create a framework to assess the legality and validity of the decision, inevitable demanding *just* algorithms and automated decision-making systems.

5.3. Desired Properties (Desiderata) For Explanations Of Automated Individual Decision-Making Systems

5.3.1. Technical Desiderata[467]

Whereas interpretability is an inner characteristic of the model, explainability requires an external action. This external action usually, if not always, is carried out by another model, the *explainer*. Still, the information provided by the explainer will be interpreted and adjusted by the user to the particular needs and requirements in a case by case basis. The two models are independent in so far as the first is the model making the predictor, and the second is the one trying to explain the outcome of the first. This independence is important because when trying to explain the inner workings of an ADM system, we would rather for that external explanation to be as close as possible to reality. The technical ways to do so vary, but the objective remains. That is why in the academic paper *How should my explanation be? A mapping of technical and legal desiderata of explanations for machine learning models*, we summarise the properties that are used from the technical perspective to measure the quality of explanations about ML models. The five desiderata were identified on the basis of three renowned publications,[468] but the list is necessarily incomplete since the state of the art of XAI methods is in constant development and, with it, the amount of literature related to it[469]. Likewise, the new legal requirements developed around the world -special emphasis on the EU legal development of explainability and interpretability for AI systems- push and shape the needs and objectives XAI methods are intended to achieve to

[467] Section 3.1. *Technical desiderata* is heavily based on the academic article *How should my explanation be? A mapping technical and legal desiderata of explanations for machine leaning models*. I acknowledge the major authorship of the article's section *Technical Desiderata AI* to Laura State, one of the co-authors of the mentioned academic article. However, the wording and content of section 3.1 has been adapted and edited by myself to reach a non-technical audience. I also include additional references and content when considered appropriate for the purpose of this chapter.

[468] Zixi Chen and others, 'What Makes a Good Explanation?: A Harmonized View of Properties of Explanations', *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*; Riccardo Guidotti and others, 'Local Rule-Based Explanations of Black Box Decision Systems' (arXiv 2018) arXiv:1805.10820 <http://arxiv.org/abs/1805.10820> accessed 7 June 2022; Molnar (n 115).

[469] We also point the interested reader towards a couple of related survey papers Barredo Arrieta and others (n 42); Markus Langer and others, 'What Do We Want from Explainable Artificial Intelligence (XAI)? - A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research' (2021) 296 Artif. Intell. 103473; Kacper Sokol and Peter A Flach, 'Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches', *FAT** (ACM 2020); Giulia Vilone and Luca Longo, 'Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence' (2021) 76 Inf. Fusion 89.

ensure compliance with the new laws. These changes will undoubtedly determine the desirable properties to XAI explanations.

When developing a XAI model, we can quantitatively measure the appropriateness of the explanation by relying on these five properties:

- Complexity, comprehensibility or interpretability describe the understandability of the explanation to the individual looking at it. A measure that can be used to determine the complexity of an explanation is the number of premises in the explanatory rule; understanding that a premise is 'the number of conditions that need to be satisfied for the consequence to be valid, usually being the consequence a prediction'[470]. Logically, a higher number of premises entails a more complex explanation, and a lower number a more comprehensible one.

- Fidelity or faithfulness describes the real approximation of the XAI explanation to the ML model. Here, it is necessary to consider that post-hoc XAI models mimic the prediction model, so the approximation will never be perfect. In a ratio of zero to one, being one full fidelity, an XAI model will necessarily be below one but as close to it as possible if its fidelity is favoured.

- Accuracy determines the extent to which the explainer is able to predict novel data points or unseen instances. The accuracy of an algorithmic model measures the proportion of correct predictions made by the model against the total number of predictions. Accuracy is one of the many performance metrics. Both the performance of the explanation and the model can be therefore addressed using the metric of accuracy.

- Robustness, sensitivity or stability measures the similarity of the explanation for two different data points. When providing explanations for an ML model, we intuitively expect similar explanations for similar data points unless a logical or good exception applies. Similarity, nonetheless, is not unique but depends on a formalised -technical- notion that can change depending on who defined it and

[470] Bringas Colmenarejo, State and Comandé (n 67) 6.

the type of date used.

- Homogeneity evaluates the possible change in the faithfulness of explanation across different subgroups. Subgroups can be constructed on the basis of sensitive features, i.e. age, race, and gender, which in turn can represent vulnerable groups. The importance of this property relies on its connection to notions of fairness, particularly in regard to the own explanation rather than the automated decision[471]. However, XAI methods have been proposed to be used to measure the fairness of the model[472], making the homogeneity of the XAI a matter of high importance.

Although these five measures can be quantitatively measured, the context where the decision-making model will be used needs to be considered in order to determine the priority of the properties or their desired levels of accomplishment. An explanation of a decision in a context where social minorities have been traditionally discriminated against might be expected to respect higher levels of homogeneity than the explanation of a decision generally affecting the same traditionally favoured subgroup of the population. Certainly, the use of a model whose own accuracy is low -independently of why or in which context that use would be justified- would unlikely demand an explanation with high levels of accuracy. Regardless of the example that we imagine, fidelity seems to be intuitively a property which needs to be prioritised and expected at a high level. All in all, concrete measures and balances of these properties need to be made following a case-by-case approach.

[471] For example, a study (Balagopalan et al. 2022) found that explanations can be less faithful to one subgroup compared to another: such a subgroup can be constructed by separating them based on a sensitive feature, such as age or gender

[472] Aparna Balagopalan and others, 'The Road to Explainability Is Paved with Bias: Measuring the Fairness of Explanations', *FAccT* (ACM 2022).

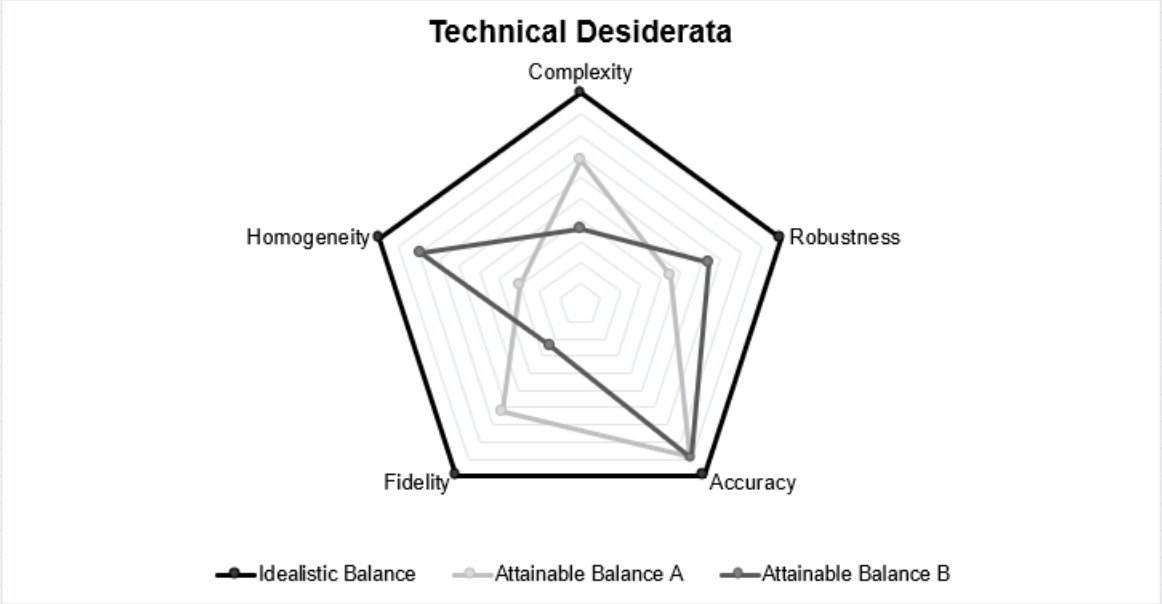

Figure 1. Example of different mappings of technical desiderata. Nodes are labelled with the five identified desiderata. Ideally, all desiderata are fully accomplished (black). In reality, explanations are not ideal. For demonstration, I depict two of such explanations (grey and light grey lines). Here, the degree to which a desideratum is satisfied by the explanation is randomly assigned.

5.3.2. Legal desiderata[473]

In this section, I present the list of five legal desiderata resulting from an assessment of the European legal framework on explainability and justificability made in Bringas, State and Comandé[474]. The desiderata were obtained as follows: we started by observing common desirable properties of legal explainability in the work of Bibal, Lognoul, De Streel, and Frenay[475]; Hacker and Passoth[476]; Lognoul[477]. To the best of our knowledge, these works are the firsts to survey and synthesise requirements on XAI systems in the European legal framework, covering both public and private law instead of limiting their

[473] Section 2.2. *Legal desiderata* is heavily based on the academic article *How should my explanation be? A mapping technical and legal desiderata of explanations for machine leaning models*. I acknowledge the major authorship of the article's section *Explainability, ML and Legal Desiderata,* although the wording and content of such article's section has been edited and adapted to the scope of this thesis. Additional content and references have been included when considered appropriate for the purpose and scope of this chapter.
[474] Bringas Colmenarejo, State and Comandé (n 67).
[475] Bibal and others (n 69).
[476] Hacker and Passoth (n 70).
[477] Lognoul (n 71).

research to a specific area of law, such as health[478], public administration[479], or law enforcement[480]. Thereupon, we reconsidered Malgieri, who stands up for *just* ADM systems, which are only possible 'through a practical justification statement and process through which the data controller proves'[481], why the AI is not unfair, not discriminatory, not obscure, not unlawful, etc. With the distinction between explainability and justificability requirements in mind, we re-examined the European laws addressing ADM systems[482] and put forward the legal desiderata. Sector-specific desiderata (e.g., for public administrations or for consumers) were not addressed in the article beyond the laws discussed in it, thus the list of desiderata shall be understood as a first approximation, which might need to be re-consider on a case-by-case basis. Although the desiderata presented in Bringas, State, and Comandé were developed with the whole European legal framework on explainability and justificability in mind[483] -the GDPR included in such framework-, they are equally and fully applicable to this thesis. To demonstrate this statement, I will connect each desideratum with the GDPR in concrete and clear terms, an addiction to the content of the academic article that cannot be found in its published version.

- Substantive Desiderata: invoke the rights, duties, obligations, and causes of action derived from legal explainability and justificability requirements.

    - Normativity: means tailoring the scope of the information that will be provided to the requirements of the law. Every decision is embedded in a context regulated by various fields of law. This norm specification needs to be determined in the explanation and justification of the decision. Articles 13(2)(h), 14(2)(g), and 15(1)(h) of the GDPR require information about the

[478] Amann and others (n 72).
[479] Olsen and others (n 55).
[480] Raaijmakers (n 74).
[481] Malgieri (n 458) p.16.
[482] I refer to Bringas Colmenarejo, State and Comandé (n 67). for a review of the EU laws addressing explainability requirements for ADM systems.
[483] The scope of ibid. Differs to the scope of this thesis in that the normative framework of the former includes all the European laws including transparency and explainability requirements, whereas this thesis's scope is limited to the GDPR. In ibid. we examine the GDPR, the Artificial Intelligence Act, the Regulation (EU 2019/1150) on promoting fairness and transparency for business users of online intermediation services, The Modernisation Directive (EU 2019/2161), the Proposal for a regulation on preventing the dissemination of terrorist content online.

existence of ADM and at least meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject. In the case of Article 22(3) of the GDPR, it demands information to context the particular automated decision in the form of an explanation. Referring to the minimum threshold of compliance presented in *Chapter 3 section* 3.4.2. The spectrum of compliance – minimum and maximum thresholds, a normative explanation shall provide the basic decision-making process's rules and consequences, as well as the characteristics of the data subject that were used as the main criteria to reach the decision. While these information requirements can seem quite straightforward at first sight, interpreting these formulations in a manner that agrees with technical concepts and approaches towards explainability can pose a great challenge. Indeed, one needs to acknowledge the XAI approaches and the distinction between technical interpretation and explanation set forth in the technical definitions presented above. Additionally, ADM do not operate in siloed environments but in situations affected by multiple laws (e.g., data protection, consumer law, finance products, etc.). Thus, the legal interpretations, which are put in relation to the different legal rules, need to be considered more granularly before defining the specific legal desiderata in any given case. This implies that legal desiderata need to reflect the actual legal requirements functionally. The information offered in such scenarios has to comply with multiple requirements whose coordinated interpretation is a preliminary requirement requesting appropriate legal skills.

- Purpose: refers to various legal objectives and interests aimed to be achieved through explainability and justificability requirements. These requirements are determined by several factors concerning the decision-making process, such as the degree of automation, who is the decision-maker (a public authority or a private firm)[484], who is the individual receiving the information (a user, deployer, or an affected person), in which context the decision is taken,

[484] Bibal and others (n 69).

and which are the potential effects and risks for the individuals and the society[485]. The combination of these factors in an algorithmic decision-making process brings forth different explainability and justificability requirements which respond to various legal objectives and interests, among which Sovrano, Sapienza, Palmirani, and Vitali[486], Hacker and Passoth[487], and Bibal et al.[488] have identified the protection of individuals towards potential risks and harms, the provision of enabling actions and rights, the compliance with the relevant obligations, the building and increase of trust in ML algorithms, the enhancements of market's and sectors' functioning, and the improvement of regulatory oversight. These purposes need to be satisfied by explanations and justifications. As we have argued throughout this thesis, Article 22's right to an explanation becomes a pre-requisite for the effective exercise of the right to contest an automated decision, hence being contestability possibly the main objective to be achieved through it, i.e. enabling data subjects' actions and rights toward the automated decision. Similarly, the information requirements set in Articles 13, 14, and 15 of the GDPR, among other things, look to justify the lawfulness and legality of the existence of ADM, intending to offer data subjects enabling rights to information and explanations[489]. All these provisions undoubtedly have the purpose of protecting data subjects from the use of ADM in a way that can result in some of the challenges and risks presented in *Chapter 2* and *Chapter 4* of this thesis[490].

[485] Hacker and Passoth (n 70); Lognoul (n 71).
[486] Francesco Sovrano and others, 'Metrics, Explainability and the European AI Act Proposal' (2022) 5 J 126.
[487] Hacker and Passoth (n 70).
[488] Bibal and others (n 69).
[489] For instance, Advocate General De la Tour argued that "generally speaking, it is apparent from the case law of the Court that the right of access provided for in Article 15 of the GDPR must enable the data subject *to ensure that the personal data relating to him or her is correct and that they are processed in a lawful manner" C-203-22 Dun & Bradstreet Austria GmbH - Request for a preliminary ruling from the Verwaltungsgericht Wien - Opinion of Advocate General Richard de la Tour* (n 327) para. 44. See also *Rb. Amsterdam - C / 13 / 742407 / HA RK 23-366* (n 465) para. 4.25 & 4.27.
[490] Advocate General De la Tour highlights that "According to the Court, the enhanced requirements laid down by the GDPR as to the lawfulness of automated decision-making and the additional information obligations of the controller and the related additional rights of access of the data subject are explained by the purpose pursued by Article 22of that regulation consisting of protecting individuals against the particular risks to their rights and freedoms represented by the automated processing of personal data,

- Procedural / Formal Desiderata[491]: specify the rules and the methods used to ensure explainability and justificability rights and obligations.

    - Truthfulness refers to the need for the information provided to be accurate, truthful, and complete. Explainability and justificability requirements are rightfully constrained by intellectual property rights and legitimate business interests. Likewise, explanations and justifications should be appropriate to the particular ADM and specific to the norm. These conditions, however, do not excuse the manipulation of the information to achieve scheming or cunning purposes. Upon the definition established in Article 2(1) of the 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, most ADM models fell under the protection of a trade secret. So, it could be argued that information about them shall be disclosed only to the competent authority and court. A strict interpretation of this provision would severely limit the data subjects' right to information and an explanation about ADM. Consequently, both interests should be balanced when providing information on the requirements set in the GDPR. The General Advocate Opinion on the Dun and Bradstreet advised that when the provision of information in accordance with Article 15(1)(h) of the GDPR is likely to result in an infringement of a trade secret – within the meaning of Article

including profiling" *C-203-22 Dun & Bradstreet Austria GmbH - Request for a preliminary ruling from the Verwaltungsgericht Wien - Opinion of Advocate General Richard de la Tour* (n 327) para.50. See also, *C-634/21 SCHUFA Holding (Scoring) - Request for Preliminary Ruling from the Verwaltungsgericht Wiesbaden* (n 169) para. 57, 58, and 59.

[491] Actually, in this case, desiderata correspond to specific existing functional legal requirements. For instance, Article 22 of the GDPR imposes to set up a procedure to address the exercise of the right to an explanation. Likewise, Articles 5 and 12 of GDPR indirectly demand the truthfulness and intelligibility of the explanations that could be used to prove conformity with the law. The same could be said, for instance, in regard to Articles 4, 13, and 14 of the Artificial Intelligence Act, which sought to ensure the interpretability and understandability of the AI system for third parties involved in its use.

2(1) (1) of the Directive on the Protection of Trade Secrets[492],

> That information must be disclosed to the competent authority or court so that the latter can weigh up, in full knowledge of the facts and in accordance with the principle of proportionality and the confidentiality of that information, the interests involved and determine the extent of the right of access that must be granted to that person[493].

- Intelligibility: concerns the language and formulation used, and the presentation chosen for the explanation and justification (text, graphs, images, figures) needed to ensure their understandability and plain clarity. This property also relates to the tension between -technical explanations'- accuracy and interpretability, meaning that a complex, information-rich explanation (i.e. accurate or precise explanation) is often hard to understand for the end-user (i.e. intelligible or comprehensive explanation) and, therefore, fails to fulfil its main purpose. Therefore, an explanation must navigate a trade-off between being easily understandable but sufficiently detailed[494]. Article 12 (1) of the GDPR clearly states that the information provided to the data subjects for the exercise of their rights in regard to the processing of their personal data (i.e. Articles 13, 14, 15 and 22) needs to be 'concise, transparent, intelligible and easily accessible form, using clear and plain language'. Furthermore, Article 29 Working Party Guidelines recommended that 'instead of providing complex mathematical explanations about how algorithms or machine-learning work, the controller should consider using clear and comprehensive ways to deliver information to the data

[492] 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 (Directive on the Protection of Trade Secrets) 2016 (OJ L157/1).
[493] *C-203-22 Dun & Bradstreet Austria GmbH - Request for a preliminary ruling from the Verwaltungsgericht Wien - Opinion of Advocate General Richard de la Tour* (n 327) para. 94.
[494] Malgieri and Comandé (n 47).

subject'[495]. For instance, visualizations or interactive techniques that provide 1) the categories of information used in the automated decision, 2) why those categories were pertinent, 3) how the profile was built, 4) why the profile was relevant for the automated decision, and 4) how was used for the concrete decision.

