



UPPSALA
UNIVERSITET

Department of Law
Spring term 2025

Master's Thesis in European Union Law
30 ECTS

Hiding in plain sight: Algorithmic Transparency and Explainability under the GDPR

Legal and Technical Challenges in Applying the
GDPR's Provisions on Transparency and
Explainability to Algorithmic AI Systems

Author: Helena Huang

Supervisor: Professor Anna-Sara Lind



Preface

Denna uppsats är inte bara resultatet av många timmars arbete, utan också ett bevis på det stöd jag har fått längst vägen. Därför vill jag ta tillfället i akt och tacka de människor som har gjort detta möjligt. Först och främst, min älskade familj. Tack Elin, och tack Max – utan er, inget jag. Tack mamma och pappa för att ni gett mig alla möjligheter ni själva aldrig hade. Er kärlek och uppoffring har varit min största drivkraft, och jag lovar att göra något meningsfullt med det förtroende ni gett mig.

Till mina vänner. Ni har inte bara delat glädjen i framgångarna, utan också burit mig genom tvivlen. Tack för alla roliga minnen och för ert ovärderliga, ovillkorliga stöd. Ett särskilt tack till Ellen, Evelina och Mariam. Utan er hade juristprogrammet varit betydligt tuffare att ta sig igenom.

Till min älskade Henrik. Tack för att du inte låter mig tvivla på mig själv och för att du tankar mig med kärlek inför varje utmaning. Din orubbliga tro på mig har smittat av sig.

Jag vill även rikta ett varmt tack till min handledare, Anna-Sara Lind. Dina skarpa frågor och kloka råd har varit oerhört värdefulla för mitt skrivande.

Och så Uppsala – tack. För insikterna, människorna, minnena. Det är med en känsla av vemod och tacksamhet som jag vänder blad. Nu väntar nya äventyr!

Helena Huang

Stockholm, 28 Februari 2025

Table of Contents

List of abbreviations	1
1 Introduction	3
1.1 Background.....	3
1.2 Problem statement	4
1.3 Purpose and research questions	5
1.4 Scope and delimitations.....	6
1.5 Method and material	7
1.6 Outline	10
2 Theoretical and conceptual frameworks	12
2.1 Introduction	12
2.2 Artificial Intelligence, Machine Learning and Deep Learning.....	12
2.3 Algorithmic decision-making.....	14
2.4 The concept of transparency in the context of AI.....	15
2.5 The concept of explainability in the context of AI	17
2.6 The importance of transparency and explainability	19
3 Algorithmic transparency and explainability under the GDPR	22
3.1 Introduction	22
3.2 GDPR: An overview.....	22
3.3 Transparency under the GDPR.....	24
3.4 Explainability under the GDPR.....	25
3.4.1 <i>Is there a “right to explanation” in the GDPR?</i>	25
3.4.2 <i>When should the explanation be disclosed?</i>	28
3.4.3 <i>The nature and scope of “meaningful information”</i>	30
4 Technical and legal challenges of algorithmic transparency and explainability	33
4.1 Introduction	33
4.2 The “Black box” problem.....	33
4.3 Trade-offs between performance and explainability	36
4.4 Additional explainability and interpretability challenges.....	38
5 Explaining algorithmic AI systems	40
5.1 Introduction	40
5.2 The implementation of the “right to explanation”	40
5.3 The concept of explainable AI (XAI).....	42
5.4 The case of algorithmic governance.....	45
5.5 Potential legal remedies.....	47
6 Final remarks.....	51
Bibliography	53

List of abbreviations

AI	Artificial Intelligence
AI Act	Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence and amending Regulations (Artificial Intelligence Act)
AI HLEG	High-Level Expert Group on Artificial Intelligence that was set up by the European Commission in 2018.
Charter	Charter of Fundamental Rights of the European Union
CJEU	Court of Justice of the European Union
Commission	European Commission
EDPB	European Data Protection Board
EDPS	European Data Protection Supervisor
EU	European Union
GDPR	Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with regard to the Processing of Personal Data and on the Free Movement of Such Data, and

	repealing Directive 95/46/EC (General Data Protection Regulation)
SvJT	Svensk Juristtidning
TEU	Treaty on the European Union
TFEU	Treaty on the Functioning of the European Union
WP 29	Article 29 Data Protection Working Party
XAI	Explainable AI

1 Introduction

1.1 Background

Artificial Intelligence (AI) has become an integral part of modern life, influencing and shaping critical sectors such as finance, healthcare¹, education², national security³, and criminal justice⁴, among others.⁵ Moreover, society is transitioning to what could be described an “algorithmic era”⁶, where automated systems and agents hold considerable influence over social and economic matters.⁷ AI, once limited to controlled environments now plays a vital role in real-world applications, as our society becomes increasingly dependent on algorithmic outputs.⁸ In today’s society, modern AI is frequently used in processes of decision-making, recommendation and prediction, often in sensitive scenarios that carry significant social and moral weight.⁹ This dependence on AI systems across a wide spectrum of applications has raised profound ethical challenges, particularly concerning the transparency of these systems.¹⁰

In the last decade, machine learning has emerged as a transformative technology, changing the world as we know it.¹¹ The rapid progress in this field have enabled machine learning algorithms to perform with remarkable precision, often surpassing human capabilities in solving highly complex problems. Central to this technological evolution are deep learning models, a subset of machine learning. These models are often described as “black boxes” due to their non-linear configurations that make them inherently difficult to interpret. This transparency presents a critical challenge, not least in domains where decision-making must be interpretable and explainable to ensure that they are trustworthy. As a result, there has been a growing focus on research aimed at making deep learning models more comprehensible through methods of visualization, explanation, and interpretation.¹²

¹ Rashid & Kausik, 2024, p. 1.

² Chaudhary, 2024, p. 93.

³ Rashid & Kausik, 2024, p. 1.

⁴ Shah et al., 2024, p. 19.

⁵ Rashid & Kausik, 2024, p. 1.

⁶ Paik, 2023, p. 1-2.

⁷ Chaudhary, 2024, p.100.

⁸ Shah et al., 2024, p. 19.

⁹ Walmsley, 2021, p. 585.

¹⁰ Franzoni, 2023, p. 118.

¹¹ Wang, 2023, p. 8.

¹² Samek & Müller, 2019, p. 5.

When responsibly designed and implemented, algorithmic AI systems have the potential to enhance human rights and democratic values.¹³ It also carries the potential to yield positive impacts for the business sector, such as reduced costs, increased productivity, as well as improved accuracy.¹⁴ However, if recklessly designed, these system could, at worst, contribute to the spread of misinformation¹⁵, enable more sophisticated cyber threats¹⁶, increase the wage gap¹⁷ and create ethical dilemmas, such as the development of autonomous weapon systems.¹⁸ In addition, the lack of transparency in AI decision-making also increases the risks of bias, discrimination, and other unintended outcomes.¹⁹ In practice, this can manifest in various forms, including fatal accidents involving self-driving cars and discriminatory practices in AI-driven employment systems. In the medical field, AI-generated diagnoses, have in some cases been discredited due to flawed analytical processes.²⁰ The demand for transparency is thereby of utmost importance, not least for its crucial role in promoting trust, ensuring fairness, and increasing accountability.²¹ Transparent systems assures that these technologies are not only effective and efficient in providing solutions but also ethically sound and trustworthy.²²

1.2 Problem statement

The concern of ensuring algorithmic transparency is fundamentally tied to the question of how decision-making processes within AI systems can be made more understandable and comprehensible to the individuals impacted by their outputs. This involves addressing the critical issue of making it clear why a given algorithm produces a certain output.²³ Different transparency mechanisms have been proposed to tackle this challenge. These tactics include conducting thorough audits and offering systematic explanations of how the system operates.²⁴ It is evident that the implications of these issues extend beyond private interests or technical challenges. Rather, they are a matter of societal and legal

¹³ Auliya et al., 2024, pp. 1-2.

¹⁴ Dignum, 2019, p. 47.

¹⁵ Monteith et al., 2024, p. 33.

¹⁶ Kaloudi & Li, 2020, p. 3.

¹⁷ Pfeiffer et al., 2023, p. 209.

¹⁸ Garcia, 2024, pp. 28-29.

¹⁹ Pfeiffer et al., 2023, p. 210.

²⁰ Benois-Pineau & Petkovic, 2023, p. 1.

²¹ Franzoni, 2023, pp. 118–119.

²² Chaudhary, 2024, p. 93.

²³ Hogan, 2015, p. 105.

²⁴ Mittelstadt, 2016, p. 4994.

priorities. As AI algorithmic systems continue to expand its impact on decision-making in the public and the private sphere, their transparency and explainability become matters of global concern.

This thesis will examine the General Data Protection Regulation (GDPR), its provisions and their implications for algorithmic transparency, considering both the strengths and limitations of these rules in addressing the opacity of deep learning models. This opacity is often described as a characteristic of “black box” systems due to their lack of interpretability and raises significant concerns about accountability and trustworthiness in automated systems²⁵ – a concern this thesis aims to address from a legal perspective.

1.3 Purpose and research questions

AI systems, particularly deep learning models, poses several challenges in ensuring transparency and accountability, both of which are fundamental principles upheld by the European Union (EU).²⁶ While the GDPR addresses data processing transparency, its application to algorithmic AI systems remains ambiguous. The purpose of this thesis is therefore to examine how the GDPR’s transparency and explainability obligations under Articles 13-15 and 22 are interpreted and applied to deep learning AI system. To achieve this objective, three research questions are addressed. By exploring the answers to these questions, the thesis aims to provide an assessment of the legal and technical challenges in ensuring algorithmic transparency and explainability. It also seeks to contribute to the ongoing discussions on how transparency and explainability can be achieved in practice and examine the regulatory framework under the EU law in terms of their efficiency. Ultimately, it aspires to bridge the gap between legal principles and technical realities, proposing remedies to ensure transparency and harmonize GDPR compliance with the opaque nature of deep learning AI systems. The first question is as follows:

1. *How does the GDPR, specifically Articles 13-15 and 22, regulate transparency and explainability in algorithmic AI systems?*

The GDPR contains specific provisions aimed at transparency and explainability. Article 13-15 establish the requirement to inform data subjects about how their data is being

²⁵ Chaudhary, 2024, p. 100.

²⁶ Varošaneć, 2022, p. 96.

processed, including the logic behind the automated decision-making. Furthermore, Article 22 restricts certain automated decision-making practices that significantly affect individuals, requiring specific safety standards. The thesis will examine the regulatory framework and its explicit and implicit requirements for transparency and explainability and the extent to which they create enforceable obligations. A thorough analysis of the scope, interpretation and application of these articles to algorithmic AI systems will be conducted. The second question is as follows:

2. What are the legal and technical challenges posed by the Black-box problem?

The Black box problem is a fundamental technical challenge for deep learning AI systems, characterized by their opaque decision-making processes, which can make compliance with GDPR transparency and explainability requirements particularly challenging. The purpose of this research question is to investigate whether the technical challenges of deep learning systems lead to potential liability gaps or enforcement barriers. Lastly, the third research question reads:

3. How do the provisions mentioned in research question 1 apply to deep learning algorithms?

Furthermore, the thesis seeks clarify how these provisions are applied to AI systems based on deep learning algorithms, which often defy comprehensible explanations due to their opaque nature. This involves analyzing whether the GDPR's standards are realistically implementable in such systems. In particular, it explores whether current technical solutions for explainability, such as explainable AI (XAI), can adequately satisfy these obligations or if there are gaps that undermine the regulation's effectiveness in practice.

1.4 Scope and delimitations

As mentioned before, the focus of the thesis is the GDPR's transparency and explainability requirements in the context of deep learning AI systems. To ensure the scope of this research is both meaningful and manageable due to the limited space and time frame, some delimitations have been established to align with the thesis' objectives.

According to Article 3(1) of the newly adopted AI act, the definition of an “AI system” is “a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.” However, “AI system” is a broad term that encompasses various technologies and applications, including machine learning, deep learning, and natural language processing.²⁷ Thus, some delimitations regarding what type of AI used in the thesis are required. The thesis narrows its focus to the regulatory frameworks within the EU that apply to deep learning models. These complex AI models have been selected due to their unique challenges concerning transparency and explainability, particularly in light of their opaque black box nature. Other AI methodologies, such as rule-based systems or more traditional machine learning models, are excluded, as they have different, less complex structures that fall outside the scope of the thesis. Furthermore, the thesis is limited to the regulatory framework of the EU.

At the heart of this thesis lies the exploration of transparency and explainability in the context of deep learning models. These principles are examined through the lens of the requirements as outlined in the GDPR. Other ethical concerns, such as ethical values of fairness or accountability and enforcement measures under these regulations, are secondary and discussed only when they intersect with the primary focus on transparency and explainability. By narrowing the focus in this way, the research seeks to provide a nuanced understanding of the legal challenges and opportunities associated specifically with implementing transparency and explainability in algorithmic AI systems.

1.5 Method and material

To fulfill the aim of this thesis, a legal dogmatic method has been applied as the primary methodology. This methodology is fundamentally concerned with systematically examining and applying legal sources to resolve a certain legal issue. It relies on an established hierarchy of legal norms, ensuring that legal interpretation aligns with the legal framework in question. Legally binding sources such as EU law and national constitutions hold the highest authority, followed by statutory law that serves as a primary legal foundation. In addition, preparatory legislative works and case law provide

²⁷ Soori et al., 2023, p. 54.

important interpretative guidance, as it offers insight into the legislative intent of the law. Furthermore, legal doctrine, consisting of writings by legal scholars and experts, contributes to the legal discussion by offering various theoretical perspectives that may help interpret the legal provisions.²⁸

The precise definition of the legal dogmatic method is a subject that have been heavily discussed in legal literature.²⁹ While some legal scholars argue that it is a practical method for resolving legal problems that arise in specific contexts by applying established legal rules, others consider it a more theoretical method, aiming to reconstruct³⁰, organize, and better understand the structure of the legal system as a whole.³¹ Although there is a disagreement regarding the definition, the essence of the legal dogmatic method is easier for scholars to agree upon. At its core, the legal dogmatic method seeks to explain and clarify the current state of the law as it exists today – *de lege lata*.³² This involves an in-depth examination of legislation, preparatory works, case law, and doctrine.³³ Moreover, some scholars argue that the method can be applied to explore the development of legal norms, offering insights into how the law might evolve or should evolve – *de lege ferenda*.³⁴ Furthermore, there is another debate concerning whether a critical analysis lies within the scope of the legal dogmatic method³⁵, or if it belongs to a broader framework often referred to as the legal analytical method.³⁶ According to Kleineman, the legal dogmatic method includes a critical analysis of the existing law.³⁷ When the method is applied critically, it often evaluates the extent to which the legal framework is coherent, systematic, and capable of appropriately balancing competing societal interests.³⁸

The research objectives of this thesis call for a critical assessment of the current legal framework. In light of the above mentioned, the thesis will adopt a legal dogmatic method contextualized within the EU legal framework. The legal dogmatic method, typically associated with national legal systems, is particularly useful for analyzing the GDPR in an EU context. Given the objective of this thesis – to assess principles embedded in the GDPR with respect to their adequacy in addressing challenges posed by algorithmic AI –

²⁸ Peczenik, 1995, pp. 35-42.

