Alma Mater Studiorum - Università di Bologna
in cotutela con UNIVERSIDAD
POLITÉCNICA DE MADRID

DOTTORATO DI RICERCA IN

Law, Science and Technology

Ciclo 35°

Settore Concorsuale: 12/H3 – Filosofia Del Diritto
Settore Scientifico Disciplinare: IUS/20 – Filosofia Del Diritto

Governing Algorithms in the Big Data Era for Balancing New Digital Rights

Designing GDPR Compliant and Trustworthy XAI Systems

Presentata da: Orhan Gazi Yalcin

Coordinatore Dottorato
Prof. Monica Palmirani

Supervisore
Prof. Giovanni Sartor

Co-supervisore
Prof. Javier Bajo

Co-supervisore
Prof. Ugo Pagallo

Esame finale anno 2023

Abstract. This thesis investigates the legal, ethical, technical, and psychological issues of general data processing and artificial intelligence practices and the explainability of AI systems. It consists of two main parts.

In the initial section, we provide a comprehensive overview of the big data processing ecosystem and the main challenges we face today. We then evaluate the GDPR's data privacy framework in the European Union. The Trustworthy AI Framework proposed by the EU's High-Level Expert Group on AI (AI HLEG) is examined in detail. The ethical principles for the foundation and realization of Trustworthy AI are analyzed along with the assessment list prepared by the AI HLEG. Then, we list the main big data challenges the European researchers and institutions identified and provide a literature review on the technical and organizational measures to address these challenges. A quantitative analysis is conducted on the identified big data challenges and the measures to address them, which leads to practical recommendations for better data processing and AI practices in the EU.

In the subsequent part, we concentrate on the explainability of AI systems. We clarify the terminology and list the goals aimed at the explainability of AI systems. We identify the reasons for the explainability-accuracy trade-off and how we can address it. We conduct a comparative cognitive analysis between human reasoning and machine-generated explanations with the aim of understanding how explainable AI can contribute to human reasoning. We then focus on the technical and legal responses to remedy the explainability problem. In this part, GDPR's right to explanation framework and safeguards are analyzed in-depth with their contribution to the realization of Trustworthy AI. Then, we analyze the explanation techniques applicable at different stages of machine learning and propose several recommendations in chronological order to develop GDPR-compliant and Trustworthy XAI systems.

Keywords: Explainable Artificial Intelligence, XAI, explainability, the right to explanation, Trustworthy AI, GDPR, digital ethics, black-box systems

List of Abbreviations

Abbreviation	Definition
AA	: Algorithmic Auditing
AC	: Automated Compliance
AI	: Artificial Intelligence
DG	: Data Governance
DL	: Deep Learning
DLT	: Distributed Ledger Technologies
DP	: Differential Privacy
FL	: Federated Learning Approaches
GDPR	: General Data Protection Regulation
HE	: Homomorphic Encryption
ML	: Machine Learning
MPC	: Secure Multiparty Computation
RAT	: Risk Assessment Tools
RtE	: Right to Explanation
SP	: Sticky Policies
SSI	: Self-sovereign Identity
XAI	: Explainable Artificial Intelligence

List of Figures

List of Tables

List of Equations

List of Appendices

1 Introduction

Data-driven technologies play a crucial role in decision-making processes in the contemporary era. The foundations of data-oriented research fields, such as data science, artificial neural networks, deep learning, and machine learning, were established by researchers in the 20th century. Despite these early developments, practical implementation was impeded by limitations in processing power, storage capabilities, and available data. With advancements in hardware technologies, researchers now have access to high computing power, inexpensive storage solutions, and abundant data sources. The advent of the Internet has further accelerated the flow of information, enabling researchers to implement their research in real-world and collaborate with their peers. Therefore, the 21st century has seen a rise of AI applications.

Thanks to these AI applications, we are on a course to building Level 5 autonomous vehicles[1] that do not require a driver. We can create computer vision systems that can detect objects with minimal error margins. We can convert audio data-to-text and text data-to-audio. We can successfully translate a text from one language to another using machine translation. The examples are countless.

The development of AI applications demands a substantial volume of data, commonly referred to as "big data." The increasing use of data-intensive AI applications has led to the creation of tools and techniques for collecting large amounts of personal data, referred to as "big personal data." Initially, data collection practices were inadequately regulated, allowing companies to collect data as they saw fit. However, as the potential dangers of such practices became apparent, governments began to enact data protection regulations.

As the advancements in data processing and AI technologies have a direct impact on society, the regulation of these technologies is not a purely academic matter. The subject has attracted widespread attention, with media outlets and bloggers taking a keen interest. Internet activists frequently voice their concerns and call on governments to address specific issues. Law and ethics researchers also publish articles evaluating these matters. Additionally, institutional publications and research groups aim to provide perspectives on these issues, while tech companies have established advocacy groups to self-regulate their practices and secure representation before regulatory bodies.

Therefore, all the stakeholders push for different policies on regulating these technologies. While a business in the field aims to provide efficient services with high value, an AI engineer aims to create robust systems that can achieve high accuracy. A minimal amount of regulation is usually the best policy for these stakeholders since it would give them the freedom to develop their services and products. On the other hand, for these technologies' users, adopting a strict data protection policy is more favorable for protecting their fundamental rights & freedoms. Ethical principles set the frontiers for extensive protection of individual rights & freedoms and the societal order. While the business and technical stakeholders push for neo-liberal policies, ethicists usually push for strict regulations.

[1] SAE International: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems: J3016_202104 (2021)

However, a healthy regulation does not prioritize a single stakeholder and disregards the needs of other interest groups. Therefore, lawmakers must find a balance between these two clusters so that fundamental rights & freedoms are well protected, yet there is still room for sustainable innovation. Researchers need to carefully analyze each stakeholder's needs and views and provide guidance to the lawmakers. Therefore, this research aims to be the melting pot of all these stakeholders and tries to find a consensus among the different views by asserting guidance.

In the initial section of this paper, the focus will be on the overarching topic of Governing Algorithms in the Big Data Era for Balancing New Digital Rights. The study will examine the current challenges related to data processing in the European Union and evaluate state-of-the-art in accordance with ethical principles and positive law. The European Union has established a robust data protection regulation with the General Data Protection Regulation (GDPR) 2016/679, which supersedes the Data Protection Directive 95/46/EC. This paper will evaluate state-of-the-art with reference to the GDPR and assess existing privacy-preserving technologies and organizational measures that comply with the GDPR's requirements. Notable Horizon2020 projects will be considered to detect these technologies and measures. Following the analysis, existing trends and patterns will be revealed, and areas requiring additional attention will be identified.

In the subsequent part, the focus will shift from a general scope to one of the primary components of a general data processing challenge: The Explainability of AI systems to address the ethical and societal implications of AI and data processing systems. The widespread utilization of AI systems in daily life has raised ethical and societal concerns, particularly regarding the use of deep learning models with black-box characteristics that pose accountability, fairness, and transparency challenges. The source of the explainability problem will be explored, followed by a clarification of XAI's goals, term definition, and brief analysis of cognitive models for explanations. The right to explanation framework of the GDPR will be analyzed, along with the contribution of legal safeguards to the Trustworthy AI principles. These findings will assist us to detect the necessary technical solutions to create legally compliant and ethically viable AI systems. The development lifecycle, algorithmic structure, and technical specifications of machine learning and deep learning systems will be discussed, with a focus on increasing the explainability property of AI systems at each stage of the AI lifecycle. The multidisciplinary recommendations offered will guide the development of GDPR-compliant and Trustworthy XAI systems.

2 Methodology

This thesis aims to understand the workflow of data processing systems and their utilization in big data and AI applications. We aim to properly review the legal framework in the European Union to understand how these data processing systems must be designed. Besides, apart from the positive law rules provided by the EU legislation, we try to understand the underlying ethical principles to identify the reasoning for the legal rules regulating this area. After the technical analysis of data processing systems and

legal & ethical norms regulating this area, we review some notable projects that aim to provide frameworks and solutions for GDPR-compliant data processing and AI systems. These solutions will be compared based on the common themes established by the GDPR and the ethical framework. After this benchmark analysis, we propose recommendations to develop reliable and socially conscious data processing and AI systems. The final goal of this research is to improve the state of the art in data processing and AI applications by making them more privacy-friendly based on the GDPR rules and ethical principles. The methodical approaches to achieve the research goals may be listed as follows:

- An interdisciplinary approach to cover all the issues in different fields related to AI and data processing practices,
- A Pragmatic Approach based on effective GDPR rules without disregarding ethical principles, and
- A comparative approach to uncover the data processing issues to address by evaluating the existing solutions.

2.1 An Interdisciplinary Approach

Two fields at the heart of this research are data protection and big data & AI. The field of data protection is an interdisciplinary one. It is a field developed to regulate data processing; therefore, it is deeply connected to fields such as Information Technology, Law, and Ethics. Thus, the literature in these fields must be carefully reviewed to thoroughly analyze a data protection issue. Big data & artificial intelligence are statistical and technical fields whose methods and techniques are used by various domains. Besides, the appropriate regulation of these technology fields can only be achieved with comprehensive legal and ethical analysis. Therefore, the scope of this thesis necessitates an interdisciplinary approach, and therefore, the multidisciplinary approach is the most critical component of our methodology.

2.2 A Pragmatic Approach with a Hint of Ethics

Another vital component of the research methodology is the pragmatic approach. In this research, we propose tangible recommendations based on the enforceable legal norms and real-life necessities derived from the GDPR provisions. Therefore, the practical approach will be the norm. However, we aim to strengthen this approach with a hint of ethical considerations. Instead of solely focusing on the GDPR provisions, we will mention ethical principles that dominate the field of data protection to make logical inferences about the legal norms instead of blindly following them. For the interpretation of the legal norms, ethical principles may be valuable facilitators. Therefore, by keeping the ethical principles in mind, our evaluations serve the purpose of complying with the GDPR regime and protecting fundamental rights & freedoms.

2.3 A Comparative Approach

The two initial components, interdisciplinary and pragmatic approaches, will help to create a list of desired qualities that we look for in data processing systems. These qualities will be the result of legal and ethical reviews as well as technical considerations. After creating a list of desired qualities, we review and analyze notable existing data processing projects. The comparative approach allows us to see the existing solutions' competencies and what they lack in terms of GDPR compliance and ethical principles. Not only will we be able to compare the notable PPT projects, but we will also be able to point out what these solutions should do to improve themselves further. Therefore, the comparative approach will allow us to compare the existing solutions and propose improvements.

Then, in the second part of the thesis, we will focus on Explainable AI, explainability, the right to explanation, and Trustworthy AI principles. We will cover why we need explainable AI systems. Then, we will analyze how explanations can help achieve the protection of the right to explanation and realization of Trustworthy AI. We will compare alternative methods for explainability and machine learning models.

In summary, the flow of this research will be as follows:

- Understanding the technical details of big data processing and AI systems,
- Analyzing the GDPR's provisions to understand its regime and requirements,
- Exploring the ethical principles which are discussed in the data protection literature to understand the legal norms better,
- Comparative analysis of the notable GDPR-compliant PPT projects; and
- Proposing improvements on the existing solutions for more sustainable and GDPR-friendly data processing systems designed with privacy in mind.

Then, we will move on to the specific topics of Explainable AI, which are listed as follows:

- Understanding why explainability is essential and how Explainable AI methods can contribute to the AI system explainability,
- Identifying the goals of Explainable AI and conducting a term clarification analysis,
- Analyzing the cognitive models of human and machine reasoning and defining what explanation is,
- Identifying the scope of the right to explanation and Trustworthy AI principles concerning the explainability of automated decisions,
- Understanding the technical structure of machine learning and deep learning algorithms and their lifecycle,
- Clarifying how to strengthen the explainability properties of AI systems at each step of the ML lifecycle, and
- Proposing how to develop explainable AI systems in a systemic and sustainable manner.

3 Governing Algorithms in the Big Data Era for Balancing New Digital Rights

3.1 Data Processing, Big Data, and Data Protection

In this chapter, we will go over the fundamental concepts in the world of data processing and their development processes throughout the 20th and 21st centuries. The issues we observe in data processing are highly relevant to technical concepts such as data lifecycle, big data, and finally, artificial intelligence. Therefore, these concepts and their relationships to data processing will be briefly covered in this part of the thesis.

Data Processing and Data Lifecycle: Data processing can be defined as the collection and manipulation of items of data to produce meaningful information. The field of data processing started with manual processing activities such as bookkeeping. With the U.S. Census 1890, automatic data processing practices started with processes such as punched cards. With the advancement in computer technologies, the first electronic data processing was done with UNIVAC I in the U.S. Census 1950.[2] The term data processing is often used with a broader term, "information technology." In fact, after the 1980s, the term data processing lost its popularity in academic books, and the term information technology gained popularity.[3] However, since the term information technology has a very broad meaning, we will use the term data processing in this report.

Data processing contains several data-related operations, such as generation, collection, processing, storage, management, analysis, visualization, interpretation, and more.[4] With these operations, the data completes a cyclical process called "data lifecycle." A data lifecycle is a set of processes that covers all the stages of data, from its collection to its storage and reuse.[5] Input, processing, output, and storage are the four major stages of data processing, and these stages are expanded and adjusted based on the use case. There are several frameworks that propose data lifecycle schemas, which contain similar characteristics in many aspects. However, they offer additional elements to the fundamental operations.[6]

Big Data and Data Processing: With the advancements in data processing and information technologies, the total size of the available data has increased tremendously. According to one estimation, we start to record 1.7 MB of data per second for every

---

2 Rosenthal, M. D. (2000). Striving for Perfection: A Brief History of Advances and Undercounts in the U.S. Census. Government Information Quarterly, 17(2), 193–208, 202. https://doi.org/10.1016/S0740-624X(00)00027-7.

3 Google Books Team, T. (2020). Google Ngram Viewer. https://bit.ly/2LnR4pn.

4 Wing, J. M. (2019). The Data Life Cycle. Harvard Data Science Review. https://doi.org/10.1162/99608f92.e26845b4.

5 Data Lifecycle | NNLM. (n.d.). National Library of Medicine. Retrieved January 5, 2021, from https://nnlm.gov/data/thesaurus/data-lifecycle.