- Accessibility: indicates the way by which information with explainability or justificability aims must be easily, prominently, and adequately available. In general, obtaining such information should not be hindered or obstructed, although it can be directly and publicly accessible or only accessible to interested parties by default or upon request. For instance, Article 12 (3) states that the provision of information, as referred to in Articles 13, 14, 15, and 22, shall be made 'without undue delay and in any event within one month of receipt of the request. That period may be extended by two further months where necessary, taking into account the complexity and number of the requests'. Information shall be provided free of charge and in electronic means when possible or unless requested otherwise. Likewise, the mere existence of information requirements in the mentioned Articles requires de facto the creation of appropriate channels to provide the specific information. This property is not inherent to the explanation (as most of the other desiderata) but a consideration on a higher level that might necessitate other technical means (e.g., provide a web interface to the explanation, write an easy-to-understand introduction on the web page). Further, it closely ties to considerations about intellectual property rights, customer rights and the intention a provider of an explanation is pursuing. The General Advocate Opinion on the Dun & Bradstreet Austria GmbH case offers a clear example of how the accessibility of the information and explanation about an automated decision can be questioned due to the data controller interests, i.e.

[495] Article 29 Data Protection Working Party, 'Guidelines on Transparency under Regulation 2016/679' (2017) p.31.

protection of trade secrets[496]. Hence, the foreseen procedure of a two-way request of information between the data subject and the data controller is disrupted by the introduction of a third actor, such as the Court or the Data Protection Office, whose main role is to ensure the accessibility of the information through an alternative process in which it -the third actor- acts as an intermediary for the information provision assessing and balancing all the interest at hand.

Contrary to our judgment of technical desiderata, we consider legal desiderata intrinsically qualitative desiderata that do not permit a quantitative analysis or evaluation and require context to determine their degree of compliance. Accordingly, the proposed desiderata should be considered as principles that need to be assessed in each specific scenario.

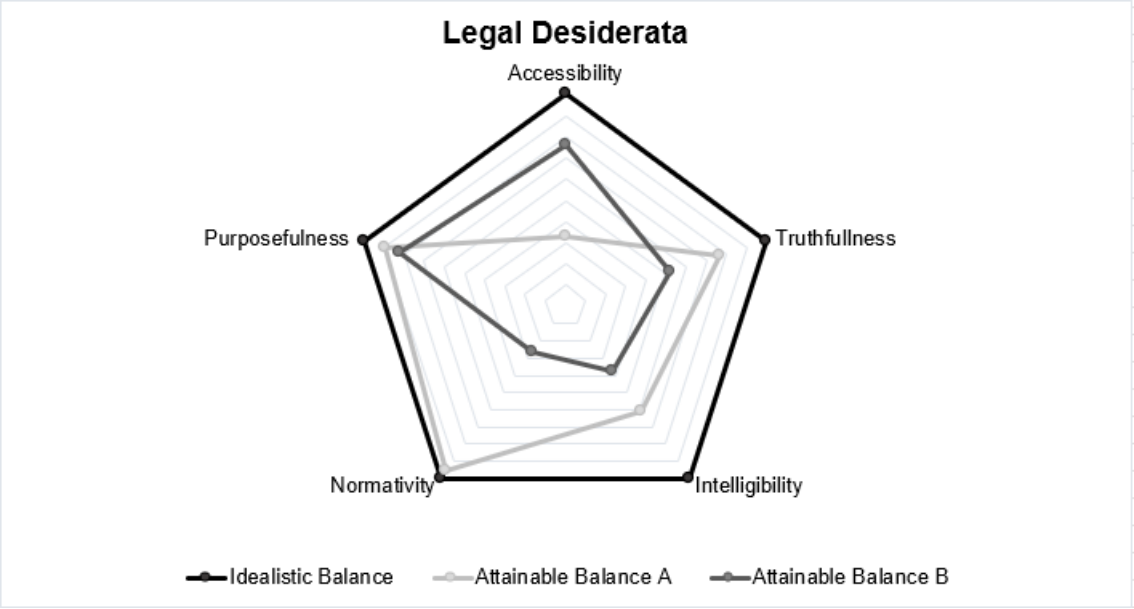

Figure 2. Example of different mappings of legal desiderata. Nodes are labelled with the five identified desiderata. Ideally, all desiderata are fully accomplished (black curve). In reality, explanations are not ideal. For demonstration, we depict two of such explanations (grey and light grey lines). Here, the degree to which a desiderata is satisfied by the explanation is randomly assigned.

[496] *C-203-22 Dun & Bradstreet Austria GmbH - Request for a preliminary ruling from the Verwaltungsgericht Wien - Opinion of Advocate General Richard de la Tour* (n 327).

5.4. Discussion[497]

After offering a comparison between the technical and legal approach to explainability, and the desiderata intended to be reached through the explanations and information provided as per each perspective, I rather answer the previously posed question of *How does technical explainability match with requirements on transparency and explainability?* by asserting that they do not, in fact, fully match.

In a nutshell, technical explainability focuses on statistical and quantitative measures of AMD models' understandability and transparency, and hence, its performance. Contrary to that underlying objective, legal explainability addresses the human comprehension of the decision-making process from a more qualitative perspective and with it the possibilities that opened up from that knowledge in terms of individuals' exercise and enjoyment of freedoms, rights and liberties.

Deepening the previous statement, I assert that technical and legal explainability, and their respective desiderata, differ in how comprehensive they are. Legal desiderata encompass both the (automated) decision and the (automated) decision-making process, while technical explainability -in many cases- concerns only the model used to make a decision within a larger (automated) decision-making process. In other words, legal explainbility covers a broader range of circumstances affecting the ADM than technical explainability. Consequently, the overlay between legal and technical desiderata is not identical, nor does it aim to be, as the object addressed by each desideratum is intrinsically different.

Technical explainability can serve the actors involved with the ADM to understand its internal functioning of the model and so even measure its performance and fairness. Technical explainability can also be used to respond to the legal requirement on explainability and justificability that shall be fulfilled -e.g., as referred to in the GDPR-.

[497] The *Discussion* is heavily based on Bringas Colmenarejo, State and Comandé (n 67).. The authorship of the article's sections *Analysis: How should an explanation be?* and *How can an explanation be?* belong to Bringas and State, and Comandé, however; wording and content of the section has been edited and adapted to the scope of this thesis. Additional references and content were included when considered appropriate for the purpose and scope of this chapter.

Technical explainability can offer insights regarding the internal logic of the ADM model and so responds partly to the legal explainability requirements as it is necessary to understand the logic of the decision-making process. Likewise, some of the information provided through technical means can also unravel some of the intentions or motives behind the decision as well as offer the appropriate reasons to justify the fairness, lawfulness, and correctness of the process and the final decision. However, other relevant information as can be the consequences for the individual or the facts known and used by the controller would not be exposed merely using technical XAI methods. Furthermore, the information obtained as per technical explainability will need to be translated to ensure its understanding for an average individual. Upon this, one should consider that even in the implausible scenario of finding the perfect balance between the proposed technical desiderata, thus the perfect technical explanation about an ADM model, legal explainability and justificability desiderata would remain incomplete. Therefore, it would be necessary to assess which other information -for example, information about the fairness of the decision-making process- would be required to offer the appropriate explanations and justifications.

In essence, the ideal balances presented in the figures above (see Figure 1 and Figure 2) are not more than unfeasible property mappings that would not respond to any real case. When considering the balance between technical and legal desiderata, real scenarios will end up offering a final image where some desiderata prevail over others, as seen in the attainable examples shown in the figures.

If I had to answer the question: How can we develop [technical] explanations about ADM systems that are compliant with the law? I would say that we cannot do it if we intend to only use technical explainability to provide us with all the information necessary to comply with the law.

Chapter 6: How Should It Be an Explanation about an Automated Decision but How Can It Really Be?

6.1. Introduction

This chapter is partially based on the co-authorship article *The explanation dialogues: an expert focus study to understand requirements towards explanations within the GDPR*[498]-hereafter *The explanation dialogues*-. The article has been published in *Artificial Intelligence and Law* and presented at the European Workshop on Algorithmic Fairness[499]. Specific declarations of authorship will be included in footnotes throughout the chapter.

Although further detail will be presented in 6.3.1, it is appropriate to offer here a small summary of *The explanation dialogues*, so the structure and content of this Chapter is more easily understood. In this academic article, my co-authors and I attempt to uncover the expectations, reasoning, and rules of legal experts and practitioners regarding XAI[500]. To do so, we conducted a series of online questionnaires and follow-up interviews with a group of legal experts[501] to whom we presented four different explanations of a fictional automated decision in the credit domain. The four explanations were obtained using three different post-hoc explainability methods: SHapley Additive exPlanations (SHAP), DIverse Counterfactual Explanations (DiCE), and LOcal Rule-based Explanations (LORE).

The explanation dialogue aims to close the gap between technical and legal explainability by inquiring legal experts about the legal compliance of the former with respect to the information and explainability requirements established in the GDPR.

---

[498] State and others (n 52).

[499] Alejandra Bringas Colmenarejo and others, 'The Explanation Dialogues: Understanding How Legal Experts Reason About XAI Methods', *European Workshop on Algorithmic Fairness: Proceedings of the 2nd European Workshop on Algorithmic Fairness* (2023).

[500] The legal experts that participate in the project all share the criteria of being legal experts on the GDPR, particularly on explainability and interpretability of automated decision-making systems. The project aimed to gather the expertise and knowledge of academics and professionals with reputable and renowned careers in legal matters and compliance with AI systems. For this reason, thirty participants were contacted, including academics, researchers, and professors.

[501] We obtained nine validly filled questionnaires and six follow-up interviews

Furthermore, the project attempts to assess the capability of XAI methods to make automated decisions understandable to humans and evaluable with regard to their lawfulness and fairness. In other words, *The explanations dialogues* seeks to provide some clarity in relation to the affinity between technical XAI methods and the legal notions of explanability, including explanations and justifications[502]. Hence, it was quite straightforward to me that the technical assessment of the rights to information and an explanation carried out in this thesis will be strengthened by including part of the results obtained in *The explanation dialogues*.

The following are the project's research questions;

> *RQ1* How do legal experts reason about explanations for automated decision-making systems, and how do they judge the legal compliance of existing methods? Some aspects to consider for a presented explanation:
>
> *(a)* Is the explanation complete or incomplete with regard to the expectations of the legal scholars, and is some information given by the XAI more relevant than others?
>
> *(b)* Is the explanation compliant with the GDPR, and is there a preference towards a specific method or presentation type?
>
> *(c)* Does the legal reasoning change when presented with the explanation of a true positive/false positive?
>
> *RQ2* Do legal experts understand and trust explanations for automated decision-making systems, and what are the steps identified to go forward? Some aspects to consider are:
>
> *(a)* How well are the presented explanations understood?
>
> *(b)* Which gaps in presented explanations are identified? How can the presented explanations be improved?

[502] The academic papers were, nonetheless, develop independently.

These research questions that lead *The explanation dialogues* complement or rather go into detail about the third research questions of this thesis, i.e. what is the potential development of the exercise of the rights to information and an explanation in the context of ADM? *The explanation dialogues* asks this same question under the premise of four concrete type of technical explanations for an automated decision, offering a befitting case study for this thesis. Instead of theorising on how XAI methods in general could be used to ensure compliance with the rights to information and an explanation, *The explanation dialogues* offers four concrete and tangible technical explanations about an automated decision that I can confront against the findings of the previous chapters of this thesis.

Furthermore, the four explanations can be also exposed to the questions presented in previous Chapter 5 regarding: How can we develop [technical] explanations about ADM that are compliant with the law? *The explanations dialogues* already presents a case study in this regard under a set of legal premises, which coincide[503] with the conclusions reached in doctrinal and normative frameworks presented in Chapter 3 and Chapter 4 of this thesis,

- The right to contest an automated decision, as referred to in Article 22 (3) of the GDPR, 'serves to perfect more substantive rights of fairness and justice and to preserve the rule of law values, by correcting errors, preventing or changing unjust outcomes, and enhancing predictability and consistency of decisions'[504].
- Article 22 of the GDPR establishes a transparency framework for legally or similarly significant solely automated decisions on the grounds that the effective implementation of the rights to contestation, information, and an explanation inevitably requires making the decision-making process and the reached decision understandable -to some or other extent- to the data subject[505].

[503] This coincidence was not accidental. The scope of *The explanation dialogues* was agreed to be delimited to the rights to information and an explanation as per the GDPR with the aim of ensuring that the academic paper fits in the scope of this thesis.
[504] Margot E Kaminski and Jennifer M Urban, 'THE RIGHT TO CONTEST AI' (2022) 121 COLUMBIA LAW REVIEW 93.
[505] Bayamlıoğlu (n 268).

- Making an ADM understandable and contestable, therefore, implies some level of technical interpretability or explainability for the ADM model.

Chapter 6 aims to assess the level of compliance XAI methods can reach when use to fulfil the requirements of information and explanations about an ADM. To do so, I start by offering a conceptual taxonomy of XAI methods. I examine the different types of methods available to make understandable the functioning of algorithmic models and the main reasons behind their outputs. I also explain the type of explanations that can be obtained through the use of these methods. This approach looks to identify the objectives behind the design, development, and application of some of the most commonly used and technically developed explanability methods in the field of XAI and appraise whether their rationale coincides with the rationale behind legal explainability.

In *Section 6.3* I expound on the perceptions, expectations, and reasoning offered by a group of legal experts and practitioners on legal explainability when questioned about four concrete explanations about an automated decision. In *Section 6.4* I support the final conclusions reached in *The explanations dialogues* with my own assessment of the four explanations that combine the Spectrum of *Compliance* presented in *Chapter 3.4.2.* and the *Technical and Legal Desiderata* analysed in *Chapter 5.3.*.

6.2. A Review Of Technical Explainability Methods

6.2.1. A conceptual taxonomy of eXplainability methods

The field of XAI aims to offer different methods that clarify the technical functioning of non-interpretable systems and provide tools for individuals to interact with or deal with their predictions and decisions. In particular, post-hoc explainability methods aim to explain the functioning of a non-interpretable model after it has been trained. Post-hoc explainability attempts to explain or present the model's functional and physical variables, structures, and processes that intervene in transforming an input into an

output[506]. These explanation methods can be classified according to the following distinctions.

Post-hoc explanability methods can refer either to global 'aiming at the overall logic of a black-box model' or local 'aiming at the reasons for the decision of a black-box model for a specific instance'[507]. The former aims to explain the general patterns of the model applicable at any input and instance. The latter focuses on the reasons for a specific decision and is only suitable for a particular input or instance. Additionally, post-hoc explanability methods can be used to interpret any type of black-box model, so-called model-agnostic, or can only be used to explain a specific type of black-box model, so-called model-specific -e.g. being specific for random forest, gradient boosted machines or neural networks. Model-agnostic methods seek to offer some insight into the function of the underlying model regardless of the type of model involved, either by approximating their function to another model -surrogate models- or by decomposing the importance of each variable or feature in the function of the model -feature importance[508]. By contrast, model-specific methods are designed based on the specific model architecture and structure and exploit the internal structure of the model to provide information and explanations about its logic and reasons[509].

Regardless of the above mentioned taxonomy of XAI methods, we can also differentiate between model-centric or subject-centric post-hoc explanations. This classification does not share a common ground with the taxonomy presented above as it refers to the explanation of the ADM, rather than to post-hoc explainability methods. However, this taxonomy of explanations could not be understood without the previous conceptual taxonomy of XAI methods, since the latter are necessary to obtain part of the information that will be included in the explanation. Model-centric explanations refers to explanations that provide broad information about a model that is not a decision or

[506] Carlos Zednik, 'Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence' (2021) 34 Philosophy & Technology 265 see that algorithmic systems cannot currently provide answers to these questions because they are not built/designed to explain to the general public nor policy makers. See also ; Ferrer and others (n 103).
[507] Bodria and others (n 80) p.4.
[508] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, 'Model-Agnostic Interpretability of Machine Learning' (arXiv, 16 June 2016) <http://arxiv.org/abs/1606.05386> accessed 22 September 2022.
[509] Edwards and Veale (n 89).

input specific -an explanation as global- while the latter are built on and around the basis of a concrete input record -an explanation as local-. Model-centric explanations can provide information regarding the model's setup, the training metadata, the performance metrics, the estimated global logic, and the process information. They can offer information such as the intentions behind the modelling process, the family model, which are the parameters used in the setup as well as the input, output and classifications used and predicted during the training, or the rate of success on specific salient subcategories of data, the variable importance score, and how the model was tested, trained and screened. By contrast, subject-centric explanations can give meaningful information related to what changes in the input data would have made the decision different, which training data was most similar to the input, who has also received a similar treatment to the subject, which was the erroneous and misclassification rate along with different groups and individuals during training[510].

A large number of XAI methods have been proposed in the technical literature. As explained above, they are distinguished along two axes: a) whether they are valid only for the data instance in focus (local) or apply to the full model (global), and b) whether they are tailored to a specific model (model-specific) or can be used to explain any (model-agnostic)[511]. Here below, I presented three well-known model-agnostic XAI methods: SHapley Additive exPlanations (SHAP), DIverse Counterfactual Explanations (DiCE), and LOcal Rule-based Explanations (LORE). DICE and LORE provide explanations with a local scope, whereas SHAP can provide explanations both with a global and local scope. In *The explanation dialogues* we used these concrete methods to develop four types of explanations -three local and one global- about a fictional ADM for credit scoring. The selection of methods was not accidental, but responded to their high level of development (state-of-the-art), their common use, and their support from the legal community[512]. Thus, I present them in this thesis because they are three XAI

[510] ibid 55–58.
[511] State and others (n 52) see also ; Guidotti and others (n 115); Molnar (n 115).
[512] DICE and LORE are both a type of XAI method that provide contrastive explanations. Although explained in more detail below, contrastive or counterfactuals explanations about an automated decision were claimed to be the most suitable type of explanations to comply with the GDPR's information and explanation requirements by Wachter, Mittelstadt and Russell (n 299).

methods that can be actually use in high-consequence ADM to provide information and explanations as per the GDPR's requirements on transparency and explainability.

The description included in this thesis of these methods is based on *The explanation dialogues* and the images that pair with each description are actual visualizations of how each particular method explains an ADM model or automated decision[513] as per the rationale of each type of XAI method.

Including a description and analysis of these particular three XAI methods in this thesis does not merely serve an exemplary purpose. With this overview, I resolve to offer some insights into what exactly the technical perspective of explainability for ADM entails, i.e. what is the type of technical information that can be provided about a non-interpretable ADM model or what are the objectives that are intended to be attained through XAI. Hence, this overview can show how the black-box problem might not be easily solved through XAI methods since the own XAI method used to overcome it can also suffer it. XAI methods can introduce another layer of obscurity and lack of neutrality to the ADM because they themselves suffer from the black-box problem .

This overview is undoubtedly incomplete, but it still presents four renowned and commonly used XAI methods, which shall be sufficient to fulfil the purpose mentioned above.

[513] The description of the selected XAI methods is heavily based on academic article *The explanation dialogues: an expert focus study to understand requirements towards explanations within the GPRR*. I acknowledge the main authorship of the article's section *2.2. Explainable AI* to Dr. Laura State. The description of each type of explanation comes with its proper visualization. The technical development of such explanation was made by Dr. Laura State, who gave permission for these to be included in this thesis as visual examples. I edited, expanded, and adapted the wording and content of the article's *section 1.2. A selection of post-hoc eXplainability methods: SHapley Additive exPlanations (SHAP), DIverse Counterfactual Explanations (DiCE), and LOcal Rule-based Explanations (LORE)* to reach a non-technical audience. Additional references and content was included when considered appropriate for the purpose of this chapter.