²⁹ Kleineman, 2018, p. 37.

³⁰ Jareborg, 2004, p. 2.

³¹ Kleineman, 2018, p. 21.

³² Sandgren, 2021, pp. 45-51.

³³ Kleineman, 2018, p. 21.

³⁴ Kleineman 2018, pp. 40-45.

³⁵ Hjertstedt, 2019, p. 167.

³⁶ Sandgren, 2021, p. 45-51.

³⁷ Kleineman, 2018, p. 36.

³⁸ Hjertstedt, 2019, p. 171; Sandgren, 2021, pp. 43-47.

a critical examination of existing legislation is both necessary and appropriate. With that said, the legal dogmatic method provides a systematic approach to interpreting and applying legal norms while still taken into consideration the uniqueness of EU law. Moreover, since this thesis explores the complex relationship between law and technology, a domain characterized by tension, the methodology adapted must align with these complexities that follows from this law-technology interplay. To ensure a comprehensive analysis, the method will be applied the way Kleineman argues it is intended to. Through critical evaluation of legal arguments, doctrine, and legislative preparation, the thesis aims to assess the robustness and limitations of the existing framework that is the GDPR.

Furthermore, the regulation of transparency and explainability under the GDPR cannot be effectively analyzed without considering the methodologies that guide the interpretation and application of EU law. Within the EU legal framework, there is no universally applicable method that governs all fields of law. Instead, the approach taken often varies depending on the specific legal context.³⁹ This diversity stems from the multifaceted nature of EU law, which concerns a wide range of policy areas, each with its own challenges and perspectives. Consequently, any legal analysis should be tailored to what is methodologically appropriate for the specific subject matter. Since this thesis is focused on EU law, it is only fitting that an EU legal method is applied alongside the legal dogmatic method. The EU legal method employed can be described as the use of EU legal sources when interpreting and analyzing the application of EU legislation.⁴⁰ Central to this method is the interpretative method developed by the Court of Justice of the European Union (CJEU). The CJEU has established an interpretative strategy that adopts a teleological approach, which emphasizes the purpose and objectives underlying the legislative text. By prioritizing the broader context and the intention behind the legislation, the CJEU ensures that its interpretations better align with the goals and principles of the EU legal order, fostering a coherent application of the legal framework.⁴¹

The materials used for this thesis are mainly legislation and doctrine. Worth mentioning in this context is the structured legal hierarchy in the EU legal order. In the domain of EU law, a distinction is made between primary and secondary law. Primary law refers to the constitutional instruments of the EU, including the Treaty on EU (TEU), Treaty of the Functioning of the EU (TFEU), and the Charter of Fundamental rights

³⁹ Reichel, 2018, p. 110.

⁴⁰ Reichel, 2018, pp. 109-112.

⁴¹ Hettne & Otken Eriksson, 2005, pp. 30-31; Reichel, 2018, p. 122.

(Charter), among others. These legal documents make up the foundational legal structure of the EU. Secondary law, on the other hand, includes regulations, directives, and additional legislations that aim to implement the goals articulated in primary law. Central to secondary law are regulations, which are directly applicable across all EU Member States as stated by Article 288 of the TFEU. This direct applicability distinguishes regulations from other legal instruments, such as directives, and emphasizes their binding nature within the EU legal framework.⁴² The GDPR is an example of such a regulation.

As the GDPR is a relatively recent legal framework, it provides limited precedent in terms of judgments from the CJEU directly addressing the subject matter of this thesis. Consequently, non-binding interpretative guidance, commonly referred to as “soft law”, take on a greater significance in clarifying the application and intent of EU legal provisions. Soft law takes the form of a variety of non-legislative instruments, such as guidelines, policies, and reports, often written by experts.⁴³ In the thesis, particular weight is placed on the former agency of the primary advisory body under the GDPR, namely the Article 29 Working Party (WP 29). This entity has provided important interpretative guidelines⁴⁴ that help bridge the gaps where judicial clarity is strictly limited. Additionally, the thesis will also give special attention to Recitals 58 and 71 of the GDPR. These recitals provide valuable insights into the legislative intent and contextual interpretation of provisions concerning transparency in explainability in algorithmic AI systems. Although recitals lack direct legal force, they are indispensable for a nuanced understanding of the legislative framework.

1.6 Outline

Given the duality of the thesis topic, connecting the areas of GDPR and the technological dynamics of algorithmic AI, each chapter of the thesis is meant to function independently. To facilitate navigation, each chapter begins with an introduction that gives a brief context to its content and positions within the broader thesis framework.

Chapter 1, now concluded, serves a foundational purpose by introducing the subject matter, articulating the research questions and presenting the objectives of the essay.

Chapter 2 focuses on the conceptual framework by exploring the foundational concepts of AI, transparency, and explainability in the context of algorithmic AI systems.

⁴² Lehrberg, 2022, pp. 114-116.

⁴³ Reichel, 2018, pp. 127-129.

⁴⁴ WP 29, 2018.

Chapter 3 analyzes the legal framework provided by the GDPR concerning transparency and explainability requirements for algorithmic AI systems. The focus is on Articles 12-14, as well as Article 22, of the GDPR. Research question 1 is answered in this chapter.

Chapter 4 presents the technical challenges posed by deep learning systems in complying with GDPR's transparency and explainability requirements. Research question 2 is answered in this chapter.

Chapter 5 aims to examine the application of GDPR provisions to deep learning systems in details, addressing the challenges of aligning the requirements of Articles 13-15 and 22 GDPR with the technical realities of these systems. It also zooms in on the potential legal remedies and recommendations aimed at bridging the gap between legal and technical perspectives on algorithmic AI systems. Research question 3 is answered in this chapter.

Chapter 6 wraps up this thesis with some final remarks. This will include considerations regarding a future legislative approach to algorithmic governance.

2 Theoretical and conceptual frameworks

2.1 Introduction

This chapter is dedicated to establishing the conceptual framework essential for understanding the interplay between AI, transparency, and explainability, within the context of algorithmic AI systems. Initially, the chapter begins with a brief description of AI, with a particular focus on the characteristics of machine learning and deep learning technologies. The focus then shifts to the principles of transparency and explainability, addressing the conceptual nuances that distinguish the two. This includes providing detailed definitions, as well as contrasting their distinct features. Finally, the chapter highlights the importance of these principles in the development of AI governance, as it fosters trust, effective oversight and accountability.

2.2 Artificial Intelligence, Machine Learning and Deep Learning

AI is a term that captures a wide variety of technological innovations designed to replicate human cognition and behavior. Fundamentally, AI equips systems with the ability to learn from data and prior interactions, engage in complex problem-solving, and operate with a degree of autonomy.⁴⁵ These systems are often designed to determine the most effective course of action based on carefully predefined parameters, in order to achieve specific goals.⁴⁶ In short, AI can be loosely described as incorporating human intelligence into machines. To achieve this, programmers construct extensive lines of code, often consisting of complex rules and so-called decision trees.⁴⁷

Machine learning is a sub-field of AI, which has achieved accomplishments across various domains.⁴⁸ In contrast to symbolic AI, where rules and data are pre-programmed, machine learning operates by utilizing a combination of raw data to autonomously generate rules.⁴⁹ In other words, machine learning systems refine their internal configurations during training, enabling them to detect nuanced patterns and connections within large-scale datasets. As a result, these systems can apply generalized knowledge

⁴⁵ Samoili et al., 2020, p. 8.

⁴⁶ Milossi et al., 2021, p. 58455.

⁴⁷ Fergus & Chalmers, 2022, p. 4.

⁴⁸ Agrawal, 2021, p. 2.

⁴⁹ Fergus & Chalmers, 2022, p. 6.

from previous training data and apply that understanding to analyze and interpret new data⁵⁰ to facilitate a decision.⁵¹

Further, deep learning is a subset of machine learning⁵² that relies on neural networks composed of multiple layers.⁵³ A deep learning architecture typically consist of three layers: the input layer, the hidden layers, and the output layer. The input layer refers to the initial stage where raw data is introduced, while the output layer delivers the final decisions or outcomes. Between these two layers exists the hidden layers, which carry out the essential process that enable the algorithm to learn from the data.⁵⁴ These layered neural networks consist of a series of interconnected layers, each contributing to the system's ability to process, categorize, or predict data accurately.⁵⁵ In fact, one of the most distinguishing features of deep learning is its capacity to process unsupervised learning from unstructured, raw data.⁵⁶ Unlike machine learning, deep learning models possess the inherent capability to identify, refine, and learn features autonomously during the training phase. While traditional machine learning algorithms may struggle with the volume and complexity of the large-scale datasets, deep learning algorithms excel in identifying nuanced relationships and patterns within such data.⁵⁷

When people mention AI nowadays, chances are they are referring to deep learning algorithms. As a matter of fact, deep learning is deeply integrated in modern society, operating seamlessly in the background of everyday life. Though often overlooked by the average individual, this field of AI fundamentally supports, optimizes and enhances everyday tasks across various domains. For instance, the influence of deep learning extends into critical sectors such as healthcare, national safety, and finance.⁵⁸ Today, deep learning has a wide range of applications, including but not limited to pattern recognition⁵⁹, signal processing⁶⁰, control and automation.⁶¹ Some examples of real life

⁵⁰ Avci et al., 2021, pp. 5-6.

⁵¹ Fergus & Chalmers, 2022, p. 6.

⁵² Avci et al., 2021, p. 6.

⁵³ Rashid & Kausik, 2024, p. 5.

⁵⁴ Fergus & Chalmers, 2022, pp. 143–148.

⁵⁵ Zhao & Flenner, 2019, p. 30.

⁵⁶ Avci et al., 2021, p. 6.

⁵⁷ Fergus & Chalmers, 2022, pp. 7-8.

⁵⁸ Fergus & Chalmers, 2022, p. 141.

⁵⁹ Zhang et al., 2019, p. 1.

⁶⁰ Pouyanfar et al., 2019, p. 20.

⁶¹ Soori et al., 2023, p. 55.

applications of deep learning are chatbots and smart speakers, such as Apple's Siri or Amazon's Alexa.⁶²

2.3 Algorithmic decision-making

The notion of automated decision-making is broadly defined as taking a decision without human intervention. The GDPR provides a specific definition in Article 22(1), stating that “automated individual decision-making” is “a decision based solely on automated processing”. This does not mean that humans cannot contribute indirectly, such as by feeding data into the system or by interpreting the outcome after a decision has been reached. However, these forms of interventions are often minimal, and sometimes entirely automated. Furthermore, the impact of automated decision-making depends on the significance of its consequences for individuals. When such decision-making processes are non-binding and do not infringe upon the rights of individuals, they can be categorized as low-impact and are thus unlikely to require strict legal oversight. Although the situation becomes different when decisions bear significant consequences for individuals. Examples of these situations include determinations of an individual's access to financial services, tax return, or employment opportunities. When these decisions are made, there should be sufficient legal safeguards in place to protect the individual's rights.⁶³

Closely related to the concept of automated decision-making is the concept of algorithmic decision-making, as algorithms serve as the operational backbone of most automated decisions nowadays. These algorithms may be defined as “a set of steps to accomplish a task that is described precisely enough that a computer can run it.”⁶⁴ With the rise of big data, the reliance on algorithmic systems has increased drastically. In today's society, they have taken a role as an indispensable tool for decision-making that far exceeds human capabilities.⁶⁵ This rapid advancement of these algorithmic systems has sparked discussions in both the legal and academic realm, with growing demands for increased “algorithmic transparency”⁶⁶ and “algorithmic accountability”.⁶⁷

⁶² Fergus & Chalmers, 2022, pp. 241-242.

⁶³ Brkan, 2019, pp. 93–94.

⁶⁴ Cormen, 2013, p. 1.

⁶⁵ Brkan, 2019, pp. 94–95.

⁶⁶ Brkan, 2019, p. 117.

⁶⁷ Caplan et al., 2018, p. 4.

2.4 The concept of transparency in the context of AI

The interpretation of the concept “transparency” varies significantly across different domains.⁶⁸ Given this diversity in meaning, it is important to define the specific terminology before moving forward to discuss the topic. Transparency, in the legal domain, encompasses a wide set of functions, including its role in facilitating accountability for the developers and operators of AI systems and in promoting public confidence and trust in oversight mechanisms.⁶⁹ Within the AI Act, transparency is a requirement that extends beyond mere disclosure. The act uses terminology such as “transparency” and “comprehensible“, mandating that high-risk AI systems ought to be designed and developed in such a manner that their operation is “sufficiently transparent” to enable users to accurately interpret the system’s outputs. This requirement is explicitly codified in Article 13 AI Act, and further reinforced by Recital 66 and 72 concerning the same Act.⁷⁰ However, beyond the AI Act’s use of the term, transparency is often understood as a broad and multifaceted concept used in multiple different contexts.⁷¹ In the legal doctrine, for instance, it is often understood as the capability to describe, inspect and track and the mechanisms by which AI systems generate decisions.⁷² Another notable example of this broader definition of the word can be found in the Ethics guidelines of Trustworthy AI, provided by the European Commission’s independent High-Level Expert Group on AI (AI HLEG) in 2018. According to these guidelines, transparency includes not only an understanding of how AI systems process data, but also broader principles such as traceability, explainability, and effective communication with stakeholders.⁷³

In the context of AI, many argue that the concept of transparency is directly linked to the principle of accountability. Accountability ensures that responsibility for the actions and outcomes of these systems is clearly assigned to their designers and developers. A sufficiently transparent AI framework enables the tracing of decisions to specific system components, which is crucial for addressing negative outcomes that may arise from the use of AI. The concept of transparency in AI systems plays a central role in identifying and addressing biases embedded in data sets or algorithms. By making AI systems more transparent, stakeholders gain the ability to detect unintended biases that may reside in

⁶⁸ Panigutti et al., 2023, pp. 1139–1140.