6 Ball, A. (2012). Review of Data Management Lifecycle Models. University of Bath, 2. http://opus.bath.ac.uk/Thisversionismadeavailableinaccordancewithpublisherpolicies.

human being.[7] This corresponds to 30GB of data for each person every year, and this number is growing with an extreme velocity. Although there are three original defining V properties (i.e., volume, velocity, and variety) of big data originated by Dough Laney,[8] the number of Vs has increased over the years. In fact, there are at least 19 different defining Vs mentioned in different publications.[9] Below, we list some of the defining characteristics of big data with their brief definitions:

- **Volume**: Volume is a characteristic that represents the amount of data generated within a given time span.[10] The "big" in big data refers to the volume of the data.
- **Velocity**: Velocity refers to the speed of data processing and storage. Big data applications must operate at high velocity since they deal with high volumes of data. There are different approaches to data processing, such as real-time and batch processing, which are part of the velocity characteristic of big data.[11]
- **Variety**: Variety refers to the number of sources that the data is collected from. These sources range from in-house devices to GPS data from mobile devices or social media content extracted from platforms. It also refers to the forms, types, and sources in which data is recorded. [12] A general rule about the relationship between data variety and data privacy is that the more varied sources are, the more complex data protection issues we tend to encounter.
- **Variability**: Apart from the data variety, variability refers to the changes within a single data source, such as the meaning of a single feature for the entire analysis.[13]
- **Veracity**: Veracity characteristic refers to the quality of the collected data. Collecting a high volume of data does not always correspond to high-quality data. Clarity and accuracy of the data are also important components of data veracity.[14]
- **Visualization**: Visualization is one of the new characteristics used to define big data applications.[15]
- **Value**: Value is the final characteristic of big data. After a company used its resources to satisfy the previous 6 Vs, the final version of the big data must be

---

7 Domo. (2018). Data never sleeps 6.0: how much data is generated every minute?, 2. https://web-assets.domo.com/blog/wp-content/uploads/2018/05/18_domo_data-never-sleeps-6verticals.pdf.

8 Cartledge, C. *How Many Vs are there in Big Data*, 1. http://www.clc-ent.com/TBDE/Docs/vs.pdf.

9 Cartledge, *Vs in Big Data*, 2-3.

10 Conocimiento, I. de I. del. (2016). Las 7 V del Big data: Características más importantes - IIC. Instituto de Ingeniería Del Conocimiento. https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/.

11 Analytics, B. I. G. D. (2019). Big Data Analytics, 7. https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf.

12 Conocimiento, *Las 7 V del Big data*.

13 Understanding the 7 Vs of Big Data. (2016, April 7). Impact. https://impact.com/marketing-intelligence/7-vs-big-data/.

14 The five Vs of big data | BBVA. (n.d.). BBVA. Retrieved January 5, 2021, from https://www.bbva.com/en/five-vs-big-data/.

15 Conocimiento, *Las 7 V del Big data*.

valuable. Therefore, value refers to the monetary value of the processed big data.[16]

The increasing volume of available data and powerful data processing systems & architectures have caused an explosion in the number of big data applications. Although these applications may be intended for completely irrelevant areas, every big data processing system foresees a similar (big) data lifecycle. The explosion in the number of applications is an important factor that has increased the significance of data privacy and protection.

Artificial Intelligence and Data Processing: The Encyclopedia Britannica defines artificial intelligence as "*the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings."*[17] AI systems can mimic the functionalities of intelligent beings, and they can range from expert systems that fulfill professional tasks to robots that behave like animals.[18] This wide range of applications is possible thanks to big data. Without large volumes of high-value data and efficient data processing practices, we would not have today's artificial intelligence applications. Therefore, the field of data processing and big data are highly related to artificial intelligence. Furthermore, as soon as the nature of the AI system includes a personalized output, the fields of data privacy and data protection become relevant as well. In fact, in any event, where personal data is collected, stored, or used in any way, then we enter the realms of GDPR. In the upcoming sections, we will cover the field of Artificial Intelligence in more detail.

3.2 Legal Framework for Data Processing in the EU

Advancements in data processing technologies led to a massive accumulation of data which created the field of big data. Access to high volume and quality data enabled us to materialize artificial intelligence applications. As these technologies started to play a greater role in our daily lives, several concerns related to fundamental rights & freedoms surfaced, such as surveillance, manipulation, discrimination & algorithmic bias. These developments contributed to the development of the "personal data" concept and the field of data privacy.

In the European Union jurisdiction, the main regulation that deals with data privacy issues is the General Data Protection Regulation. Therefore, in this study, we will analyze GDPR to clarify the terminology and the solutions to data privacy issues. According to Art. 4 (1) of the General Data Protection Regulation, personal data refers to "any information relating to an identified or identifiable natural person."[19] Identifiability

---

16 BBVA, *The five Vs of big data*.

17 Copeland, B. J. (n.d.). Artificial Intelligence | Definition, Examples, and Applications. Britannica. Retrieved January 5, 2021, from https://www.britannica.com/technology/artificial-intelligence.

18 Boston Dynamics. (2019). Spot® | Boston Dynamics. In Boston Dynamics. https://www.bostondynamics.com/spot.

19 See. General Data Protection Regulation, Art. 4 (1).

characteristics of information should be interpreted broadly. Identifiable information does not have to contain the person's name. Therefore, identifiability may be direct and indirect.[20] As long as we can establish a reasonable likeliness of identifiability using available tools or external data, the data should be regarded as personal data.[21]

3.2.1 The Key Actors under GDPR

In a data processing ecosystem, there are several stakeholders defined under GDPR. While some of the stakeholders collect and store personal data, some of them benefit from the stored data. Depending on their role, the actions of these stakeholders are regulated under GDPR.[22] We can talk about seven main actors that are defined under GDPR and/or affect the data privacy ecosystem, which may be listed as follows:

- Data Subject
- Controller
- Processor
- Data Protection Officer
- Third Parties
- Recipient
- Lawmakers, policymakers, and governmental bodies

It is important to properly understand these key actors' characteristic features along with their roles in the data privacy ecosystem:

Data Subject: Data subject can be defined as any natural person whose personal data is collected and processed. Pursuant to GDPR, only natural persons are the only beneficiaries of the data protection rules. GDPR defines personal data as any information relating to an identified or identifiable natural person. In addition, CoE law (i.e., Modernized Convention) also defines personal data in line with GDPR. This should not be interpreted as that legal persons do not enjoy any data protection. The ECtHR case law contains several judgments on the violation of legal persons' right to protection against their data. However, these protections find their basis in European Convention on Human Rights and other legislative pieces.[23] In other words, GDPR does not apply to the protection of the legal persons' data.

Controller: The controller is the main actor that decides the "purpose and means of processing personal data."[24] A controller can be a legal or natural person as well as a

---

20 Giakoumopoulos, C., Buttarelli, G., & O'Flaherty, M. (2018). Handbook on European data protection law 2018, 83. In Luxembourg: Publications Office of the European Union. https://doi.org/10.2811/58814.

21 Sharma, S. (2019). Data Privacy and GDPR Handbook. In Data Privacy and GDPR Handbook. Wiley, 47. https://doi.org/10.1002/9781119594307.

22 Sharma, *GDPR Handboo*k, 68.

23 Giakoumopoulos & Buttarelli, *Handbook on GDPR*, 83-84.

24 Sharma, GDPR Handbook, 52.

public authority or body. However, the legal status of the controller is irrelevant, and the decision-making power with respect to data processing is the main element in classifying an entity as a controller. A controller controls the direction and purpose of personal data processing. Some of the factors that indicate that the actor in question has the controller capabilities can be listed as follows[25]:

- The capability of making independent data processing decisions,
- The capability of using the collected data in business decision-making processes,
- Having the responsibility for the data processing activities,
- The capability of integrating the collected data with internal databases, and
- Having a legal or contractual relationship with the data subject for data collection.

Based on the factors listed above, the decision-making power of an entity may be measured, and this measure may be used to determine the controller status of an entity.

Processor: A processor is an actor responsible for processing personal data on behalf of the controller. A processor can be a natural person, legal person, public authority, public body, or other body. By using the term "other body," GDPR leaves room for previously unforeseen occasions where the processor has a sui generis legal status. Just as for controller, the legal status of the processor is irrelevant if it satisfies the following conditions:

- Having a separate identity from the controller; and
- Processing personal data on behalf of the controller.[26]

Controller and processor are often mentioned together under GDPR since their activities often collide and overlap. Although they have similar liabilities and duties, they may differ from one another depending on the data processing activity.[27] For example, a cloud computing service provider can enter into a contractual relationship with an e-commerce company to process the e-commerce company data by offering its hosting services. In this event, the cloud computing service provider is regarded as the processor. However, in any event where the cloud computing service provider exceeds its mission and takes on a role in determining the use and purpose of the data processing, this processor is started to be regarded as a controller.[28]

Data Protection Officer (DPO): The role of a data protection officer is to be a medium between controllers/processors and data subjects. Pursuant to GDPR Art. 37, the controller and processor must appoint a data protection officer when:

---

25 Giakoumopoulos & Buttarelli, *Handbook on GDPR*, 104.

26 Voigt, P., Wessing, T., & Bussche, A. (2018). The EU General Data Protection Regulation (GDPR) - A Practical Guide. In Irish Medical Journal (Vol. 111, Issue 5), 20. https://www.springer.com/gp/book/9783319579580.

27 Sharma, GDPR Handbook, 53.

28 Voigt & Wessing, A GDPR Practical Guide, 20.

- The data processing is carried out by a public authority or body except as an enforcement of the judicial decisions,
- The data processing is carried out by controllers and processors, which regularly and systematically monitor the data subjects as part of their core activities, and
- The processed data is part of special categories of data pursuant to Art. 9 of GDPR or personal data relating to criminal activities indicated in Art. 10 of GDPR.[29]

The main responsibility of the DPO is to protect the data of the data subjects. A DPO can be (i) an individual who works for the controller or the processor, (ii) a legal or natural person contractor hired by the controller or the processor. A DPO can represent multiple data processors and controllers as long as he/she remains easily accessible.[30] Therefore, professional service providers can offer DPO services to data processors and controllers.

Third Parties: A third party is any natural or legal person, public authority, agency, or body that is authorized to process personal data, but not a data subject, controller, processor, or a person under the authority of the controller or processor. [31] In other words, they have the authority to process personal data; however, they do not fit into one of the definitions above. [32] For example, payment processors for an e-commerce transaction are considered third parties.

Recipient: The recipient is used as a broader term than the term "third party." Pursuant to GDPR Art. 4 (9), a recipient is any natural or legal person, public authority, agency, or another body, to which the personal data are disclosed. This definition applies to any entity that receives personal data regardless of their third-party status.[33] Therefore, all the third parties are recipients under GDPR, yet the opposite does not hold true. The only exception for being classified as a recipient is in the case of public authorities receiving personal data for a specific purpose under Union or Member law. [34]

It is important to note that the employee of the controller or processor cannot be third parties, but they can be recipients. Therefore, making this distinction is very important for lawful data disclosure. A third party is not authorized to use the personal data that the controller possesses unless there is a contractual or legal ground. However, the employee of the controller can automatically use personal data without additional legal requirements.[35]

---

[29] General Data Protection Regulation, Art. 37 (1).

[30] Sharma, S. (2019). Data Privacy and GDPR Handbook. In Data Privacy and GDPR Handbook, 53. Wiley. https://doi.org/10.1002/9781119594307.

[31] See. General Data Protection Regulation, Art. 4 (10).

[32] Sharma, GDPR Handbook, 53.

[33] See. General Data Protection Regulation, Art. 4 (9).

[34] See. General Data Protection Regulation, Art. 4 (9).

[35] Giakoumopoulos & Buttarelli, *Handbook on GDPR*, 110-111.

Lawmakers, policymakers, and governmental bodies: One final group of actors that can be included in the GDPR-actors is the umbrella group for the lawmakers, policymakers, and governmental bodies. When lawmakers, policymakers, and governmental bodies receive information as part of a GDPR schema, they will be considered a recipient. On the other hand, when they process personal data, they may be regarded as data controllers, processors, or third parties. Apart from the data processing activities, lawmakers, policymakers, and governmental bodies have additional roles such as regulation development, preparing lower-level administrative guidelines, and conducting research activities and surveys. Therefore, lawmakers, policymakers, and governmental bodies can still act as important actors without undertaking data processing activities, and these lawmakers, policymakers, and governmental bodies that are not processing data should be regarded as the seventh group of data privacy actors.

3.2.2 Technical and Organizational Measures.

Article 32 of GDPR requires data controllers and processors to implement technical and organizational measures to safeguard the personal data they process. The wording of the GDPR Art. 32 requires data processors and controllers to satisfy the following goals to achieve the appropriate data protection level:

- the pseudonymization and encryption of personal data,
- the ability to ensure the ongoing confidentiality, integrity, availability, and resilience of processing systems and services,
- the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident,
- a process for regularly testing, assessing, and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing.

In addition to GDPR Art. 32, other responsibilities are laid out in their corresponding GDPR articles, as listed below:

- Keeping records of processing activities in writing (see GDPR Art. 30(3)),
- Conducting routine data protection impact assessment (see GDPR Art. 35(1)),
- Appointing a data protection officer (DPO) (see GDPR Art. 37),
- Following the principles of Privacy by Design and Privacy by Default (see GDPR Art. 25),
- Notifying relevant authorities in case of a Personal Data Breach incident, including accidental or unlawful destruction, loss, alteration, unauthorized disclosure of, or access to, personal data (see GDPR Art. 4(12)), and
- Implementing self-regulation procedures such as Codes of Conduct, Certifications, and Seals (see GDPR Art. 40-43).

There is a large variety of technical and organizational measures that organizations can undertake to achieve the minimum standards listed above. A non-exhaustive list of these technical and organizational measures is as follows:[36]

- Minimizing the personal data processing,
- Irreversible anonymization and pseudonymization of data,
- Developing solutions that the data subject can utilize to monitor his personal data,
- Developing solutions to improve the security property of the data processing activities,
- Developing privacy-preserving technologies with Privacy by Design and Privacy by Default principles in mind,
- Implementing irreversible encryption solutions for safe data transfer,
- Physical measures to prevent unauthorized access to personal data such as secured rooms,
- Undertaking routine risk assessments on data processing systems, and
- Enforcing codes of conduct and legal, ethical & technical standards.

Although the Data Protection Directive of 1995 did not articulate an explicit accountability clause, the GDPR introduces the principle of accountability in Art. 5(2). According to Art. 5(2), a data controller has the responsibility to ensure compliance with the GDPR and the burden of proof with regards to compliance. The accountability principle requires data controllers to implement appropriate technical and organizational measures before initiating data processing activities. Failure to do so may result in administrative fines up to EUR 20 million or up to 4% of the total worldwide annual turnover of the data controller.[37]

Developing and implementing these technical and organizational measures can help address some of the big data and data processing challenges identified in the upcoming section. However, the challenges covered in the next section show that simple privacy-preserving techniques are often vulnerable to adversarial attacks. Therefore, another section after the challenges is dedicated to state-of-the-art privacy-preserving technical technologies and organizational measures that may be effective against the risks and challenges identified in this report.