6.2.2. A selection of post-hoc eXplainability methods: SHapley Additive exPlanations (SHAP), DIverse Counterfactual Explanations (DiCE), and LOcal Rule-based Explanations (LORE)[514].

6.2.2.1. Feature relevance methods

On the one hand, feature relevance methods provide a measure of how relevant a feature is to the decision outcome. A feature with a high value, therefore, means that the feature is highly relevant to the decision outcome. Methods vary by how this measure is calculated and, therefore, by the exact meaning of the measure. As a way of simplification, feature relevance methods will highlight the features that were more relevant, for instance, to classify an individual as a good or bad debtor, e.g. the person's income, years of stable work experience, or personal assets.

[514] Although it might not look intuitive at first sight, the ADM model used in the fictional case-scenario presented in *The explanation dialogues* was designed to identify the probability of default (or stablished creditworthiness) of a given individual. This is to say the individual's ability or willingness to pay a credit. In *The explanation dialogues*, we fix a threshold of creditworthiness that is set to be optimal and the individual would be either rejected or non-rejected. Hence, the explanations presented below refer to an output of rejection that was correctly predicted, meaning that the individual was correctly predicted to be a high-risk creditor. In a real case scenario, though, the label of true positive (defaulters that have been assigned a high probability of default) would not be accessible as they were not actually given a credit so it would not be possible to check and known. This problem is called the *rejected inference problem*. One of the main downfalls created by the prediction of creditworthiness is that you would never know whether the individuals the system predicted as high-risk (and therefore were rejected their application) would ended up being a bad creditor. In essence, the only way to explain the functioning of the systems for "bad creditor" instead of high-risk score it is to work based on the false negative outputs, hence those who were predicted to have a low risk-score but ended up not paying. This situation creates an observer bias, meaning that you only know how actually those who you granted credit behave. A quite straightforward example applies to the female population who has been historical excluded from credit access. These lack of information about their credit behaviour, left them to be usually consider high-risk, no matter their circumstances. For interested reader I recommend Sebastián Maldonado and Gonzalo Paredes, 'A Semi-Supervised Approach for Reject Inference in Credit Scoring Using SVMs' in Petra Perner (ed), *Advances in Data Mining. Applications and Theoretical Aspects* (Springer 2010); Nikita Kozodoi and others, 'Shallow Self-Learning for Reject Inference in Credit Scoring' in Ulf Brefeld and others (eds), *Machine Learning and Knowledge Discovery in Databases* (Springer International Publishing 2020); Zhiyong Li and others, 'Reject Inference in Credit Scoring Using Semi-Supervised Support Vector Machines' (2017) 74 Expert Systems with Applications 105.

6.2.2.1.1. Shap

SHAP is a unified framework[515] for interpreting predictions based on the cooperative game theory's concept of 'Shapley values'[516]. The method provides an importance value to each feature considered in a prediction. The relevance of a feature for the prediction is indifferent to whether it is a positive or negative influence.

The average feature importance (calculated as an aggregated Shapley) is obtained as the absolute average of individual SHAP explanations of any set of applications determined by the developer (e.g. those application occurred in the last three years, or all application ever occurred in the bank). For example, during development a bank may only use individual SHAP explanations of training data whereas in production a bank may use the applications for credit of the past ten years. For instance, if the targeted population changed for unexpected reasons, a bank might be interested in modify the set or even in comparing the Global SHAP explanations obtained for the applications included in the training set and the applications occurred in the last five weeks.

The first image shows a (local) SHAP explanation for a prediction of a bad debtor whose request for credit was rejected on the basis of their high-risk. Features in blue represent the particular features that contribute (negatively or positively) to the prediction of the individual as good debtor, and in red, the ones contributing (negatively or positively) to the prediction of the individual as bad creditor. The first explanation shows the importance of the features for determining the individual as a bad creditor, whereas the second explanation displays the features that would determine the individual to be a good debtor. The added magnitude of the features classifying the individual as a bad debtor was higher that the respective added magnitude of the features for the opposite classification. Hence, the individual was classified as a bad creditor. The order in which

[515] Scott M Lundberg and Su-In Lee, 'A Unified Approach to Interpreting Model Predictions' in Isabelle Guyon and others (eds), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (2017) <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>.

[516] Shapley values is a solution concept in cooperative game theory introduced by Lloyd Shapley in 1851. The solution assigns a unique distribution (among players) of a total value or payoff to each cooperative game. The value represents a fair share or payout to each player, considering their marginal contributions to all possible coalitions.

the features are displayed and their length show their greater or lesser importance for the final outcome.

A downside of local SHAP explanations, as presented below, is that is it not possible to gather whether the influence of a feature in the final decision was positive or negative. Other type of visualisation of the explanations could, nonetheless, provides more clearance on that regard.

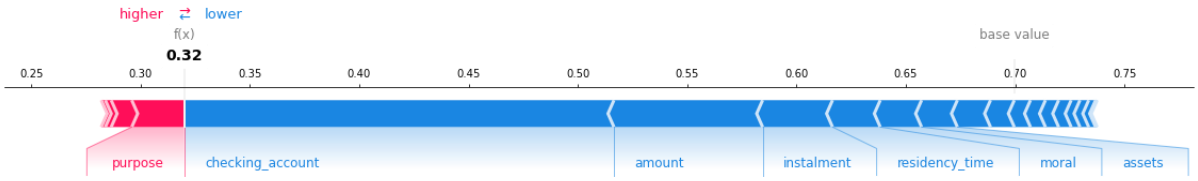

Figure 4: (Local) Shap explanation – Features' importance for determining the individual as a bad creditor.

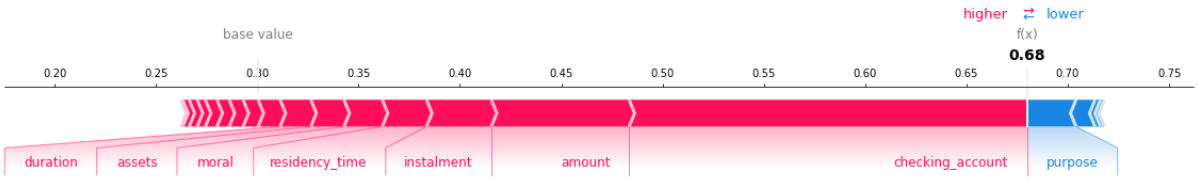

Figure 5: (Local) Shap explanation- Features' importance for determining the individual as a good creditor.

The image below shows the average importance of a feature in a model's predictions. In order to showcase the inner working of the model at the time of development the plot shows the SHAP values with respect to the training data. Therefore, we can have a feature importance assessment regarding the true label[517] (good or bad creditor shown in blue and red respectively), instead of regarding the high-risk or low-risk (predicted label) credit score. This way we have a better understanding of the model behaviour on

[517] In the contest of credit scoring we refer to true label as the observed default status of a given customer, in essence whether the client has repaid their debt or not. On the other hand, we refer to predicted label as the one we have estimated via modelling.

the bad creditors that due to their small number will be overshadowed by the good ones[518].

In the case of credit scoring the distinction between true label and predicted label is of special relevance for two distinct reasons. The first one being that the true label is not immediately known and will take some time to show, depending on the type of credit this can go from weeks to years. The second is that, once an individual is predicted to be high-risk, the creditor will reject the credit, and hence the individual will disappear from the database. As a consequence, the true label is never observed due to a selection bias. This selection bias is a consequence of the inner workings of the industry and it is unavoidable, meaning that the training set -portrayed in the plot below- will inevitably suffer from the previously mentioned selection bias as it has been generated from accepted applications from which we have the true label. We deemed appropriate that the best approach to show the global SHAP was to show them in true labels since it is expected that the mix of good and bad creditors is more even after deployment. This is because whereas the set of applications received does not suffer from the selection bias, the set of approved applications do suffer from it.

The plot below is showing the mean of the absolute Shapley for each of the true labels. This means -in the case of the bad ones- taking all the applications to which we know the true label to be bad (one), obtaining and summing (in absolute form) the SHAP values for each of the features and then dividing it over the number of bad creditors in the sample.

However, it is worth noting that Global SHAP explanations can be obtained both from the training set and from the population subjected to the predictor, even when the true label is not known.

$$aggregated\ SHAP = \frac{\sum_{i=0}^{n} |Shap_i|}{N}$$

[518]Credit scoring is an unbalance classification problem, meaning by this that the number of bad creditors is considerably smaller than the good creditor. Hence, any statistic in this case, the average feature importance will unfairly represent the effect on the bad creditor.

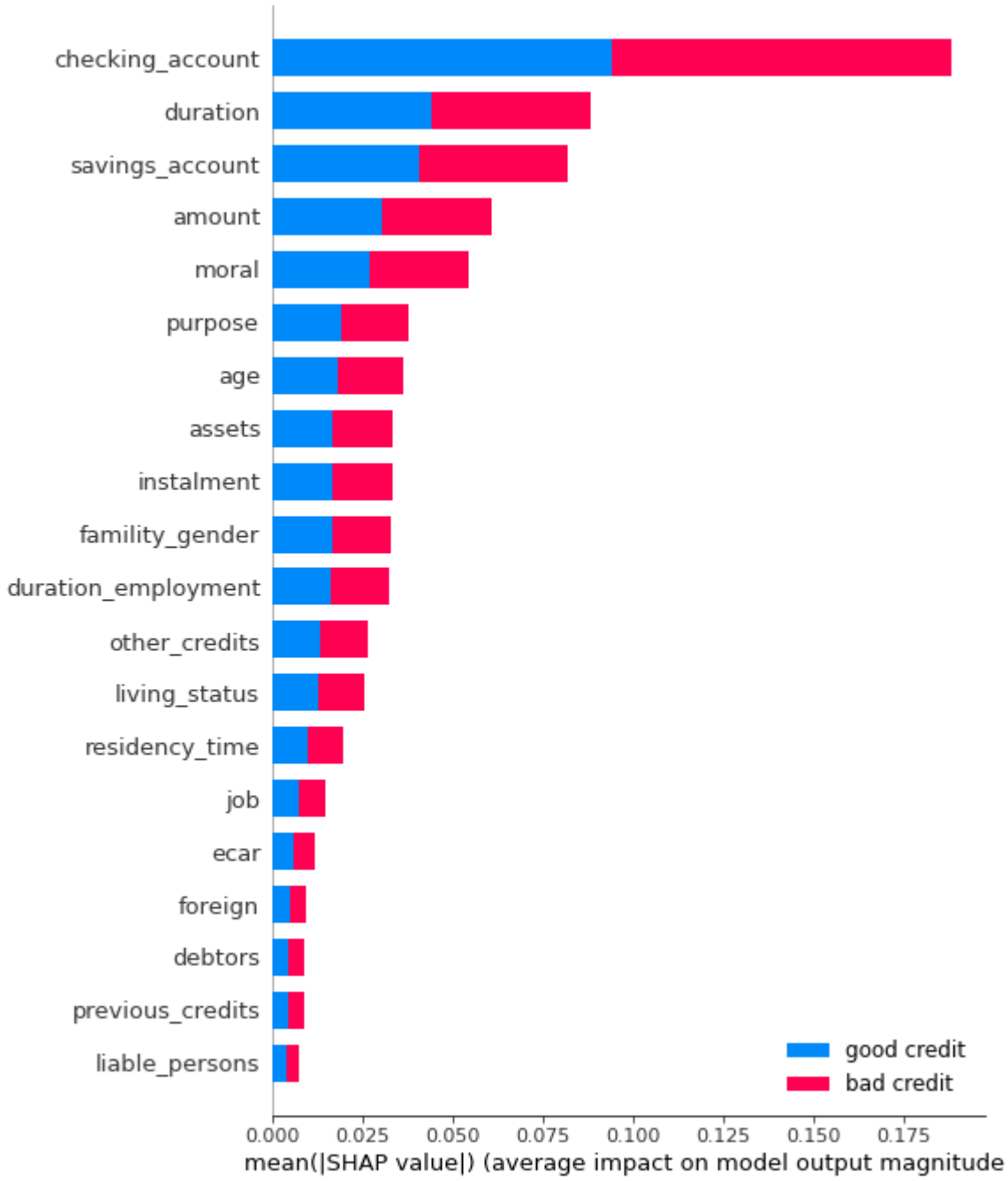

Figure 6: (Global) SHAP explanation – Features' importance in the model's prediction.

519

[519] Explanation as described in *The explanation dialogues*: The figure displays global SHAP values, which average the local SHAP values over a set of instances. The higher a value, the more relevant it is an instance for the model's prediction, with the attribute "checking account" being most relevant, followed by "duration" and "savings account".
The color differentiates the relevance with respect to the predicted class, i.e. good or bad credit predicted

6.2.2.2. Contrastive explanations

On the other hand, contrastive explanations[520] highlight the difference between two or more outputs (predictions) of a model, i.e. offers for a single instance the opposite -contrastive- outcome. They attempt to clarify why a particular prediction happened in contrast with another prediction outcome and so the difference between the contrastive and the original outcome. Contrastive explanations provide an answer to the question: How should the input data look like in order to obtain a different output? In other words: How should the current input change to obtain a different output? Or to answer what-if questions ('What happens to the output if the input changes that way?'). Contrastive explanations were introduced into the field of XAI by Wachter, Mittelstadt, & Russell[521] under the name of 'counterfactual'. In their work, Wachter, Mittelstadt, & Russell proposed the argument by which each automated decision shall be accompanied by an statement of 'how the world has to be different for a desirable outcome to occur'[522]. Counterfactuals do not necessarily need to be singular since multiple desirable outcomes could be possible or advisable depending on the circumstances of the individual. The most attractive counterfactual shall be found on the basis of the 'closest possible world(s)', understood as 'the smallest change to the real world that can be made to obtain the desirable outcome'[523]. In other words, the most attractive - feasible and attainable- counterfactual would be the one requiring the smallest of changes in the real world, for instance, a reduction in the amount of credit requested, rather than an increase in the perceived annual income of the individual. Sometimes, instead of the 'closest possible world', the best counterfactuals are the 'close possible worlds', which provide counterfactuals with diverse and relevant changes in the attributes of the individuals or their circumstances. Logically, the most relevant counterfactuals for a particular prediction are quite context-specific as they

[520] Contrastive explanations are related to the concept of counterfactuals as understood in the statistical causality literature. However, these concepts are not the same. To avoid confusion, we therefore use "contrastive".
[521] Wachter, Mittelstadt and Russell (n 299).
[522] ibid 844.
[523] ibid 845.

depend on the facts that led to the decision. Hereafter, I focus on only two contrastive XAI methods, DiCE and LORE.

6.2.2.2.1. DICE

DiCE provides hypothetical examples of how to obtain a different prediction than the one received. This method looks for the perturbation or changes that would lead to a different outcome. The proposed counterfactual are constrained by the properties of diversity and feasibility in so far as an effective and actionable counterfactual is one for which one or several input features are modified to change the model's decision while respecting the user's context and constraints (feasibility) and providing users with different ways of changing the outcome (diversity)[524]. The feasibility feature highly relates to the concept of the 'closest possible world'[525] but incorporates other user-defined constraints that are chosen in a case-by-case way.

In the image below, we see in the first column the all the attributes/features that are used by the model to predict the creditworthiness of the individual. influencing the prediction of bad creditworthiness that lead to the rejection of credit. The second column shows the category that each feature obtained based on the particular circumstances of the individual's request for credit -i.e. the data point-. The purpose declared by the individual for the credit was *furniture/equipment,* which was categorised as 3. We can imagine that other purposes, such as housing, holidays, or means of transport, would be located in other categories, for instance, 1, 2, 5, or 9. The third and fourth columns provide the counterfactuals for the individual to obtain a *low-risk* score and so be granted credit. For this particular scenario, the first closed possible world requires the individual to have a negative balance in a checking account and three or more liable persons. Hence the feature of the *checking-account* would need to change to category 2 (...<0) (checking account with negative balance) instead of remaining in 1 (no checking account) and the feature of *liable-person* would need to

[524] Ramaravind K Mothilal, Amit Sharma and Chenhao Tan, 'Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) p.609 <https://dl.acm.org/doi/10.1145/3351095.3372850> accessed 15 October 2024.
[525] Wachter, Mittelstadt and Russell (n 299).

change to category 1 (3 or more). The second closed possible world presented would require the individual to also have a checking account in negative balance and have a guarantor. In other words, the checking-account feature would need to change category as in the other counterfactual, the feature debtor would need to change from 1 (none) to 3 (guarantor). Surprisingly, both counterfactuals provide the two closest possible worlds where the individual needs to have at least a checking account open. This circumstance could be explained for numerous reasons, but a very logical explanation is the high importance of such a feature for the granting of credit. Banks will consider individuals with no checking accounts very unreliable. So, no matter what other changes could be made to their credit application, individuals will -most likely- be considered *high-risk* unless they open a checking account and the associated category changes.

Attribute name	**Original Data Point** Level (Associated category)	**Counterfactual 1** Level (Associated category, if changed)	**Counterfactual 2** Level (Associated category, if changed)
checking_account	1 (no checking account)	2 (... < 0 EUR)	2 (... < 0 EUR)
duration	24	24	24
moral (whether the past credits were paid on time, delayed etc)	2 (no credits taken/all credits paid back duly)	2	2
purpose	3 (furniture/equipment)	3	3
amount	1282	1282	1282
savings_account	2 (... < 100 EUR)	2	2
duration_employment	1 (unemployed)	1	1
installment	4 (< 20)	4	4
familty_gender	2 (female : non-single or	2	2

	male : single)		
debtors	1 (none)	1	3 (guarantor)
residency_time	2 (1 <= ... < 4 yrs)	2	2
assets	3 (building soc. savings agr./life insurance)	3	3
age	32	32	32
other_credits	3 (none)	3	3
living_status	2 (rent)	2	2
previous_credits	2 (2-3)	2	2
job	2 (unskilled - resident)	2	2
liable_persons	2 (0 to 2)	1 (3 or more)	2
Electric car	2 (yes)	2	2
foreign	2 (no)	2	2

Figure 7: DICE explanation – Table of features for the prediction and two counterfactual.

6.2.2.2.2. LORE

Finally, LORE is a method that aims to provide the rule behind a particular decision and a set of counterfactual rules proposing the conditions that shall be changed to alter the outcome[526]. LORE paid special importance to the neighbourhood of the data point (the particular decision), meaning that instead of looking at the whole data space (the immense number of possible other data points), it looks for a data point in the vicinity of the particular decision which, however; pertained to the other side of the decision boundary (would obtain a reverse outcome)[527]. This logic follows, again, the rule of 'closest possible world'. Still, LORE offers two types of explanations in one, first, a

[526] Riccardo Guidotti, 'Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking' (2024) 38 Data Mining and Knowledge Discovery 2770.
[527] ibid.

factual rule responding to the question of 'Why did the model predicted this?' and, secondly, a contrastive rule highlighting 'How can a different prediction of the ML model be achieved?'.