⁶⁹ Shah et al., 2024, p. 21

⁷⁰ AI act, Recital 66 & 72.

⁷¹ Almada, 2023, pp. 9-10.

⁷² Franzoni, 2023, p. 119.

⁷³ AI HLEG, 2019, p. 14.

AI-generated decisions. Such transparency ensures that users affected by the decisions receive fair and unbiased information, emphasizing the ethical use of these technologies.⁷⁴ The correlation between the concepts of accountability and transparency allows us to discuss transparency within the framework of three fundamental levels.⁷⁵ Each of these levels builds upon the other, forming a layered structure of AI accountability.

The first level, implementation, involves an AI systems' capability to transform input data into a predictable output, guided by technical principles and the associated parameters. The Commission has underscored the value of transparency at this level by comparing "white box" models, characterized by clarity and transparency, to "black box" models, which remain obscure to us. Additionally, the second level addresses specification. This involves the disclosure of the underlying objectives, tasks, and datasets that together shaped the AI system during its decision-making process. This too, offers insights into the choices made in the design phase, and links to the third level, that refers to interpretability. This last layer seeks to unravel the fundamental mechanisms and reasoning embedded within the AI algorithm. Transparency at this stage ensures a coherent understanding of the logic and context behind the certain outputs and conclusions the algorithm arrives to. This is especially important since transparency at this stage is closely linked to fairness in decisions, as the lack of clarity can lead to biases and unfair outcomes.⁷⁶ According to the Commission, current AI systems rarely meet the criteria of interpretability at this level.⁷⁷ While efforts can be made in achieving transparency at the implementation and specification levels, interpretability seems to be the most persistent challenge of them all.

Adding to the challenge of algorithmic transparency is the inherently complex nature of transparency within the legal domain. While there seems to be somewhat a consensus around its central function as a form of disclosure, significant variations exist regarding the specifics of the content disclosed. The substance of what must be disclose, the timing of disclosure, and the intended audience are all aspects subject to discussion.⁷⁸ For regulators responsible for implementing legal frameworks, it is imperative to account for this issue. The legal expectations concerning transparency extend far beyond mere technical openness and require careful consideration of its implications for accountability

⁷⁴ Franzoni, 2023, p. 119.

⁷⁵ Hamon et al., 2020, pp. 11-12.

⁷⁶ Sinha & Dunbar, 2022, p. 5.

⁷⁷ Hamon et al., 2020, pp. 11-12.

⁷⁸ Busuioc et al., 2023, p. 86.

and public trust. Although transparency alone is not likely a guarantee for achieving accountability, the ability to access and understand algorithmic processes is a necessity for facilitating it, as it provides insights into the function of AI's decision-making processes.

2.5 The concept of explainability in the context of AI

The concepts of “transparency” and “explainability” are frequently brought up in the same contexts, even though their meanings vary greatly depending on the specific setting.⁷⁹ As aforementioned, transparency in the context of AI is typically understood as the ability to comprehend mechanisms or processes through which an algorithmic AI system arrives at its decisions. Explainability, on the other hand, extends beyond this step. It refers not only to an understanding of the decision-making process but also a requirement for justification. This additional dimension involves providing an explanation that clarifies why a certain outcome was reached by the algorithmic AI system.⁸⁰ According to the High-Level expert group on AI, the principle of explainability is essential to the responsible deployment of AI technologies. Explainability, in their view, refers to “the ability to explain both the technical processes of an AI system and the related human decisions”.⁸¹ In the legal doctrine, however, explainability has been described as the extent to which AI-generated outcomes are comprehensible and interpretable for the general public.⁸²

In the doctrine, the concept of explainability has been divided into two perspectives. One is the model-centric explainability, which focuses on uncovering the internal workings of the algorithm itself and offering an explanation of the underlying logics of the AI system's outputs. Edwards and Veale suggests that these explanations should include various technical details, such as which type or family the AI model belongs to, the input data it relies upon and the process through which the model has been validated.⁸³ The second perspective, named subject-centric explainability, is more task-oriented, addressing how specific decisions are reached within a particular context or domain.⁸⁴ According to Edwards and Veale, this should in include counterfactuals that clarify the

⁷⁹ Almada, 2023, p. 9.

⁸⁰ Brkan & Bonnet, 2020, p. 27.

⁸¹ AI HLEG, 2019, p. 18.

⁸² Brkan & Bonnet, 2020, p. 27.

⁸³ Edwards & Veale, 2017, pp. 55–56.

⁸⁴ de Bruijn et al., 2022, p. 2.

changes in specific inputs that would lead to a different outcome for the individual concerned. Additionally, they argue for providing insights into the attributes of other individuals who have been classified similarly.⁸⁵ An example of how subject-centric explainability is applied in practice is when looking at the possible harm caused by decisions to a certain individual and group, to potentially reverse those decisions.⁸⁶ In other words, explainability could be said to manifest in various forms. Yet the most evident manifestation involves providing a justification of the significant features that influence an AI algorithm and its decisions. For instance, disclosing the criteria or factors considered in determining an action or decision may enhance user understanding. Additionally, clarifying which criteria have not been satisfied, would fill the same function, as it offers a justification for the decision made. This approach is particularly useful in involving classification, such as scholarship applications, where it may be necessary to explain why a student's application is denied.⁸⁷ Other examples include a denial of a credit loan, or the rejection of an unemployment insurance or a medical claim.⁸⁸ In all these scenarios, the applicant might demand to know not only how the decision was reached but also a why that specific outcome was reached. In other words, the individual may seek a coherent justification behind the rejection.⁸⁹

However, it is important to note that there are a few factors that complicates the assessment of a valid explanation. To fully understand a certain decision, it is essential to dissect and explain each of the sequential steps leading to its conclusion. To begin with, one must receive an explanation of how the algorithmic system examines the situation based on the input data. This initial step sets the foundation for the decision-making process, as the data creates the framework for the system's evaluation. Following this, the explanation must account for the factors that influence the decision, taking into consideration to their alignment with the system's overall objectives. Thereafter, an in-depth explanation of each factor involved in the decision-making must be provided, in order to paint a picture of the complex dynamics at play. Finally, the decision itself, should be explained. It is insufficient to only explain one of these stages, as each phase provides context for understanding the final outcome.⁹⁰ Although the challenge of explanation is

⁸⁵ Edwards & Veale, 2017, pp. 57-58.

⁸⁶ de Bruijn et al., 2022, p. 2.

⁸⁷ Hamon et al., 2020, pp. 12-13.

⁸⁸ Selbst, 2021, p. 5.

⁸⁹ Brkan & Bonnet, 2020, p. 27.

⁹⁰ Brkan & Bonnet, 2020, p. 32.

a task that presents a significant challenge, it cannot be overlooked. This is because explainability is an essential pillar in the quest to navigate the complex landscape of effective AI, not least because of its close link to the principle of responsibility. AI experts are expected to maintain a thorough understanding for their systems and bear a duty to explain and justify the ways in which their algorithms influence individuals. This responsibility is not merely technical but extends into the realm of ethics and morals. On a deeper level, the combination of explainability and responsibility reveals a strong moral imperative.⁹¹

2.6 The importance of transparency and explainability

The rapid advancements of AI have cemented its critical role in decision-making across multiple domains, influencing not only the operations of private and public entities, but also individual lives. Thanks to its unique capabilities, AI facilitates the development of tailored solutions in areas such as international relations, warfare, law enforcement, criminal justice, and healthcare.⁹² This wide range of applications underscores the authority AI has gained in shaping critical decisions across society. As AI tools increase decision-making processes across various sectors, it is essential to address its ethical implications proactively.⁹³ In the midst of this discussion stands algorithmic transparency and explainability, which involves making the operations of AI systems comprehensible and clarifying how they arrive at decisions.⁹⁴ Explanations play a key role in human education and learning, functioning as a tool to facilitate deeper understanding. For instance, teachers often clarify solutions to students to help them grasp complex concepts. Similarly, in the realm of healthcare, doctors regularly provide detailed explanations of medical procedures to patients. This is not only to ensure that patients are informed, but also to enhance the legitimacy and acceptance of the doctor's decisions. Although the patient does not have the capacity to verify the decisions in these cases, because of their lack of medical knowledge, receiving an explanation often makes us feel included in the decision-making process. By fostering inclusivity and transparency, we also gain trust and acceptance. Against this backdrop, the design of AI systems that interacts with human users, should be explainable.⁹⁵

⁹¹ Coeckelbergh, 2020, p. 2052.

⁹² Garcia, 2024, p. 25.

⁹³ Wang, 2023, p. 8.

⁹⁴ Chaudhary, 2024, p. 93.

⁹⁵ Samek & Müller, 2019, p. 8.

In everyday applications, algorithmic AI models are often judged primarily based on their efficiency, while transparency and explainability are less prioritized. Occasional errors, such as smart phone's failing to recognize a person's face or an online translator delivering grammatically incorrect sentences, typically lead to minor consequences that do not undermine the acceptance of the technology itself. In such cases, the stakes are generally low, and consequently, the need for transparency and trust is less important. In contrast, the application of algorithmic AI models in high-stake decision-making processes presents a completely different set of challenges. In these cases, the opacity of AI systems can limit their use, sometimes entirely. This is particularly relevant in decision-making processes when wrong outcomes can potentially endanger human life or health.⁹⁶ Consequently, algorithmic transparency is especially critical in areas such as employment, credit evaluation, and criminal justice, where AI systems can profoundly affect individuals' lives, sometimes permanently. In such context, the decisions often have long lasting personal and societal consequences. Examples of these scenarios include self-driving cars and healthcare systems, where human lives may depend on the decisions made by the AI systems. Put simply, the absence of such transparency poses challenges to fairness and non-discrimination, which raises concerns about equal treatment and justice.⁹⁷ In reality, algorithmic transparency plays a vital role in ensuring accountability. If the internal workings of an AI algorithm are opaque, it becomes incredibly difficult to trace errors or identify the root causes of unintended outcomes.⁹⁸ When AI is employed in decision-making, and the stakes are high, principles such as accountability, fairness, and the safeguarding of rights are highlighted. Ensuring that AI systems operate within a transparent framework is essential, necessitating the development as well as enforcement of legal standards that address algorithmic transparency.⁹⁹

In other words, transparency facilitates for individuals to detect potential errors that may affect their rights or interests. It empowers the affected individuals to challenge decisions and safeguard their right to legal action and in some cases, adequate compensation for unjust outcomes.¹⁰⁰ An individual cannot meaningfully evaluate whether a decision is justified, or straight up unlawful, without insight into the reasoning

⁹⁶ Samek & Müller, 2019, pp. 6–7.

⁹⁷ Chaudhary, 2024, pp. 93–94.

⁹⁸ Chaudhary, 2024, p. 100.

⁹⁹ Tzimas, 2023, p. 386.

¹⁰⁰ Doshi-Velez et al., 2019, p. 6.

behind the said decision. The lack of knowledge hinders their ability to seek recourse¹⁰¹ through legal contest or appeal.¹⁰² Against this backdrop, it is safe to say that transparency and explainability are important cornerstones required to achieve procedural justice.¹⁰³ Without transparency, there is no way to assess whether these systems are treating individuals equally fairly and equally without prejudice and bias.¹⁰⁴ As a consequence, the risks of unfairness and unaccountability may severely undermine trust in AI's potential. The demand for transparency is thus both urgent and unavoidable.

In the context of transparency and trustworthy AI, the EU has positioned itself as a global leader in shaping the future use of AI. One of their most prominent initiatives being the creation of Ethics Guidelines for trustworthy AI by the European Commission's AI HLEG. These guidelines reflect the overall principles established in the EU treaties, the Charter, and the GDPR, emphasizing human dignity and transparency as some of the top priorities.¹⁰⁵ Human dignity is understood as an intrinsic worth people possess simply by being human. This worth cannot be neglected, compromised or repressed, not even by the most advanced technology. When shaping the ethical use of AI, individuals right to control their lives and exercise autonomy must therefore be one of the main focuses. Autonomy is in turn a precondition for fully enjoying democracy, justice and equality.¹⁰⁶ Transparency is closely linked to the notions of dignity and autonomy, since it allows for individuals to exercise their rights. According to the Ethics guidelines for trustworthy AI, transparency should be understood as AI systems that "are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights".¹⁰⁷ All these concepts - dignity, autonomy, transparency, and explainability – are included in Recital 27 of the AI Act, which further illustrates the emphasis of transparency in EU's legal framework.

¹⁰¹ Barocas et al., 2020, p. 82.

¹⁰² Selbst, 2021, p. 5.

¹⁰³ Solum, 2004, pp. 74.

¹⁰⁴ Chaudhary, 2024, p. 93.

¹⁰⁵ Milossi et al., 2021, p. 58456.

¹⁰⁶ McCrudden, 2008, pp. 679-680.

¹⁰⁷ AI Act, Recital 27.

3 Algorithmic transparency and explainability under the GDPR

3.1 Introduction

This chapter of the thesis analyzes the legal framework provided by the GDPR concerning transparency and explainability requirements for algorithmic AI systems. Specifically, the focus is on Articles 13-15, as well as Article 22, of the GDPR. The chapter examines the scope, interpretations, and enforceability of these provisions. Furthermore, the challenges of implementing these provisions are addressed, including the practical difficulties of providing “meaningful information” to data subjects and whether a “right to explanation” exists. Research question 1 is thereby answered in this chapter. Additionally, the section examines the current practices of regulatory bodies in interpreting and enforcing these obligations, identifying areas of ambiguity and inconsistency.

3.2 GDPR: An overview

The EU’s GDPR came into force in May 2018.¹⁰⁸ Article 1 of the act lays down the subject matter and objective of the regulation. It declares that the regulation enshrines the protection of personal data as a fundamental human right, emphasizing the need to grant individuals greater control over the way their personal information is handled, processed, and shared. This is done by establishing and imposing obligations on entities both within and beyond the border of the EU. This extraterritorial reach extends to any organization, regardless of its geographic location, that provides goods or services to individuals situated in the EU. Put simply, the regulation applies to all organizations involved in the processing of data belonging to EU data subjects, irrespective of the location of the data processor or the target audience of the organization.¹⁰⁹

As one of the first ever legislative frameworks to address the interplay between AI and personal data rights, the GDPR serves as a foundational regulatory framework in this regard.¹¹⁰ It establishes a structure for the governance of personal data processing,

¹⁰⁸ Kaminski, 2019, p. 192.