3.3 Information Ethics, Data Processing, and Trustworthy AI

Since regulations have a bidirectional relationship with ethical principles, there is a strong relationship between the legal regime of data protection and information ethics. While hard ethics may shape the regulations and governance, soft ethics set the moral values within the borders that the regulations draw.[38] Therefore, it is important to identify the ethical principles dictating the data protection sphere to clarify the goals aimed

---

36 Voigt & Wessing, A GDPR Practical Guide, 38-39.

37 Voigt & Wessing, A GDPR Practical Guide, 31-32.

38 Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379-380 (2133). https://doi.org/10.1098/rsta.2018.0081.

with the legal norms and, therefore, the reasoning behind the necessity to protecting an ethical principle and what technical or organizational measure to use to protect it.

The field of information ethics is a relatively new field, and however, the number of ethics frameworks in artificial intelligence and data protection has already reached several dozens, nearing hundreds. However, the increasing number of proposed ethics frameworks is rather a sign of infancy of the ethical side of information technologies. These frameworks should go through a comprehensive consolidation process to be reliable. However, this process can take several years to complete, and until then, the best strategy seems to be to focus on the frameworks that are published by the official authorities. We will focus on the Trustworthy AI Framework proposed by the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG). The principles covered under AI HLEG's Trustworthy AI framework will be further strengthened with the other studies to derive a better understanding of the common themes of information ethics on artificial intelligence and data protection.[39]

3.3.1 Trustworthy AI Framework

With its Communications dated back to 2018, the European Commission defined the European vision for AI in three pillars:

- *"Increasing Public and Private Investments in AI to boost its uptake"*
- *"Preparing for Socio-economic changes"*
- *"Ensuring an appropriate ethical and legal framework to strengthen European values."*[40]

Following these communications, the European Commission established AI HLEG, consisting of 52 experts to deliver AI Ethics Guidelines and a List of Policy and Investment Recommendations.[41] In March 2019, the European Commission's AI HLEG completed the first public draft of the AI Ethics Guidelines. This document can be seen as a quasi-official response of the European Commission to the ever-growing importance of artificial intelligence and or the north start[42] for the European communities. While designing the Trustworthy AI Framework, AI HLEG aimed at maximizing the benefits of AI while minimizing its risks. While this goal sounds very intuitive, in many cases, there is a trade-off between the benefits and the associated risks, which requires AI system developers to make difficult decisions. For instance, developing AI applications trained on extensive personal data including special categories of personal data can help data controllers to develop state-of-the-art AI systems; however, they can also breach the right to privacy of the data subjects due to the difficulty of complying with consent

---

[39] Fjeld & Achten, Principled AI, 5.

[40] High-Level Expert Group on Artificial Intelligence (European Commission). (2019). Ethics Guidelines for Trustworthy AI. In European Commission, 4. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[41] AI HLEG, Trustworthy AI, 4.

[42] AI HLEG, Trustworthy AI, 4.

requirements. On the other hand, severely limiting data controllers' capability to develop AI systems in the name of data privacy may hinder innovation in artificial intelligence. Another principle that the AI HLEG takes into account when designing the Trustworthy AI Framework is the human-centric approach to Artificial Intelligence. With a human-centric approach in mind, AI HLEG points out that the goal should be to increase human well-being instead of the AI itself. Therefore, with a human-centric approach and benefit maximization approach in mind, the Trustworthy AI framework list recommendations to design ethically responsible and technically robust AI systems.[43]

Trustworthiness is an important feature of the AI systems that people and societies pay attention to its realization. Without the trustworthiness of the AI systems, people may never embrace the AI technologies, especially in sensitive areas. If this scenario occurs, they will not be able to benefit from the vast social and economic opportunities the AI technologies offer. The trustworthiness of an AI system should be present in different stages of AI systems such as its development, deployment, and use. In addition, trustworthiness of all the stakeholders and processes that are associated with an AI system is also an inseparable part of overall trustworthiness of an AI system.[44]

Trustworthy AI Framework has three main components to deem a system trustworthy, which are (a) lawfulness, (b) ethicality, (c) robustness of the system. Lawfulness of AI systems refer to the compliance to the applicable laws at every stage of the AI systems, including the development, deployment, and usage stages. Lawfulness requires the compliance to the EU primary law, EU secondary law, EU member states laws as well as international treaties and conventions. Finally, domain-specific legislations should also be taken into account when developing, deploying, and using AI systems. Ethicality of the AI systems refers to respecting ethical principles when developing, deploying, and using AI systems. Ethicality is particularly important where legal norms cannot keep pace with the advancement in AI technologies. Therefore, with the ethicality of the AI systems, creating an untrustworthy environment can be prevented. Robustness of AI systems refers to the AI systems' operability in a safe, secure, and reliable manner. Robustness should be maintained both from a technical perspective and from a societal perspective. Technical robustness refers to the robustness in system's development, deployment, and usage lifecycle and domain applicability whereas social robustness refers to the robustness related to the context and the environment in which the system operates.

In the original framework, the trustworthiness of the systems is covered under three parts. The first part focuses on the ethical purpose of Artificial Intelligence by identifying the relevant fundamental rights and principles and the applicable regulation. The second part deals with the realization of Trustworthy AI by focusing on harmonizing the ethical purpose and technical robustness of AI systems. In this part, AI HLEG lists the fundamental ethical principles and technical and organizational measures that can contribute to this realization. In the third chapter, AI HLEG proposes a non-exhaustive assessment list for Trustworthy AI. Compared to other ethical AI frameworks, AI

[43] AI HLEG, Trustworthy AI, 4.
[44] AI HLEG, Trustworthy AI, 5.

HLEG's Trustworthy AI framework aims to offer tangible and practical recommendations to develop Trustworthy AI Systems.[45] The relationship between these chapters can be shown in Fig 1, as follows:

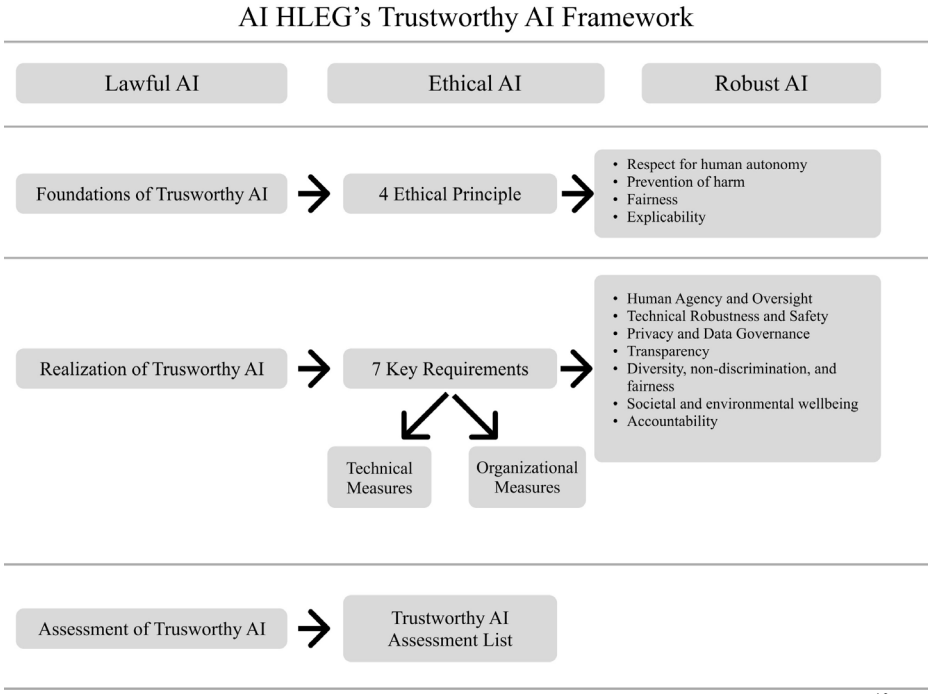


Fig 1. Trustworthy AI Framework Proposed by European Commission's AI HLEG[46]

3.3.2 Foundations of Trustworthy AI

In this part, AI HLEG describes the foundations of Trustworthy AI, which lays on fundamental rights and is reflected by four ethical principles. While fundamental rights fall under the first component of Trustworthy AI (lawfulness), ethical principles fall under the second component of Trustworthy AI (ethicality). Although fundamental rights are legally binding at the highest level, they may not always provide remedies for every case due to their generic nature. However, the lower-level laws that are enacted in reference to fundamental rights can provide remedies for specific cases. Therefore, these fundamental rights embody important checks for the Trustworthiness of AI systems. AI HLEG list a number of significant fundamental rights for the foundation of Trustworthy AI, which are (i) respect for human dignity, (ii) respect for human dignity, (iii) respect for democracy, justice, and the rule of law, (iv) equality, non-discrimination, and solidarity, (v) citizens' rights.[47]

[45] AI HLEG, Trustworthy AI, 4.
[46] AI HLEG, Trustworthy AI, 8.
[47] AI HLEG, Trustworthy AI, 10.

3.3.2.1 Fundamental Rights in the Context of AI Systems

Respect for human dignity refers to the idea that every human being possesses an "intrinsic worth", which cannot be reduced or diminished under no circumstances, and AI is no exception. Human dignity is often regarded as the most sacred fundamental right and requires the treatment of human beings as subjects, not objects. When developing, deploying, and using AI systems, the stakeholders should be involved in these processes with respect to human dignity and protects human beings' physical and mental integrity and identity. **Respect for human dignity** refers to the ability of human beings to make decisions independently. This freedom encapsulates the freedom from the influences of other individuals, institutions as well as sovereign powers. Freedom of individuals necessitates maximizing human autonomy and minimizing the threats to mental health, unlawful surveillance, deception, and manipulation. Therefore, the freedom of individuals also requires AI systems to strengthen the freedom of the individual, not only protect it. **Respect for democracy, justice, and the rule of law** refers to the commitment of AI stakeholders to comply with the legal norms and also aims to strengthen the democratic process with the power of technology and automation. **Equality, non-discrimination, and solidarity** refer to showing equal respect for the moral worth of all human beings. This set of fundamental rights prohibits AI systems to deliver biased outputs that may adversely affect individuals. It also requires vulnerable groups and minorities (e.g., women, workers, minors, consumers, people with disabilities, and ethnic and religious minorities) to have equal access to AI Systems. **Citizens' rights** refer to the human beings' fundamental rights that were born out of being a citizen of a country. These rights include a wide range of rights, such as the right to vote, the right to access public information, the right to petition. AI systems should strengthen the exercise of these rights when there is an opportunity.[48]

3.3.2.2 Ethical Principles in the Context of AI Systems

Although the fundamental rights mentioned above can provide guidance on how to comply to the existing norms, there might be cases where ethical norms can strengthen the existing protective shield. In addition to legal response to what we should do, ethical principles can help providing answers to what we can do with artificial intelligence. Therefore, fundamental rights and ethical principles can co-exist in harmony to cover each other's weaknesses. AI HLEG identifies four ethical principles in the context of AI systems, which are (i) respect for human autonomy, (ii) prevention of harm, (iii) fairness, and (iv) explicability.

In parallel with the respect for human dignity, the respect for human autonomy principle refers to ensuring that human beings have freedom and autonomy when they make decisions. Therefore, when human beings interact with AI systems, they should have full and effective self-determination, and AI systems should impair their judgments with coercion, deception, manipulation, or subordination. Finally, this principle also

[48] AI HLEG, Trustworthy AI, 10-11.

requires human oversight on the decisions made by AI systems. The principle of prevention of harm requires that AI systems neither cause nor inflict harm on human beings and that they prioritize the mental and physical integrity of human beings and human dignity. In combination with the equality principle, it necessitates that AI systems pay special attention to vulnerable groups and ensure their protection in a non-discriminatory manner. The principle of fairness should be regarded highly when developing, deploying, and using AI systems. The substantive dimension of the fairness principle refers to the equal and just distribution of the benefits and costs of artificial intelligence by individuals, and the AI systems are free from bias, discrimination, and stigmatization. On the other hand, the procedural dimension of the fairness principle refers to the ability to contest and seek effective remedies for automated decisions made by AI systems. Finally, the principle of explainability refers to the transparency of the AI systems' development, deployment, and usage processes. The capabilities and purpose of the AI systems should be clearly communicated to the people who might be affected by their processes directly or indirectly. In cases where AI systems use black-box models, the developers should use proper explicability measures to create a more trustworthy environment for the people who are affected by the automated decisions.

After identifying the fundamental rights and ethical principles that had a foundational nature for Trustworthy AI, AI HLEG lists several ethical principles for the realization of Trustworthy AI. Additionally, AI HLEG also lists several technical and organizational measures that provide practical importance in realizing Trustworthy AI.

3.3.3 Realizing Trustworthy AI

In this part, AI HLEG list recommendations to translate the fundamental rights and ethical principles into concrete requirements. These requirements should be fulfilled by all the relevant stakeholders that are part of the development, deployment, or usage of AI systems. These stakeholders include developers, data subjects, data controllers, domain experts, policymakers, and public and private organizations. While developers should implement the requirements at the design and development stages, deployers should conduct audits to ensure that the requirements are met. Finally, the members of society should be informed about these requirements and should demand compliance. AI HLEG lists seven ethical principles in a non-exhaustive manner. The principles that are crucial to realize Trustworthy AI are as follows:

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, non-discrimination, and fairness
- Societal and environmental wellbeing
- Accountability

3.3.3.1 Human Agency and Oversight

To achieve trustworthiness, AI Systems should strengthen human autonomy and human decision-making instead of completely replacing or diminishing them. Therefore, AI Systems should allow human oversight and agency to create a more democratic society that respects fundamental rights and freedoms.

AI systems have the capacity to both damage and strengthen fundamental rights and freedoms. In sensitive areas of AI use, an impact assessment should be made, and ideal remedies should be proposed to remedy the risks.

The first component of the human agency and oversight principle is the human agency part. In case of the existence of an AI system, they should be informed about and given appropriate tools for these systems so that they can properly interact with and self-assess them. It is important for data subjects to be at the center of the AI systems so that they are not regarded as mere objects in the AI lifecycle, but instead, they are an active participant to them. In addition, human autonomy also requires not to be subject to automated decision-making when they don't give appropriate consent. GDPR Art. 22 defines safeguards to protect human autonomy and reflects these ethical issues by prohibiting being subject to automated decision-making and profiling without consent.