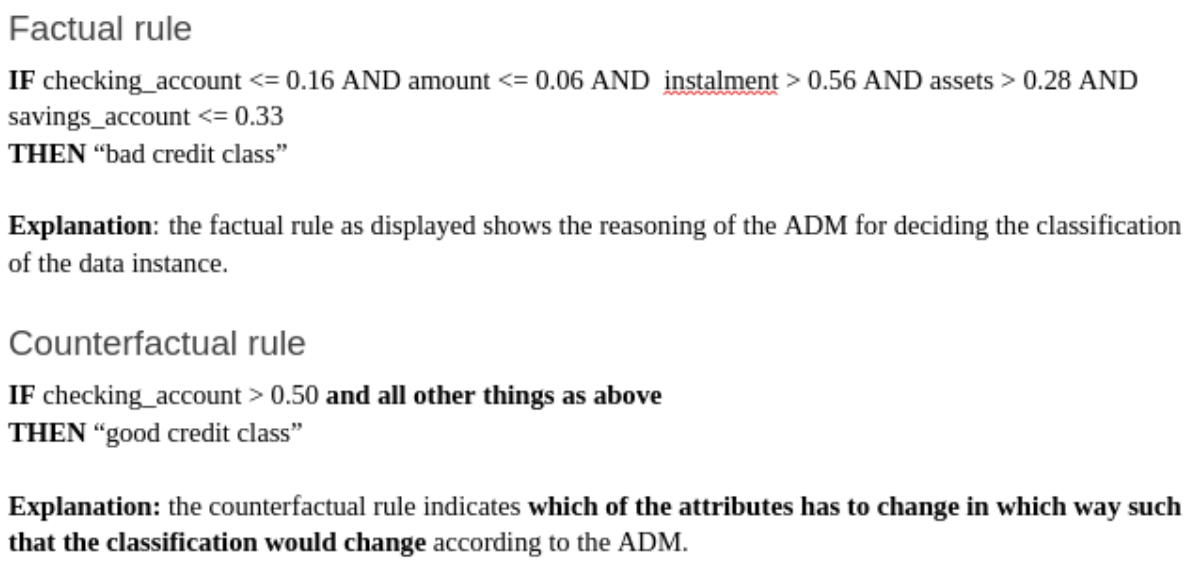

Figure 8: LORE explanation – Prediction's factual rule and counterfactual.

6.2.2.3. Discussion

The methods presented above seek to clarify, in one way or another, the functioning of the black-box model and provide explanations to individuals about their functioning and the reasons behind the particular automated decision so they can understand and act upon them. The XAI methods try to approximate the logic behind the ADM model, either in regard to its general functioning or a particular decision. The *fidelity* property of technical explainability pinpoints the impossibility of obtaining a perfect approximation of how exactly the ADM model works. The methods presented above show different strategies used in the design of XAI methods to leverage this unattainability or, at least, to partly provide relevant information about the ADM model. For instance, SHAP focuses on the weight each feature has in the final decision or their average importance in the general logic of the model. In contrast, DICE only focuses on which feature should change to obtain a different decision; the particular feature's importance or the logic rule followed by the model to reach a decision are not relevant for the way this method provides explanations.

These three XAI methods are merely three concrete types of explainer systems, but in essence, they exemplify the interests and objectives aimed to be achieved through XAI. It is very relevant to ponder whether these objectives aligned with the legal requirements on explainability and to what extent. For this reason, the following expounds on the perceptions, expectations, and reasoning offered by a group of legal experts and practitioners on legal explainability when questioned about the these same explanations. In turn, in *Section 6.4.* I provide an assessment on how these post-hoc explanation methods respect the legal transparency and explainability requirements proposed in Chapter 3.4.2. *The spectrum of compliance – minimum and maximum thresholds.* This third section also serves as a conclusive discussion for this thesis through a concrete case study.

6.3. A Legal Experts' Reflection On Post-Hoc Explanations – Shap, Lore, And Dice

The purpose of this section is to answer the question: How should it be an explanation but how can it really be? As stated above, the information and explanations about non-interpretable ADM models that can be provided nowadays to comply with the explainability and transparency requirements as established in the GDPR are highly restricted by the current state of the art of XAI methods. The interest of this section relies on the possibility of examining real explanations about an ADM model and its particular decision and assess the real compliance of XAI methods to the information and explainability requirements of the GDPR.

6.3.1. Introduction to the Explanation Dialogues Project[528]

The Explanation Dialogues was tied to a specific application scenario of an ADM system that determined whether credit was granted or not and that was provided by a private actor situated in the EU. The loan application scenario involved a fictional bank, an internal consultant of the bank (the participant-expert), and a fictional customer who is applying for credit. The creditworthiness of the consumer was assessed using an ADM

[528] This thesis section is heavily based on the work of State, Bringas, and Beretta for the sections *Background and Related Work* and *The explanations dialogues* in State and others (n 52).

model, which was trained to predict a risk score. Three XAI methods were used to provide an explanation to the customer about the approval or rejection of the application. The fictional receiver of the explanation was a layperson (the *average* bank customer). The bank was a private actor and, thus, must comply with the GDPR.

In the questionnaire, the participants -legal experts- were explicitly asked to answer the questions about the explanations from the perspective of the bank's internal legal consultant. The participants had to assess three types of explanations: 1) an explanation common to all participants, which includes basic information about the ADM and its use (i.e. model information, the data set and splits, and the performance and confusion matrix of the ML model), 2) a global explanation, common to all participants as well, obtained using global SHAP, and 3) two local explanations obtained through two different explanainability methods and randomly selected for each participant from Local SHAP, LORE, and DICE explanations. The assessment of the set of explanations was done twice by each participant, once for a true positive and once for a false positive output, which means that one output of the ADM model correctly predicted the bad creditworthiness of the individual while the other did not. The same set of questions followed each presented explanation, and at the end of each case (false or true positive), a group of comparison questions were also asked. After the online questionnaire, Laura and I interviewed each interested participant. The role of the participants in the interview was much broader. We referred back to the loan application scenario and asked the participant to answer general and specific questions from the previously assigned role of an internal consultant of the bank. Although we had a set of shortlist questions, when the reflections and knowledge of the participants opened unplanned but relevant paths of questioning we slightly drifted towards that direction.

The questionnaire responses were analysed after the follow-up interviews. Still, we identified two preliminary relevant factors from a superficial reading of the participants' questionnaire responses done before the interviews: 1) explanations' understandability and legal compliance were interconnected, and 2) there was a potential lack of a conflict of interest between the parties involved in the ADM in regard to explainability. Hence, the interview script was developed in such a manner that the participants could

offer further details, arguments, and discussion over two problems: 1) the challenge to understand the explanations shown due to the amount of (or lack thereof) information provided and the lack of helpfulness of such explanations for individuals to understand the lawfulness of the decision and contest it if deemed appropriate, and 2) the possible lack of concerns regarding the interest of the data subject and the interest of the bank[529].

Hereafter, I present first the results from the questionnaire and later a summary of the questionnaire and interview analysis in the form of answers to the research questions of *The Explanation Dialogues*. I refrain from presenting the analysis of both the questionnaire and the interviews that were obtained using grounded theory methodology[530] due to space and relevance constraints in regard to this thesis. I consciously decided to exclude that part of *The Explanation Dialogue* analysis from this thesis as its inclusion will also require an in-depth exposition of the project's design, technical details, evaluation methodology, and results. Such content, although of high relevance and novelty, exceeds the purpose of this chapter and so this thesis. Therefore, I limit the content included in this section to 1) the summary of the legal experts' reflections on each type of explanation -creating a straightforward and clear comparison with the structure followed in the immediately preceding section *Brief overview of eXplainability methods*- and 2) the summary of the whole project results, provided in the form of concrete answers to the research questions.

[529] The questions that lead the experts' interview script partly coincides with two of the questions asked by the Administrative Court of Vienna in the Request for Preliminary Ruling in the Dun & Bradstreet Case; 1) the concept of 'meaningful information about the logic involved' in automated decision-making and 2) the balancing the rights of the data subject against the rights and freedoms of others.

[530] Grounded theory is a qualitative research method designed to generate new theories that are not rooted in the qualitative data collected during the research process. Using this methodology, data (in *The Explanation Dialogues*, the questionnaire responses and the interview transcripts) is coded through a coding process that serves to distil and categorise data, providing a structured framework that facilitates comparisons with other segments of information.

6.3.2. Perceptions, expectations, and reasoning towards XAI explanations[531]

6.3.2.1. Feature Relevance Methods

6.3.2.1.1. SHAP (global)

For our participants, the global explanations presented did not contribute to the understanding of the explanation. Global explanations were also considered lacking in information in general. For example, one of our participants highlighted the necessity to include information regarding 'the data that train the model, [or] the final percentage [for being considered high-risk]', but special attention was made to the lack of information regarding 'reference [to the customer] case rather than a global description'. Particularly, participants highlighted the need to include 'any information pertaining to the individual', such as 'the characteristics of the single features', 'the variable weight on the final result', or 'factors that actually apply to the case'. Likewise, the graph provided through this method was considered misleading. Global explanations were described as being difficult to understand by an average consumer and not appropriate to exercise individuals' rights, e.g. 'as a user, I would need additional training to understand and interpret the explanation', 'in trying to build a case for why an individual is creditworthy [...] it is unclear to me how this explanation helped. It does not seem possible to build up a coherent argument solely on this explanation'.

Global SHAP explanations were generally considered to be unclear and non-understandable for an average individual. In particular, this type of explanation was deemed insufficient in providing enough information to data subjects for them to understand the reasons and motives behind the particular decision affecting them. Likely based on this same basis, Global SHAP explanations were not found adequate to allow data subjects to verify the lawfulness and fairness of the automated decision affecting them nor to effectively exercise their right to contest [if deemed appropriate].

[531] This section of the thesis is heavily based on the work of Bringas and State for the section *Results – Questionnaire* in State and others (n 52).

6.3.2.1.2. SHAP (local)

Local SHAP explanations were disputed in terms of their helpfulness and relevance to understand the overall decision affecting the individuals. Participants expressed their doubt about their relevance and their partially or limited usefulness, e.g. the SHAP graph seems 'somewhat useful', 'uninformative', or merely 'too difficult for me'. Some participants coincided in the overall understanding of the model itself but requested more information regarding the single features affecting it. In this regard, participants differed on the appropriateness of local SHAP's delivery format and method; a participant affirmed that SHAP explanations allow the understanding of the model, whereas two other participants perceived the explanations as 'possibly cognitively misleading'. Furthermore, the format of local SHAP explanations, concretely its graphic design and plot, was perceived as confusing and not easy to understand. For example, a participant recommended more detail in the explanation or the use of examples as 'I usually understand more written text better than graphs and schedules'. Furthermore, SHAP explanations were perceived as not entirely nor directly understandable for the average consumer since 'a plain written text instead of the plot would probably be more intelligible [for an average consumer]'. Finally, local SHAP explanations were described as partially suitable for customers to contest the decision as 'they [customer] know where to further inquire', admitting that 'they can contest the decision even if they do not understand the explanation'.

In essence, local SHAP explanations were perceived with a high level of neutrality. Participants almost equally agreed and disagreed (or strongly disagreed) on the clarity and understandability of local SHAP explanations. This type of explanation was, by a small majority, reported insufficient in providing information about the reason and motives of the automated decision, although they were equally considered to allow and disallow the data subject to effectively exercise her right to contest the automated decision. However, they were strongly perceived as inadequate in allowing data subjects to verify the lawfulness and fairness of the automated decision in regard to other sectorial laws applicable to the case.

6.3.2.2. Contrastive explanations

6.3.2.2.1. DICE

DICE explanations were considered to contribute to the understanding of the decision and were well appreciated due to their provision of information regarding the single features defining the final decision. However, participants were conflicted on whether contrastive explanations were more or less easy to understand, e.g. 'generally I can understand well the features because I have much more information regarding them but at the same time I have difficulty to understand the explanation model itself'. In concrete terms, participants agreed on the benefits of explaining in more detail what a contrastive explanation is and how the provided explanations were selected. They also acknowledged the positive side of pairing them with a narrative box. That said, participants found the actionability of DICE explanations controversial as 'the consumer is empowered to verify that their data are entered correctly', but 'non-actionable counterfactual [contrastive explanation] do not adhere to my intuition of what an explanation is. There is information there that allows the customer to gain insight into the algorithm, but not a tremendous amount'[532]. A participant further unfolded

> The given explanation seems great to make a customer happy; they are informed about how they can improve their creditworthiness, and can reapply having improved these factors. While I find it hard to imagine ways to enable an individual to assess discrimination risk in any productive way, at least this explanation didn't help.

[532] Critics on counterfactuals are not unique to their considerations as explanations. The counterpart of providing several possible counterfactuals relies on the possibility to "be able to both exclude and include desired features when multiple counterfactuals are available". Malicious actors could use this possibility to choosing the counterfactual that best suits their interest, including for instance fairwashing an unfair machine learning model. Dieter Brughmans, Lissa Melis and David Martens, 'Disagreement amongst Counterfactual Explanations: How Transparency Can Be Misleading' (2024) 32 TOP 429. See also Solon Barocas, Andrew D Selbst and Manish Raghavan, 'The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) <https://dl.acm.org/doi/10.1145/3351095.3372830> accessed 13 September 2022; Amir-Hossein Karimi, Bernhard Schölkopf and Isabel Valera, 'Algorithmic Recourse: From Counterfactual Explanations to Interventions', *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (2021); Dylan Slack and others, 'Counterfactual Explanations Can Be Manipulated' (2021) 34 Advances in neural information processing systems 62.

The participants more heterogeneously perceived DICE explanations. For the most part, DICE explanations were regarded as unclear and non-understandable for the average consumer. However, no compromise was reached regarding their adequacy to provide sufficient information for the data subjects to understand the reasons and motives of the automated decisions. Additionally, DICE explanations were, by a small majority, deemed suitable for data subjects to exercise their right to contest, even though they were heavily considered inadequate to verify the lawfulness and fairness of the decision affecting the individuals.

6.3.2.2.2. LORE

LORE explanations were perceived quite differently among our participants in terms of their relevance to the overall understanding of the decision. They were described as 'hardly helpful' but also 'fairly intelligible' and 'more clear and understandable [than the other type of explanations presented to this participant]'. All participants, nonetheless, agree on the lack of clearness regarding the appropriateness of the delivery method, particularly highlighting the lack thereof regarding how a change in the individual circumstances besides the one suggested by LORE would change the benchmark of the decision. Likewise, a change in the delivery format to improve the intelligibility of the explanation was proposed in the form of including 'text in natural language'. Participants stand out how the average consumer would have high difficulties understanding LORE explanations, e.g. 'This is too difficult for a random bank consumer'.

LORE explanations were considered both clear and understandable for the average consumer and the opposite in equal parts. However, they were considered insufficient to provide information on the motives and reasons for the automated decisions. An equal number of participant disagreed on the adequacy of LORE explanations to allow data subjects to exercise their right to contest or were hesitant about such suitability [neither agreeing nor disagreeing with the pertinent statement]. Opinions were varied regarding whether LORE explanations would allow the verification of automated decisions' lawfulness and fairness.

6.3.2.3. Significance of the decision's output with respect to explainability requirements

One factor to note in the analysis of the questionnaire's answers is that, with only a couple of exceptions, all participants agreed on the lack of significance of the decision´s outcome with respect to the amount of information that needs to be provided about the automated decision. In other words, participants were indifferent to whether the decision affected positively or negatively the individual, underlining the importance of how the model functions, not its negative or positive outcome, e.g. 'if the outcome was a true positive, I need to know the same information. It's a matter of the model, not the result. I need to know how the model works, doesn't matter strictly the outcome itself'. That said, one of our participants clarified that the significance of the decision's outcome 'would only change since contesting a positive decision seems unlikely' in the sense that, as another participant pointed out, 'in case of a positive assessment, I would be less demanding'. We could not establish whether the correctness of the decision was relevant for the participants' assessment of the significance of the decision and, thus, the significance of the decision's outcome.

6.3.2.4. Summary of all questionnaire results

The results we gathered from the questionnaire recount that for the majority of the participants, the explanations provided in our case-study were not helpful to understand the decision nor suitable to assess its lawfulness, i.e. do not help in terms of legal explainability nor justificability. Particularly, we found that our explanations were generally considered difficult to understand, and incomplete and lacking in relevant and legally required information. Explanations were also asserted to not allow individuals to effectively exercise their rights; they were not found suitable nor adequate for individuals affected by an automated decision to understand it, verify its lawfulness and fairness, and contest it if deemed appropriate. These perceptions towards the explanations we provided were made irrespective of whether the decision was a true or false negative and the positive or negative impact they had on the bank's customers.

It is also worth to mention that no concerns were shown regarding possible conflicts of interests that could arise between the bank and its customers when providing the later with XAI explanations about the automated decisions affecting them.

6.4. Assessing Post-Hoc Explanations' Respect For Legal Explainability Requirements And Desiderata[533]

6.4.1. Explanations understanding, reasoning and compliance

The explanations we presented to the legal experts were not complete in the information they displayed. Global SHAP explanations were deemed lacking as they were focused on the global description of the model logic rather than the particular case of the individual affected. To being complete, Global SHAP would need to be accompanied at least by a local explanation providing information pertinent to the individual and the specifics of the actual case. However, the local XAI methods used in the project were also portrayed incomplete to some degree. DICE explanations were good appreciated as they seemed to clarify the features deciding the particular case, but they were treat with some caution as for the participants is was not easy to understand how the XAI method works nor it seemed to offer actionable information beyond the possible change of behaviour or adjustment on the personal conditions. Local SHAP explanations might not be the best addition to its global counterpart as they were perceived too complex in their format (graphical) and missing information regarding the single features determining the decision, which was among the missing information in Global SHAP. LORE explanations, although received a positive response, coincided in the same weakness as the two other local XAI methods. They were unclear in their format and lacked information about the decisive features' benchmark and the effect of a change in circumstances for the decision.

Transposing these findings to the legal desiderata framework, I infer some conclusions. An automated decision can be explained through the sum-up of a general statement

[533] The content included in this section was in its majority extracted from the work of Bringas and State for *The explanation dialogues*, section *Discussion*. However, all mentions and connections to Chapters 3, 4 and 5 are additional content original to this thesis.

about the system type of algorithmic model and used dataset, a model-centric XAI method, and a local-centric XAI method, and still lacks normativity. Combined XAI methodology can still fail to offer the particular information the law requires. The scenario presented in *The explanation dialogue* can be situated in the scheme offered by Article 13 (2) (f) and Article 22 of the GDPR, as I assumed the personal data used in the automated decision was collected from the data subject by the bank at the moment of the credit request. Hence, the pertinent information to be provided entails at least; 'the existence of automated decision-making, including profiling', 'meaningful information about the logic involved', 'the significance and the envisaged consequences' and 'enough explanation of the decision reached after such assessment to contest the decision'. For the reflections offered by the legal experts, I argue that the minimum threshold of compliance – see Chapter 3.4.2. *The spectrum of compliance – minimum and maximum thresholds*- was not reached in our fictional scenario, nor the explanations showed an acceptable level of normativity -see Chapter 5.3.2. Legal desiderata.