¹⁰⁹ GDPR, Article 3; GDPR, Recital 15.

¹¹⁰ Wulf & Seizov, 2020, pp. 625-626.

including cases involving AI.¹¹¹ Of particular importance is the regulation's material scope and its inclusion of automated and algorithmic processes. This is explicitly stated in Article 2(1) of the GDPR, which provides that the regulation is applicable to processing carried out through "algorithmic means", ensuring that automated systems falls within the regulatory scope of the GDPR. Adding to this, articles 15 and 22 GDPR further addresses automated decision-making processes. These provisions impose certain obligations on data controllers, such as the duty to inform individuals about the existence of automated decision-making, the underlying logic concerning these processes, and their potential consequences.¹¹² It is stated in Article 22 GDPR that "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" that "significantly affects him or her". By including these articles, the regulation aims to strike a balance between technological innovation¹¹³ and preserving the individual's right to the protection of data.¹¹⁴ While these measures do offer a baseline for safeguarding individuals against the risk of algorithmic opacity, the vague language of its provisions, combined with the many exceptions it allows, has faced strong criticism for its lack of enforceability and practical impact.¹¹⁵

Through its provisions, the GDPR facilitates algorithmic impact assessments, particularly by requiring evaluations of potential effects on data protection rights. A key feature of these provisions is the requirement for transparency, obliging data controllers to inform individuals about the existence of automated decision-making and "provide meaningful information" about their underlying logic.¹¹⁶ To fulfill this obligation, the principles of algorithmic operations need to be translated into language that is comprehensible to non-experts.¹¹⁷ This challenge, however, raises important legal and practical questions. For instance, does the duty to provide meaningful information include an obligation to explain AI-driven decisions after they occur? And what kind of information must be provided? Do post hoc explanations fall within the scope of the legal obligations of data controllers? A post hoc explanation is one that seeks not to provide a complete explanation of a model's inner workings, but rather to justify decisions made

¹¹¹ Goodman & Flaxman, 2017, pp. 2–3.

¹¹² Wulf & Seizov, 2024, p. 235.

¹¹³ GDPR, Recital 6.

¹¹⁴ GDPR, Recital 1.

¹¹⁵ Wulf & Seizov, 2024, p. 235.

¹¹⁶ GDPR, Articles 13(2)(f) & 14(2)(g).

¹¹⁷ Chaudhary, 2024, p. 110.

by the model by explaining how the certain outcome was reached.¹¹⁸ Such an obligation, if confirmed, would no doubt impose great burdens on data controllers, especially in cases involving highly complex AI models and algorithms.

3.3 Transparency under the GDPR

One critical dimension of the discussions concerning the GDPR's "right to explanation" is the regulation's explicit commitment to ensure a high level of transparency in the processing of personal data. Central to the GDPR's framework is the principle that the processing of personal data should be transparent to the individuals whose data is "collected, used, consulted or otherwise processed", as stated in Recital 39 of the GDPR. For this reason, we can assume that the principle is enshrined in the text of the regulation, emphasizing the commitment to safeguard individual's rights in a fair manner. This should be taken into consideration when interpreting all the articles within the framework.

At a closer look, it is clear that the principle of transparency makes up one of the core elements upon which the GDPR is founded.¹¹⁹ Within the GDPR, multiple provisions can be said to directly or indirectly support the notion of algorithmic transparency. Among these, Article 5(a) GDPR establishes a set of fundamental principles applicable to all data processing activities. These principles are lawfulness, fairness and transparency, and together they form the essence of the EU's data protection. Moreover, the principles in Article 5(a) GDPR are not isolated principles in a vacuum; they are deeply interconnected with the other principles listed in the same article. These include purpose limitation, data minimization, accuracy, storage limitation, integrity, confidentiality, and accountability. Each of these principles is supported by transparency on one side, and accountability on the other,¹²⁰ forming a cohesive framework for ethical and lawful data processing.

Furthermore, Article 12 GDPR elaborates on the principle of transparency, establishing that the information provided to the data subject must meet several criteria. These criteria are that the information given to the data subject "in a concise, transparent, intelligible and easily accessible form, using clear and plain language". Adding to this, recital 60 of the GDPR states that "the principles of fair and transparent processing require that the data subject be informed of the existence of the processing operation and its purposes". Against this backdrop, we can conclude that transparency is not merely a

¹¹⁸ Mourali et al., 2025, p. 2.

¹¹⁹ Chaudhary, 2024, p. 100.

¹²⁰ Goddard, 2017, p. 703.

principle, but also a mechanism through which the other principles are implemented and assessed. The mechanism ensures that the data subject is properly informed and thereby able to engage meaningfully with the processing activities concerning their personal data. By receiving knowledge about how their data is being used and processed, individuals are better positioned to make informed decisions and take steps to safeguard their rights and interests provided by the GDPR.¹²¹

3.4 Explainability under the GDPR

3.4.1 *Is there a “right to explanation” in the GDPR?*

The alleged “right to explanation” of automated decisions within the framework of the GDPR has long been a subject of heated debate among legal scholars.¹²² In particular, this debate concerns whether data subjects are entitled to an explanation of the underlying logics behind automated decision-making processes. Goodman and Flaxman initiated this discussion by interpreting Articles 13 and 14 of the GDPR in a way that extended such a right, namely from the obligation to provide “meaningful information about the logic involved.”¹²³ Their interpretation was then widely echoed, not least by the WP 29, gaining support amongst other scholars.¹²⁴ However, it did not take long until dissenting opinions emerged in the debate. Some people, including Wachter et al., argue that the GDPR’s requirements merely required a general, ante hoc explanation of system functionality rather than a detailed, post hoc justification for specific decisions. More importantly, they highlighted that the GDPR does not provide a right to explanation of algorithmic decision.¹²⁵ This position was then challenged by Selbst and Powles who firmly reject the restrictive view of Wachter et al. by arguing for a robust “right to explanation” rooted in Articles 13-15 of the GDPR.¹²⁶ Casey, Farhangi, and Vogl support this claim, arguing that the GDPR unambiguously establishes this right.¹²⁷ Edward and Veale further explore the debate by acknowledging the potential for a “right to explanation”, but emphasizing the operational challenges posed by the complexity of machine learning systems.¹²⁸ Needless

¹²¹ Chaudhary, 2024, pp. 100-101.

¹²² Brkan, 2019, p. 91.

¹²³ Goodman & Flaxman, 2017, p. 6.

¹²⁴ WP 29, 2018, pp. 25-26.

¹²⁵ Wachter et al., 2017, pp. 76–83.

¹²⁶ Selbst & Powles, 2017, p. 242.

¹²⁷ Casey et al., 2019, pp. 187–188.

¹²⁸ Edwards & Veale, 2017, pp. 81–82.

to say, this has been a topic around which scholars seem to lack consensus, and significant ambiguities remain. Despite all these contributions, key questions persist. If the “right to explanation” exist, how should it be implemented? How detailed should the explanation be, and what specific information should be disclosed, and when should it be disclosed? Addressing these questions is essential for interpreting GDPR provisions in a manner that aligns with its overarching objectives to ensure transparency and accountability.

Central to the GDPR is the data subject’s right to be informed about the collection and use of their data, as enshrined in Articles 13 and 14 of the regulation. According to these provisions, data controllers must notify individuals about key aspects of processing, such as the purpose of data collection and the mechanisms involved. More specifically, Articles 13(2)(f) and 14(2)(g) GDPR obligate data controllers to disclose “the existence of automated decision-making”, including “meaningful information about the logic involved”, as well as the “expected consequences of such processing”. In addition, Article 15(1)(h) GDPR explicitly grants individuals the right to access their personal data and obtain relevant information about its processing. Article 22(3) GDPR further enhances this right in the context of automated decisions. According to the article, data controllers must “take adopt measures to protect the rights, freedoms, and legitimate interests of the data subject, including at least the right to get human intervention from the controller, to express his or her point of view, and contest to the decision”. These safeguards allow individuals to express their opinions, contest automated decisions, and seek redress. Recital 71 further reinforces these rights by emphasizing the necessity of providing information about the logic behind an automated decision. It implies a right to “specific information” and the right “to obtain an explanation of the decision reached after such assessment and to challenge the decision”. While the recital is non-binding legally¹²⁹, it still provides interpretative guidance and emphasizes the necessity for safeguards in cases of automated decision-making under Article 22 GDPR. In addition, the recital serves as a tool for understanding the legislative intent behind the regulatory framework.

It becomes evident when reading Article 22 GDPR, which regulates automated decision-making, that it does not explicitly enshrine a “right to explanation”. The same goes for Articles 13-15 GDPR, which govern individual notice and access. Despite the GDPR’s emphasize on transparency, the concept of a “right to explanation” in the GDPR appears to rest primarily on Recital 71, which lacks binding legal force as part of the

¹²⁹ Klimas & Vaiciukaite, 2008, pp. 15-33.

preamble. Thus, it is necessary to assess whether the actively debated “right to explanation” is established as a right within the GDPR or merely an interpretative extension of the “right to information” in Article 13-15 GDPR, read together with Article 22 GDPR. In my opinion, the derivation of a “right to explanation” under the GDPR is not a question of possibility, but rather a necessity. Without the existence of such a right, the broader rights enshrined in the regulation cannot be properly safeguarded. Articles 13(2)(f), 14(2)(g), and 15(1)(h) GDPR explicitly require that data controllers provide “meaningful information about the logic involved” in automated decision-making processes. My views align with scholars like Goodman and Flaxman, who argue that this obligation inherently supports the existence of a “right to explanation”. Furthermore, I believe that the right to access and the right to contest under Article 22(3) GDPR implicitly necessitate an understanding of the decision-making process. While Wachter et al. claim that the GDPR only requires ante hoc explanations of system functionality, if such a right would even exist, my belief is that this take undermines the commitment of transparency GDPR’s broader objectives. Without a robust explanation of how algorithmic decisions are made, data subjects are very unlikely to exercise their rights effectively, not least because of the practical and technical challenges posed by opaque algorithms as noted by Edwards and Veale. In the absence of a clear understanding of the underlying reasons for the decision made, a data subject’s right to challenge the decision would be next to meaningless. Transparency in this context requires insight into the data inputs used in the decision-making process, as well as a coherent justification of the outcome. Moreover, I support Selbst and Powles’ view, as they emphasize the practical necessity of deriving a “right to explanation” from Articles 13-15 of the GDPR. Although, in my opinion, these rights should be read together with Article 22 of the GDPR and through the lens of GDPR’s overarching objective to ensure transparency. In other words, there are several different layers of explanations of algorithmic decision-making established within the GDPR.

With this said, it is possible that this right is difficult to implement in practice due to the ambiguity of the GDPR’s provisions. To illustrate, we can use the example of Article 22 GDPR that aims to ensure safeguards for individuals subjected to decisions based “solely” on automated processing that produce “significant effects” for the individual involved. Yet, there are no clues in the text that clarifies the meaning of the terms “solely automated” or “significant effects”, which leaves us uncertain. This lack of clarity can create loopholes and increase the risk of certain automated processes to evade the

intended safeguards entirely. Additionally, it can lead to challenges in ensuring uniform application in the EU and may undermine the broader objectives of transparency. As a result of this vagueness, it is far from clear what the right includes. If interpreted in a narrow sense, the obligation to provide an explanation may apply only after an automated algorithmic decision has already been made and affected the data subject. On one hand, such an post hoc explanation would arguably reflect the GDPR's intent to prioritize remedies for data subjects, facilitating the realization of their rights to contest the decisions. On the other hand, this approach risk placing the main focus on the remedies in the aftermath, and thus, limit the potential for systemic transparency improvements in the AI system's design. In contrast, a more extensive interpretation of the "right to explanation" advocates for more proactive means that inform individuals at earlier stages of the decision-making process. However, this is rather difficult to achieve given the nature of modern deep learning models. The explanations generated by these systems may lack the coherence and comprehensibility needed to fulfill the requirements of the "right to explanation". To summarize, no matter how we choose to interpret the transparency and explainability provisions, we are faced with complicated issues. Addressing these concerns requires a coordinated system of governance that involves perspectives from different groups of stakeholders. Regulatory bodies cannot operate independently to achieve transparency and explainability. Instead, they must work in cooperation with industry leaders, researchers, civil society and data engineers. By fostering collaboration among stakeholders, we have a chance at increasing a sufficient level of transparency and explainability when it comes to algorithmic AI systems.

3.4.2 When should the explanation be disclosed?

Another question that is left unanswered in the text of the GDPR is when the explanation should be disclosed. It is safe to assume that this must occur at some stage in the lifecycle of the algorithmic AI systems. Such specification can take place during the design phase or be during its use. As mentioned before, a way to categorize explanations is based on when they are provided. There are ante hoc explanations and post hoc explanations. This distinction is of great importance for determining appropriate transparency mechanisms required in each given context. Post hoc explanations are arguably more relevant for the discussion of GDPR's transparency requirements, as they offer individuals meaningful insights into the decisions that directly impact them.

Regarding this topic, Wachter et al. have held that only an ante hoc explanation would be required, if a “right to explanation” existed in the GDPR.¹³⁰ However, this is challenged by Brkan who further examines this question by attempting to interpret the language of the provisions within the GDPR. According to her, Articles 13(2)(f) and 14(2)(g) of the GDPR refer to the “existence” of automated decision-making, a phrase that suggests such decision-making has already taken place when the information should be disclosed to the data subject. Brkan argues that if the legislative intent had been to require prior notification of intended automated decisions, it would have been phrased accordingly by using language such as “intended” automated decision-making or similar expressions. Consequently, the language used would support the argument that post hoc explanations are mandated according to the provisions of the GDPR. Moreover, Brkan aligns her interpretation with the opinion of WP 29¹³¹, which advocates for the necessity of providing sufficiently detailed information to the data subject, in order for them to grasp the rationale behind the automated decision.¹³² This perspective underscores that the GDPR, while fostering transparency and accountability, focuses on justifications concerning decisions that have been made, rather than explanations about the functionalities of the system.