Apart from the ability to opt out of an automated decision, another important component of the human agency and oversight principle is the human oversight of automated decisions[49] There are several concepts developed around the principle of human agency and oversight, such as (i) a human-in-the-loop (HITL), (ii) human-on-the-loop (HOTL), and finally, (iii) human-in-command (HIC). The concept of HITL proposes the most involvement of humans in the automated decision-making process. Where HITL approach is adopted, humans can be part of each decision cycle. Especially in high frequency decision making, adopting HITL may not be technically feasible or possible. HOTL refers to human involvement in which system design and the operating of the system can be overseen by designated individuals so that they can intervene at any time. The concept of HOTL refers to a human-enabled review process, usually attained by a subject-matter expert. The HOTL concept can be very effective in industries such as manufacturing, where there are scenarios that can only be detected by human experts. Apart from industries where personal data does not play a vital role, HOTL can be also useful for more sensitive fields such as medicine or law enforcement. For example, medical operations conducted by AI-enabled robots should be overseen by a human surgeon. On the other hand, HIC refers to a broader concept where designated individuals oversee how an AI system is affecting socially and economically. HIC concept also deals with when and in which context to use AI systems.

3.3.3.2 Technical Robustness and Safety

Since they are built on scalable business models that rely on data processing, AI systems have the potential to cause harm on a large scale, intentionally or unintentionally.[50] As a result, the technical robustness and safety of data processing and AI systems are hotly debated ethical issues, given the scale of harm they can cause. To minimize

[49] Fjeld & Achten, Principled AI, 53.
[50] Fjeld & Achten, Principled AI, 37.

the risk of unintentional and unexpected harm, developers of these systems should design them with a focus on preventing harm, taking steps to avoid any potential hazardous outcomes.

Adversarial attacks can manipulate the behavior of AI systems, produce unexpected outputs that impact others, compromise the model infrastructure, and contaminate the data. Therefore, a number of security concerns must be addressed when designing and operating these systems. In other words, data processing and AI systems must be designed to be resilient to attacks and must have effective security measures in place to protect against vulnerabilities that could be exploited by adversaries.[51]

On the other hand, there may be instances where adversarial attacks cannot be prevented, and the outcomes of automated decision-making by data processing and AI systems may impact individuals. To be prepared for such scenarios, organizations should have contingency, fallback, and general safety plans in place. For example, if an AI system is compromised by an adversary and unable to operate properly, the system operator should have the capability to switch to an alternative model or a manual, human-controlled environment. Additionally, organizations must ensure that security properties such as accuracy, reliability, and reproducibility are strengthened. Accuracy of the AI systems, regardless of security threats, is a critical property and individuals should only be subject to decisions made by these systems if they can provide a sufficient level of accuracy. In the event of an adversarial attack, the behavior of the AI system should also be reproducible, so that the behavior can be traced and evaluated for any potential issues. Organizational measures, such as replication files, can serve as useful tools for analyzing and testing the behavior of AI systems.[52]

3.3.3.3 Privacy and Data Governance

The principle of privacy and data governance is a crucial ethical issue that is thoroughly addressed by data protection regulations. Some key themes related to privacy that are covered under GDPR (General Data Protection Regulation) include consent, control over personal data, the ability to restrict processing, the right to erasure, and privacy by design.[53]

It's essential for data processing systems to ensure privacy and data protection throughout the entire data lifecycle, starting from the moment a data subject provides their personal information and continuing through every interaction with the data processing and AI system. During this lifecycle, the personal data of the data subject may contain sensitive information, such as ethnic and religious identity, sexual orientation, age, gender, or political views. The data controller or AI system operator must never use this sensitive information for discriminatory purposes.

Moreover, the data controller and AI system operator have a responsibility to ensure that the data they process and possess is of high quality and free from biases, inaccuracies, errors, and imbalances. The personal data provided by the data subject, as well as

[51] AI HLEG, Trustworthy AI, 16.
[52] AI HLEG, Trustworthy AI, 17.
[53] AI HLEG, Trustworthy AI, 21.

the data inferred over time through data analysis and artificial intelligence, must be protected. AI systems should not be trained on datasets of low quality or integrity.[54]

The entire lifecycle of AI systems, including planning, training, testing, and deployment, must be thoroughly documented. Access to personal data should be regulated within the organization, with clear policies outlining the list of personnel who have access and the extent of that access. Only authorized individuals should be able to view sensitive information about data subjects.[55]

3.3.3.4 Transparency

With advancements in data processing and AI systems, automated decision making has become increasingly prevalent in society. The use of automated decision systems that are trained on historical data that may contain patterns of discriminatory practices can result in harmful outcomes. This highlights the importance of transparency in the algorithms used in these systems, which is a critical aspect of ethical discussion. The transparency principle is tied to the principle of explainability and refers to the transparency of AI system components such as the model, data, and business logic. The transparency principle has three pillars: (i) traceability, (ii) explainability, and (iii) communication.

Documenting the AI system lifecycle is crucial to strengthening the traceability property. It is important to document the operations of data gathering, data labeling, model development, and prediction processes in detail, as this can help experts trace any issues that negatively impact data subjects.

Explainability, within the context of Trustworthy AI, refers to providing explanations for both the technical processes of AI systems and relevant human decisions. Thus, its scope extends beyond mere model explainability and includes the decisions made by business managers to use AI systems in a particular field. The scope of technical explanations also extends beyond model explainability to include pre-modeling operations such as data collection, processing, and storage. Model benchmarking and management-level explainability are also part of technical explainability. At the modeling stage, there is a trade-off between the accuracy and explainability of machine learning algorithms. Transparent and explainable algorithms often have lower accuracy performance, while "black box" models such as neural networks can achieve higher accuracy. Therefore, it is important to address the issues of transparency and explainability with novel methods and more complex explainable algorithms, especially in sensitive areas.[56]

Finally, the last pillar of transparency, communication, involves informing data subjects that they are interacting with an AI system and not with a human being. This principle is already reflected in the GDPR Art. 13-15, which gives data subjects the

---

[54] AI HLEG, Trustworthy AI, 17.

[55] AI HLEG, Trustworthy AI, 17.

[56] Yalcin, O. G. (2020). *Examination of current AI systems within the scope of right to explanation and designing explainable AI systems*, 2-3. CEUR Workshop Proceedings, 2598. https://ceur-ws.org/Vol-2598/paper-11.pdf.

right to obtain information about the existence of automated decisions, as well as information about the limitations, accuracy, and inner logic of the AI systems.[57]

3.3.3.5 Diversity, non-discrimination, and fairness

To achieve trustworthiness, the controllers of AI systems must empower all data subjects who belongs to different groups in a society. They must keep the inclusiveness and diversity in mind when they develop these systems. The AI systems should ensure equal access as well as equal treatment for the members of all groups. diversity, non-discrimination and fairness is closely associated with the principle of fairness defined under Art. 5 of GDPR. According to the principle of fairness, the development, deployment, and use of data processing and AI systems must be fair.[58] Although there are different interpretations for the scope of fairness, the most common themes revolving around fairness are non-discrimination, equality, inclusiveness, representative and high-quality data.[59] To achieve true fairness, these themes must be thoroughly considered and assessed before data processing. The processed data may be used to identify a data subject's sensitive information such as sexual orientation, age, gender, and religious and political views.

These datasets used by AI systems can include historical biases, imbalances, and imperfections, which may lead to unintentional or intentional discriminatory or exploitive practices. Therefore, AI system developers have a positive obligation to clean the data from all these historical biases. Furthermore, it is important to utilize privacy-preserving technologies and organizational measures to prevent unlawful or unfair use of personal data against data subjects.60 In addition, including people from different backgrounds in the development and testing teams can also contribute to the development of more inclusive AI systems. This strategy can even increase the likelihood of equal distribution of technology-related wealth and prevent the formation of technology-poor and information-poor groups.[61]

On the other hand, another pillar of the diversity, non-discrimination, and fairness principle is ensuring accessibility and universal design. AI systems should have a user-centric design instead of an enterprise-centric or one-fits-all design to allow data subjects to use the AI system regardless of their gender, age, political view, sexual orientation, and ethnic identity. AI systems should be accessible by the person with disabilities. When designing these systems, the AI developers should not only include the majority groups to maximize profitability. Instead, they should actively aim for more accessibility and inclusiveness.

Finally, when designing AI systems, the developers should consult with the stakeholders that may be affected by the system directly or indirectly. This mechanism should start at the planning stage and continue even after the system outputs a decision.

---

57 AI HLEG, Trustworthy AI, 18.
58 AI HLEG, Trustworthy AI, 12.
59 Fjeld & Achten, Principled AI, 47.
60 AI HLEG, Trustworthy AI, 17.
61 Fjeld & Achten, Principled AI, 60-61.

Additionally, this consultation should be a continuous one and create a long-term sustainable relationship between the AI developers and all the system stakeholders to ensure the realization of Trustworthy AI.

3.3.3.6 Societal and environmental wellbeing

With the increasing prevalence of AI systems in social applications, the effects of Artificial Intelligence are becoming more observable in broader society, on the environment, and on other living beings.[62] There are several issues that need to be addressed to ensure the societal and environmental well-being of Trustworthy AI. Firstly, AI systems should be designed in a sustainable and environmentally friendly manner. The development, deployment, usage, and data collection and processing operations of AI systems should all be designed with sustainability in mind. Developers should opt for optimized AI algorithms that use less energy and minimize their long-term energy footprint. Additionally, the supply chain of AI systems should be environmentally friendly to enable continuous improvement in a sustainable manner. Secondly, as the popularity of social AI applications grows, the social impact of AI systems is becoming increasingly visible. With the widespread adoption of AI systems, there may be negative impacts on the mental and physical health of data subjects. To prevent these adverse effects, the developers of AI systems should ensure that their systems do not harm individuals. This can be achieved through careful and continuous monitoring of the systems using technical and organizational measures.[63] Finally, AI systems should not compromise the well-being of society and the integrity of democratic institutions. AI systems can have positive impacts on public and private institutions, but they can also have negative effects on democratic processes. For instance, they could manipulate political communities and interfere with electoral outcomes. Therefore, it is important for developers to consider the potential consequences of their AI systems on society and democracy and take steps to ensure that these systems do not have negative impacts.

3.3.3.7 Accountability

With its addition to GDPR, accountability has become an even more important principle in the field of data protection. Some of the themes that are usually associated with accountability are verifiability, replicability, monitoring, available remedies, impact assessments,[64] auditability, minimization and reporting of negative impact, trade-offs, and redress.[65] In addition to general principle of accountability, professional responsibility is an ethical issue closely related to accountability. Instead of the responsibility of the entities, professional responsibility focuses on the people and the teams that develop data processing and AI systems. Professional responsibility should be considered

---

62 Fjeld & Achten, Principled AI, 37.
63 AI HLEG, Trustworthy AI, 14.
64 Fjeld & Achten, Principled AI, 28.
65 AI HLEG, Trustworthy AI, 14.

as part of the accountability. In addition, some of the themes that are related to professional responsibility can be listed non-exhaustively as accuracy, responsible design, considering long-term effects, multi-stakeholder collaboration, and scientific integrity.[66] AI HLEG selects and prioritizes four main themes among this wide variety of topics: (i) auditability, (ii) minimization and reporting of negative impacts, (iii) trade-offs, and (iv) redress. Auditability refers to the enabling periodic assessment of pre-modeling activities such as data collection, cleaning, and storage, modeling activities such as model selection, model training, and hyperparameter tuning, and post-modeling activities such as predicting and integrating the trained model to the AI system. In addition, the audibility also includes the auditability of the non-technical parts of the AI systems such as the business logic. However, the auditability principle does not require constant assessments of the sensitive information. Instead, it refers to periodic assessments by internal and external auditors who respects the AI system's controllers trade secrets and intellectual property rights. The importance of the principle of auditability becomes more important as the risk-level of the AI system increases. Furthermore, since AI systems has become more and more suitable to cause negative impacts, the AI system developers should actively work on minimizing the negative impacts of these systems. They should set predefined standards to allow reporting of these negative effects. The accountability principle requires data controllers to implement appropriate technical and organizational measures to ensure the minimization and allow the reporting of these negative impacts even before initiating data processing activities. While implementing these technical and organizational measures, there might be inevitable trade-offs between system performance and ethical requirements. When a trade-off issue occurs, the developers should always prioritize the ethical requirements and do not violate any fundamental rights for the sake of system performance and utility guarantees. When the system developers detect a trade-off, they should document the components of the trade-off and periodically review the documentation to continuously ensure that the development decisions never violate fundamental rights and freedoms under no circumstances. In case where a negative impact of AI system takes place, the AI system should project and set appropriate redress to mitigate all the existing and upcoming negative effects. While predefining the measures to redress these negative effects, the system developers should pay special attention to vulnerable groups. Only by fulfilling their obligations, data controllers may be relieved from extensive liability and accountability burdens if they can provide evidence that they have taken appropriate measures for data processing and AI activities.[67]

3.3.4 Technical and non-technical methods to realize Trustworthy AI

To comply with all the ethical principles and cover their related themes, the AI system developers must employ technical and organizational measures, which are also defined under GDPR Art. 32 and other relevant articles. The process that started with the foundational values of Trustworthy AI including the relevant fundamental rights and ethical

---

66 Fjeld & Achten, Principled AI, 56.

67 Voigt & Wessing, A GDPR Practical Guide, 31-32.

principles explained above. Then, it continuous with the AI system developers efforts to comply with the seven umbrella principles that together help with the realization of Trustworthy AI. To comply with these principles, the AI system developers must use the technical and non-technical measures and safeguards. With the implementation of these measures, a cyclical period of evaluation and justification starts. This process comprises of using, analyzing, re-designing, and development activities to continuously improve the compliance with the Trustworthy AI principles. The technical and organizational measures vary to a great extent and can be complimentary or alternative to each other. For example, homomorphic encryption can be a technical measure to comply with privacy and data governance requirements as well as distributed ledger technologies (**DLT**). Therefore, they can be alternatives to or complement each other. Explainable AI can offer measures to comply with both transparency and diversity, non-discrimination, and fairness requirements. Depending on the nature of the measures, they can be integrated into the AI systems at the design, development, use, or management stages. The technical and organizational measures will be analyzed in greater detail in Section3.5 with real-life examples.

3.3.5 Assessing Trustworthy AI

After defining the principles for the foundation and the realization of Trustworthy AI and the technical and organizational measures to be used for these goals, AI HLEG proposes a non-exhaustive list to operationalize Trustworthy AI, which is referred to as the "Trustworthy AI assessment list." AI HLEG explicitly states that the scope of the Trustworthy AI assessment list does not include Lawful AI and, therefore, does not ensure compliance with the legal requirements set out in GDPR and other relevant laws. The list is developed for the AI systems that directly interact with the data subjects and mainly provide guidance to the developers and designers of these systems. The list is projected to be designed in a horizontal manner, providing guidance across all industries.