Further, the XAI methods used appear to not provide enough and suitable information in terms of individuals' rights actionability. Being it either for the lack of particular information about the concrete case, for the complexity in their format and delivery method and its subsequent limited understanding for an average individual, or because they are not considered strictly explanations as far as the intent of the GDPR is concerned. On that last account, a controversial point was brought up in the questionnaire regarding whether contrastive explanations, which provide information about what features would need to change in order to get the opposite decision, shall or shall not be considered suitable explanations to contest a decision or assess its lawfulness and fairness. Hence, the legal explainability and particularly the justificability of contrastive counterfactual was put under question. Contrastive explanations -counterfactuals- seem to be focused on helping the individuals to achieve their aspirations or ambitious (e.g. obtain a credit), rather than in helping them understand the decision affecting their rights and freedoms. From the experts' responses I gather that the compliance of XAI explanations to the pertinent law -here the GDPR- strongly depends on whether the exercise of the rights associated with such

information and explanation requirements are allowed – and ensured – thanks to those same explanations. Thus, I argue that contrastive explanations might help individuals to achieve their life expectations, but they might not be enough to assure the GDPR's reasons and motives behind the rights to information and an explanation are attained – see **Chapter 4** -.

In The Explanation Dialogue, we deduced that a combination of global and local XAI methods is preferable over the use of a single XAI method, as multiple methods could provide a more complete and actionable explanation about the ADM model and the automated decision. However, besides the information straightforwardly provided by the XAI methods, a text or narrative deemed highly necessary, particularly explaining the XAI graphs and tables and addressing the motives and reasons behind the decision. We could not assert whether a specific part of the explanation was more relevant than another, but we could state that the experts' reasoning did not significantly change between a true positive or a false positive case, i.e., the correctness of the outcome did not have any special impact on the answers.

In this regard, I conclude that the information provided in *The Explanation Dialogue* through XAI methods do not comply with the legal explainability and justificability requirements established in Article 13 (2) (f) nor Article 22(3) of the GDPR -nor would do for Article 14 (2) (g) and Article 15 (1) (h)-. The technical explanations offered to the legal experts did not help them -and by extension what would be the data subject- to understand the ADM and the final automated decision reached, nor to assess the lawfulness, legality and validity of them both. In essence, the information provided failed in its purposefulness insofar as it was considered insufficient and unsuitable to ensure the assessment of the ADM and the decision, and contest them if deemed appropriate.

I find interesting that participants did not give particular relevance to the positive or negative effects of the ADM's outcome for the individual. From this, I infer that the threshold of compliance of the right to information and an explanation is independent of the outcome of that decision and the positive or negative effects it might have in the individual. In this regard, individuals affected by an automated decision might be less

demanding when a positive decision affects them, just like they could found redundant the exercise of their individuals rights. However, they could still wish to verify the lawfulness, fairness, and accuracy of such decision for what clear, complete and understandable explanations are necessary. In essence, I gather that the normativity and purposefulness of the information and explanations provided in accordance with the GDPR shall not be undermine for a possible positive or negative impact on the individuals' rights and freedoms.

The GDPR explicitly mentioned 'legal effects concerning him or her or similarly significantly affects him or her'. In Chapter 4 of this thesis, I argued that the importance of information and explanation rights in the context of data protection relies on the high impact ADM -as employed nowadays- have on our rights and freedoms and our participation in society. In that regard, the right to information and an explanation for automated individual decision-making seek to ensure that the challenges and risks provoked by these type of technologies can be overseen and the lawfulness and legality of their use assessed. In the case of the GDPR, the negative or positive effects on our live is not as relevant as the mere existence of the specific processing of ADM and the impacts and effects it can cause by putting the individual in a situation resembling Kafka's *The Trials*[534].

Accordingly to Article 12(1) of the GDPR, information about automated decisions and ADM shall be concise, transparent, intelligible and easily accessible. However, XAI methods and techniques provide – in general – quite technical and complex information about the model and the decisions reached. Thus, if technical explanations about the model and the reached decision could be difficult to understand for an expert in explainability, for an average individual their understanding could be even more challenging. For this reason, I argue that the intelligibility of the explanations and

[534] Different relevance is given to the positive or negative effects of a AI system in our life in the case of the new Artificial Intelligence Act. Article 86 of the Regulation establishes a right to an explanation "for any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system [...] and which produce legal effects or similarly significant affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights". Article 86 further clarify that this right to a "clear and meaningful explanation of the role of the AI system in the decision-making procedure and the main element of the decision taken" "shall apply only to the extent that the right to an explanation is not otherwise provided for under Union law".

information provided in The Explanation Dialogue- as referred to in *Chapter 5. 2.3.* Legal desiderata- is highly doubted. The legal experts perceived the technical explanations to be quite technically specific and lacking a clarification in natural text. From this, I infer that even if the technical explanations were rich and complex in information about the model and the particular decision, such information will get lost in translation for the data subject. The trade-off between easily understandable and sufficiently detailed was certainly inclined towards the latter, at least in what respect to technical information. The lack of intelligibility impacts on the purposefulness of the information, since low levels of understandability undermine the capability of data subjects to understand and assess the automated processing and decision.

6.4.2. Explanations' integrity and trust

Some of the explanations provided to participants were describe as possible cognitively misleading and confusing, what demonstrates a certain level of untruthfulness. At the same time, however, such explanations were seemed as an opportunity to the data controller to play fair and exploit their information duties, and centre the needs of the data subject. Furthermore, correctly designed and intended explanations were perceived to be able to prevent any potential conflict of interest between controllers and subjects, anticipating legal disagreements and complaints. In that regard, although our explanations gained some hesitation and scepticism from our participants, better constructed explanations can be positively been received with more confidence and trust.

It is interesting to acknowledge that the legal experts expressed how the explanations caused feelings of mislead and confusion. The explanations provided were not manipulated with such intention, nor were shaped in any way to offer (or not to offer) a specific information. In truth, the explanations showed to the legal experts were the visual image of what each type of XAI method was designed and intended to provide. I extract from this that XAI methods are not neutral. In essence, they are also algorithm models designed and developed with -usually- the only intention of making a black-box system understandable. In *The Explanation Dialogue* such understandability was aimed at providing information to comply with the rights to information and an explanation, in

another scenario it can be aimed at helping the developers and users to assess the performance and fairness of the ADM model. In any case, the primary intention behind an XAI method is to open-up the black-box, even if the means to do it is another model that also suffers from the black-box problem. This act creates a cycle of inscrutability, systematicity, and homogeneity that does not end in the decision-making process as such, but in the machinery developed around it. The issue is relevant because one of the main request of the legal experts was the inclusion of an explanation of the XAI method in natural language. It is a logical request if we have in mind the explanations were deemed extremely technical for an average consumer, but it also arises attention to the inclusion of another layer of human intervention in the provision of information. Whilst it is not my intention to offer an extremely negative view on this regard, it is important to recognise the -high- relevance of the truthfulness property. Using XAI methods to make an ADM models and its final decision understandable shall not excuse the manipulation of the information.

Corporate secrecy, and intellectual and industry rights can raise problems in regard to ADM's transparency, as they may impede and obstruct the possibility of providing information about the model and their decisions. Secrecy – understand as a umbrella terms –, thus, can limit the amount of information the data controller is compelled to provide, but such limitation shall be justified. The possible burden of providing information could be, nonetheless, beneficial for data controllers if they show their predisposition to stretch its own secrecy limits. For example, putting in place ethical benchmarks or transparency codes. In any case, the rights to contest and to obtain an explanation shall not be unbearably undermined on the name of secrecy or the latter would deemed unjustifiably. The General Advocate Opinion for the Dun & Bradstreet Austria GmbH may offer a straightforward solution for this situation. A Court or Data Protection Office can act as the intermediary who balance the interest of both parties and decide the exact information about the ADM model and the automated decision that shall be provided to the data subject, both to ensure the protection of any trade-secret and third party interests as well as the rights and interest of the data subjects in accordance with the pertinent provisions of the GDPR.

6.5. Discussion

In *The Explanation Dialogues* we found that the presented state-of-the-art XAI methods face both shortcomings in terms of their understandability, presented information, and suitability to exercise the rights with regard to the data subject and the controller. We also discussed issues that may arise from possibly different interests of the data controller and subject. Furthermore, the outcomes of our user study would not be sufficient to argue that any of the presented methods fully comply with the GDPR. However, we could assert that the perceived conformity of XAI explanations of an ADM model with regard to the GDPR is closely connected to how they allow individuals to exercise their rights.

We further found that while the interviewed legal experts are well informed as regards explainability, they may have some knowledge gaps regarding the technical properties of explanations.

I have to acknowledge that although *The Experts Dialogue* offers significant insights on the connections between existing state-of-the-art literature and debates on explainability, its scope was limited to the reasoning of a small group of legal experts on three concrete XAI methods. Therefore, there are plenty of opportunities for future research. That said, in our study we identified different positions towards XAI methods, ranging from sceptical stances towards optimism, mirroring the current debate in this research area.

Upon the results obtained from *The Explanation Dialogues* and my own conclusion on the matter, if I had to respond to the question posed in previous Chapter 5: How can we develop [technical] explanations about ADM systems that are compliant with the law? I would answers that we possibly could not.

On the one hand, what can be considered an explanation about the ADM model and the particular decision in technical terms -e.g., a SHAP, LORE, or DICE explanation- might not seem like it in legal terms. It could be, nonetheless, considered the *raw* material from which data controllers can develop the explanations and information to be provided to the data-subject. I lack confidence in an explanation as the ones showed in

The Explanation Dialogues to be considered compliant with the information and explanation requirements of the GDPR in real life.

On the other hand, none of the XAI methods seems to offer all the information that is required by the GDPR. It is not only a matter of what is the aimed information to be provided by it, but the difference in the scope of the information requested. For instance, LORE offers both a logical rule behind the particular decision and a counterfactual, but the method itself does not explain what are the consequences of the automated decision in the individual. This argument seems quite simple, but it hides a greater discussion behind it. Explainability from a technical perspective is limited to the technical functioning of the ADM model and the output provided by it, explainability from a legal perspective extends to the normativity of automated individual decision-making processing -see *Chapter 2* and *3*-. By failing to assert this distinction, we can failed to understand the motives and reasoning behind the rights to information and an explanation. Thus, returning to the question: Would you [user] intend for just the [technical] explanation to provide all the information required under the law? I suggest to reconsider this approach and concede that to comply with the rights to information and an explanation a further exercise of understanding the black-box problem and the aggregated risk behind the ADM -the automated individual decision-making processing- is necessary. The processing of personal data by automated means for the purpose of decision-making and the particular decision affecting an individual need to be explained and justified to the data subject and any technical XAI method is able to do that on its own just for the single reason that their aim is to explain the algorithm but not the decision-making process as a whole nor the decision.

The user study carried out in *The Explanation Dialogues* was not designed to uncover the exact kind of information an explanation about an ADM system should have to pass the minimum threshold of compliance of the GDPR -see Section 3.4.2 The spectrum of compliance – minimum and maximum thresholds. *The Explanation Dialogues*, in that sense, was envisaged to showcase the attitudes, behaviours, and perspectives of legal experts towards XAI explanations. Knowledge that could be framed as 1) logical rules to be used in approached to technical explainability and interpretability and/or 2) legal

requirements for explanations and information about ADM systems as referred to in the GDPR.

Referring to Chapter's 3 **Discussion**, the minimum threshold of compliance of the rights to information and an explanation would require sufficiently detailed explanations of the method used to calculate the score and the reasons for a certain result. Data subjects would need to be offered general information, notably on factors taken into account for the decision-making process and on their respective weight on an aggregate level. To meet the minimum threshold of compliance, these rights would need to enable data subjects to ensure that the processing of their personal data was lawful, fair, accurate and transparent and that the personal data processed was correct. In other words, if data subjects are not provided with enough information and explanations about the decision-making process and the particular decision to exercise their rights if deemed necessary and to confirm the normativity of the process and decision affecting them, it could not be said that data controllers comply with their duties of information and explainability.

However, the exact information required to meet the minimum threshold of compliance may vary depending on the rights data subjects intend to exercise and the decision they want to contest. To my understanding, there is not a concrete kind of technical explanations and information that need to be provided, but a set of multiple possibilities to be determined in a case-by-case basis. Therefore, while to my understanding there is no one-size-fits-all type of solution, the following table is a good starting point of the kind of the content that can be disclosed when providing information and explanations compliant with the GDPR.

Minimum Threshold	- Factors taken into account for the decision-making process, respective weight on an aggregate level. - Intentions and logic behind the automated processing. - Description of method and the rules used to calculate the score - Characteristics of the data subject that were used as the main criteria to reach the decision. - Specific reasons for a certain result (particular decision).

	-Features and inferences particular to the decision.
Extensive approach	-Necessary changes in the input data to make the decision different. - Family model, - Parameters used in the setup, input, and output, - Classifications used and predicted during the training, - Rate of success on specific salient subcategories of data, - Variable importance score, - Model's fairness and accuracy metrics, - Model's performance during its learning and testing, - Decisions and changes made during its design and development - How the model was tested, trained and screened. - Who has also received a similar treatment to the subject. - Which was the erroneous and misclassification rate along with different groups and individuals during training.

Chapter 7: Conclusion

7.1. Final discussion

In this thesis, I have explored the obligations of information and explanation established in the General Data Protection Regulation for automated individual decision-making processing used in high-consequence decisions of our everyday. The question that led my thesis was: Can the right to information and to explanation applicable to automated decisions affecting individuals adequately address the problems arising from their inscrutability and lack of neutrality? After the carried outanalysis, I would dare to answer that they can succeed in doing so if the right circumstances are met.

However, in order to go into detail in my answer, I first need to clarify that, in my understanding, there is a right to information and a right to an explanation for automated individual decision-making processing. A data subject has a right to access to personal data in accordance with Articles 13(2)(f), 14(2)(g), and 15(1)(h) that grant, in turn, a grant to information about the existence of automated individual decision-making processing, meaningful information about the logic involved in the processing, and the significance and the envisaged consequence of such processing for him or her. Article 22 (3) grants the data subject three concrete safeguards towards automated individual decision-making processing: the right to obtain a human intervention, to express their own point of view and to contest the automated decision. From these safeguards, and particularly from the right to contest, the data subjects are entitled to receive an explanation of the decision that allows them to exercise their rights. To effectively do so, the data subjects need to understand the decision affecting them and assess the lawfulness, fairness, and accuracy of the decision-making process and the decision itself.

The rights to information and an explanation are distinct, but not necessarily completely independent. As I argued in *How should an explanation be?* An explanation of an automated individual decision-making process includes information about the connection between the inputs and the final decision and the intention and objectives that motivated the decision. Hence, an explanation of automated decision-making includes information about the logic followed in decision-making processing and its

significance and consequences for the individual. Although considered an explanation, that information is closer to the GDPR wording of the right to information than the right to an explanation. In turn, a justification of an automated individual decision-making process encompasses the provision of enough information to demonstrate the legality and lawfulness of the automated processing and decision in accordance with the pertinent law, here the GDPR. Although it could seem that this is distinct from an explanation, what it is, to assess whether the automated processing and decision, for example, are compliant with the principles and values established in the GDPR for the processing of personal data, -as referred to in Article 5 and 6-, a data subject will also need to understand the logic of the decision-making process and the motives and consequences of the decision-making process.

In essence, justifying and explaining an automated individual decision-making process has originally different goals, but the information provided can coincide. In this sense, the right to information and the right to an explanation may differ not exactly in the goal they want to achieve -justify or explain- but in the moment they occur and to what they refer. The right to information affects the automated individual decision-making processing as a whole. It shall be used to ensure the individual knows an automated decision-making process can take place or has already happened. By contrast, the right to an explanation activates when an automated decision has already been made and has affected the rights, freedoms, and legitimate interests of the data subject. This does not mean that the right to an explanation only involves the automated decision, because to challenge a decision, a person also needs to understand the process through which it has been made.

These claims serve to answer some sub-questions of this thesis: What objectives are to be achieved, and what is exactly to be provided through those explanations and information? Meanwhile, the last argument provided above brings me back to the research sub-questions: Why does society ask for explanations and information about automated individual decision-making processing? Moreover, what is the legal rationale behind the rights to information and an explanation about automated decisions?

To my understanding, automated individual decision-making processing demands specific information and explanation obligations because the technologies used in the processing of personal data for such purpose pose significant risks and threats to the rights, freedoms, and legitimate interests of the individuals. It is not merely the fact that this type of processing is being increasingly used to decide the allocation of private and public products and services indispensable for the enjoyment of our lives, and it would be understandable to introduce such types of rules or safeguards to ensure that they comply with our society values and principles. It is also that the technologies -algorithms- used to automate the process naturally have a set of characteristics, i.e., inscrutability and lack of neutrality, that when introduced in a decision-making process could transform it into something pernicious, dehumanising, arbitrary, and systematic There is an inherent challenge in the processing of our personal data. Still, the reason automated individual decision-making demands specific information and explanation obligations is that this type of processing can put the individual in a situation of vulnerability and defencelessness that concur with the metaphor found in Kafka's *The Trials*. The rights to information and an explanation act as safeguards and measures against the risks that automated individual decision-making processes intrinsically have and can potentially create. There might be no better way to act against that vulnerability and defencelessness than granting individuals rights that resemble due process safeguards and that, traditionally, aim to ensure the lawfulness, fairness, and accuracy of the process and decision affecting individuals.

Importantly, the rights to information and an explanation concern the processing of personal data with the aim of automating a decision-making process. To my understanding, they are not rights to receive information and explanation about the technology used in the processing. Undoubtedly, to explain and justify the processing that has been done using algorithms, it is necessary to understand the inner workings of the algorithm and assess whether the algorithm has processed the personal information in a lawful, accurate, and fair manner. Consequently, the type of algorithm used to process the individual's personal data determines how easily or difficult is to understand its inner workings and so provide information about it as a necessary element of the processing process. These distinctions are relevant to answer the last

research sub-question: What are the potential development of the exercise of the rights to information and an explanation taking into account the state-of-the-art of post-hoc explainability methods?

When addressing the risks and challenges of automated individual decision-making processing, I made no distinction between interpretable and non-interpretable algorithms. The mere use of algorithms for such a process introduces the algorithm's black-box problem, i.e., its normative characteristics of complexity and lack of neutrality. However, when considering the understanding of the data processing and the provision of information, there is assuredly a difference between technical black and white boxes. The case study offered in *The Explanation Dialogues* shows that the technical methods designed and used to open up the black box can be useful in an attempt to understand the functioning and logic of algorithms technically. Still, they do not attain by themselves the objectives behind the rights to information and an explanation. These methods can help data controllers to understand the technologies they use in the automated individual decision-making process, but they are not suitable for explaining and justifying the processing and the final decision to the data subjects. In my opinion, explainability methods are tools that, used correctly, can help to understand how algorithms work. However, the technical information they provide needs to be adjusted and complemented with more information about the actual decision-making process affecting the individuals. Explainability methods do not allow individuals to assert their rights in accordance with the GDPR, at least not if they are intended to be used in isolation and without any other intervention from the data controller.

Returning to the principal question of this thesis, I argue that the right to information and to explanation applicable to automated decisions affecting individuals can adequately address the problems arising from their inscrutability and lack of neutrality if they address the automated individual decision-making process in its entirety. The rights to information and an explanation shall need to be understood as safeguards to the aggregated risks posed by these types of processing, not uniquely posed by algorithm models. The model is an intrinsic part of the processing, but it is not everything that the decision-making process implies.