Brkan uses a similar line of reasoning when examining Article 15(1)(h). This article too is inherently ambiguous concerning the timing of the disclosure of information. In this respect, she highlights that the Article lacks a specific timeline, which would suggest that the data subject’s right to access is triggered after an automated decision has been made. Without the explanation for the decision, the right would be highly ineffective. Worth mentioning is that this interpretation is further confirmed by the WP 20, which asserts that data subjects must be able to understand the reasons for decisions affecting them to be able to exercise their rights effectively. Additionally, Brkan claims that the term “envisaged consequences” in said article does not undermine but rather supports the above interpretation. After an automated decision has been made, the data controller cannot be aware of all the possible consequences of the outcome. At best, the controller can only explain the consequences anticipated at the time of the disclosure. To illustrate this point, Brkan paints a scenario involving a university admission. An automated rejection may have minimal impact on an applicant with several other opportunities, but

¹³⁰ Wachter et al., 2017, p. 78.

¹³¹ W P29, 2018, p. 25.

¹³² Brkan, 2019, pp. 113–114.

still, it could be devastating for a disabled applicant with that relied on this university due to its accessibility features. This example underscores the limitations of the controller's inability to foresee all the possible consequences that the decision could lead to, hence justifying the use of the term "envisaged consequences".¹³³

I find Brkan's interpretation of Articles 13(2)(f), 14(2)(g), and 15(1)(h) of the GDPR persuasive, as it offers a balanced approach to the rights of data subjects, while taking into consideration the practical challenges faced by data controllers. I agree that the phrasing "existence" of automated decision-making would imply that there is a post-decision context. To me, this phrasing reflects a conscious choice made by the legislator to focus on the transparency after the decision has been reached, rather than imposing a potential obligation that could lead to confusion. If terms such as "intended" or "proposed" automated decision-making would be used instead, it would be different. Additionally, the fact that Brkan's opinion aligns with WP 29's statement further convinces me. By requiring that data controllers provide sufficient information for data subjects to understand the justification behind certain automated decisions, the GDPR ensures meaningful transparency. As Brkan argues, the provision is most effective when interpreted as obliging the data controller to provide a post-decision explanation. Without such an explanation, the provision would quickly lose its significance, particularly in scenarios where the data subject's need for information arises only after being affected by the decision. Regarding Brkan's analysis of the term "envisaged consequences", I agree as well. As aforementioned, automated decision-making processes sometimes involve unintended or unpredictable outcomes. Controllers can only be expected to share information about the consequences they are able to foresee at the time of the disclosure. Higher expectations would not be fair, nor proportional, in my opinion.

3.4.3 The nature and scope of "meaningful information"

When it comes to the question of what the "right to explanation" should entail in practice, the text in the GDPR gives little to no guidance. The absence of a clear and explicit "right to explanation" within the framework opens up the risk of misinterpretation of what the legislator intended. An explicit acknowledgement of this right would have contributed significantly to the principle of legal clarity, an essential cornerstone of effective democracy and regulation. However, since we have concluded that the "right to

¹³³ Brkan, 2019, p. 114.

explanation” is derived indirectly from various provisions in the GDPR, combined with the recitals, we will have to closely interpret the text of these provisions. Of particular importance is the term “meaningful information about the logic involved” found in Article 13 of the GDPR. Understanding the term “meaningful information” in the context of automated decision-making processes is vital for the discussion of a “right to explanation”. When interpreting this term, I find it important to keep in mind the primary aims of the GDPR; to safeguard the data subject’s rights, as well as the overarching objective of transparency that permeates the legal framework.

The task of defining and satisfying the requirement for “meaningful information” in the context of deep learning algorithms presents several practical challenges. The fact that many algorithms and AI systems today are inherently opaque and complex does not make it easier. In addressing this challenge, one must consider the perspective of data subjects. What matters is that the information provided is meaningful to them. The disclosure of highly detailed underlying codes and mechanisms behind the decision-making process is unlikely to make sense to the data subjects. These forms of disclosure are often inaccessible to individuals without advanced, and sometimes no, technical expertise. Consequently, they fail to facilitate the data subject’s understanding for the logic behind the automated process.¹³⁴ Instead, the word “meaningful” could be understood as information that is understandable and significant.¹³⁵ However, there are multiple ways this could be applied in practice.

Scholars Selbst and Powles have proposed a framework to better understand the concept of “meaningful information”. In this respect, they identified four criteria that should be fulfilled in order to categorize information as meaningful. First and foremost, since Article 13-15 GDPR relate to the right of the data subject, meaningful information should be understood and evaluated from the perspective of the data subject. This person should be assumed to be someone without any broader technical knowledge. Second, the information should be functional in its nature. It should have a direct correlation with the rights of the data subject and enable them to act on their rights, such as contesting an automated decision under Article 22(3) GDPR. The values of the explanations should be both intrinsic and instrumental. While the intrinsic value serves the purpose of autonomy and self-determination, the instrumental value focuses on the practical outcomes. Third, the authors argue that meaningfulness should meet a minimum threshold of detail. For

¹³⁴ Kuner et al., 2017, p. 2.

¹³⁵ Malgieri & Comand , 2017, p. 257.

instance, explanations must include enough detail to function as an empowerment for the data subject to understand whether an actionable claim, such as discrimination, potentially exists. This criterion is closely linked to Article 5 GDPR and its requirements for lawful, fair and transparent data processing, as well as Article 12 and its emphasis for intelligibility and accessibility. Lastly, the fourth criteria refer to flexibility. It is important to highlight that rigid standards concerning explanations could slow down innovation in the field of AI. Demanding detailed explanations of decision-making processes, might render some efficient models impractical, not least if they are made of complicated neural networks. To prevent this from happening, the term “meaningful explanation” should be interpreted in a functional and flexible manner.¹³⁶

In my opinion, Recital 58 of the GDPR further supports the point made by Selbst and Powles. According to the recital, the principle of transparency requires that any communication directed toward either the public or the data subject must be “concise, easily accessible and easy to understand” and that “clear and plain language” should be used. In addition, visualization is also encouraged to aid understanding when appropriate. This phrasing adds to the belief that the language used when explaining the automated decision-making processes ought to be comprehensible for the average person. It also shows that the transparency requirement places a special focus on the roles of the “data subject” and the “public” as the main recipients of that information. In other words, the disclosed information must not merely exist, but also hold substantive meaning for its intended recipients, namely the public. However, it is certainly difficult to draw the line of how detailed the information should be for it to be considered meaningful.

¹³⁶ Selbst & Powles, 2017, pp. 236–237.

4 Technical and legal challenges of algorithmic transparency and explainability

4.1 Introduction

The “right to explanation” under the GDPR brings about a number of practical issues. Among the most pressing ones is the determination of the scope and nature of what information should be disclosed, especially given context of AI-driven decision-making. From a theoretical perspective, the challenges associated with explaining the logic behind AI’s operational mechanics are enhanced in the context of algorithmic AI models. These difficulties are widely acknowledged and can be attributed to numerous factors. For instance, the extensive scale of data involved, combined with the nature of algorithms, creates an exceptionally complex landscape. This chapter aims to answer research question 3. It focuses on the technical challenges posed by deep learning systems in complying with GDPR’s transparency and explainability requirements. I will be providing an in-depth exploration of why deep learning models are inherently opaque and how this makes it more difficult to achieve compliance. The legal implications of these technical challenges are analyzed, particularly their potential to create liability gaps or enforcement limitations under the GDPR.

4.2 The “Black box” problem

As algorithmic decision-making becomes increasingly prevalent, the need to demystify the inner workings of AI systems has become critical.¹³⁷ The rapid advancement of algorithmic technologies has introduced several new challenges, one of them being the growing opacity of operational mechanisms within automated AI systems. Traditional analytical and predictive models, such as linear regressions or simple correlations, are interpretable, allowing for us to understand and replicate it. In contrast, modern methods, such as neural networks, operate within what is often described as a “black box”.¹³⁸ These techniques, especially those with extensive non-linear parameters, are far less comprehensible to human understanding. As the complexity of these parameters increase, so does the difficulty in deciphering the model’s decision-making process, further

¹³⁷ Busuioc et al., 2023, pp. 79-80.

¹³⁸ Wulf & Seizov, 2024, p. 235.

complicating efforts to explain such models to stakeholders, let alone non-experts.¹³⁹ In general terms, the black box problem can be described as an inability to fully comprehend an AI system's decision-making process, as well as the inability to foresee the AI's outcome.¹⁴⁰ The methods many AI systems operate with today are inherently opaque, which makes it difficult to provide clear explanations or traceable information for the generated content, as the reasoning behind the decisions are usually incomprehensible.¹⁴¹ Critics have highlighted a crucial tension in this respect; the same attributes that make these AI systems highly effective, such as large-scale neural networks in deep learning models, also make them highly opaque.¹⁴²

The construct of deep neural networks is based on a mathematical model known as the artificial neuron. Initially, this construct was based on the neurons in human and animal brain. However, it is not intended to function as a computational replica of a biological neuron. Rather, the primary goal of the artificial neuron lies in replicating the ability to learn from its past experience, similar to the biological neuron.¹⁴³ Worth mentioning is that multi-layered networks of this multitude consisting of interconnected artificial neurons became a reality once advancements in this field were made in the mid-1980s. At that time, researchers rediscovered and refined methods enabling the training of such networks.¹⁴⁴ These breakthroughs then led to the development of deep neural networks, which consist of multiple interconnected layers of artificial neurons, working together to identify patterns to establish logical connections between different types of data. By relying on the cooperative functioning of multitude of neurons, the AI system arrive at a decision.¹⁴⁵

Grasping the inner working of algorithmic AI systems have been compared to the intellectual challenge of studying a non-human highly intelligent species. Such a species would presumably possess senses and abilities alien to ours, making it a unique challenge to understand its thought process that is most likely profoundly different from human cognition.¹⁴⁶ This is not far from how we perceive AI systems. While inputs and outputs of AI systems can be tracked, the underlying neural pathways and their contributions

¹³⁹ Chaudhary, 2024, p. 106.

¹⁴⁰ Bathaee, 2018, p. 6.

¹⁴¹ Burrell, 2016, pp. 1–4.

¹⁴² Bathaee, 2018, pp. 14-15.

¹⁴³ Bathaee, 2018, p. 5.

¹⁴⁴ Widrow & Lehr, 1990, pp. 1415–1416.

¹⁴⁵ Bathaee, 2018, p. 5.

¹⁴⁶ Bathaee, 2018, p. 2.

remain unknown to us. In other words, the connections responsible for producing certain decisions is concealed. For this reason, engineers themselves often struggle to determine which variables hold the greatest weight for a given decision.¹⁴⁷ Some authors have compared this complexity to how individuals learn complex tasks, such as riding a bicycle. The ability to balance and steer a bike is not something that can be explained to you by given instructions. Instead, people learn by repeated attempts and the gradual development of an intuitive understanding. Similarly, the internal mechanisms of AI systems can rarely be reduced to formal rules or instructions¹⁴⁸, since it learns experientially.¹⁴⁹ Nor can one reliably identify specific neurons or neuron clusters responsible for highlighting particular types of data.¹⁵⁰ Instead, the unique abilities of these systems derive from “connectionism”, the notion that complex systems can consist of numerous simple components.¹⁵¹ This complexity is the same one that gives rise to the black box problem¹⁵², making it difficult to achieve algorithmic transparency. If AI systems surpass human understanding entirely, it will give rise to a fundamental dilemma for regulatory bodies. The assumption that these systems can be regulated using human-centric concepts is directly dependent on the fact that we can understand them, at least partly.¹⁵³

The idea of “algorithmic accountability” rests on the belief that understanding the mechanisms by which algorithms function is a necessity for their effective oversight.¹⁵⁴ This notion is directly linked to the pursuit of algorithmic transparency, which aims to make the algorithmic processes fair, accurate and unbiased.¹⁵⁵ However, as technological advancements result in more sophisticated and complex algorithms, achieving such transparency grows increasingly difficult.¹⁵⁶ In practice, many organizations use AI models with black box techniques, where the logic behind the decisions are concealed. These techniques have become more common in numerous high-stakes domains, such as finance, education, employment, and law enforcement. The results generated by such systems are characterized by a lack of clarity, posing significant challenge for

¹⁴⁷ Castelvechi, 2016, p. 21.

¹⁴⁸ Bathaee, 2018, p. 5.

¹⁴⁹ Bathaee, 2018, p. 2.

¹⁵⁰ Bathaee, 2018, p. 5.

¹⁵¹ Goodfellow et al., 2016, p. 36.

¹⁵² Bathaee, 2018, p. 5.

¹⁵³ Bathaee, 2018, p. 14.

¹⁵⁴ Diakopoulos, 2016, pp. 60-61.

¹⁵⁵ Islam et al., 2021, p. 12.

¹⁵⁶ Ananny & Crawford, 2018, p. 981.

accountability.¹⁵⁷ As a matter of fact, even the organizations responsible for deploying these technologies often unable to provide coherent explanations for their decisions.¹⁵⁸ In the absence of clear insight into how these algorithms operate, it becomes inherently difficult to audit their decisions, address harms, or guarantee fair treatment. Consequently, the demand for transparency has emerged as both a strictly ethical imperative, as well as a foundational requirement for the responsible integration of AI into society. As algorithms become more difficult to explain, the calls for explanations to the public grow.¹⁵⁹ The transition to so called “glass box” AI systems, designed with interpretability and comprehensibility in mind, has become a priority.¹⁶⁰

4.3 Trade-offs between performance and explainability

The recent increase of interest in deep learning and other models classified as highly opaque has brought about a hype, often overshadowing simpler yet effective interpretable models. In practice, these interpretable models have frequently demonstrated their ability to achieve good results in various important tasks¹⁶¹, which raises questions about the necessity of relying on less transparent and interpretable models. However, these models lack the predictive power¹⁶², particularly in scenarios requiring deep analysis of vast amounts of datasets.¹⁶³ It is still not proven possible for interpretable models to match the performance of their opaque counterparts in all fields.¹⁶⁴ Instead, one could argue that the increased use of complex machine-learning based models implies the opposite – that these systems are not only tools for convenience but rather that they offer genuine practical advantages over interpretable models. If this is the case, mandating interpretability as a universal standard could come at a cost. At worst, it could hinder technological advancements or decrease the accuracy of AI systems. Forcing developers to prioritize high standards of transparency might result in unwanted outcomes, as less complex architectures may fail to deliver the desired functionality. In this sense, regulations that impose strict transparency requirements risk limiting innovation and the progress of AI

¹⁵⁷ Rudin, 2019, p. 1.