For the assessment list to be effective, its implementation should be embraced both at the managerial level and the operational level. The management of the organization should undertake comprehensive planning by taking the Trustworthy AI requirements into consideration. In addition, it should assign the relevant employees with the appropriate tasks with the required authority to implement the assessment list requirements. Compliance and legal departments should monitor the implementation of the assessment list and its effectiveness in compliance with the legal framework laid out in the relevant laws. They should also develop appropriate internal policies as organizational measures for the realization of Trustworthy AI. At the product and service development level, the issues covered under the assessment list must be analyzed, and the results should be shared with the management for their approval or refusal of the current version of the AI system. In parallel with the product and service development specialists, the quality assurance team must check the results of the assessment list issues and notify the management if the requirements are not satisfied. As a secondary issue, the human resources teams must ensure that the operation teams have enough diversity to check against the principles of Trustworthy AI. Furthermore, they should provide training for

the team members and raise awareness about the importance of Trustworthy AI. Procurement specialists should ensure that the products and services they procure follow Trustworthy AI requirements.

One of the most challenging problems when implementing an assessment list is the lack of diversity of skills and competencies in the team. Achieving Trustworthy AI requires the fulfillment of several principles that require legal, ethical, and technical knowledge, as well as product development skills. Therefore, to truly achieve Trustworthy AI, the AI system developers should put together competent people with complimentary, often multidisciplinary, backgrounds. In addition to the competency of the internal team, the AI system developer should also involve the other stakeholders that may be directly or indirectly affected by the AI system in line with the Trustworthy AI principles. Data subjects, third parties, governmental bodies, market participants, and trade unions, as well as labor unions, should be part of the Trustworthy AI development process. Finally, the assessment list should not be considered as a single source for overall legal and ethical compliance, but it should complement the other legal and ethical compliance tools to further strengthen the Trustworthiness of the system.

3.4 Challenges to security and privacy in Big Data

In the previous sections, we covered state of the art in big data and artificial intelligence from technical, legal, and ethical perspectives. We identified the main actors, values, and fundamental norms. After the technical and legal analysis, we analyzed -perhaps- the most important ethical framework proposed in the European Union: AI HLEG's Trustworthy AI Framework. In this section, we will explain the data privacy and protection challenges sitting at the intersection of all the legal, ethical, and technical fields that are identified in the literature.

3.4.1 The Contradiction between Big Data innovation and Data Protection

Perhaps, the most noticeable challenge to security and privacy in Big Data is the contradiction between big data innovation and data protection. In the European Union, GDPR acts as a protector for the protection of ethical and societal values. To ensure this protection, it regulates and heavily limits the processing of personal data. Although the introduction of this regulatory practice is easily justified with ethical and societal concerns such as privacy and security, regulation inevitably creates challenges for innovation. Therefore, proposing technical solutions that ensure sustainable innovation together with effective data protection is a necessity. For example, the introduction of explicit consent makes it difficult to collect and share data; however, if privacy-preserving technologies provide enough warranties for privacy awareness, secure workflows that include usage/access control, transparency, and compliance verification to the data subject, obtaining data subjects' consent becomes easier. Therefore, value and

knowledge creation can still be achieved where the personal data of the data subject is effectively protected.[68]

3.4.2 Societal and ethical implications of big data technologies:

The societal and ethical implications of big data technologies is rather an umbrella term that contains a number of sub-challenges. Big data technologies can be used to limit free will and manipulate data subjects with unethical profiling practices. They can create systematic unfairness, especially in sensitive areas, which may damage ethical values such as equality, non-discrimination, and digital inclusion. Especially when automatic decision-making is utilized in these areas, explainability mechanisms must be integrated into the systems to guarantee transparency, accountability, and trustworthiness. Designing big data solutions that respect societal and ethical values such as human welfare, autonomy, non-maleficence, justice, accountability, trustworthiness, privacy, dignity, and solidarity is an important and difficult challenge that developers face. To achieve this level of advancement in these solutions, algorithmic auditing must be introduced along with ethical guidelines and technical standards. Bias in data must be eliminated with technical methods and tools while the use of black-box models should be discouraged in sensitive areas.[69]

3.4.3 Secure and trusted personal data sharing:

In the current state of the art, once the data is shared with a third party, there is almost no guarantee for the secrecy of the data. Therefore, most data collectors prefer to keep their datasets strictly private, which hampers the efforts to create a data market and economy. Lack of the trusted and secure platforms and privacy-aware analytics methods for secure data sharing is one of the most important challenges in big data technologies, which limits the potential of the field. To create a healthy data market, technical standards, quality levels, and legal approaches must be adopted, and new solutions allowing the secure transfer of data must be developed. Technical solutions developed for this purpose should not only enable secure data transfer and privacy-aware analytics in a costly manner. For the mass adoption of these solutions, the computational costs must be minimized, and integration with the existing systems must be ensured.[70]

3.4.4 Processing sensitive data

According to Article 9 of GDPR, personal data that reveals "racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and

---

[68] Home - SPECIAL. (n.d.). Retrieved January 30, 2021, from https://www.specialprivacy.eu/.

[69] Custers, B., La Fors, K., Jozwiak, M., Esther, K., Bachlechner, D., Friedewald, M., & Aguzzi, S. (2018). Lists of Ethical, Legal, Societal and Economic Issues of Big Data Technologies. *SSRN Electronic Journal*, 19. https://doi.org/10.2139/ssrn.3091018.

[70] Markopoulos, I. (2020). Industry specific requirements analysis, definition of the vertical E2E data marketplace functionality and use cases definition I, 11. https://trusts-data.eu/.

the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation" are considered as the special categories of personal data. The personal data that belongs to one of these categories is often referred to as sensitive data. Processing sensitive personal data is subject to stricter rules pursuant to GDPR. However, sensitive data can be extremely useful to provide more personalized service and improve the life quality of the data subject, such as by predicting and preventing a future disease by using health data of the data subject. In addition to complying with the additional legal restrictions for the processing of sensitive personal data, the data subject must also be given additional privacy warranties to give consent for the processing of his sensitive personal data. Achieving a satisfactory level of privacy and security for processing sensitive personal data is one of the challenges of big data solutions. Utilizing distributed technologies in a scalable manner can be useful to overcome this challenge.[71]

3.4.5 Limits of anonymization and pseudonymization:

Although privacy-preserving techniques for anonymization and pseudonymization are effective solutions to ensure effective data protection, many of these techniques have flaws that can be exploited. Big data system designers must ensure the irreversibility of the privacy-preserving technologies. Ensuring the reliability of the anonymization and pseudonymization practices is a difficult challenge to overcome since these operations are open to adverse attacks and can be reversed with malicious attacks unless they are designed properly. On the other hand, achieving irreversibility of the anonymization often require removal of connection between data sources which may adversely affect the service quality and customer satisfaction. Therefore, too rigid anonymization methods may cause data to lose its value, whereas light methods may not be reliable since attackers can deanonymize the data that are anonymized with light methods. The data processing systems must provide data subjects with the options to select the level of privacy they can sacrifice for enhanced service quality. These options can only be effective if the data subjects can analyze the trade-off between the loss of privacy and service quality.[72]

3.4.6 Dealing with multiple data sources and untrusted parties:

There is an ever-increasing and accelerating trend in data generation. Accumulation of a high volume of data and careful analysis provides opportunities for knowledge and value creation. As the volume of data grows, the number of data sources also increases.

---

[71] Rizzo, A. (2017). MHMD Project Presentation. In My Health My Data, 4. http://www.myhealthmydata.eu/deliverables/D11.2-MHMD-Project-Presentation.pdf.

[72] European Big Data Value Association. (2017). Strategic Research and Innovation Agenda. European Big Data Value, 4(October), 66. https://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf.

These sources must be integrated, and the processors should be encouraged and provided with the correct solutions to share data across organizations for better services. However, efficiency and security issues are some of the obstacles which discourage data sharing across organizations. Therefore, dealing with multiple data sources and untrusted parties is one of the challenges to overcome for knowledge and value creation. To overcome this challenge, several privacy-preserving analytics solutions and multi-party computation techniques must be introduced. These solutions should provide opportunities to make data available for encrypted processing instead of sharing the data without proper control over it.[73]

3.4.7 A general, easy to use and enforceable data protection approach:

One of the challenges identified by the European Big Data Value Association is the need for a more general, easy-to-use, and enforceable data protection approach., particularly for large-scale commercial processing as the most important area where data privacy is sought. The development of mechanisms that enables data subjects to define the boundaries for the collection, processing, and sharing of their personal data contributes to the enforceability of data protection regulations. These mechanisms should also facilitate the exercise of fundamental rights, such as the right to be forgotten. While designing these mechanisms, developers must ensure that they are easy to use and understandable by all the relevant stakeholders. In addition to these mechanisms, technical measures must be introduced to examine the conformity of the data processing systems to the data subjects' consents and the limitations for the processing of their personal data. These measures must be designed for both general and case-by-case auditing of the data processing systems.[74]

3.4.8 Maintaining robust data privacy with utility guarantees:

One of the other important challenges that data processing systems face is to ensure reliability and efficiency of the services while maintaining robust data privacy—for example, being able to analyze data to provide highly personalized services while implementing privacy solutions such as encryption, anonymization, and pseudonymization. Besides keeping the services efficient and reliable, implementing these solutions cause another related challenge: Scalability of the solutions. For instance, encryption solutions such as multiparty computation or homomorphic encryption are not applicable to large-scale databases due to performance issues. Therefore, one of the challenges that the system designers must overcome is to find efficient ways to encrypt and anonymize personal data. [75]

---

[73] Veeningen, M. (2020). SODA - Scalable Oblivious Data Analytics. SODA Project. https://soda-project.eu/.

[74] European BDVA. *Strategic Research and Innovation Agenda*, 66.

[75] European BDVA. *Strategic Research and Innovation Agenda*, 66.

3.4.9 Risk-based approaches calibrating data controllers' obligations:

Under GDPR, one of the responsibilities of the data controller and processor is to ensure the safety of the processed data and assess the risk of security issues. Therefore, one of the challenges in big data processing is to calibrate data controllers' obligations with a risk-based approach. Especially when dealing with multiple datasets and data from multiple sources, the risks associated with data processing reaches a higher level where data controllers must approach the risks with utmost significance. For example, using adverse attacks, malicious users can gather anonymized, pseudonymized, and public datasets to identify data subjects by creating relations between these sources. The risk is much greater in private datasets that are not protected with any privacy-preserving technology. Therefore, developing and utilizing tools to assess and prevent these risks is one of the challenges that the data processing and AI community faces.[76]

3.4.10 Combining different techniques for end-to-end data protection

General Data Protection Regulation requires data controllers to utilize several materials, technical, and organizational measures for end-to-end data protection. These measures include technical measures such as secure hardware enclaves, secure multiparty computation, encryption, and anonymization, as well as organizational measures such as IT awareness training, auditing, and certification. Integrating all these solutions might be very costly and cause performance overheads. Therefore, achieving end-to-end data protection is a challenge that can only be overcome with careful planning and efficient optimization. In a world where cloud and fog computing has led us to a very complex and dynamic ecosystem for data processing systems, the integration of these systems must be adaptive and robust. Failure to develop proper solutions for integrating these solutions with minimal cost and maximum efficiency can cause legal problems such as violating data privacy legislation as well as technical problems such as slowness and interruptions in the systems. [77]

3.5 Privacy-Preserving Technologies and Organizational Measures for the Challenges in Big Data

In the previous section, we covered the main data privacy and protection challenges to be addressed by the big data community. There are many potential solutions in the form of privacy-preserving technologies or organizational measures that can address these challenges. In this section, we will examine these solutions in more detail. These solutions, then, will be linked to the challenges it potentially addresses.

---

[76] European BDVA. *Strategic Research and Innovation Agenda*, 67.

[77] Timan, T., Mann, Z. (2021). *Data Protection in the Era of Artificial Intelligence: Trends, Existing Solutions and Recommendations for Privacy-Preserving Technologies*. The Elements of Big Data Value, 7-8. https://doi.org/10.1007/978-3-030-68176-0_7.

3.5.1 Explainable AI

Explainable AI refers to a subfield of artificial intelligence that aims to propose methods and techniques to develop artificial intelligence systems that "produces details or reasons to make its functioning clear and easy to understand."[78] While the simple machine learning models such as linear regression or decision trees are easy to interpret without additional methods, complex models such as neural networks or ensemble methods are black-box models in their nature. Although it is difficult to explain the reasoning of the black box models, they tend to overperform the simple models. Apart from data privacy, one of the data protection properties laid out in GDPR is having access to "meaningful information about the logic involved"[79] to maintain transparency and fairness in the case of automated decision-making.[80]

Explainable AI methods and techniques are useful techniques that can provide information about significant issues such as the existence of an automated system, the logical process to generate a particular decision, data used for a particular decision, and even the security measures used to protect data subject's personal data, and the overall decision algorithm. On the other hand, Explainable AI methods also pose privacy vulnerabilities since valuable information can be retrieved with an adverse attack (e.g., membership inference attacks or reverse-engineering attacks) using explanation methods.[81] The Explainable AI techniques should be implemented with high security measures to protect the data subjects' personal data from adverse attacks.

At the model level, Explainable AI aims to develop techniques to provide local and global explanations. A set of the proposed explainability techniques are model agnostic and post-hoc and, therefore, can be applied to any machine learning model. The others are model-specific techniques and, therefore, try to utilize specific model architectures to provide explanations.[82]

On the other hand, the scope of Explainable AI should not be limited to model explainability. Explainability of the data used to train the AI system is part of the overall system explainability. By using different standardization, visualization, and exploration techniques, the developers can create more explainable dataset, which can be used with explainable models. In addition, the scope of Explainable AI also include an explanation interface. Explanation interface is an adaptable user-centric user interface designed with a presentation logic that can use the plain explanations generated by the AI models to create meaningful explanations tailored for the needs of a specific user. A properly structured explanation interface can adjust the granularity of the system explanations for the users who might have different level of knowledge in each AI use case. Finally, apart from the technical side, Explainable AI offers a number of organizational measures that can be used to strengthen the overall explainability of the AI systems.

---

78 Budig, T., Herrmann, S., Dietz, A., Pandl Supervisor, K., & Sunyaev, A. (n.d.). Trade-offs between Privacy-Preserving and Explainable Machine Learning in Healthcare, 5. Retrieved February 1, 2021, from www.kit.edu.