7.2. Limitations and future research

The claims presented in this thesis can be affected by the Judgment of the European Court of Justice in the Dun & Bradstreet Austria GmBH Case. Although I consider my arguments to be coherent with the existent jurisprudence and literature, the Court can interpret the wording of Article 15(1)(h) in regard to 'meaningful information about the logic involved' in a manner that differs from the reasoning followed in this thesis. This circumstance stresses the significance and relevance of this thesis insofar as my research has been developed in parallel with the Courts and Data Protection Offices resolving issues related to the interpretation and exercise of the rights to information and explanation. Just as it demonstrates the originality of my thesis, it also highlights one of its limitations, which is that my interpretation of the motives and objectives of both rights does not coincide exactly with the court's interpretation. This would mean that my interpretation of what their exercise would entail could also be different. Even so, the analyses proposed in this thesis coherently support the conclusions I have reached.

It is worth noting that the case study carried out in *The Explanation Dialogues* and Chapter 6 is limited. Future research can benefit from a bigger pool of participants as well as from a more varied range of backgrounds. An interesting path of further research can be found in carrying out a group study with average individuals as participants instead of experts on legal explainability. Very interesting and useful results could be obtained if individuals who could well be real people affected by an automated decision were confronted with different types of explanations about the automated individual decision-making process obtained through the use of technical methods. Likewise, new explainability methods are constantly being presented, so carrying out a case study similar to the one we did but with new methods could shed more light on the area, regardless of what type of participant is involved in it. Perhaps the most promising research is one that presents information and explanations about decisions that are not only technical but follow the arguments presented in this thesis. In this sense, a single

type of comprehensive information and explanation may not be enough to ensure that the rights and safeguards of individuals are respected.

It would be very interesting to develop and investigate in more detail how the two rights are exercised in different contexts and affect different rights, freedoms and interests of the individual. In this regard, this thesis offers a high-level analysis of the rights to information and an explanation for automated individual decision-making processing in high-consequence contexts. Despite being an advantage insofar as the arguments presented can be applied more generally in different contexts and at different levels, it also has a limitation in that it does not take into account the particularities that may arise in specific situations. For the individual, it may not be the same as an automated decision affecting them financially as affecting their health. I have sought to emphasise this fact by offering a review of existing case law and by demonstrating that, depending on the rights and interests affected by the automated decision-making processing, individuals will be able to assess the lawfulness, fairness, and accuracy of the process and challenge the decision on the basis of one reason or another. Future research can be done with a more limited approach to the GDPR or with national and sectorial laws in mind.

Inevitably, the subject of this thesis has been affected by the recent enactment of the Artificial Intelligence Act. Not only is there a right to an explanation in the GDPR concerning automated individual decision-making processing, but now we also have a right to an explanation for decisions which were taken on the basis of the output from a high-risk artificial intelligence system. Whereas the GDPR's right to explanation covers automated decisions with legal or similarly significant effects for the data subject, the Artificial Intelligence Act's right to explanation is limited to decisions that, to the opinion of the individual, negatively impact on their health, safety and fundamental rights. Be that as it may, the existence of both rights demonstrates how important it is that the decision-making processes, which are impacted by the use of algorithms, are understandable to the individual and that the individual is not left defenceless at moments of great importance for their personal development.

Given that throughout this thesis I have emphasised the importance of the decision-making process and the effects and impacts algorithms have on it, rather than focusing solely on the technology itself, the conclusions drawn in this thesis may serve to shed light on the future interpretation of the Artificial Intelligence Act's right to an explanation. All this is without denying that specific research can be carried out in the context of artificial intelligence systems instead of data processing. After all, the algorithmic model is not the same as the artificial intelligence system, and therefore, the risks and impacts that these introduce into decision-making systems are also different.

Glossary of Terms

Algorithms – the concrete steps and processes a computer needs to follow or employ to complete a task or solve a problem.

Artificial Intelligence (AI) - The technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision-making, creativity and autonomy.

Automated decision – a decision made only by technological means.

Automated decision-making processing (ADM) – the process where decisions are made by automated means without a meaningful human involvement, based on the processing of factual data or profiling.

Black box – algorithms that are very difficult to interpret, even for human experts in functional domains, and require post-hoc explanation methods to achieve some level of understanding of the functioning of the system.

Contrastive explanations – explainability method that highlights the difference between two or more outputs (predictions) of a model.

Explainability - the actions taken to make the inner workings of non-interpretable systems clear to humans in a manner that allows them to comprehend and literally explain the mechanisms that drive the learning of the model.

Feature relevance explainability method – explainability method that provides a measure of how relevant a feature is to the decision outcome.

Global post-hoc explanability method – explainability method that aims at the overall logic of a black-box model'

Interpretability - the ability of an algorithmic model to describe [explain or present] the internals of a system in a way that is understandable to humans.

Local post-hoc explanability method – explainability method that aim at the reasons for the decision of a black-box model for a specific instance

Machine Learning (ML) – algorithms designed to discover correlations and seek patterns through statistical inferences, measurements and analytics that would otherwise be difficult to identify.

Model-agnostic explainability method – explainability method designed to offer insights into the function of the underlying model regardless of the type of model involved.

Model-centric explanations – explanations that provide broad information about a model.

Model-specific explainability method – explainability method designed based on the specific model architecture and structure.

Post-hoc explanability method – method designed to explain the functioning of a non-interpretable model after it has been trained.

Profiling – the analysis of the aspects of an individual's personality, behaviour, interests and habits to make predictions or decisions about them.

SHapley Additive exPlanations (SHAP)

Subject-centric explanations – explanations that provide information about the basis of a concrete input record.

White box – algorithms that are interpretable to humans.

List of References

Table of Cases

Judgments of the EU

Dun & Brdstreet Austria GmbH - Request for preliminary ruling from the Verwaltungsgericht Wien (C-203/22) [2022] ECJ OJ C 222, 7.6.2022

Dun & Bradstreet Austria GmbH - Request for a preliminary ruling from the Verwaltungsgericht Wien - Opinion of Advocate General Richard de la Tour (C-203-22) [2024] ECJ ECLI:EU:C:2024:745

Österreichische Datenschutzbehörde and CRIF - Request for a preliminary ruling from the Bundesverwaltungsgericht (C-487/21) ECJ EU:C:2023:369

TNT Express Nederland BV v AXA Versicherung AG Reference for a preliminary ruling: Hoge Raad der Nederlanden (C-503/08) [2010] ECJ ECLI:EU:C:2010:243

SCHUFA Holding (Scoring) - Judgement of the Court (C-634/21)[2023] ECJ OJ C, C/2024/913

SCHUFA Holding and Others (Scoring) - Opinion of Advocate General Pikamäe (C-634/21) [2023] ECJ ECLI:EU:C:2023:220

SCHUFA Holding (Scoring) - Request for Preliminary Ruling from the Verwaltungsgericht Wiesbaden (C-634/21) [2023] ECJ OJ C, C/2024/913

Judgements of other Jurisdictions

Auto 72/2021 [2021] Audiencia Provincial Penal de Barcelona Seccion 9 Rec 840/2021 ECLI: ES:APB:2021:1448A

Beslut efter tillsyn enligt dataskyddsförordningen - Klarna Bank AB [2022] Integritetsskyddsmyndigheten DI-2019-4062

C/13/692003/ HA/RK 20-302 [2021] Rechtbank Amsterdam ECLI:NL:RBAMS:2021:1018

C/13/742407 / HA RK 23-366 [2024] Rechtbank Amsterdam ECLI:NL:RBAMS:2024:4019

Case 8937120 CV EXPL 20-22882 [2021] Rechtbank Amsterdam ECLI:NL:RBAMS:2021:5029

Case C/13/689705/HA RK 20-258 [2021] Rechtbank Amsterdam (Amsterdam District Court) ECLI:NL:RBAMS:2021:1019

Cass sez lav n 1663/2020 (Foodora) (Corte Suprema Di Cassazione, Sezione Lavoro)

Cass sez lav n 2949/2019 (Deliveroo Italia SRL) (Tribunale Ordinario di Bologna Sezione Lavoro)

Civile Ord Sez 1 Num 14381 (Corte Suprema di Cassazione)

Computer say no - BInBDI ("Berliner Beauftragte für Datenschutz und Informationsfreiheit)

Ordinanza ingiunzione nei confronti di Deliveroo Italy s.r.l 9685994 (Il Garante per la Protezione dei Dati Personali)

Ordinanza ingiunzione nei confronti di Foodinho s.r.l n 9675440 (Il Garante per la Protezione dei Dati Personali)

Rb Amsterdam - C / 13 / 742407 / HA RK 23-366 [2024] Rechtbank Amsterdam (Amsterdam Court) ECLI:NL:RBAMS:2024:4019

Sentenza Foodora - n 778 (Tribunal di Torino)

Sentenza n 3570/2020, causa civile n 7283/2020 (Tribunal di Palermo Sezione Lavoro)

STS 2924/2020 [2020] Tribunal Supremo, Sala de lo Social 805/2020 ECLI: ES:TS:2020:2924

Tilladelse til behandling af biometriske data ved brug af automatisk ansigtsgenkendelse ved indgange på Brøndby Stadion (Datatilsynet meddeler)

Vinnsla Creditinfo Lánstrausts - 2020010592 (Persónuvernd)

Legislation

Charter of Fundamental Rights of the European Union [2012]

Consolidated Version of The Treaty on the Functioning of the European Union (TFUE) [2016] OJ C202/1

Council Directive Proposal 19990/0287/COD Concerning the Protection of Individuals in Relation to the Processing of Personal Data [1990] COM(90) 314 final – SYN 287 90/C 277/03, 29

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 (Directive on the Protection of Trade Secrets) [2016] OJ L157/1

Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 Amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as Regards the Better Enforcement and Modernisation of Union Consumer Protection Rules. (Modernisation Directive) [2019] OJ L 328/7

European Commission Proposal for a Regulation 2012/0011 of 25 January 2012 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (European Commission Proposal GDPR) [2021] 2012/0011 (COD)

European Parliament and Council Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data 1995 (Data Protection Directive) [1995] (OJ L281/31)

European Parliament and Council Legislative Resolution 2012/0011 of 12 March 2014 on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (European Parliament and Council Proposal GDPR) [2012] (COM(2012)0011)

Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 2017 (A29WP Guidelines on ADMs and Profiling) [2017] WP251rev01

Ley 12/2021, de 28 de septiembre, por la que se modifica el texto refundido de la Ley del Estatuto de los Trabajadores, aprobado por el Real Decreto Legislativo 2/2015, de 23 de octubre, para garantizar los derechos laborales de las personas dedicadas al reparto en el ámbito de plataformas digitales (Ley Rider) [2021] BOE-A-2021-15767

Opinion 1/98 Platform for Privacy Preferences (P3P) and the Open Profiling Standard (OPS) [1998]

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC 2016 (General Data Protection) [2016] OJ L 1191/1

Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services (Online Intermediary Service Regulation) [2019] OJ L186/57

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU)2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU)2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L, EU/2024/1689

Bibliography Secondary Sources

Aarnio A, *The Rational as Reasonable: A Treatise on Legal Justification*, vol 4 (Springer Science & Business Media 1986)

Adadi A and Berrada M, 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)' (2018) 6 IEEE Access 52138-52160 <ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8466590> accessed 23 January 2025

Adams-Prassl J and others, 'Regulating Algorithmic Management: A Blueprint' [2023] 14 European Labour Law Journal (forthcoming) <ssrn.com/abstract=4373355> accessed 23 January 2025

Agarwal A, Singhal C and Thomas R, 'AI-Powered Decision Making for the Bank of the Future' [2021] McKinsey & Company.–2021.–March.–URL: mckinsey. com/~/media/mckinsey/industries/financial% 20services/our% 20insights/ai% 20powered% 20decision% 20making% 20for% 20the% 20bank% 20of% 20the% 20future/ai-powered-decision-making-forthe-bank-of-the-future. pdf (дата обращения 15.04. 2021)

Alasadi SA and Bhaya WS, 'Review of Data Preprocessing Techniques in Data Mining' (2017) 12 Journal of Engineering and Applied Sciences 4102 ISSN:1816-949X

'Algorithm, n. Meanings, Etymology and More | Oxford English Dictionary' (*Oxford English Dictionary*) <oed.com/dictionary/algorithm_n> accessed 12 January 2025

'Algorithmic Language' (*Oxford Reference*) <oxfordreference.com/display/10.1093/oi/authority.20110803095402323> accessed 5 December 2023

Allen Qc RA and Masters D, 'Technology Managing People – the Legal Implications' (AI Law Consultancy) <tuc.org.uk/sites/default/files/Technology_Managing_People_2021_Report_AW_0.pdf> accessed 23 January 2025

Almada M, 'Automated Decision-Making as a Data Protection Issue' [2021] SSRN Electronic Journal <ssrn.com/abstract=3817472> accessed 28 November 2024

Alvarez M and Way J, 'Reasons for Action: Justification, Motivation, Explanation.', *The Stanford Encyclopedia of Philosophy* (edited by Edward N Zalta&Uri Nodelman, Metaphysics Research Lab, Stanford University 2024) <plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/.> accessed 23 January 2025

Amann J and others, 'Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective' (2020) 20 BMC medical informatics and decision making 1 <doi.org/10.1186/s12911-020-01332-6> accessed 23 January 2025

Ananny M and Crawford K, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973 <doi.org/10.1177/1461444816676645>accessed 23 January 2025

Anderson LR and J, 'Theme 7: The Need Grows for Algorithmic Literacy, Transparency and Oversight' (*Pew Research Center: Internet, Science & Tech*, 8 February 2017) <pewresearch.org/internet/2017/02/08/theme-7-the-need-grows-for-algorithmic-literacy-transparency-and-oversight/> accessed 12 December 2023

Aneesh A, *Technologically Coded Authority: The Post-Industrial Decline in Bureaucratic Hierarchies* (2002) <web.stanford.edu/class/sts175/NewFiles/Algocratic%20Governance.pdf> accessed 23 January 2025

Angwin J and others, 'Machine Bias - There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.' (*ProPublica*, 23 May 2016) <propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Aristotle, *The Politics* (Sinclair T and Saunders TJ (trans), Penguin UK 1981)

——, *Aristotle's Art of Rhetoric* (Robert C Barlett ed, University of Chicago Press 2019)

Article 29 Data Protection Working Party, 'Guidelines on Transparency under Regulation 2016/679' (2017)

Astromskė K, Peičius E and Astromskis P, 'Ethical and Legal Challenges of Informed Consent Applying Artificial Intelligence in Medical Diagnostic Consultations' (2021) 36 AI & SOCIETY 509 <doi-org.soton.idm.oclc.org/10.1007/s00146-020-01008-9> accessed 23 January 2025

Ayre LB and Craner J, 'The Baked-in Bias of Algorithms' (2028) 10 Collaborative Librarianship <digitalcommons.du.edu/collaborativelibrarianship/vol10/iss2/3> accessed 23 January 2025

Aziz LA-R and Andriansyah Y, 'The Role Artificial Intelligence in Modern Banking: An Exploration of AI-Driven Approaches for Enhanced Fraud Prevention, Risk Management, and Regulatory Compliance' (2023) 6 Reviews of Contemporary Business Analytics 110 <researchberg.com/index.php/rcba/article/view/153> accessed 23 January 2025

Balagopalan A and others, 'The Road to Explainability Is Paved with Bias: Measuring the Fairness of Explanations', *FAccT* (ACM 2022) <doi.org/10.1145/3531146.3533179> accessed 23 January 2025

Baldi P and others, 'Assessing the Accuracy of Prediction Algorithms for Classification: An Overview' (2000) 16 Bioinformatics 412 <doi.org/10.1093/bioinformatics/16.5.412> accessed 23 January 2025

Barocas S, Hardt M and Narayanan A, 'When Is Automated Decision Making Legitimate?', *Fairness and Machine Learning - Limitations and Opportunities* (MIT Press 2023) <fairmlbook.org> accessed 23 January 2025

——, 'Fairness and Machine Learning. Limitations and Opportunities. When Is Automated Decision Making Legitimate?' (*Fairmlbook*, 13 December 2023) <fairmlbook.org/legitimacy.html> accessed 23 January 2025

Barocas S and Selbst AD, 'Big Data's Disparate Impact' [2016] SSRN Electronic Journal <ssrn.com/abstract=2477899> accessed 20 May 2022

Barocas S, Selbst AD and Raghavan M, 'The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) <dl.acm.org/doi/10.1145/3351095.3372830> accessed 13 September 2022

Barredo Arrieta A and others, 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI' (2020) 58 Information Fusion 82 <doi.org/10.1016/j.inffus.2019.12.012> accessed 23 January 2025

Barros Vale S and Zanfir-Fortuna G, 'Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities' (Future of Privacy Forum 2022) <fpf.org/blog/fpf-report-automated-decision-making-under-the-gdpr-a-comprehensive-case-law-analysis/> accessed 23 January 2025

Bayamlıoğlu E, 'The Right to Contest Automated Decisions under the General Data Protection Regulation: Beyond the So-called "Right to Explanation' (2022) 16 Regulation & Governance 1058 <dx.doi.org/10.1111/rego.12391> accessed 23 January 2025

Becerra SD, 'The Rise of Artificial Intelligence in the Legal Field: Where We Are and Where We Are Going' (2018) 11 Journal of Business, Entrepreneurship and the Law 27 <digitalcommons.pepperdine.edu/jbel/vol11/iss1/2> accessed 23 January 2025

Bharadiya JP, Thomas RK and Ahmed F, 'Rise of Artificial Intelligence in Business and Industry' (2023) 25 Journal of Engineering Research and Reports 85 <doi.org/10.9734/jerr/2023/v25i3893> accessed 23 January 2025

Bibal A and others, 'Legal Requirements on Explainability in Machine Learning' (2021) 29 Artificial Intelligence and Law 149 <doi.org/10.1007/s10506-020-09270-4> accessed 23 January 2025

Binns R and Veale M, 'Is That Your Final Decision? Multi-Stage Profiling, Selective Effects, and Article 22 of the GDPR' (2021) 00 International Data Privacy Law <http://dx.doi.org/10.1093/idpl/ipab020> accessed 23 January 2025

Blume P, 'Data Protection and Privacy – Basic Concepts in a Changing World' Scandinavian Studies In Law <scandinavianlaw.se/pdf/56-7.pdf> accessed 23 January 2025

Bodria F and others, 'Benchmarking and Survey of Explanation Methods for Black Box Models' (arXiv, 25 February 2021) <http://arxiv.org/abs/2102.13076> accessed 13 September 2022

Bohr A and Memarzadeh K, 'The Rise of Artificial Intelligence in Healthcare Applications' [2020] Artificial Intelligence in Healthcare 25 <doi.org/10.1016/B978-0-12-818438-7.00002-2> accessed 23 January 2025

Bringas Colmenarejo A and others, 'The Explanation Dialogues: Understanding How Legal Experts Reason About XAI Methods', *European Workshop on Algorithmic Fairness: Proceedings of the 2nd European Workshop on Algorithmic Fairness* (2023)

Bringas Colmenarejo A, State L and Comandé G, 'How Should an Explanation Be? A Mapping of Technical and Legal Desiderata of Explanations for Machine Learning Models' (2025) International Review Law, Computers and Technology <doi-org.soton.idm.oclc.org/10.1080/13600869.2025.2497633>