¹⁵⁸ Ananny & Crawford, 2018, p. 981.

¹⁵⁹ de Bruijn et al., 2022, p. 1.

¹⁶⁰ Franzoni, 2023, p. 118.

¹⁶¹ Rudin, 2019, pp. 3–4.

¹⁶² Adadi & Berrada, 2018, p. 52145.

¹⁶³ Wang, 2023, p. 8-9.

¹⁶⁴ Puiutta & Veith, 2020, p. 82.

technologies.¹⁶⁵ This dilemma underscores the trade-offs between efficiency and transparency¹⁶⁶, as well as the difficulty of balancing the need for transparency with the innovation of technology.

There is no doubt that we must strike a balance between explainability and performance, but the question is how. Implementing strict transparency requirements will most likely have an effect on AI design. By mandating transparency and explainability as a prerequisite for regulatory compliance, legislators limit the design choices available to algorithmic system developers. This limitation may give designers strong incentives to prioritize interpretability and explainability over performance, leading to them creating less accurate AI systems. By doing so, regulators risk becoming unintentional gatekeepers of technological innovation – a role they are not fit to fulfill. This problem can be avoided to some extent by regulators having an active conversation with developers and experts in the field to establish the best practices for accountability without hindering innovation more than absolute necessary.

Additionally, the trade-off dilemma becomes even more relevant in high-stake contexts where decisions carry heavy ethical and social weight. In my opinion, the necessity for transparency transcends the need for operational efficiency and performance in these situations. For instance, the need for explainability and justification in the health care, social justice and employment is not merely desirable, but rather an essential safeguard. For instance, opacity in the judicial systems could undermine fundamental principles of due process and equal treatment, whereas opacity in medical diagnostic tools could compromise a patient's trust for the system as a whole. In other words, these domains demand a higher threshold of accountability, as the consequences of algorithmic decisions have direct and profound impact on individual rights and their trust in society. Without transparency, the logic behind algorithmic decisions remains inscrutable, leading to difficulty to demand accountability. Opacity is thus not merely an inconvenience in these fields, but a valid and direct threat to the legitimacy of outcomes. For this reason, I believe the demand for transparency should be seen as an obvious component in the contextual frameworks of the field in which the algorithmic AI system operates.

Furthermore, the notion of efficiency could do well with some reforming. In high-stake contexts, transparency and performance should not be put against each other, as that would paint the false narrative that they are mutually exclusive. Transparency, when

¹⁶⁵ Bathaee, 2018, pp. 14-15.

¹⁶⁶ Wang, 2023, p. 8.

embedded in an algorithm's design, can serve as a catalyst for an AI system's functionality and efficiency. By enabling and facilitating stakeholder's understanding for the underlying logic of the system, we can foster trust and easier align systems functions with desirable societal values. To summarize the argument so far, the ethical and legal issues concerning the trade-off between explainability and performance demand a contextual analysis that include the specific standards of the specific industry involved. By considering the context, we not only address immediate accountability and transparency concerns, but also lay the foundation for systems that are both efficient and aligned with human values and societal expectations.

4.4 Additional explainability and interpretability challenges

The challenge of ensuring explainability for AI systems depends on both the technical aspect of explanation and the audience's ability to understand. In order to achieve sufficient explainability, the intended audience must possess a baseline understanding of basic technical concepts. However, as algorithms continue to grow more complicated and accurate, the level of required knowledge also increases. The problem arises when the level of knowledge needed is too much to ask from the average individual. This trend risks excluding a huge part of the general public who might not comprehend the explanations, and thus, undermining the very objective of an explanation. On top of this, the terms "transparency" and "explainability" are interpreted differently by different types of groups, further complicating the matter. An explanation deemed sufficient for one group of individuals, may fall short to another. For instance, for some, technical clarity might be a priority, while other value moral or ethical framing over the technical aspects. This dilemma is a good reminder that the pursuit of explainability is not merely technical; it is just as much about navigating deeply integrated cultural and societal differences.¹⁶⁷ As algorithmic AI systems start making crucial decisions in fields with significant societal consequences, its regulation must take the subjective characteristics of an explanation into consideration.

Another obstacle arises from the concept of decontextualization and the dynamic nature of algorithmic AI. This mainly occurs when an algorithm is initially designed for one purpose but then repurposed for entirely different objectives or applied in a different

¹⁶⁷ Selbst et al., 2019, pp. 60–62.

contextual framework.¹⁶⁸ The dynamic nature of algorithms is a consequence of their autonomous characteristics and ability to mimic certain cognitive patterns. This implies that algorithms are not static but rather evolve over time and may depend heavily on the specific conditions under which they are used. Unlike static systems, algorithmic AI systems evolve drastically, changing their outputs as a result of new inputs. This creates an issue of “lack of explanation robustness”, as it is difficult to maintain consistency, clarity and accuracy when giving explanations.¹⁶⁹ An explanation that was once deemed sufficient in one context, may lose its accuracy and validity in a different scenario. Moreover, the aforementioned black box nature of modern AI systems further complicates the issue, complicating the efforts to achieve consistent and stable explanations.¹⁷⁰ This too could undermine the coherence and consistency of transparency efforts customized for the original purpose.

The concept of AI transparency and explainability, is frequently framed, at least partially, as a technical challenge. This narrative is valid, as several technical factors inherently contribute to the opacity of AI systems, making them incomprehensible to the general public.¹⁷¹ Even some of the simpler AI systems are, by their very nature, a complex mystery for the average human mind, due to their mathematical algorithms and programming code. These are not likely to be comprehensible to those lacking specialized knowledge.¹⁷² In fact, the possession of advanced technical expertise does not guarantee clarity or transparency either.¹⁷³ Experts themselves may face significant challenges trying to demystify the inner working of an AI system¹⁷⁴, especially when it includes vast volumes of data.¹⁷⁵ Given these challenges, it is understandable that regulatory frameworks emphasize the importance of addressing technical opacity in algorithmic AI systems.

¹⁶⁸ Selbst et al., 2019, p. 66.

¹⁶⁹ Panigutti et al., 2023, p. 1143.

¹⁷⁰ Dombrowski et al., 2019, p. 2.

¹⁷¹ Burrell, 2016, pp. 1–4.

¹⁷² Kolkman, 2022, p. 93.

¹⁷³ Caplan et al., 2018, p. 15.

¹⁷⁴ Castelvechi, 2016, p. 21.

¹⁷⁵ de Bruijn et al., 2022, p. 2.

5 Explaining algorithmic AI systems

5.1 Introduction

In this chapter, research question 3 is answered. I aim to examine the application of GDPR provisions to deep learning systems in details, addressing the challenges of aligning the requirements of Articles 13-15 and 22 GDPR with the technical realities of these systems. The discussion evaluates whether existing solutions, such as explainable AI (XAI), can adequately address the transparency and explainability provisions under the GDPR or whether gaps remain. In other words, the feasibility of implementing the GDPR's safeguards for algorithmic decision-making in deep learning systems is assessed from a critical perspective. Finally, to tie it all together, this chapter zooms in on the potential legal remedies and recommendations aimed at bridging the gap between legal and technical perspectives. It suggests ways to harmonize GDPR compliance with the opaque nature that follows from deep learning models. This is done by proposing regulatory clarifications or guidelines, as well as advancing XAI techniques.

5.2 The implementation of the “right to explanation”

The implicit “right to explanation”, embedded in the GDPR, is without a doubt a significant achievement in the global quest of promoting transparency and accountability in algorithmic decision-making. This right, although not explicitly articulated in a specific provision, can be derived from Articles 13-15 GDPR, read alongside Article 22 GDPR. Its objective is that individuals directly affected by algorithmic decision-making are entitled to receive meaningful information about the underlying logic, significance, and consequences of such decisions. Particularly in high-risk domains, transparency measures must be sufficient to minimize the risk for discrimination and other forms of harm. AI-driven systems used in areas such as healthcare, employment, military warfare, and law enforcement require higher standards of transparency. Only then can the individuals affected receive the proper tools to contest and understand the algorithmic decisions they are subjected to. However, the implementation of the “right to explanation” does not come without complications, as it involves an interplay of legal, technical, and ethical issues. Modern AI systems, particularly those supported by deep learning methodologies, are characterized by non-linear and complex neural networks, making them highly opaque.

Consequently, translating these algorithmic processes to something comprehensible for the average person is a challenge in itself. Furthermore, the objective of transparency has to be carefully balanced against the interests of embracing innovation.

Given the vagueness of the “right to explanation” in the GDPR, it is difficult to interpret the provisions supporting such a right. There is a serious lack of guidance and uncertainty regarding the principles of transparency and explainability, leading to issues with the implementation. Since the provisions can be interpreted in multiple different ways, stakeholders cannot prioritize where to place their efforts. Some argue for the prioritization of inherently transparent and explainable models, claiming that these systems are less prone to opacity. However, others advocate for post hoc explanations to disclose the reasoning behind an AI system’s decisions for the individuals involved. The lack of consensus on a definitive strategy to achieve increased transparency complicates the efforts to create comprehensive regulatory guidelines, as various stakeholders interpret the provisions differently within the legal framework of the GDPR.

Several suggestions have been made for achieving qualified and sufficient transparency in algorithmic AI systems. Firstly, the notion of explainable AI (XAI) is a crucial tool in making algorithmic AI systems more comprehensible to non-specialist users. It involves the development of systems capable of articulating the logic behind their outputs. This concept will be explored further in the following section. Second of all, the adoption of interactive dashboards and data visualization tools have been proposed for users to actively engage with algorithmic outputs. These tools make it possible for individuals to assess the correlation between data inputs and the final decisions, enabling a more transparent use of the AI systems.¹⁷⁶ A third strategy involves conducting algorithmic impact assessments. This could systematically help with identifying the risks to fundamental rights and thereby functioning as a safeguard for proactive compliance.¹⁷⁷ Another suggestion for facilitating compliance is clear communication with users regarding their rights and the functionality of the algorithmic AI systems. By making the information accessible, users can effectively understand and assert their legal rights. Transparency is not only a procedural requirement; it is a cornerstone of trust in both private and public contexts where algorithmic AI systems are used.¹⁷⁸

¹⁷⁶ Chaudhary, 2024, p. 111.

¹⁷⁷ Selbst, 2021, p. 2.

¹⁷⁸ Chaudhary, 2024, p. 111.

5.3 The concept of explainable AI (XAI)

In response to the need for transparency and explainability in algorithmic AI models, Explainable AI (XAI) was developed as a technical solution.¹⁷⁹ In recent years, XAI has become an increasingly discussed topic, drawing attention from both academic researchers and industry practitioners. Despite the growing interest for this notion, however, there is still no reached consensus about the definition of the term “XAI”. Instead, XAI is characterized by its overarching objective, which is to make AI systems more explainable and comprehensible to the general public. In the legal doctrine, XAI has been described as “the movement, initiatives and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept.” XAI thus addresses an urgent need: it enables the existence of coherent justifications for algorithmic AI systems, particularly in case of unexpected outcomes. By doing this, XAI allows stakeholders to verify that decisions comply with ethical and procedural standards, contributing to fairness and accountability. As a consequence, people might feel more confident in automated systems and learn to trust them.¹⁸⁰ This is of particular importance in high-stakes fields, such as healthcare and finance, where the demand for interpretability cannot be stressed enough. Specialists operating in these domains require not only assistance in facing complex challenges, but also outputs that are explainable, in order to foster trust for algorithmic decisions.¹⁸¹ Equally important is the capacity to proactively identify risks, such as algorithmic bias and discrimination. For this reason, XAI systems should be designed with the principles of fairness, accountability and transparency in mind.¹⁸² These principles ensure that the explanations disclosed to the stakeholders and the public not only meet technical and legal criteria but also align with societal expectations of justice.

There are numerous XAI methodologies used to effectively interpret AI systems. One frequently used method is the attempt to “open” black boxes. This refers to the challenge of gaining insight into the functions of algorithms, that are often invisible to us.¹⁸³ When dealing with black boxes, explainability is often achieved through post hoc methods.¹⁸⁴ These methods do not aim to explain the underlying functions of the AI system but rather

¹⁷⁹ Busuioc et al., 2023, p. 89.

¹⁸⁰ Adadi & Berrada, 2018, pp. 52140-52142.

¹⁸¹ Ali et al., 2023, p. 2.

¹⁸² Islam et al., 2021, p. 12.

¹⁸³ Rudin, 2019, p. 1.

¹⁸⁴ Panigutti et al., 2023, p. 1143.

to generate explanations after the decision is made to facilitate understanding.¹⁸⁵ A common method for post hoc explanations is the construction of explanatory models, or algorithms. These algorithms aim to provide simplified representations of a certain AI system, making the behavior of black box algorithms more comprehensible to human understanding.¹⁸⁶ In doing so, they allow AI engineers to convey information about algorithmic decision-making without having to disclose the inner working of the black box model.¹⁸⁷ This is especially valuable when providing insights into a deep learning model's behavior. These systems often rely on completely raw or minimally processed datasets, making it difficult to trace their decision-making pathways in a way that is comprehensible.¹⁸⁸ Furthermore, various techniques can be used to achieve such comprehensibility. Among them are natural language explanations, which translate the AI system's rationale into language that we understand; visualizations, which help illustrate patterns or decision criteria used in categorization; and example-based explanations, which draw parallels to similar cases to give context.¹⁸⁹

The many benefits of XAI gives rise to a compelling question: why has the implementation of XAI not yet become a universal practice? Put differently, why has XAI not been adopted by all stakeholders despite its evident benefits? The answer lies in the technical challenges with achieving transparency and explainability in algorithmic AI systems.¹⁹⁰ One challenge is the issue of scalability, which arises when the need for explanations increases on a large scale.¹⁹¹ Another major challenge, as mentioned before, is the robustness of generated explanations. In cases where robustness is lacking, instabilities may arise, leading to inconsistency and incoherence. This instability might, in turn undermine the trustworthiness of the system and complicate the efforts of achieving accountability. In addition, post hoc XAI-explanations often serve as approximations rather than precise replications of the model's decision-making logic. Scholars are still discussing whether the explanations produced by current methods accurately represent the internal working of algorithmic AI systems.¹⁹² In fact, studies have shown that attention weights within certain models do not align with gradient-based

¹⁸⁵ Lipton, 2018, p. 15.