79 See. General Data Protection Regulation, Art. 15(1)(h).

80 Timan & Mann. *Data Protection in the AI Era*, 11.

81 Budig & Herrmann, *XAI in Healthcare*, 12-13.

82 Budig & Herrmann, *XAI in Healthcare*, 5-7.

Explainability audits, collaborative R&D frameworks, and cooperated policy developments are the main organizational measures that can be used at the managerial level. The details of these technical and organizational measures will be shared in the upcoming chapters.

3.5.2 Secure Multiparty Computation

Secure multiparty computation (MPC) is a popular privacy-preserving technology that can be utilized for data protection. A secure MPC protocol allows multiple parties to compute a joint function without having access to each other's data. In an MPC protocol, generally, sensitive information is distributed to each party as a share. These shares do not reveal the actual sensitive information by themselves. Therefore, the privacy of sensitive information is achieved without a trusted third party. [83] By accumulating their shares, each party can compute their part, and the results from these parties are aggregated to conclude the process. Thus, the calculation on sensitive information is completed by the parties without revealing each other their sensitive information.[84]

Although multiparty computation offers great opportunities for privacy-enhanced collaborative data analysis, scaling these solutions comes with high overheads. Therefore, further improvements must be made to implement scalable MPC solutions. In addition, maintaining fairness between the parties (in terms of access to each other's data) is another issue that requires careful design considerations. Overcoming these issues may help address several challenges identified above, such as maintaining robust data privacy with utility guarantees, securing trusted personal data sharing, and dealing with multiple data sources and untrusted parties.[85]

3.5.3 Self-sovereign Identity (SSI) Management

Self-sovereign identity (SSI) management is a privacy-preserving solution that is based on the concept that the users should be the sole owners of their identity data. With self-sovereign identity, the data subjects do not have to rely on an intermediary to verify their identity on a digital platform and create their own verifiable credentials. Implementation of self-sovereign identity management has become more feasible with the advancements in distributed ledger technologies such as blockchain.[86] In recent years,

---

[83] Domingo-Ferrer, J., Blanco-Justicia, A. (2020). Privacy-Preserving Technologies. In International Library of Ethics, Law and Technology (Vol. 21, pp. 279–297), 285-286. Springer Science and Business Media B.V. https://doi.org/10.1007/978-3-030-29053-5_14.

[84] Choi, J. I., Butler, K. R. B. (2019). Secure Multiparty Computation and Trusted Hardware: Examining Adoption Challenges and Opportunities. *Security and Communication Networks*. https://doi.org/10.1155/2019/1368905.

[85] Domingo-Ferrer & Blanco-Justicia, *Privacy-Preserving Technologies*, 293.

[86] Mühle, A., Grüner, A., Gayvoronskaya, T., & Meinel, C. (2018). A survey on essential components of a self-sovereign identity. In Computer Science Review (Vol. 30, pp. 80–86), 80. Elsevier Ireland Ltd. https://doi.org/10.1016/j.cosrev.2018.10.002.

from an environment where each service provider maintains its own identity management system, they have started to rely on a number of large identity providers such as Facebook (Facebook Connect) and Google (Google Sign-In). Although this process streamlined the identification process of the service providers, the underlying economic interests of these large identity providers raise privacy and security concerns.[87]

Managing the identification process with a distributed ledger technology has the potential to ensure security, controllability, and portability of the personal data of the data subject.[88] Therefore, self-sovereign identity management solutions can be helpful in overcoming big data challenges such as secure and trusted personal data sharing.

3.5.4 Homomorphic Encryption

Homomorphic encryption is one of the popular privacy-preserving technologies that can be used for outsourced storage and computation. Homomorphic encryption is a form of encryption that allows calculations on encrypted data. In other words, homomorphic encryption is an encryption method providing additional evaluation capability for computing over encrypted data without a secret key. Therefore, both the input and the output of the computation process are encrypted that strengthen the privacy of the data transferred to third parties. Alternative homomorphic encryption methods are categorized based on the computations that they can perform over encrypted data. For example, partially homomorphic methods support only one type of operation, such as addition or multiplication. On the other hand, more complex methods such as somewhat, leveled-fully, and fully homomorphic encryption methods provide higher computational capabilities. However, the computational complexity comes with an increased cost; therefore, scaling systems with fully homomorphic encryption capacities poses a much bigger challenge. In addition to the difficulty of scaling, homomorphic encryption schemes are inherently malleable, and therefore, they have security vulnerabilities.[89]

3.5.5 Differential Privacy

Public release of datasets carries risks of identification even though they are carefully anonymized. For example, in 2007, researchers were able to identify 99% of the data subjects in the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, with their political opinion.[90] They achieve this de-anonymization by matching the dataset observations with the available user data on

---

87 Mühle & Grüner, *Self-Sovereign Identity*, 81.

88 Allen, C. (2016). The Path to Self-Sovereign Identity. Life With Alacrity. http://www.lifewithalacrity.com/2016/04/the-path-to-self-soverereign-identity.html.

89 Domingo-Ferrer & Blanco-Justicia, *Privacy-Preserving Technologies*, 286.

90 Narayanan, A., Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. Proceedings - IEEE Symposium on Security and Privacy, 111–125. https://doi.org/10.1109/SP.2008.33.

IMDB. One useful privacy-preserving solution for the public release of datasets is differential privacy. Differential privacy improves privacy by perturbing the values (*N*) in the dataset by maintaining a balance between utility and privacy. Perturbing can be done by adding random noise (*L*) to the values based on a distribution algorithm (*N*+*L*).[91] The added noise should neither be too big nor too small so that the dataset can still carry valuable insight, whereas the identifiability risk of the data subjects is minimized.

3.5.6 Document Sanitization and Redaction

Document sanitization and document redaction are two similar privacy-preserving methods developed to ensure that only the intended information can be accessed from a document. While document redaction consists of removing or blacking out sensitive information in a text document (e.g., AIDS →****), document sanitization consists of replacing sensitive information with more generic information (e.g., AIDS → disease). Sanitization is a more desirable method since it (i) still preserves the utility of the text and (ii) does not raise awareness of the sensitivity of the information to potential attackers as opposed to redaction.[92]

In both methods, two process consists of two steps. Firstly, sensitive information should be detected and marked. Secondly, the marked information should either be blacked-out or replaced with less sensitive content. For large datasets with free text content, the process might be tedious since they are mostly done manually. However, in recent years, automated redaction and sanitization methods have been proposed to accelerate and streamline the process.[93]

Sanitization and redaction can be used with anonymization methods and provide additional privacy against de-anonymization. Therefore, they shine out as complementary remedies against the limitations of anonymization and pseudonymization.

3.5.7 Federated Learning Approaches

Federated learning is a machine learning technique that offers additional privacy-preserving properties. Federated learning approach proposes the decentralized model training across multiple edge devices or servers using their local data samples without any access to external data sources. The difference between federated learning and distributed learning is that in federated learning, data is stored locally and not shared with other servers for training purposes, which strengthens the privacy property, whereas in distributed learning, the main goal is to parallelize the computer power.[94] To achieve

---

91 Domingo-Ferrer & Blanco-Justicia, *Privacy-Preserving Technologies*, 289.
92 Domingo-Ferrer & Blanco-Justicia, *Privacy-Preserving Technologies*, 290.
93 Domingo-Ferrer & Blanco-Justicia, *Privacy-Preserving Technologies*, 290.
94 Budig & Herrmann, *XAI in Healthcare*, 4-5.

privacy, federated learning approaches usually take advantage of other privacy-preserving technologies such as homomorphic encryption,[95] secure multiparty computation (MPC), and differential privacy.[96]

3.5.8 Distributed Ledger Technologies and Blockchain

A distributed ledger is a consensus of replicated, shared, and synchronized digital data stored in a distributed manner across multiple devices or nodes. They are not controlled by a central administrator.[97] On the other hand, blockchain is an implementation of distributed ledger technology and can be described as "*a public ledger distributed over a network that records transactions executed among network participants. Each transaction is verified by network nodes according to a majority consensus mechanism before being added to the blockchain.*" Distributed ledger technologies such as bitcoin can achieve overcome some of the challenges identified in this report, such as developing generic, easy-to-use and enforceable data protection solutions, processing sensitive data with enhanced privacy, and maintaining robust data privacy without damaging the utility of the personal data.[98]

3.5.9 Sticky Policies

The free flow of data is one of the driving power of Big Data innovation. However, once the data is released or shared with a third party, it is very difficult to control how the released data is being used. One privacy-preserving solution that we can address this problem is sticky policies. With sticky policies methods, machine-readable policies can be added to the released data in a standard format (e.g., XML or JSON) to improve the privacy and terms of use.[99] They are named "sticky" policies due to the fact that they are attached to the data and travel with it along its life cycle. [100] Sticky policies can regulate the formats that data can be accessed, the way it can be used throughout its life cycle, limitations of its use and share. [101]

---

[95] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. In ACM Trans. Intell. Syst. Technol (Vol. 10), 12:4. https://doi.org/.

[96] Truex, S., Steinke, T., Baracaldo, N., Ludwig, H., Zhou, Y., Anwar, A., & Zhang, R. (2019). A hybrid approach to privacy-preserving federated learning. Proceedings of the ACM Conference on Computer and Communications Security, 1–11, 1. https://doi.org/10.1145/3338501.3357370.

[97] Walport, M. (2015). Distributed ledger technology: Beyond block chain. Government Office for Science, 1–88, 21. https://youtu.be/4sm5LNqL5j0.

[98] European BDVA. *Strategic Research and Innovation Agenda*, 66.

[99] Pearson, S., Casassa-Mont, M. (2011). Sticky policies: An approach for managing privacy across multiple parties. Computer, 44(9), 60–68, 60. https://doi.org/10.1109/MC.2011.225

[100] Miorandi, D., Rizzardi A. Sticky Policies: A Survey, 3-4. Retrieved February 1, 2021, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8807248.

[101] Miorandi & Rizzardi*, Sticky Policies*, 1.

Sticky policies can provide assurance to the data processor regarding how the data they release is used, which would increase the circulation of the data in a privacy-enhanced fashion. Therefore, they can strengthen the pace of the big data innovations by providing additional privacy properties.

3.5.10 Algorithmic Auditing

Auditing is an effective measure to identify the legal and ethical risks that the big data systems might carry.[102] Algorithmic auditing is an effort to develop solutions that automatically evaluate these systems and identify the risks in a streamlined fashion.[103] By automating and standardizing the auditing process, data processors may analyze their services from ethical & legal standards and can remedy without breaching data subjects' rights as soon as relevant issues (e.g., algorithmic bias, illegitimate profiling, and discrimination) are detected. Algorithmic auditing can be a powerful privacy-preserving technology to overcome the challenges related to societal and ethical implications of big data technologies.[104]

3.5.11 Risk Assessment Tools

In addition to algorithmic addition, utilization of a set of risk assessment tools can shield the data processor from outstanding risks with their early detection mechanism. Risk assessment tools are a perfect answer to the risk-based data protection principle since they can measure the level of compliance with data privacy regulations (e.g., GDPR), identify privacy and cybersecurity risks (e.g., the reversibility of the anonymization mechanisms), recommends mitigations against these risks (e.g., differential privacy), and demonstrates accountability.[105] Risk assessment tools can be fully automated, semi-automated, or fully manual, depending on the sophistication of the tool and the nature of the assessed risk.

3.5.12 Automated Compliance

The feasibility and applicability of techno-regulation is a hotly debated issue. In other words, scholars often idealize computerized regulations published under a standard format (e.g., XML) and a fully automated compliance. While the widespread adoption of

---

[102] Deborah Raji, I., Smart, A., White Google Margaret Mitchell Google Timnit Gebru Google Ben Hutchinson Google Jamila Smith-Loud Google Daniel Theron Google Parker Barnes Google, R. N., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing ACM Reference Format, 1. https://doi.org/10.1145/3351095.3372873.

[103] Kassir, S. (n.d.). Algorithmic Auditing: The Key to Making Machine Learning in the Public Interest. The Business of Government, 1–4. http://www.businessofgovernment.com/sites/default/files/Algorithmic Auditing.pdf.

[104] Custers & La Fors, *Issues of Big Data*, 19.

[105] European BDVA. *Strategic Research and Innovation Agenda*, 67.

techno-regulation is under question, some fields offer unquestionable opportunities for techno-regulation implementations.[106] Therefore, especially for these fields, automated compliance can be a powerful solution to address some of the challenges we face today, such as measuring and reducing data controllers' obligations.[107]

Implementation of techno-regulations can streamline and automate the legitimacy verification of all the processes within the lifecycle of personal data.[108] With the advancement in these solutions, several protective protocols can be integrated into data processing systems, which requires compliance with the existing laws and valid certificates for processing. The main issue related to techno-regulation and automated compliance is to identify the limitations of the machine-readable policies for the protection of the data subjects' rights.[109]

3.5.13 Data Governance

Data governance can be summarized as defining the rules for accessing and sharing personal data by taking into account the privacy and data protection issues. With the ever-increasing volume, velocity, and variability of personal data, one of the challenges faced by the big data processing systems is to keep the quality of the data. The privacy property is one of the components of high-quality big data, and one of the main tasks undertaken by the data governance is to ensure proper data privacy and protection procedures are in place.[110] Data governance deals with the standardization of these procedures for sharing metadata, defining terms between stakeholders, providing guidance on the use of privacy-preserving technologies[111] such as encryption, pseudonymization, and anonymization for better privacy protection. [112]

3.5.14 Ethical and Technical Standards, Guidelines, Laws, and Codes of Conduct

The development and widespread adoption of big data processing systems inevitably raised several ethical and societal issues. Although many of these concerns can be resolved with advanced privacy-preserving technologies, pure technology implementation without proper standards will fail to deliver the expected results. In addition to privacy-preserving technologies, one of the important components of a successful data protection policy is to set and follow ethical and technical standards and guidelines. By involving the complete value chain of big data stakeholders, organizations can agree on and adopt standards and guidelines that reflect common ethical and societal values. The values that are already identified in existing AI guidelines (e.g., transparency, justice &

---

[106] Timan & Mann. *Data Protection in the AI Era*, 14.
[107] Timan & Mann. *Data Protection in the AI Era*, 11.
[108] Timan & Mann. *Data Protection in the AI Era*, 15.
[109] European BDVA. *Strategic Research and Innovation Agenda*, 55.
[110] European BDVA. *Strategic Research and Innovation Agenda*, 55.
[111] European BDVA. *Strategic Research and Innovation Agenda*, 26.
[112] European BDVA. *Strategic Research and Innovation Agenda*, 88.

fairness, non-maleficence, responsibility, privacy, beneficence, freedom & autonomy, trust, sustainability, dignity, solidarity) can be evaluated further to create applicable and widely adopted frameworks.[113] The enhanced dialogue between data subjects and other stakeholders in light of these standards and guidelines can improve the confidence of data subjects towards big data technologies and create a more reliable and sustainable big data ecosystem.[114]

3.5.15 Integration of Approaches, Toolboxes, Overviews, and Repositories of PPT

As listed above, there are dozens of privacy-preserving technologies, tools, and approaches addressing different challenges identified by the big data community. Although they provide valuable solutions to specific problems, without the uncoordinated implementation of these solutions may not provide the desired effect on the challenges. Therefore, one of the organizational measures that data processors must undertake is to coordinate the implementation and the integration of technical data protection measures, which involve all the relevant approaches, toolboxes, overviews, and repositories of privacy-preserving technologies. By doing so, data processors can achieve end-to-end data protection Combining different techniques for end-to-end data protection combining different techniques.[115]

3.6 Existing Solutions for Responsible Data Processing and ICT Systems

In the previous two sections, we identified the most significant challenges in big data and the most promising solutions proposed to address these challenges. In this section, we will list some of the projects that utilize the aforementioned solutions to address the challenges listed above. There are other very successful and significant projects addressing the big data challenges all around the world. However, since the focus in this report is on GDPR and its application, we limit the geographical scope of the analysis with European projects. To further narrow it down, we mainly focused on the projects funded as per the Horizon2020 program. To prepare this report, we identified and analyze 24 notable projects that adopt at least one of the solutions identified above.