Brkan M, 'AI-Supported Decision-Making under the General Data Protection Regulation', *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law* (ACM 2017) <dl.acm.org/doi/10.1145/3086512.3086513> accessed 4 January 2025

Brożek B and others, 'The Black Box Problem Revisited. Real and Imaginary Challenges for Automated Legal Decision Making' [2023] Artificial Intelligence and Law <link.springer.com/10.1007/s10506-023-09356-9> accessed 30 November 2023

Brughmans D, Melis L and Martens D, 'Disagreement amongst Counterfactual Explanations: How Transparency Can Be Misleading' (2024) 32 TOP 429 <doi.org/10.1007/s11750-024-00670-2> accessed 23 January 2025

Buhmann A, Paßmann J and Fieseler C, 'Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse' (2020) 163 Journal of Business Ethics 265 <link.springer.com/article/10.1007%2Fs10551-019-04226-4> accessed 23 January 2025

Burgess P, 'Googling the Equivalence of Private Arbitrary Power and State Arbitrary Power: Why the Rule of Law Does Not Relate to Private Relationships' (2021) 17 International Journal of Law in Context 154 <doi.org/10.1017/S1744552321000100> accessed 23 January 2025

Burrell J, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society 2053951715622512 <doi.org/10.1177/2053951715622512> accessed 23 January 2025

Bygrave LA, 'AUTOMATED PROFILING' (2001) 17 Computer Law & Security Review 17 <doi.org/10.1016/S0267-3649(01)00104-2> accessed 23 January 2025

Bygrave LA, 'Article 22. Automated Individual Decision-Maing, Including Profiling' in Christopher Kuner, Lee A Bygrave and Christopher Docksey (eds), *The EU General Data Protection Regulation (GDPR)* (Oxford University PressNew York 2020) <academic.oup.com/book/41324/chapter/352297995> accessed 4 January 2025

Castets-Renard C, 'Accountability of Algorithms in the GDPR and Beyond: A European Legal Framework on Automated Decision- Making' [2019] 30 Fordham Intell. Prop. Media & Ent. L 91. <ir.lawnet.fordham.edu/iplj/vol30/iss1/3> accessed 23 January 2025

'Category: Article 22 GDPR' (*GDPRhub*) <gdprhub.eu/index.php?title=Category:Article_22_GDPR> accessed 23 January 2025

Cerioli A and others, 'Newcomb–Benford Law and the Detection of Frauds in International Trade' (2019) 116 Proceedings of the National Academy of Sciences 106 <doi.org/10.1073/pnas.1806617115> accessed 23 January 2025

Chander S, 'Recommendations for a Fundamental Rights-Based Artificial Intelligence Regulation - Addressing Collective Harms, Democratic Oversight and Impermissable Use.' (European Digital Rights EDRI 2020) <edri.org/wp-content/uploads/2020/06/AI_EDRiRecommendations.pdf> accessed 23 January 2025

Chen Z and others, 'What Makes a Good Explanation?: A Harmonized View of Properties of Explanations', *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022 <doi.org/10.48550/arXiv.2211.05667>* accessed 23 January 2025

Citron DK, 'Technological Due Process' 85 <openscholarship.wustl.edu/law_lawreview/vol85/iss6/2> accessed 23 January 2025

Clarke R, 'Information Technology and Dataveillance' (1988) 31 Communications of the ACM 498 <rogerclarke.com/DV/CACM88.html> accessed 23 January 2025

Collings JC, 'Using Excel and Benford's Law to Detect Fraud' (*Journal of Accountability*, 1 April 2017) <journalofaccountancy.com/issues/2017/apr/excel-and-benfords-law-to-detect-fraud.html> accessed 19 January 2025

Crawford K, 'Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms' 55 <Ssrn.Com/Abstract=2325784> Accessed 23 January 2025

Creel K and Hellman D, 'The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems' (2022) 52 Canadian Journal of Philosophy 26 http://doi.org/10.1017/can.2022.3 accessed 23 January 2025

Dalla Corte L, 'A Right to a Rule: On the Substance and Essence of the Fundamental Right to Personal Data Protection' in Dara Hallinan and others (eds), *Data protection and privacy* (Hart Publishing 2020) ISBN: 9781509932740

De Hert P and Gutwirth S, 'Data Protection in the Case Law of Strasbourg and Luxemburg: Constitutionalisation in Action' in Serge Gutwirth and others (eds), *Reinventing Data Protection?* (Springer Netherlands 2009) <http://link.springer.com/10.1007/978-1-4020-9498-9> accessed 28 November 2024

'Decision Automation' (*DevX*) <devx.com/terms/decision-automation/> accessed 10 December 2023

'Decision Management Software & Solutions | IBM' <ibm.com/decision-management> accessed 10 December 2023

Department for Business, Energy & Industrial Strategy, 'The Characteristics of Those in the Gig Economy' (2018) <assets.publishing.service.gov.uk/media/5aa69800e5274a3e391e38fa/The_characteristics_of_those_in_the_gig_economy.pdf> accessed 23 January 2025

Dessain J, Bentaleb N and Vinas F, 'Cost of Explainability in AI: An Example with Credit Scoring Models' in Luca Longo (ed), *Explainable Artificial Intelligence*, vol 1901 (Springer Nature Switzerland 2023) <link.springer.com/10.1007/978-3-031-44064-9_26> accessed 13 December 2023

Di Bello M, 'Algorithmic Fairness – ProPublica v. Northpointe' (*Marcello Di Bello*, Fall 2021) <marcellodibello.com/algorithmicfairness/handout/ProPublica-Northpointe.html> accessed 28 November 2024

Diakopoulos N, 'Accountability in Algorithmic Decision Making' (2016) 59 Communications of the ACM 56 <doi.org/10.1145/2844110> accessed 23 January 2025

Dicey AV, *Introduction to the Study of the Law of the Constitution* (10th edn, Palgrave Macmillan 1885)

Dinant J-M and others, 'Application of Convention 108 to the Profiling Mechanism' 35 <rm.coe.int/16806840b9> accessed 23 January 2025

Dordevic M, 'Council Post: How Artificial Intelligence Can Improve Organizational Decision Making' (*Forbes*) <forbes.com/sites/forbestechcouncil/2022/08/23/how-artificial-intelligence-can-improve-organizational-decision-making/> accessed 10 December 2023

Doshi-Velez F and Kim B, 'Towards A Rigorous Science of Interpretable Machine Learning' <arxiv.org/pdf/1702.08608> accessed 23 January 2025

Edwards L and Veale M, '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 <scholarship.law.duke.edu/dltr/vol16/iss1/2> accessed 23 January 2025

Elgendy N and Elragal A, 'Big Data Analytics in Support of the Decision Making Process' (2016) 100 Procedia Computer Science 1071 <doi.org/10.1016/j.procs.2016.09.251> accessed 23 January 2025

Eubanks V, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press 2018)

European Commission, 'Antitrust: Commission Fines Google €2.42 Billion for Abusing Dominance as Search Engine by Giving Illegal Advantage to Own Comparison Shopping Service - Factsheet' (*European Commission - Press Corner*, 27 June 2017) <ec.europa.eu/commission/presscorner/detail/es/memo_17_1785> accessed 23 January 2025

——, 'Commission Fines Meta €797.72 Million over Abusive Practices Benefitting Facebook Marketplace' (*European Commission - Press Release*, 14 November 2024) <ec.europa.eu/commission/presscorner/detail/en/ip_24_5801> accessed 23 January 2025

European Parliament. Directorate General for Parliamentary Research Services., *Understanding Algorithmic Decision-Making: Opportunities and Challenges*. (Publications Office 2019) <data.europa.eu/doi/10.2861/536131> accessed 28 November 2024

Fayyad U, Piatetsky-Shapiro G and Smyth P, 'From Data Mining to Knoledge Discovery in Databases' (1996) 17 AI Magazine <ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230> accessed 23 January 2025

Federal Trade Commission, 'FTC Sues Amazon for Illegally Maintaining Monopoly Power' (*Federal Trade Commission*, 26 September 2023) <ftc.gov/news-events/news/press-releases/2023/09/ftc-sues-amazon-illegally-maintaining-monopoly-power> accessed 23 January 2025

Ferrer X and others, 'Bias and Discrimination in AI: A Cross-Disciplinary Perspective' (2021) 40 IEEE Technology and Society Magazine 72 <doi.org/10.1109/MTS.2021.3056293> accessed 23 January 2025

Floridi L and others, 'AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 Minds and Machines 689 <doi.org/10.1007/s11023-018-9482-5> accessed 23 January 2025

Future of Privacy Forum, 'Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making' (2017) <fpf.org/blog/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/> accessed 23 January 2025

Gellert R, 'We Have Always Managed Risks in Data Protection Law': (2016) 2 European Data Protection Law Review 481 <doi.org/10.21552/EDPL/2016/4/7> accessed 23 January 2025

Ghahramani Z, 'Probabilistic Machine Learning and Artificial Intelligence' (2015) 521 Nature 452 <doi.org/10.1038/nature14541> accessed 23 January 2025

Gillespie T, 'The Relevance of Algorithms' in Tarleton Gillespie, Pablo J Boczkowski and Kirsten A Foot (eds), *Media Technologies* (The MIT Press 2014) <academic.oup.com/mit-press-scholarship-online/book/14976/chapter/169333383> accessed 24 November 2023

Gilpin LH and others, 'Explaining Explanations: An Overview of Interpretability of Machine Learning', *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE 2018) <ieeexplore.ieee.org/document/8631448/> accessed 15 September 2022

Godoy J, 'Meta Will Face Antitrust Trial over Instagram, WhatsApp Acquisitions' (*Reuters*, 13 November 2024) <reuters.com/legal/meta-will-face-antitrust-trial-over-instagram-whatsapp-acquisitions-2024-11-13/> accessed 23 January 2025

Goodman B and Flaxman S, 'European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"' (2017) 38 AI Magazine 50 <doi.org/10.1609/aimag.v38i3.2741> accessed 23 January 2025

Grochowski M and others, 'Algorithmic Transparency and Explainability for EU Consumer Protection: Unwrapping the Regulatory Premises' (2021) 8 Critical Analysis of Law 43 <ssrn.com/abstract=3826415>

Guidotti R, 'Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking' (2024) 38 Data Mining and Knowledge Discovery 2770 <doi.org/10.1007/s10618-022-00831-6> accessed 23 January 2025

——, 'A Survey Of Methods For Explaining Black Box Models' <doi.org/10.1145/3236009> accessed 23 January 2025

——, ‘Local Rule-Based Explanations of Black Box Decision Systems’ (arXiv 2018) arXiv:1805.10820 <http://arxiv.org/abs/1805.10820> accessed 7 June 2022

Gunning D and others, ‘XAI—Explainable Artificial Intelligence’ (2019) 4 Science Robotics eaay7120 <doi.org/10.1126/scirobotics.aay7120> accessed 23 January 2025

Gutwirth S, ‘Privacy, Data Protection and Law Enforcement. Opacity of the Individual and Transparency of Power’ (2021) 18 http://dx.doi.org/10.11117/rdp.v18i100.6200 accessed 23 January 2025

(eds), *Reinventing Data Protection?* (Springer Netherlands 2009) <http://link.springer.com/10.1007/978-1-4020-9498-9> accessed 28 June 2024

Hacker P and Passoth J-H, ‘Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond’, *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers* (Springer 2022) <doi.org/10.1007/978-3-031-04083-2_17> accessed 23 January 2025

Haleem A and others, ‘Artificial Intelligence (AI) Applications for Marketing: A Literature-Based Study’ (2022) 3 International Journal of Intelligent Networks 119 <doi.org/10.1016/j.ijin.2022.08.005> accessed 23 January 2025

Hänold S, ‘Profiling and Automated Decision-Making: Legal Implications and Shortcomings’, *Robotics, AI and the Future of Law* (Springer Singapore 2018) http://link.springer.com/10.1007/978-981-13-2874-9 accessed 23 January 2025

Hayek FA, *The Road to Serfdom. Text and Documents.* (Bruce Caldwell ed, 2007)

Hellman D, ‘Measuring Algorithmic Fairness’ (2020) 106 Virginia Law Review 811 <virginialawreview.org/articles/measuring-algorithmic-fairness/> accessed 23 January 2025

Henin C and Le Métayer D, ‘A Framework to Contest and Justify Algorithmic Decisions’ (2021) 1 AI and Ethics 463 <link.springer.com/article/10.1007/s43681-021-00054-3> accessed 23 January 2025

Hildebrandt M and Gutwirth S (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008) <http://dx.doi.org/10.1007/978-1-4020-6914-7> accessed 23 January 2025

Hogan -Doran Dominique, ‘Computer Says “No”: Automation, Algorithms and Artificial Intelligence in Government Decision-Making’ 13 The Judicial Review: Selected Conference Papers: Journal of the Judicial Commission of New South Wales 345 <http://dx.doi.org/10.53300/001c.87776> accessed 23 January 2025

‘How SCHUFA Works - Our Principle: Reciprocity: SCHUFA and Its Contractual Partners’ (*SCHUFA*) <schufa.de/en/ueber-uns/schufa/schufa-works/> accessed 21 January 2025

Huq A, ‘A Right to Human Decision’ 106 Va. L. Rev.611 <https://heinonline.org/HOL/LandingPage?handle=hein.journals/valr106&div=16&id=&page= > accessed 15 September 2025

Independent High-Level Expert Group On Artificial Intelligence Set Up By The European Commission, ‘Assessment List for Trustworthy Artificial Intelligence (ALTAI)’ (European Commission 2020) <doi.org/10.2759/002360> accessed 23 January 2025

Ivanova Y, 'The Role of the EU Fundamental Right to Data Protection in an Algorithmic and Big Data World' (2021) 13 Data Protection and Privacy, Volume 13: Data Protection and Artificial Intelligence 145 <dx.doi.org/10.2139/ssrn.3697089> accessed 23 January 2025

Jacobs AZ and Wallach H, 'Measurement and Fairness', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) <dl.acm.org/doi/10.1145/3442188.3445901> accessed 20 May 2022

Johnson' J, 'Mixing Humans and Nonhumans Together: The Sociology of a Door-Closer' <doi.org/10.2307/800624> accessed 23 January 2025

Kafka F, *The Trial: A New Translation Based on the Restored Text*

Kaminski ME, 'The Right to Explanation, Explained' <lawcat.berkeley.edu/record/1128984> accessed 28 November 2024

Kaminski ME and Urban JM, 'The Right To Contest AI' (2022) 121 Columbia Law Review 93 <ssrn.com/abstract=3965041> accessed 23 January 2025

Kamiran F and Calders T, 'Data Preprocessing Techniques for Classification without Discrimination' (2012) 33 Knowledge and information systems 1 <doi.org/10.1007/s10115-011-0463-8> accessed 23 January 2025

Kampourakis I, Taekema S and Arcuri A, 'Reappropriating the Rule of Law: Between Constituting and Limiting Private Power' (2023) 14 Jurisprudence 76 <ssrn.com/abstract=4238712> accessed 23 January 2025

Karimi A-H, Schölkopf B and Valera I, 'Algorithmic Recourse: From Counterfactual Explanations to Interventions', *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (2021) <doi.org/10.1145/3442188.3445899> accessed 23 January 2025

Kasirzadeh A and Smart A, 'The Use and Misuse of Counterfactuals in Ethical Machine Learning', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021) <doi.org/10.1145/3442188.3445886> accessed 23 January 2025

Kaur D and others, 'Trustworthy Artificial Intelligence: A Review' (2023) 55 ACM Computing Surveys 1 <doi.org/10.1145/3491209> accessed 23 January 2025

Kayser-Bril N, 'Austria's Employment Agency Rolls out Discriminatory Algorithm, Sees No Problem' (*Algorithm Watch*) <algorithmwatch.org/en/austrias-employment-agency-ams-rolls-out-discriminatory-algorithm/> accessed 23 January 2025

Khakurel J and others, 'The Rise of Artificial Intelligence under the Lens of Sustainability' (2018) 6 Technologies 100 <doi.org/10.3390/technologies6040100> accessed 23 January 2025

Kim TW and Routledge BR, 'Informational Privacy, A Right to Explanation, and Interpretable AI', *2018 IEEE Symposium on Privacy-Aware Computing (PAC)* (IEEE 2018) <ieeexplore.ieee.org/document/8511831/> accessed 28 November 2024

Kleinberg J and others, 'Algorithmic Fairness', *Aea papers and proceedings* (American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2018) ISSN 2574-0768

Kozodoi N and others, ‘Shallow Self-Learning for Reject Inference in Credit Scoring’ in Ulf Brefeld and others (eds), *Machine Learning and Knowledge Discovery in Databases* (Springer International Publishing 2020) <arxiv.org/abs/1909.06108v1> accessed 23 January 2025

Langer M and others, ‘What Do We Want from Explainable Artificial Intelligence (XAI)? - A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research’ (2021) 296 Artif. Intell. 103473 <doi.org/10.1016/j.artint.2021.103473> accessed 23 January 2025

Lee MK and others, ‘Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers’, *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM 2015) <dl.acm.org/doi/10.1145/2702123.2702548> accessed 11 November 2023

Li Z and others, ‘Reject Inference in Credit Scoring Using Semi-Supervised Support Vector Machines’ (2017) 74 Expert Systems with Applications 105 <doi.org/10.1016/j.eswa.2017.01.011> accessed 23 January 2025

Lisboa PJG, ‘Interpretability in Machine Learning – Principles and Practice’ in Francesco Masulli, Gabriella Pasi and Ronald Yager (eds), *Fuzzy Logic and Applications* (Springer International Publishing 2013) <doi.org/10.1007/978-3-319-03200-9_2> accessed 23 January 2025

Locke J, *Two Treatises of Government* (Peter Laslett ed, 3rd edn, Cambridge University Press)

Logins A, *Normative Reasons: Between Reasoning and Explanation* (Cambridge University Press 2022)

Lognoul M, ‘Explainability of AI Tools in Private Sector: An Attempt for Systemization’ [2020] SSRN Electronic Journal <ssrn.com/abstract=3685906> accessed 17 January 2022

Loi M, Ferrario A and Viganò E, ‘Transparency as Design Publicity: Explaining and Justifying Inscrutable Algorithms’ (2021) 23 Ethics and Information Technology 253 <doi.org/10.1007/s10676-020-09564-w> accessed 23 January 2025

Loyola-González O, ‘Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View’ (2019) 7 IEEE Access 154096 <http://dx.doi.org/10.1109/ACCESS.2019.2949286> accessed 23 January 2025

Lundberg SM and Lee S-I, ‘A Unified Approach to Interpreting Model Predictions’ in Isabelle Guyon and others (eds), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (2017) <proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html> accessed 23 January 2025

MacCarthy M, ‘Fairness in Algorithmic Decision-Making’ [2019] The Brooking Institution’s Artificial Intelligence and Emerging Technologies (AIET) 14 <brookings.edu/articles/fairness-in-algorithmic-decision-making/> accessed 23 January 2025

Macenaite M, ‘The “Riskification” of European Data Protection Law through a Two-Fold Shift’ (2017) 8 European Journal of Risk Regulation 506 <doi:10.1017/err.2017.40> accessed 23 January 2025