¹⁸⁶ Rai, 2020, p. 138.

¹⁸⁷ Busuioc et al., 2023, p. 89.

¹⁸⁸ Lipton, 2018, p. 20.

¹⁸⁹ Lipton, 2018, p. 15.

¹⁹⁰ Adadi & Berrada, 2018, p. 52144.

¹⁹¹ Saeed & Omlin, 2023, p. 9.

¹⁹² Panigutti et al., 2023, p. 1143.

feature attributions, indicating that the model might be focusing on features other than those identified by attention mechanisms. This gap between interpretative and actual features underscores the challenge of achieving genuine transparency with the help of XAI.¹⁹³

Worth mentioning in this respect is that the approach of understanding black boxes has its critics. There are those who argue that we should focus on designing so called white boxes, or glass boxes, instead of trying to convert black boxes to something more comprehensible.¹⁹⁴ For instance, Rudin suggests that rather than focusing on explanations of opaque models, it may be more sensible to prioritize models that are designed to be interpretable to begin with.¹⁹⁵ This would benefit not only users, but also developers. Developers, who are responsible for the accuracy and reliability of the systems, benefit from inherently explainable systems, especially when errors are found. Should a model produce an error, it is easier to investigate the source of failure and solve it.¹⁹⁶ Moreover, using an inherently interpretable model contribute to three additional functions in AI research. It fosters the evaluation and validation of existing knowledge, it advances that knowledge through refinement, and it generates new theories.¹⁹⁷ From a functional point of view, explainability enables researchers and developers to achieve justification, control, enhancement, and discovery. A model that can be both understood and closely inspected has the ideal conditions to continuous optimization.¹⁹⁸ The implantation of inherently interpretable models, however, comes with its own challenges. It is not proven that all decision-making scenarios will be equally accurate and productive using such models. In some cases, performance may be compromised by a less complex system structure.¹⁹⁹ Consequently, the emphasis on transparency can sometimes result in reduced predictive accuracy.²⁰⁰ In addition, there are factors beyond the technical limitations that should be considered. For instance, the potential success with this approach is heavily reliant on the willingness of both private and public sector agencies to make deliberate choices that prioritize transparency and explainability.²⁰¹

¹⁹³ Jain & Wallace, 2019, p. 1.

¹⁹⁴ Ali et al., 2023, p. 3.

¹⁹⁵ Rudin, 2019, p. 1.

¹⁹⁶ Ali et al., 2023, p. 2.

¹⁹⁷ Rieg et al., 2020, p. 12.

¹⁹⁸ Adadi & Berrada, 2018, pp. 52142–52143.

¹⁹⁹ Puiutta & Veith, 2020, p. 82.

²⁰⁰ Adadi & Berrada, 2018, p. 52145.

²⁰¹ Rudin, 2019, pp. 10–11.

In light of the foregoing considerations, it is evident that despite significant advancements in AI, the current state of XAI remains underdeveloped in terms of end-to-end integration. Part of the reason for this is the large emphasis on the technical aspects of AI systems in current research. While this focus is crucial for system development, it easily overlooks the role of communication, which is essential to foster trust in AI applications. Without mechanisms for user interaction and engagement, such as systems that offer detailed explanations and opportunities for feedback, AI systems risk being perceived as unreliable. Consequently, the development of interactive and user engaging XAI frameworks is foundational to the successful integration of XAI.²⁰² By designing transparency and accountability mechanisms into the systems, society can learn to trust these technologies and navigate the challenges they pose with greater confidence.²⁰³ Furthermore, facing these challenges necessitates a collaborative approach that integrates the research, insights, and expertise of diverse fields.²⁰⁴ Computational science offers technological insights, while the humanities bring the ethical aspects of XAI. Effective collaboration between experts in these fields will be required to meet the complex demands of transparent and explainable AI systems.

5.4 The case of algorithmic governance

Given the aforementioned impact of algorithmic decision-making on modern society, legal frameworks are necessary for confronting the complex challenges of transparency and explainability. However, the discussion of potential legal remedies must be preceded by a discussion about the governance structures that oversee algorithmic systems.²⁰⁵ Governance is central to minimizing algorithmic harms, fostering public trust, and ensuring stability across various social institutions. Regulations, when designed carefully and considerately, do not limit innovation. Instead, they support innovation by guiding the creation of robust and clear legal frameworks. In addition, policy initiatives must address how human responsibility should be assigned and maintained in systems that are increasingly driven autonomously.²⁰⁶ As decision-making processes become more and more automated, the legal frameworks governing such processes must evolve to ensure that human oversight is properly integrated and enforced.

²⁰² Ali et al., 2023, p. 41.

²⁰³ Chaudhary, 2024, pp. 93–94.

²⁰⁴ Adadi & Berrada, 2018, p. 52145.

²⁰⁵ Taeihagh, 2021, p. 147.

²⁰⁶ Theodorou & Dignum, 2020, p. 10.

The overarching objective of algorithmic governance is to safeguard societal well-being while ensuring that AI's integration into various domains aligns with principles of sustainability and responsibility. This process requires an understanding of AI as a component of a socio-technical system that does not exist in isolation. Instead, it is a part of a socio-technical system consisting of multiple human and institutional actors. All these actors together contribute to shaping the ethical and functional frameworks of algorithmic AI systems. Consequently, the quality or trustworthiness of AI systems cannot be attributed to the technological artefact itself. The responsibility for the consequences of an AI system's decision-making and actions rather lies with the human and institutional actors involved in their development and deployment. The argument that the AI system itself should bear responsibility for its outputs is flawed, as it risks diminishing the accountability of individuals and organizations that need to act in accordance with established legal frameworks.²⁰⁷ Furthermore, the implementation of governance mechanisms for AI must be based on collaborative efforts involving multiple stakeholders, for it to be effective. Institutional bodies, private sector actors, academia, and civil organizations should all be included to achieve transparency, explainability and accountability for AI systems. Together they bring to the table various perspectives on potential AI risks and opportunities, creating a necessary discussion to establish governance structures that balance innovation with risk mitigation. However, the process of designing a regulatory framework for AI remains a major challenge. The inherent unpredictability of AI-driven algorithms, combined with their nonlinear and complex characteristics, makes it rather difficult to create precise policy objectives.²⁰⁸ Regulatory efforts must therefore be both dynamic and adaptable, leaving room for ongoing research and evaluations.

Moreover, a well-structured governance framework cannot be limited to the national jurisdictions, as AI is a highly global phenomenon. The international nature of AI development, characterized by a competitive landscape²⁰⁹, requires global efforts. Without such an approach, regulatory inconsistency and jurisdictional conflicts risk undermining legal clarity, as well as the main objectives of AI governance. For reference, the current global AI regulatory landscape remains fragmented, with various national strategies creating inconsistencies and ambiguities. This approach is not equipped for AI,

²⁰⁷ Theodorou & Dignum, 2020, p. 10-11.

²⁰⁸ Gasser & Almeida, 2017, pp. 58–60.

²⁰⁹ Garcia, 2024, p. 30.

given that the technology does not recognize national borders. Instead, enhanced international cooperation and universal standards are required to establish harmonized and unified AI governance at a global level. Such an effort would ideally be pursued by the United Nations (UN), given its well established role in fostering AI governance.²¹⁰ Standards play a key role in ensuring stability and enabling comparability across jurisdictions.²¹¹ The creation of globally recognized standards for algorithmic transparency and explainability would not only promote regulatory coherence, but it would also reduce the compliance burden for corporations.²¹² However, I believe the notion of algorithmic governance requires the establishment of robust legal frameworks that go beyond mere encouragement of ethical guidelines. The implementation of binding legal remedies is essential to achieve widespread compliance, instead of voluntary compliance. This is essential for safeguarding fundamental rights, ensuring corporate accountability and fostering trust in AI systems. Only by creating a comprehensive legal framework, embedded within a robust system of governance, can we hold stakeholders accountable for the deployment and consequences of algorithmic AI systems. For this reason, standards and guidelines should be perceived as secondary complements that enhance, rather than replace the existence of enforceable and binding regulatory obligations.

5.5 Potential legal remedies

The enforcement of legal remedies plays a key role in fostering ethical and responsible frameworks for AI. A fundamental aspect of legal remedies in the field of algorithmic AI is the implementation of transparency and explainability obligations on the developers and deployers of the AI systems. The level and scope of transparency and explainability requirements may differ depending on multiple factors, such as the characteristics of the algorithm, the consequences of the decision, and the degree of risk associated with its use.²¹³ Regulations can put pressure on organizations to disclose the methodologies employed by AI systems to arrive at specific decisions, in a clear and comprehensible manner.²¹⁴ This is of particular importance when algorithmic decisions are made in high-stake domains, potentially having vast impacts on specific individual's rights. The

²¹⁰ Tzimas, 2023, pp. 395-396.

²¹¹ Selbst, 2021, p. 21.

²¹² Tzimas, 2023, p. 395.

²¹³ Hogan et al., 2021 p. 8.

²¹⁴ Mittelstadt et al., 2016, p. 7.

transparency obligations in question may take various forms, as long as they include the key variables and parameters influencing the algorithmic decisions.²¹⁵ To facilitate understanding, regulations may oblige that stakeholders provide full disclosure regarding the datasets used to train the AI systems, as well as the potential biases embedded in the data.²¹⁶ By implementing these legal measures, regulators can effectively empower individuals with tools to critically assess the fairness of AI-driven decisions that directly affect them.

In this context, there is an existing debate concerning the scope and nature of the explanatory obligations that should be imposed on algorithmic AI systems. More specifically, should legal provisions focus on rendering AI systems themselves transparent by disclosing the inner functions of the systems, or should they focus on explaining the decisions reached by the systems? While the former may appear to be the preferable approach from a theoretical point of view, the latter is more realistically achievable in practice. However, a legal framework that mandates an explanation of the decision-making rather than the algorithmic structure would still require proof of causal relationships between the inputs and outputs of the decision. This approach demands a systematic analysis of how input data, algorithmic features, and internal processes reach a specific outcome. Such an obligation, though more feasible than an explanation of the inner workings of the algorithm, comes with its own challenges.²¹⁷

The primary risk with a strict regulatory framework requiring transparency disclosures seems to be that certain AI systems, by their very nature, are not explainable. Regulatory frameworks that demand a “meaningful” and sufficient explanation could therefore force developers to entirely abandon highly efficient but opaque models in favor of less sophisticated, more interpretable ones. At worst, strict transparency obligations could dramatically halt technological innovation. Compliance with these regulations would most likely mainly affect startups and small businesses, potentially leading to an environment where only well-established technology corporations can compete. In light of these challenges, a more nuanced approach may be required when regulating algorithmic AI. The need for AI accountability needs to be taken into consideration, while recognizing its inherent and unique complexities.

²¹⁵ Bibal et al., 2021, pp. 161–164.

²¹⁶ Tzimas, 2023, p. 398.

²¹⁷ Tzimas, 2023, p. 397.

There is no doubt that the GDPR introduces a set of rights aimed to ensure that data subjects are provided adequate protection against algorithmic decision-making in the forms of transparency obligations. Among these, the “right to explanation” stands out as one of the most significant. However, the absence of detailed guidelines on what constitutes a “meaningful” explanation result in legal uncertainty and compliance inconsistencies across various jurisdictions. As scholars continue to debate the specific scope of the legal obligations imposed by the GDPR in relation to algorithmic decision-making, the lack of consensus highlights a need for further clarification and harmonization. This is not only important, but a necessity to ensure compliance and accountability. In other words, this important right is far from assured in practice. As a result, organizations are allowed to offer vague or less-than-satisfactory explanations, making the “right to explanation” nothing more than a symbolic statement. To counter this outcome, courts could take the important role of shaping this right, by interpreting it in a manner that reinforces the necessity of transparency and explainability. By emphasizing principles linked to meaningful explanations, such as trust, accountability and interpretability, courts have the power to influence regulations and push them towards a direction of transparency. However, to ensure the enforceability of the “right to explanation” in the long run, standardized guidelines on what constitutes a sufficient explanation must be developed. For instance, the European Data Protection Board (EDPB) could issue more detailed interpretative guidelines precising the details of this requirement, particularly regarding when, how and what should be disclosed. Furthermore, officially acknowledging the “right to explanation” would be a great step in the right direction. Without these measures, the “right to explanation” risks remaining a theoretical construct rather than a robust safeguard for individuals subject to algorithmic decision-making.

The rapid adoption of AI across various domains and industries indicates that the expansion is unlikely to slow down anytime soon.²¹⁸ Meanwhile, the regulation of algorithmic transparency and explainability remains a subject of intense legal, technological, and ethical debate, particularly in light of the ever-evolving landscape of AI technologies. To address the issues that arise in relation to such regulation must evolve to take nature of the AI systems into consideration. While existing legal frameworks, such as the GDPR, provide some guidelines for algorithmic transparency and explainability,

²¹⁸ Grace et al., 2018, p. 729.

they remain insufficient and unsatisfactory in addressing the risks related to modern algorithmic AI systems. Comprehensive legal interventions are therefore needed to govern AI. These may include XAI methodologies, oversight mechanisms and enforced accountability for unfair algorithmic decision outcomes. Furthermore, interdisciplinary collaboration is crucial to bridge the gap between innovation and regulatory oversight. Bridging this gap is a fundamental prerequisite for ensuring that algorithmic AI applications serve the function of benefitting society, rather than harming it. As mentioned before, legal scholars, computer scientists, ethicists, and policymakers must come together to develop effective strategies for ensuring AI transparency and explainability. Ideally, this would be done without hindering or limiting technological advancement in the field of AI. The overarching aim should be to balance transparency with innovation, ensuring that AI development remains aligned with democratic values, fundamental human rights, and societal interests.