3.6.1 Brief Descriptions of the Notable European Projects

In this section, these selected projects will be briefly introduced. For each project, the project scope, the project goal, the challenge it tries to address, and the technical and organizational measure they use will be shared. Then, collection of this information

---

[113] Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines, 7. In arXiv.

[114] e-SIDES Project. e-Sides. Retrieved February 11, 2021, from https://e-sides.eu/e-sides-project.

[115] Timan & Mann. *Data Protection in the AI Era*, 11.

will be used to create insights about the research trends and the major data privacy problems in the European Union.

BOOST: The BOOST (Big Data Value Spaces for COmpetitiveness of European COnnected Smart FacTories 4.0) project started in 2018 with the goal of improving the big data adoption in the European manufacturing industry and providing the industrial sector with the necessary tools to maximize their benefit from utilizing big data solutions.[116] Although the main theme of the project is not data protection,[117] the BOOST project aims to standardize the data modeling, storing, sharing in the manufacturing industry to increase the adoption of big data. Additionally, it aims this standardization with the integration of blockchain technology. Therefore, the BOOST project aims to address two identified challenges. By utilizing standardization efforts and blockchain adoption, they aim to deal with multiple data sources and untrusted parties. In addition, this standardization process will also support the challenge of maintaining a general, easy to use and enforceable data protection practice across different data processors.[118]

A4CLOUD: The A4CLOUD (Accountability for Cloud and other Future Internet Services) Project aims to enable cloud service providers to provide their users with a reasonable level of control and transparency over their personal data. To achieve this goal, A4CLOUD develops tools for cloud providers so that the users (i.e., data subjects) would have the confidence that their data is being processed in a legitimate manner and used appropriately. Data controllers and processors can benefit from A4CLOUD tools for risk assessment, automated compliance, and data governance. Development efforts of these tools will help to minimize the adverse societal and ethical implications of big data technologies. Automated compliance and risk assessment capabilities with up-to-date data subject consent lie in line with the risk-based approaches calibrating data controllers' obligations.[119]

GenoMed4ALL: The GenoMed4ALL (Genomics and Personalized Medicine for all through Artificial Intelligence in Hematological Diseases) project aims to improve state of the art to diagnose, treat and predict hematological diseases. To achieve this goal, GenoMed4ALL utilizes trustworthy AI algorithms that offer explainability. In addition to explainable AI algorithms, GenoMed4ALL will utilize federated learning and high-performance computing (HPC) to address the challenge of processing sensitive data.[120]

---

[116] BOOST4.0 CONSORTIUM. (2018). Boost 4.0 | Big Data for Factories. https://boost40.eu/.

[117] CONSORTIUM, B. . (n.d.). About Boost 4.0. In Boost 4.0. Retrieved March 19, 2021, from https://boost40.eu/wp-content/uploads/2018/02/boost_leaflet.pdf.

[118] Raggett, Dave. (2019). *Big Data Value Spaces for Competitiveness of European Connected Smart Factories 4.0.*

[119] About A4Cloud | Cloud Accountability Project. (n.d.). Retrieved March 19, 2021, from http://a4cloud.eu/about.html.

[120] GenoMed4All. (n.d.). About. 2021. Retrieved March 20, 2021, from http://genomed4all.eu/about/.

TRUSTS: The TRUST (Trusted Secure Data Sharing Space) project aims to develop a data-sharing platform that integrates European national digital markets and provides interoperability. To achieve this goal, the TRUST project analyzes the national legal frameworks within Europe and suggests standardization methods and ethical guidelines to the data processing ecosystem. Additionally, The TRUST project develops and offers privacy-aware analytics solutions by utilizing homomorphic encryption and multi-party computation.[121] By offering an ethical framework and standardization, TRUST deals with the adverse societal and ethical implications of big data technologies. Additionally, with the privacy-enhancing technologies it integrates into its ecosystem, it aims to achieve secure and trusted personal data sharing and deal with the limits of anonymization and pseudonymization by utilizing homomorphic encryption.

RESTASSURED: The RESTASSURED (Secure Data Processing in the Cloud) project focuses on developing end-to-end cloud architectures and methodologies to ensure secure data processing in the cloud. To achieve this goal, RESTASSURE uses homomorphic encryption. Also, it offers an implementation of sticky policies for decentralized data lifecycle management. It also introduced automated risk assessment and management tools. In this way, RESTASSURED ensures secure and trusted personal data sharing. It adopts a risk-based approach with the projected risk assessment tools.[122]

DECODE: The DECODE (Decentralized Citizens Owned Data Ecosystem) analyzes privacy-enhancing technologies and proposes a blockchain-based mobile app for accessing services privately. Additionally, the DECODE project designs a data governance framework to collect IoT data from Barcelona residents, which further strengthens the data protection practices. By applying these technologies to process the Barcelona residents' views on certain issues, DECODE shows that it also addresses the challenge of sensitive data processing. By developing a mobile app with privacy-enhancing technologies, DECODE adopts a general, easy to use and enforceable data protection approach and combines different techniques for end-to-end data protection.[123]

MUSKETEER: The MUSKETEER (Machine learning to augment shared knowledge in federated privacy-preserving scenarios) project aims to create machine learning models by taking different privacy-preserving scenarios into consideration. It aims to ensure security and robustness against external and internal threats and provide additional standardization. These efforts would potentially create a secure and scalable environment for data sharing. To achieve these goals, MUSKETEER utilizes privacy-preserving technologies such as multi-party computing, homomorphic encryption, and federated learning.[124]

---

[121] Motivation & Objectives - TRUSTS. (n.d.). Retrieved March 19, 2021, from https://www.trusts-data.eu/motivation-objectives/.

[122] FAQ Archive - RestAssured. (n.d.). Retrieved March 19, 2021, from https://restassuredh2020.eu/faq/.

[123] DECODE Tools | cryptography, authentication, anonymisation and data visualization. (n.d.). Retrieved March 20, 2021, from https://tools.decodeproject.eu/.

[124] Musketeer. (2020). ABOUT – MUSKETEER. https://musketeer.eu/project/.

SPECIAL: The SPECIAL (Scalable Policy-awarE linked data arChitecture for prIvacy, trAnsparency and compliance) project aims to address the contradiction between big data innovation and data protection compliance. To achieve this goal, it proposes a technical solution that streamlines the data sharing process with sticky policies. Thanks to sticky policies, data controllers and processors can easily share personal data with valid consent that is attached to the data shared. SPECIAL proposes a scalable and robust big data platform that complies with the GDPR and can provide feedback to the related stakeholders about how the data is used.[125]

MHMD: The MHMD (My Health – The My Data My Health) project mainly focuses on processing sensitive personal data. It aims to encourage hospitals to share anonymized medical data to research and prompt data subjects to become the real owners of their medical data. To improve the privacy of the medical data, MHMD proposes a system that utilizes privacy-preserving technologies such as multi-party computation, dynamic consent interface, and blockchain. MHMD also proposes ethical and legal guidelines to ensure more reliable and secure sensitive data processing practices.[126]

AEGIS: The AEGIS (Advanced Big Data Value Chain for Public Safety and Personal Security) project aims to create an "interlinked Public Safety and Personal Security Data Value Chain" and design a platform for curating, integrating, analyzing, and sharing big data. Apart from the technical and business goals, the AEGIS platform is set to enhance data privacy by utilizing privacy-preserving technologies such as anonymization, linked data, and blockchain.[127] In addition to the technology implementations, the AEGIS project aims to build its system after analyzing the ethical and societal implications of big data technologies.[128]

BPR4GDPR: The BPR4GDPR (Business Process Re-engineering and functional toolkit for GDPR compliance) project aims to develop a general-purpose data processing toolkit that complies with the data protection and privacy regulations in a robust and scalable manner. It aims to develop an end-to-end data processing platform by utilizing privacy-preserving technologies such as scalable anonymization & pseudonymization and risk assessment tools.[129] BPR4GDPR aims to go a step further and plans to offer automatic adaptation and transformation of processes to comply with privacy policies that served both regulatory and business goals.[130]

---

[125] SPECIAL. (n.d.). Home. Retrieved March 20, 2021, from https://www.specialprivacy.eu/.

[126] MHMD. (2019). My Health My Data. In My Health My Data. http://www.myhealthmydata.eu/.

[127] AEGIS. (n.d.). Approach – AEGIS Big Data. Retrieved March 20, 2021, from https://www.aegis-bigdata.eu/approach/.

[128] Da Bormida, M. (2019). Tackling ethical issues in a H2020 Project in the Big Data domain - AEGIS Ethics White Paper, 7-15.

[129] Dellas, N. (2019). Initial Specification of BPR4GDPR architecture, 34-36.

[130] BPR4GDPR. (n.d.). Innovation Proposal. 48. Retrieved March 20, 2021, from https://www.bpr4gdpr.eu/about/research-description/.

PAPAYA: The PAPAYA (PlAtform for PrivAcY preserving data Analytics) project aims to create a secure and trust-based ecosystem for data analytics. PAPAYA particularly aims to develop this system with scalability and robustness in mind. With the development of the proposed platform, data owners will be able to extract valuable insight from protected personal data. To achieve these goals, PAPAYA proposes several modules such as the Privacy Preferences Manager (PPM), The Data Subject Rights Manager (DSRM), and a user-centric Graphical User Interface (GUI), which comprises the Privacy Engine (PE). With the introduction of these modules, PAPAYA aims to ensure secure and trusted personal data sharing and maintain a robust data privacy compliance with utility guarantees for the data that comes from multiple data sources and untrusted parties.[131]

SMOOTH: The SMOOTH (GDPR Compliance Cloud Platform for Micro Enterprises) project aims to provide Micro-enterprises with easy-to-use and affordable tools to facilitate their compliance with the GDPR. SMOOTH develops a cloud-based platform built upon existing privacy-preserving technologies such as automated auditing and risk analysis tools. The novelty of the SMOOTH project is to integrate and combine different techniques for micro-enterprises, the most vulnerable group for involuntary data privacy violations, to achieve end-to-end data protection.[132]

PDP4E: The PDP4E (Methods and tools for GDPR compliance through Privacy and Data Protection Engineering) aims to integrate effective privacy and data protection engineering functionalities into the other tools already used by engineers. PDP4E also aims to integrate data protection methods such as LINDDUN, PRIPARE, and PROPAN into mainstream software methodologies and data processing models. Therefore, its innovation is the integration of privacy-enhancing methods into the widely adopted engineering standards and methods.[133] The four main functionalities that the PDP4E project focuses on are risk management, engineering requirements, privacy-aware design, and assurance management functionalities.[134]

DEFeND: The DEFeND (Data Governance for Supporting GDPR) project -*similar to the PDP4E project*- focuses on improving existing data processing tools and methods and developing a new integration software with the purpose of delivering an organizational data privacy platform designed with the Privacy-by-Design principle in mind. The platform offers tools and methods across three management areas, including data

---

131 Timan & Mann. *Data Protection in the AI Era*, 13.

132 SMOOTH. (n.d.). About Smooth Project. Retrieved March 20, 2021, from https://smooth-platform.eu/about-smooth-project/.

133 Martin, Y. S., & Kung, A. (2018). Methods and Tools for GDPR Compliance Through Privacy and Data Protection Engineering. Proceedings - 3rd IEEE European Symposium on Security and Privacy Workshops, EURO S and PW 2018, 108–111. https://doi.org/10.1109/EuroSPW.2018.00021.

134 Yod, S. M. (2019). PDP4E - D 2.4 Overall system requirements, 12. https://www.pdp4e-project.eu/deliverables/.

scope, data process, and data breach. DEFeND project aims to combine different techniques for end-to-end data protection with a general, easy to use and enforceable data protection approach.[135]

MOSAICrOWN: The MOSAICrOWN (Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and Owner control) project aims to enable data sharing and collaborative analytics in multi-owner scenarios. MOSAICrOWN aims to provide effective and deployable solutions allowing data owners to take control of their data with a data governance framework. In addition to the data governance framework, MOSAICrOWN also aims to develop data wrapping and data sanitization techniques to enhance data privacy further. MOSAICrOWN is one of the unique projects that aim to develop data sanitization and wrapping techniques to address the limitations of anonymization and pseudonymization. MOSAICrOWN outcomes will combine different techniques for end-to-end data protection and also help to deal with multiple data sources and untrusted parties.[136]

SODA: The SODA (Scalable Oblivious Data Analytics) project aims to address big data issues such as secure and trusted personal data sharing, secure processing of sensitive data, and dealing with multiple data sources and untrusted parties. After identifying that for reliable data analytics, the personal data does not have to be shared; instead, it should be made available for encrypted processing; it advocates the use of multi-party computation protocols for data analytics. Besides, it proposes the use of differential privacy to further decrease the re-identifiability of the personal data without harming the reliability of the results.[137]

PoSeID-on: The PoSEID-on (Protection and control of Secured Information by means of a privacy-enhanced Dashboard) project aims to develop a blockchain-based platform that facilitates the exercising data subject rights such as the right to be forgotten.[138] To achieve its goals, the PoSEID-on project analyzes the legal and ethical principles and proposes a comprehensive framework. Then, it develops a blockchain-based platform that data subjects can exercise their rights as articulated within this framework. The compliance and the associated risks as per the framework are monitored to ensure appropriate protection of the data subject's rights. The PoSEID-on project adopts a risk-based approach to calibrating data controllers' obligations and thanks to its blockchain technology. Since it aims to develop trustworthiness, sustainability, and ethics-driven

---

135 DEFeND. (n.d.). What is the Defend Project - Defend Project. Retrieved March 20, 2021, from https://www.defendproject.eu/.