Makin S, 'The Four Biggest Challenges in Brain Simulation' (2019) 571 Nature S9 <nature.com/articles/d41586-019-02209-z> accessed 23 January 2025

Maldonado S and Paredes G, 'A Semi-Supervised Approach for Reject Inference in Credit Scoring Using SVMs' in Petra Perner (ed), *Advances in Data Mining. Applications and Theoretical Aspects* (Springer 2010) <doi.org/10.1007/978-3-642-14400-4_43> accessed 23 January 2025

Malgieri G, 'Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations' (2019) 35 Computer Law & Security Review 105327 <doi.org/10.1016/j.clsr.2019.05.002> accessed 23 January 2025

——, '"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation' (2021) 1 Law and Business 16 <http://dx.doi.org/10.2478/law-2021-0003> accessed 23 January 2025

Malgieri G and Comandé G, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (2017) 7 International Data Privacy Law 243 <doi.org/10.1093/idpl/ipx019> accessed 23 January 2025

Markus AF, Kors JA and Rijnbeek PR, 'The Role of Explainability in Creating Trustworthy Artificial Intelligence for Health Care: A Comprehensive Survey of the Terminology, Design Choices, and Evaluation Strategies' (2021) 113 Journal of Biomedical Informatics 103655 <doi.org/10.1016/j.jbi.2020.103655> accessed 23 January 2025

Marr B, '10 Amazing Examples Of How Deep Learning AI Is Used In Practice?' (*Forbes*) <forbes.com/sites/bernardmarr/2018/08/20/10-amazing-examples-of-how-deep-learning-ai-is-used-in-practice/> accessed 12 December 2023

Martin K, 'Ethical Implications and Accountability of Algorithms' (2019) 160 Journal of Business Ethics 835 <link.springer.com/article/10.1007%2Fs10551-018-3921-3> accessed 23 January 2025

McCabe, '"Google Is a Monopolist", Judge Rules in Landmark Antritrust Case' (*The New York Times*, 5 August 2024) <nytimes.com/2024/08/05/technology/google-antitrust-ruling.html> accessed 23 January 2025

Mendoza I, 'The Right Not to Be Subject to Automated Decisions Based on Profiling - Applied to Examples of Online Scoring Technology, Weblining, and Behavioral Advertising' (PhD Thesis, University of Oslo, Faculty of Law UiO 2016)

Mendoza I and Bygrave LA, 'The Right Not to Be Subject to Automated Decisions Based on Profiling' in Tatiana-Eleni Synodinou and others (eds), *EU Internet Law* (Springer International Publishing 2017) <http://link.springer.com/10.1007/978-3-319-64955-9_4> accessed 28 November 2024

Miller AR, *The Assault on Privacy : Computers, Data Banks, and Dossiers* (University of Michigan Press 1971) accessed 23 January 2025

Mitchell S and others, 'Algorithmic Fairness: Choices, Assumptions, and Definitions' (2021) 8 Annual Review of Statistics and Its Application 141 <doi.org/10.1146/annurev-statistics-042720-125902> accessed 23 January 2025

Mittelstadt B, Russell C and Wachter S, 'Explaining Explanations in AI', *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM 2019) <dl.acm.org/doi/10.1145/3287560.3287574> accessed 28 November 2024

Mittelstadt BD and others, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3 Big Data & Society 205395171667967 <doi.org/10.1177/2053951716679679> accessed 23 January 2025

Molnar C, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2019) <christophm.github.io/interpretable-ml-book/> accessed 23 January 2025

Mothilal RK, Sharma A and Tan C, 'Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) <dl.acm.org/doi/10.1145/3351095.3372850> accessed 15 October 2024

Nantais J, 'Predictive Analytics in Government Decisions' (*Medium*, 31 July 2019) <towardsdatascience.com/predictive-analytics-in-government-decisions-8128ba019a77> accessed 6 December 2022

Naqvi R and others, 'The Nexus Between Big Data and Decision-Making: A Study of Big Data Techniques and Technologies' in Aboul Ella Hassanien and others (eds), *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)* (Springer International Publishing 2021) <doi-org.soton.idm.oclc.org/10.1007/978-3-030-76346-6_73> accessed 23 January 2025

Narayanan A, 'Fairness Definitions and Their Politics' [21] Youtube: Arvind Naranayan, Available online: https://www. youtube. com/watch

Neill DB, 'Using Artificial Intelligence to Improve Hospital Inpatient Care' (2013) 28 IEEE Intelligent Systems 92 <doi.org/10.1109/MIS.2013.51> accessed 23 January 2025

Nigrini MJ, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection* (John Wiley & Sons 2012) ISBN: 978-1-118-15285-0

'NoBIAS - About' (*NoBIAS*) <nobias-project.eu/about/> accessed 17 January 2025

Normandeau K, 'Beyond Volume, Variety and Velocity Is the Issue of Big Data Veracity' [2013] Inside big data <insideainews.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/> accessed 23 January 2025

Office of the Data Protection Ombudsman, 'The Data Protection Ombudsman Ordered Svea Ekonomi to Correct Its Practices in the Processing of Personal Data' (1 April 2018) <tietosuoja.fi/-/tietosuojavaltuutettu-maarasi-svea-ekonomin-korjaamaan-kaytantojaan-henkilotietojen-kasittelyssa> accessed 23 January 2025

Olsen HP and others, 'What's in the Box? The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration' <ssrn.com/abstract=3402974> accessed 23 January 2025

O'neil C, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2017) ISBN:978-0-553-41881-1

Orwell G, *Nineteen Eighty-Four (1984)*

‘Overfitting vs. Underfitting: What’s the Difference?’ (*Coursera*, 11 April 2024) <coursera.org/articles/overfitting-vs-underfitting> accessed 20 January 2025

Pagallo U, ‘Algo-Rhythms and the Beat of the Legal Drum’ (2018) 31 Philosophy & Technology 507 <doi.org/10.1007/s13347-017-0277-z> accessed 23 January 2025

Pedreschi D and others, ‘Meaningful Explanations of Black Box AI Decision Systems’ (2019) 33 Proceedings of the AAAI Conference on Artificial Intelligence 9780 <doi.org/10.1609/aaai.v33i01.33019780> accessed 23 January 2025

Pedreschi D, Ruggieri S and Turini F, ‘Discrimination-Aware Data Mining’ 9 <doi.org/10.1145/1401890.1401959> accessed 23 January 2025

Pencheva I, Esteve M and Mikhaylov SJ, ‘Big Data and AI – A Transformational Shift for Government: So, What next for Research?’ (2020) 35 Public Policy and Administration 24 <doi.org/10.1177/0952076718780537> accessed 23 January 2025

PhD SP, ‘Mitigating Algorithmic Bias in Predictive Justice: 4 Design Principles for AI Fairness’ (*Medium*, 24 November 2018) <towardsdatascience.com/mitigating-algorithmic-bias-in-predictive-justice-ux-design-principles-for-ai-fairness-machine-learning-d2227ce28099> accessed 13 December 2023

Raaijmakers S, ‘Artificial Intelligence for Law Enforcement: Challenges and Opportunities’ (2019) 17 IEEE security & privacy 74 <doi.org/10.1109/MSEC.2019.2925649> accessed 23 January 2025

Ribeiro MT, Singh S and Guestrin C, ‘Model-Agnostic Interpretability of Machine Learning’ (arXiv, 16 June 2016) <http://arxiv.org/abs/1606.05386> accessed 22 September 2022

Richards NM and King JH, ‘Three Paradoxes of Big Data’ (2013) 66 Stanford Law Review Online 41 <stanfordlawreview.org/online/privacy-and-big-data-three-paradoxes-of-big-data/> accessed 23 January 2025

Richardson J, Witzleb N and Paterson M, ‘Political Micro-Targeting in an Era of Big Data Analytics: An Overview of the Regulatory Issue’, *Big Data, Political Campaigning and the Law* (Routledge 2019)

Rieder B, ‘Big Data and the Paradox of Diversity’ (2016) 2 Digital Culture & Society 39 <doi.org/10.14361/dcs-2016-0204> accessed 23 January 2025

Ritchie KL and others, ‘Public Attitudes towards the Use of Automatic Facial Recognition Technology in Criminal Justice Systems around the World’ (2021) 16 PLOS ONE e0258241 <doi.org/10.1371/journal.pone.0258241> accessed 23 January 2025

Roig A, ‘Safeguards for the Right Not to Be Subject to a Decision Based Solely on Automated Processing (Article 22 GDPR)’ (2017) 8 <ejlt.org/index.php/ejlt/article/view/570/772> accessed 23 January 2025

Rott P, ‘Powerful Private Players in the Digital Economy: Between Private Law Freedoms and the Constitutional Principle of Equality’ (2020) 18 Baltic Yearbook of International Law Online 32 <doi.org/10.1163/22115897_01801_004> accessed 23 January 2025

Rovatsos M, Mittelstadt B and Koene A, 'Landscape Summary: Bias in Algorithmic Decision-Making: What Is Bias in Algorithmic Decision-Making, How Can We Identify It, and How Can We Mitigate It?' <research.ed.ac.uk/en/publications/landscape-summary-bias-in-algorithmic-decision-making-what-is-bia> accessed 20 May 2022

Rubinsteain P, 'Asynchronous Video Interviews: The Tools You Need to Succeed' *BBC* (Remote Control, 6 November 2020) <bbc.com/worklife/article/20201102-asynchronous-video-interviews-the-tools-you-need-to-succeed> accessed 23 January 2025

Rudin C, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (arXiv, 21 September 2019) <http://arxiv.org/abs/1811.10154> accessed 20 September 2022

Sancho D, 'Automated Decision-Making under Article 22 GDPR: Towards a More Substantial Regime for Solely Automated Decision-Making' in Martin Ebers and Susana Navas (eds), *Algorithms and Law* (1st edn, Cambridge University Press 2020) <cambridge.org/core/product/identifier/9781108347846%23CN-bp-4/type/book_part> accessed 28 November 2024

Sarra C, 'Defenceless? An Analytical Inquiry into The Right to Contest Fully Automated Decisions In the GDPR' <researchgate.net/profile/Claudio-Sarra> accessed 23 January 2025

Scanlon TM, *What We Owe to Each Other* (Harvard University Press 2000)

Schmitt M, 'Automated Machine Learning: AI-Driven Decision Making in Business Analytics' (2023) 18 Intelligent Systems with Applications 200188 <doi.org/10.1016/j.iswa.2023.200188> accessed 23 January 2025

Schmon C, 'Automated Decision Making and Artificial Intelligence – A Consumer Perspective' <beuc.eu/sites/default/files/publications/beuc-x-2018-058_automated_decision_making_and_artificial_intelligence.pdf> accessed 23 January 2025

Selbst AD and Powles J, 'Meaningful Information and the Right to Explanation' (2017) 7 International Data Privacy Law 233 <doi.org/10.1093/idpl/ipx022> accessed 23 January 2025

Shany Y, 'From digital rights to international human rights: The emerging right to a human decision maker' *Institute for Ethic in AI* (11 December 2020) https://www.oxford-<aiethics.ox.ac.uk/blog/digital-rights-international-human-rights-emerging-right-human-decision-maker> accessed 15 September 2025

Slack D and others, 'Counterfactual Explanations Can Be Manipulated' (2021) 34 Advances in neural information processing systems 62 <proceedings.neurips.cc/paper_files/paper/2021/file/009c434cab57de48a31f6b669e7ba266-Paper.pdf> accessed 23 January 2025

Sokol K and Flach PA, 'Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches', *FAT** (ACM 2020) <doi.org/10.1145/3351095.3372870> accessed 23 January 2025

Sokolova M, Japkowicz N and Szpakowicz S, 'Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation' in Abdul Sattar and Byeong-ho Kang (eds), *AI 2006: Advances in Artificial Intelligence*, vol 4304 (Springer Berlin Heidelberg 2006) <http://link.springer.com/10.1007/11941439_114> accessed 2 December 2024

Solove DJ, ‘Privacy and Power: Computer Databases and Metaphors for Information Privacy’ (2000) 53 Stanford Law Review 1393 <ssrn.com/abstract=248300> accessed 23 January 2025

Sovrano F and others, ‘Metrics, Explainability and the European AI Act Proposal’ (2022) 5 J 126 <doi.org/10.3390/j5010010> accessed 23 January 2025

Speith T, ‘A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods’, *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (2022) <doi.org/10.1145/3531146.3534639> accessed 23 January 2025

Spina A, ‘A Regulatory *Mariage de Figaro* : Risk Regulation, Data Protection, and Data Ethics’ (2017) 8 European Journal of Risk Regulation 88 <doi.org/10.1017/err.2016.15> accessed 23 January 2025

State L, ‘Logic Programming for XAI: A Technical Perspective’, *ICLP Workshops* (CEUR-WS.org 2021) <ceur-ws.org/Vol-2970/meepaper1.pdf> accessed 23 January 2025

——, ‘The Explanation Dialogues: An Expert Focus Study to Understand Requirements towards Explanations within the GDPR’ [2025] Artificial Intelligence and Law <doi.org/10.1007/s10506-024-09430-w> accessed 23 January 2025

Steen M, ‘Upon Opening the Black Box and Finding It Full: Exploring the Ethics in Design Practices’ (2015) 40 Science, Technology, & Human Values 389 <jstor.org/stable/43671241> accessed 23 January 2025

Stone HS, *Introduction to Computer Organization and Data Structures* (McGraw-Hill, Inc 1971)

Stryker C and Kavlakoglu E, ‘What Is Artificial Intelligence (AI)? | IBM’ (*IBM*, 9 August 2024) <ibm.com/think/topics/artificial-intelligence> accessed 17 January 2025

Sullivan E, ‘Facial Recognition Technology: Verification vs. Identification’ (Montana State Legislature, Economic Affairs Interim Committee 2021) <leg.mt.gov/content/Committees/Interim/2021-2022/Economic%20Affairs/Meetings/November%202021/Facial-verification-vs-facial-identification.pdf> accessed 23 January 2025

Szczepański M and others, ‘The Methods and Approaches of Explainable Artificial Intelligence’ in Maciej Paszynski and others (eds), *Computational Science – ICCS 2021*, vol 12745 (Springer International Publishing 2021) <link.springer.com/10.1007/978-3-030-77970-2_1> accessed 15 September 2022

Taipale KA, ‘Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy and the Lessons of King Ludd’ (2004) 7 Yale Journal of Law & Technology and International Journal of Communications Law & Policy 123 <ssrn.com/abstract=601421> accessed 23 January 2025

Thieme N, ‘We Are Hard-Coding Injustices for Generations to Come’ (*Undark Magazine*, 20 February 2018) <undark.org/2018/02/20/ai-watchdog-computational-justice/> accessed 6 December 2022

Thouvenin F, Früh A and Henseler S, ‘Article 22 GDPR on Automated Individual Decision-Making: Prohibition or Data Subject Right?’ (2022) 8 European Data Protection Law Review 183 <doi.org/10.21552/edpl/2022/2/6> accessed 23 January 2025

Tito J, ‘Destination Unknown: Exploring the Impact of Artificial Intelligence on Government’ [2017] Centre for Public Impact 7

Tolan S, ‘Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges’ 25 <doi.org/10.48550/arXiv.1901.04730> accessed 23 January 2025

Tosoni L, ‘The Right to Object to Automated Individual Decisions: Resolving the Ambiguity of Article 22(1) of the General Data Protection Regulation’ (2021) 11 International Data Privacy Law 145 <ssrn.com/abstract=3845913> accessed 23 January 2025

Turek M, ‘Explainable Artificial Intelligence’ (*DARPA Defense Advanced Research Projects Agency*) <darpa.mil/program/explainable-artificial-intelligence> accessed 22 September 2022

Umoja Noble S, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press` 2018)

U.S. Department of Health, Education & Welfare., ‘Record Computers and the Rights of Citizens’ (Secretary’s Advisory Committee on Automated Personal Data Systems 1973) (OS)73-94

Vedder A and Naudts L, ‘Accountability for the Use of Algorithms in a Big Data Environment’ (2017) 31 International Review of Law, Computers & Technology 206 <doi.org/10.1080/13600869.2017.1298547> accessed 23 January 2025

Vilone G and Longo L, ‘Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence’ (2021) 76 Inf. Fusion 89 <doi.org/10.1016/j.inffus.2021.05.009> accessed 23 January 2025

Vredenburgh K, ‘The Right to Explanation’ (2022) 30 The Journal of Political Philosophy 209 < 10.1111/jopp.12262> accessed 23 January 2025

Wachter S and Mittelstadt B, ‘A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI’ [2019] Colum. Bus. L. Rev. 494 <ssrn.com/abstract=3248829> accessed 23 January 2025

Wachter S, Mittelstadt B and Floridi L, ‘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 <doi.org/10.1093/idpl/ipx005> accessed 23 January 2025

Wachter S, Mittelstadt B and Russell C, ‘Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR’ [2017] SSRN Electronic Journal <ssrn.com/abstract=3063289> accessed 10 January 2022

——, ‘Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law’ [2021] SSRN Electronic Journal <ssrn.com/abstract=3792772> accessed 9 February 2022

Weber M, *Politics as a Vocation* (1921)

Wesstein E, ‘Benford’s Law’ (*Wolfram MathWorld*) <mathworld.wolfram.com> accessed 23 September 2022

Westin AF, ‘Privacy and Freedom’ [1967] New York: Atheneum <scholarlycommons.law.wlu.edu/wlulr/vol25/iss1/20> accessed 23 January 2025

‘What Are Neural Networks? | IBM’ <ibm.com/topics/neural-networks> accessed 5 December 2023

‘What Is an “Algorithm”? It Depends Whom You Ask’ (*MIT Technology Review*) <technologyreview.com/2021/02/26/1020007/what-is-an-algorithm/> accessed 4 December 2023

Whitman JQ, ‘The Two Western Cultures of Privacy: Dignity versus Liberty’ The Yale Law Journal <http://dx.doi.org/10.2139/ssrn.476041> accessed 23 January 2025

Willson M, ‘Algorithms (and the) Everyday’ (2017) 20 Information, Communication & Society 137 < https://doi.org/10.1080/1369118X.2016.1200645>accessed 23 January 2025

Wong P-H, ‘Democratizing Algorithmic Fairness’ (2020) 33 Philosophy & Technology 225 <doi.org/10.1007/s13347-019-00355-w> accessed 23 January 2025

Xavier B, ‘Biases within AI: Challenging the Illusion of Neutrality’ [2024] AI & SOCIETY 1 <doi.org/10.1007/s00146-024-01985-1> accessed 23 January 2025

Xu J and others, ‘Translating Cancer Genomics into Precision Medicine with Artificial Intelligence: Applications, Challenges and Future Perspectives’ (2019) 138 Human Genetics 109 <doi.org/10.1007/s00439-019-01970-5> accessed 23 January 2025

Yeung K and A. Bygrave L, ‘Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship’ 16 Regulation & Governance 137 <doi.org/10.1111/rego.12401> accessed 23 January 2025

Zednik C, ‘Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence’ (2021) 34 Philosophy & Technology 265 <doi.org/10.1007/s13347-019-00382-7> accessed 23 January 2025

Zerilli J and others, ‘Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?’ (2019) 32 Philosophy & Technology 661 <doi.org/10.1007/s13347-018-0330-6> accessed 23 January 2025