6 Final remarks

At the heart of today's discourse on AI lies the concepts of transparency and explainability – ideals that are frequently incentivized and encouraged but remains limited due to both technological and legal challenges. The inherent opacity, due to the black box nature of algorithmic AI systems, generates significant obstacles for legal accountability mechanisms. Some initiatives, such as XAI, have been developed in hopes of achieving transparency and explainability, but their efficiency remains limited. In practice, some of these methods have even shown to provide incorrect approximations rather than accurate explanations for a given outcome. The complexity of multi-layered neural networks in deep learning AI models makes it exceedingly difficult, if not impossible, to fully understand the causal chains of decision-making within these models. Given these limitations, legal scholars and policymakers might have to rethink their approach to algorithmic governance. While the GDPR and its “right to explanation” represents a major step forward in making AI systems more trustworthy and transparent, its effectiveness is undermined by the lack of precision in its transparency requirements. It is just as important to determine what constitutes an “explanation”, as it is to define what standards such explanations should meet in order to satisfy both legal and ethical obligations.

To ensure the GDPR achieves its said purpose, future legislative amendments or complementary guidelines must clarify the regulations. A failure to resolve this issue would make the “right to explanation” highly ineffective. Furthermore, the ability to measure the degree to which an explanation is comprehensible to the average person is not merely a theoretical issue but a necessity in ensuring compliance with transparency and explainability obligations. Further adding to this uncertainty are EU's newly adopted regulations concerning AI governance, introducing frameworks such as the AI Act, the data Governance Act, and the Data Act. It remains to be seen how these frameworks will interact with the GDPR as new AI technologies are constantly emerging.

In conclusion, it is evident that further research is needed in the field of AI, particularly deep learning-based decision-making. As AI systems increasingly influence decisions in high-stake domains, such as healthcare, finance, and employment, the need for a structured approach to understanding and regulating algorithmic decisions becomes ever more urgent. To do this, a fundamental shift in perspective is required. Instead of treating transparency as an absolute obligation, legal frameworks should pursue a more pragmatic

and realistic approach. We should strive for a meaningful level of disclosure without undermining the efficiency and accuracy of AI systems. However, achieving this balance is difficult due to the black box nature of deep learning systems. It will necessitate ongoing dialogue between legal scholars, AI developers, policy makers, and civil society. The burden of resolving this dilemma cannot fall solely on AI developers and engineers. While it is true that computer scientists must strive to develop more interpretable models, the policy makers of our society play a crucial role in shaping an environment where transparency is meaningfully integrated into existing frameworks. Only by working together, can we foster an ecosystem in which AI remains both innovative and trustworthy.

Bibliography

Literature

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160.

Agrawal, T. (2021). Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient (1 ed.). Apress.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805-.

Almada, M. (2023). Governing the Black Box of Artificial Intelligence, *SSRN Scholarly Paper No. 4587609*.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3), 973–989.

Auliya, S. F., Kudina, O., Ding, A. Y., & Van de Poel, I. (2024). AI versus AI for democracy: Exploring the potential of adversarial machine learning to enhance privacy and deliberative decision-making in elections. *Ai and Ethics*.

Avci, O., Abdeljaber, O., Kiranyaz, S., Hussein, M., Gabbouj, M., & Inman, D. J. (2021). A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. *Mechanical Systems and Signal Processing*, 147, 107077-.

Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. 80–89.

- Bathae, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent Causation. *Harvard Journal of Law & Technology*, 31(2), 889-.
- Benois-Pineau, J., & Petkovic, D. (2023). Explainable Deep Learning AI, *Academic Press*, 1-6.
- Bibal, A., Lognoul, M., de Streel, A., & Frénay, B. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29(2), 149–169.
- Brkan, M. (2019). Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond. *International Journal of Law and Information Technology*, 27(2), 91–121.
- Brkan, M. & Bonnet, G. (2020). Legal and Technical Feasibility of the GDPR’s Quest for Explanation of Algorithmic Decisions: Of Black Boxes, White Boxes and Fata Morganas. *European Journal of Risk Regulation*, 11(1), 18–50.
- Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1).
- Busuioc, M., Curtin, D., & Almada, M. (2023). Reclaiming transparency: Contesting the logics of secrecy within the AI Act. *European Law Open*, 2(1), 79–105.
- Caplan, R., Donovan, J., Hanson, L., & Matthews, J. (2018). Algorithmic Accountability: A Primer. *Data & Society Research Institute*.
- Casey, B., Farhangi, A., & Vogl, R. (2019). Rethinking Explainable Machines: The GDPR’s “Right to Explanation” Debate and the Rise of Algorithmic Audits in Enterprise. *Berkeley Technology Law Journal*, 34(1), 143–188.
- Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, 538(7623), 20–23.

Chaudhary, G. (2024). Unveiling the Black Box: Bringing Algorithmic Transparency to AI. *Masaryk University Journal of Law and Technology*, 18(1), 93–122.

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, 26(4), 2051–2068.

Cormen, T. H. (2013). *Algorithms Unlocked* (1 ed.). MIT Press.

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly*, 39(2), 101666-.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, 59(2), 56–62.

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (1 ed.). Springer International Publishing AG.

Dombrowski, A.-K., Alber, M., Anders, C. J., Ackermann, M., Müller, K.-R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. ArXiv.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2019). Accountability of AI Under the Law: The Role of Explanation.

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a “Right to an Explanation” Is Probably Not the Remedy You Are Looking for International. *Duke Law & Technology Review*, 16, 18–84.

Fergus, P., & Chalmers, C. (2022). *Applied Deep Learning: Tools, Techniques, and Implementation*. Springer International Publishing.

- Franzoni, V. (2023). From Black Box to Glass Box: Advancing Transparency in Artificial Intelligence Systems for Ethical and Trustworthy AI. In *Computational Science and Its Applications – ICCSA 2023 Workshops*, 118–130. Springer Nature Switzerland.
- Garcia, D. (2024). Algorithms and Decision-Making in Military Artificial Intelligence. *Global Society: Journal of Interdisciplinary International Relations*, 38(1), 24–33.
- Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. *IEEE Internet Computing*, 21(6), 58–62.
- Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): European Regulation that has a Global Impact. *International Journal of Market Research*, 59(6), 703–705.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. *MIT Press*.
- Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision Making and a “Right to Explanation.”, *AI Magazine*, 38(3), 50–57.
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research*, 62, 729–754.
- Hettne, J. & Otken Eriksson, I. (eds.) (2005). EU-rättslig metod, teori och genomslag i svensk rättstillämpning. Norstedts Juridik, Stockholm.
- Hjertstedt, M. (2019). Beskrivningar av rättsdogmatisk metod: om innehållet i metodavsnitt vid användning av ett rättsdogmatiskt tillvägagångssätt. In Mannelqvist, R., Ingmanson, S., Ulander-Wänman, C. (eds.), *Festskrift till Örjan Edström*, 165-173. Umeå University, Umeå.

Hogan, B. (2015). From Invisible Algorithms to Interactive Affordances: Data After the Ideology of Machine Learning. In Bertino, E., Matei, S. (eds.), *Roles, Trust, and Reputation in Social Media Knowledge Markets*, 103–117. Springer International Publishing.

Hogan, N. R., Davidge, E. Q., & Corabian, G. (2021). On the Ethics and Practicalities of Artificial Intelligence, Risk Assessment, and Race. *J Am Acad Psychiatry Law*, 49(3).

Islam, S. R., Eberle, W., Ghafoor, S. K., & Ahmed, M. (2021). Explainable Artificial Intelligence Approaches: A Survey. ArXiv.

Jareborg, N. (2004). Rättsdogmatik som vetenskap, *SvJT*, 1-10.

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. ArXiv.

Kaloudi, N., & Li, J. (2020). The AI-Based Cyber Threat Landscape: A Survey. *ACM Comput. Surv.*, 53(1), 20:1-20:34.

Kaminski, M. E. (2019). The Right to Explanation, Explained. *Berkeley Technology Law Journal*, 34(1), 189–218.

Kleineman, J. (2018). Rättsdogmatisk metod. In Nääv, M. & Zamboni, M. (eds.), *Juridisk metodlära* (2 ed.). Studentlitteratur AB. Lund.

Klimas, T. & Vaiciukaite, J. (2008). The law of Recitals in European Community Legislation. *ILSA Journal of International & Comparative Law*. 15:1.

Kolkman, D. (2022). The (in)credibility of algorithmic models to non-experts. *Information, Communication & Society*, 25(1), 93–109.

Kuner, C., Svantesson, D. J. B., Cate, F. H., Lynskey, O., & Millard, C. (2017). Machine learning with personal data: Is data protection law smart enough to meet the challenge?, *International Data Privacy Law*, 7(1), 1–2

- Lehrberg, B. (2022). *Praktisk juridisk metod* (14 ed.). Iusté. Uppsala.
- Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM queue*, 16:3, 31–57.
- Malgieri, G., & Comandé, G. (2017). Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. *International Data Privacy Law*, 7(4), 243–265.
- McCrudden, C. (2008). Human Dignity and Judicial Interpretation of Human Rights. *European Journal of International Law*, 19(4), 655–724.
- Milossi, M., Alexandropoulou-Egyptiadou, E., & Psannis, K. E. (2021). AI Ethics: Algorithmic Determinism or Self-Determination? The GDPR Approach. *IEEE Access*, 9, 58455–58466.
- Mittelstadt, B. (2016). Auditing for transparency in content personalization systems. *International Journal of Communication*, 4991–5003.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2).
- Monteith, S., Glenn, T., Geddes, J. R., Whybrow, P. C., Achtyes, E., & Bauer, M. (2024). Artificial intelligence and increasing misinformation. *British Journal of Psychiatry*, 224(2), 33–35.
- Mourali, M., Novakowski, D., Pogacar, R., & Brigden, N. (2025). Post hoc explanations improve consumer responses to algorithmic decisions. *Journal of Business Research*, 186, 1–11.
- Paik, S. (2023). Journalism Ethics for the Algorithmic Era. *Digital Journalism*, 1–27.

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). The role of explainable AI in the context of the AI Act. 1139–1150.

Peczenik, A. (1995). Juridikens teori och metod: En introduktion i allmän rättslära. (1 ed.). Fritze. Stockholm.

Pfeiffer, J., Gutschow, J., Haas, C., Möslin, F., Maspfuhl, O., Borgers, F., & Alpsancar, S. (2023). Algorithmic Fairness in AI: An Interdisciplinary View. *Business & Information Systems Engineering*, 65(2), 209–222.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2019). A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Computing Surveys*, 51(5), 1–36.

Puiutta, E., & Veith, E. M. S. P. (2020). Explainable Reinforcement Learning: A Survey. In Holzinger, A., Kieseberg, P., Tjoa, A.M, & Weippl, E. (eds.), *Machine Learning and Knowledge Extraction*, Vol. 12279, pp. 77–95. Springer International Publishing.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141.

Rashid, A. B., & Kausik, M. A. K. (2024). AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. *Hybrid Advances*, 7, 100277.

Reichel, J. (2018). EU-rättslig metod. In Nääv, M. & Zamboni, M. (eds.) *Juridisk metodlära* (2 ed.). Studentlitteratur AB. Lund.

Rieg, T., Frick, J., Baumgartl, H., & Buettner, R. (2020). Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms. *PloS One*, 15(12), e0243615–e0243615.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263, 110273-.

Samek, W. & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Vol. 11700. Springer International Publishing.

Sandgren, C. (2021). Rättsvetenskap för uppsatsförfattare: ämne, material, metod och argumentation. (5 ed.) Norstedts Juridik AB. Stockholm.

Selbst, A. D. (2021). An Institutional View of Algorithmic Impact Assessment. *Harvard Journal of Law & Technology*, ISSN: 0897-3393, 35;1, 117.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). *Fairness and Abstraction in Sociotechnical Systems*. 59–68.

Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4), 233–242.

Shah, M. U., Rehman, U., Parmar, B., & Ismail, I. (2024). Effects of Moral Violation on Algorithmic Transparency: An Empirical Investigation. *Journal of Business Ethics*, 193(1), 19–34.

Sinha, G. & Dunbar, R. (2022). Chapter 1 - Artificial Intelligence and its regulation in the European Union. In Bielicki, D. M. (ed.). *Regulating Artificial Intelligence in Industry* (1 ed.), 3-20. Routledge.

Solum, L. B. (2004). Procedural Justice. University of San Diego Public Law and Legal Theory Research Paper Series. 2.

Soori, M., Arezoo, B., & Dastres, R. (2023). Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, 3, 54–70.

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy & Society*, 40(2), 137–157.

Theodorou, A., & Dignum, V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, 2(1), 10–12.

Tzimas, T. (2023). Algorithmic Transparency and Explainability under EU Law. *European Public Law*, 29(4).

Varošanec, I. (2022). On the path to the future: Mapping the notion of transparency in the EU regulatory framework for AI. *International Review of Law Computers & Technology*, 36(2), 95–117.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7(2), 76–99.

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & Society*, 36(2), 585–595.

Wang, Y. (2023). Balancing Trustworthiness and Efficiency in Artificial Intelligence Systems: An Analysis of Tradeoffs and Strategies. *IEEE Internet Computing*, 27(6), 8–12.

Widrow, B., & Lehr, M. A. (1990). 30 years of adaptive neural networks: Perceptron, Madaline, and backpropagation. *Proceedings of the IEEE*, 78(9), 1415–1442.

Wulf, A. J., & Seizov, O. (2020). Artificial Intelligence and Transparency: A Blueprint for Improving the Regulation of AI Applications in the EU. *European Business Law Review*, 31(4), 611–640.

Wulf, A. J., & Seizov, O. (2024). “Please understand we cannot provide further information”: Evaluating content and transparency of GDPR-mandated AI disclosures. *AI & Society*, 39(1), 235–256.

Zhang, Z., Shan, S., Fang, Y., & Shao, L. (2019). Deep Learning for Pattern Recognition. *Pattern Recognition Letters*, 119, 1–2.

Zhao, Y., & Flenner, A. (2019). Deep Models, Machine Learning, and Artificial Intelligence Applications in National and International Security—Part Two. *The AI Magazine*, 40(2), 29–30.

EU publications

Article 29 Data Protection Working Party (WP 29). (2018). Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. WP 251.

Hamon, R., Junklewitz, H., Sanchez, I. (2020). Robustness and Explainability of Artificial Intelligence – From technical to policy solutions. EUR 30040, *Publications Office of the European Union*, Luxembourg.

High Level Expert Group on Artificial Intelligence (AI HLEG). (2019). Ethics Guidelines for Trustworthy AI. Technical report, *European Commission*, Brussels.

Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F. & Delipetrev, B. (2020). AI WATCH. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence. EUR 30117, *Publications Office of the European Union*, Luxembourg.