136 MOSAICrOWN. (n.d.). Research work. MOSAICrOWN. Retrieved March 20, 2021, from https://mosaicrown.eu/the-project/research-work/.

137 Timan & Mann. *Data Protection in the AI Era*, 16.

138 Riccio, G. M., Peduto, A., Iraci, F., Briguglio, L., Sartin, E., Occhipinti, C., Gutiérrez, I., & Natale, D. (2021). The POSEID-ON Blockchain-Based Platform Meets the "Right to Be Forgotten", 14. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3745516.

Technologies, it also addresses one of the major challenges: Societal and ethical implications of big data Technologies. [139]

E-SIDES: The E-SIDES (Ethical and Societal Implications of Data Sciences) project aims to identify ethical and societal issues of privacy-preserving big data technologies. This analysis is followed by the validation of the effectiveness of privacy-preserving big data technologies.[140] As opposed to the majority of the relevant projects, E-SIDES predominantly focuses on the ethical and legal issues in big data. It aims to improve the dialogue among big data stakeholders and improve the trust toward big data technologies and data markets.[141]

LINDDUN: The LINDDUN is a privacy threat modeling project that develops a methodology for analysts to systematically detect and eliminate privacy threats in big data systems. The LINDDUN project was developed after analyzing several privacy properties such as hard-soft privacy, unlinkability, anonymity, pseudonymity, plausible deniability, Undetectability and unobservability, Confidentiality, Content awareness, and Policy and consent compliance. After analyzing privacy properties and the privacy threats associated with these properties, the LINDDUN project creates threat tree patterns that can be used to adopt privacy-preserving technologies to address these threats. Therefore, the LINDDUN project helps to address the challenge of combining different techniques for end-to-end data protection.[142]

XAI: The XAI (Science and technology for the explanation of AI decision making) project aims to take the state of the art of explainable AI a step further. XAI is a Horizon project whose main goal to is to develop solutions to construct meaningful explanations of black-box ML models to empower data subjects against the negative effects of automated decision making. XAI develops algorithms to infer local and global explanations from black-box models. They aim to work on languages representing explanations in logic rules and statistical interpretation concepts. They aim to build a platform to share their findings in an open-source fashion. Finally, they plan to propose a framework to analyze the interaction between Explainable AI and the legal & ethical issues relevant to AI.[143]

---

[139] Riccio & Peduta, *The POSEID-ON*, 29.

[140] La Fors, K. (n.d.). Human-centric big data governance: responsible ways to innovate privacy-preserving technologies. Retrieved March 20, 2021, from https://e-sides.eu/resources/e-sides-lessons-for-the-responsible-innovation-of-privacy-preserving-technologies-in-the-era-of-ai-karolina-la-fors-e-sides-beyond-privacy-learning-data-ethics-14-nov-2019-brussels.

[141] e-SIDES Project. e-Sides. Retrieved February 11, 2021, from https://e-sides.eu/e-sides-project.

[142] Wuyts, K. (2014). LINDDUN : a privacy threat analysis framework. https://distrinet.cs.kuleuven.be/software/linddun.

[143] *Xai - Website*. (n.d.). Retrieved August 26, 2022, from https://xai-project.eu/research-lines.html.

NL4XAI: Similar to the XAI project, NL4XAI (Interactive Natural Language Technology for Explainable Artificial Intelligence) aims to empower data subjects against the negative effects of AI systems. NL4XAI focuses on generating automated explanations in a human-understandable and interactive fashion. To achieve its goals, NL4XAI relies on the power of natural language generation and processing as well as argumentation techniques. After testing and validating the project outcomes, these techniques will be shared in an open-source software framework. While the XAI project aims to develop XAI techniques and methods in an overall sense to enhance the protection of right to explanation, NL4XAI focuses on NLP and NLA-oriented human understandable explanation generation. Therefore, NL4XAI's focus is narrower and more specific.[144]

XMANAI: The XMANAI (Explainable Manufacturing Artificial Intelligence) project aims to bring explainability to the manufacturing industry. It prioritizes a human-centric and trustful approach to be tested in real-life manufacturing cases. Since the manufacturing industry is prioritized in this project, the main goal of the project is to achieve trust in the AI systems instead of empowering of data subjects in fundamental rights and freedoms. XMANAI aims to propose hybrid models (glass-box models that are explainable to a domain expert) that are explainable to activate a human-in-the-loop and produce value-based explanations with complex AI-enabled technologies to multiply the latent data value in a trusted manner. XMANAI uses targeted manufacturing apps to solve concrete manufacturing problems with high impact in 4 pilot cases.[145]

TAPAS: The TAPAS (Towards an Automated and exPlainable ATM System) Project aims to bring the power of explainability to ATM (Air Traffic Management) Systems. TAPAS aims to systemically explore AI solutions that can be useful in ATM scenarios, such as conflict detection and resolution in Air Traffic Control and Air Traffic Flow Management. TAPAS aims to develop XAI techniques tailored to the aviation industry's needs. To achieve these goals, they rely on data visualization techniques to enhance the explanations generated by AI systems. The main explainability goal of the project is to increase trust in the system.[146]

3.6.2 Quantitative Evaluation of the Projects

After numerically evaluating the details of these projects, we can create numerical insights. Table 1 lists the 24 projects, the challenges they address, and the privacy-preserving technologies and organizational measures they utilize:

---

144 *About – NL4XAI – Interactive Natural Language Technology for Explainable Artificial Intelligence*. (n.d.). Retrieved August 26, 2022, from https://nl4xai.eu/about/.

145 *Project – XMANAI*. (n.d.). Retrieved August 26, 2022, from https://ai4manufacturing.eu/project/.

146 *TAPAS – SESAR ER – Descripción corta del proyecto*. (n.d.). Retrieved August 26, 2022, from https://tapas-atm.eu/.

Project Abbreviation	The Challenges Addressed	PPTs and Org. Measures
BOOST 4.0	C6, C7	DLT, ETS
A4CLOUD	C2, C9	RAT, DG, AC
GenoMed4ALL	C1, C4, C6	XAI, FL
TRUSTS	C2, C3, C5	MPC, HE, ETS
RESTASSURED	C1, C3, C5, C9, C10	HE, SP, RAT, INT
DECODE	C4, C7, C10	DLT, INT
MUSKETEER	C1, C3, C7, C8, C10	FL, MPC, HE
SPECIAL	C1, C8	SP
MHMD	C3, C4, C10	DLT, INT
AEGIS	C3, C6	DLT
BPR4GDPR	C5, C7, C8, C9, C10	AC, AA, RAT, INT
PAPAYA	C1, C3, C6, C7, C8	MPC, AA
SMOOTH	C2, C7, C10	AA, RAT, INT
PDP4E	C10	AA, RAT, DG, INT
DEFeND	C7, C10	DG, ETS, RAT, INT
MOSAICrOWN	C5, C6, C10	DG, DSR, INT
SODA	C3, C4, C6	DP, MPC
PoSeID-on	C2, C9	DLT, RAT
E-SIDES	C2	ETS
LINDDUN	C7, C10	RAT, ETS, INT
XAI	C2	XAI
NL4XAI	C2	XAI
XMANAI	C2, C10	XAI, INT
TAPAS	C2, C10	XAI, INT

Table 1. Notable European PPT and Organizational Measure Projects with Their Main Focus Areas and the Challenges They Address

Privacy-Preserving Technologies and Organizational Measures

• **XAI**: Explainable AI • **MPC**: Secure Multiparty Computation • **SSI**: Self-sovereign Identity (SSI) Management • **HE**: Homomorphic Encryption • **DP**: Differential Privacy • **DSR**: Document Sanitization and Redaction • **FL**: Federated Learning Approaches • **DLT**: Distributed Ledger Technologies and Blockchain • **SP**: Sticky Policies • **AA**: Algorithmic Auditing • **RAT**: Risk Assessment Tools • **AC**: Automated Compliance • **DG**: Data Governance • **ETS**: Ethical and Technical Standards, Guidelines, Laws, and Codes of Conduct • **INT**: Integration of Approaches, Toolboxes, Overviews, and Repositories of Privacy-Preserving Technologies

The Challenges Identified in the Previous Section

• **C1**: Contradiction between Big Data innovation and data protection • **C2**: Societal and ethical implications of big data technologies • **C3**: Secure and trusted personal data sharing • **C4**: Processing sensitive data • **C5**: Limits of anonymization and pseudonymization • **C6**: Dealing with multiple data sources and untrusted parties • **C7**: A general, easy to use and enforceable data protection approach • **C8**: Maintaining robust data privacy with utility guarantees • **C9**: Risk-based approaches calibrating data controllers' obligations • **C10**: Combining different techniques for end-to-end data protection

The projects that are examined under this section have a wide range of starting dates. While the earliest of them all, A4CLOUD, started in October 2012, the most recent

project, GenoMed4ALL, started in January 2021. The median starting date for the projects is January 2018. The starting date of the project is a significant indicator of its focus area. Among these 24 projects, while the initial theme was the adoption of blockchain technology until 2018, after 2018, we see a redistribution of the themes from risk assessment tools to encryption techniques. Finally, the most recent project, GenoMed4ALL, priorities explainable AI and federated learning, which might be the starting point of a new trend. The overall privacy-preserving technology and organizational measure solutions and the total number of projects that use these solutions are shared below:

Abbreviation	PPT or Org. Measure	#Project Adopted
INT	Integration of Approaches, Toolboxes, Overviews, and Repositories of PPT	11
RAT	Risk Assessment Tools	8
DLT	Distributed Ledger Technologies and Blockchain	5
ETS	Ethical and Technical Standards, Guidelines, Laws, and Codes of Conduct	5
XAI	Explainable AI	5
MPC	Secure Multiparty Computation	4
AA	Algorithmic Auditing	4
DG	Data Governance	4
HE	Homomorphic Encryption	3
FL	Federated Learning Approaches	2
SP	Sticky Policies	2
AC	Automated Compliance	2

Table 2. The Frequency Table of the Adoption of the PPT Solutions and Organizational Measures in the Selected Projects

3.6.2.1 Integration of Existing Methods

Table 2 shows that the Integration of Approaches, Toolboxes, Overviews, and Repositories of PPTs is the most preferred solution for big data issues, with 11 of the 24 projects offered related solutions. Therefore, most of the projects choose to integrate existing privacy-preserving technologies and offer data privacy and protection solutions in a particular field such as medicine[147] or SMEs[148]. Apart from integrating existing technologies, most projects develop risk assessment tools for GDPR compliance. Additionally, we see a high adoption ratio of DLT and Blockchain technologies, especially in older projects. Although there are five projects developing solutions in "ethical and technical standards, guidelines, laws, and codes of conduct," this category does not seem saturated relative to its wide scope.

---

[147] MHMD. (2019). My Health My Data. In My Health My Data. http://www.myhealthmydata.eu/.

[148] SMOOTH. (n.d.). About Smooth Project. Retrieved March 20, 2021, from https://smooth-platform.eu/about-smooth-project/.

3.6.2.2 The Promising PPTs Not Covered under the Selected Projects

On the other hand, some of the PPT solutions are mostly disregarded. Differential privacy, document sanitization and redaction, federated learning, and multi-party computation can be useful to address the limitations of anonymization, pseudonymization, and encryption techniques. However, most of these projects fail to develop revolutionary PPT solutions. In addition, there is a significant mismatch between the significance of explainable AI and the number of projects in this category.

Abbreviation	Challenge	#Project Covered
C10	Combining different techniques for end-to-end data protection	12
C2	Societal and ethical implications of big data technologies	9
C7	A general, easy to use and enforceable data protection approach	8
C3	Secure and trusted personal data sharing	7
C6	Dealing with multiple data sources and untrusted parties	6
C1	Contradiction between Big Data innovation and data protection	5
C4	Processing sensitive data	4
C5	Limits of anonymization and pseudonymization	4
C8	Maintaining robust data privacy with utility guarantees	4
C9	Risk-based approaches calibrating data controllers' obligations	4

Table 3.The Frequency Table of the Coverage of the Data Privacy and Data Protection Challenges in the Selected Projects

3.6.2.3 Combining Different Techniques and Offering Easy-to-Use Data Privacy Solutions

For the challenge categories covered in the notable data protection projects list, in parallel with the integration efforts, most projects aim to combine different techniques for end-to-end protection. This result is in agreement with the trend that most data privacy and data protection projects aim to combine existing PPT solutions and offer a platform for businesses in a particular field. The second most popularly addressed challenge, A general, easy to use and enforceable data protection approach, also supports this thesis. Since most companies do not have in-house data protection expertise, most projects try to standardize and facilitate the adoption of GDPR-compliant data processing platforms. While some project aims to integrate PPT solutions to existing infrastructure

that are already in the market[149], others develop their own platforms by combining PPT solutions and other big data solutions.[150]

3.6.2.4 Dealing with Secure Data Transfer and Multiple Data Sources and Untrusted Parties

After these two closely related challenge categories, secure and trusted personal data sharing and dealing with multiple data sources and untrusted parties are the following two very important. After identifying that there is ever-increasing data flow, data generated, and data sources, around a third of the projects aim to address the secure transfer of data and dealing with multiple data sources. While multi-party computation and homomorphic encryption support secure data transfer and enable data analytics and machine learning among untrusted parties, sticky policies, integration efforts, and ethical & technical standards help to deal with multiple data sources.

3.6.2.5 More Emphasis Needed in Some Challenge Categories

During the examination of the 24 notable projects, the least addressed challenges were limits of anonymization and pseudonymization, maintaining robust data privacy with utility guarantees, and risk-based approaches to calibrating data controllers' obligations. Although many projects offer secondary solutions to these challenges, most projects do not prioritize these challenges. Although identifiability is identified as a significant indicator of personal data and most anonymized and pseudonymized data can be reidentified, there is a lack of data privacy-oriented projects aimed to address this challenge. Additionally, even though there are projects which aim to develop dynamic consent and sticky policy solutions, the number of solutions in this field seems limited. Finally, we identified that the number of projects with a high focus on adopting a risk-based approach is relatively low.

3.6.2.6 Addressing Societal and ethical implications of big data technologies

In our analysis, we identified that nine projects actively aim to address adverse societal and ethical implicants of big data technologies. Considering the wide variety of societal and ethical implications of big data technologies, this number can be higher. On the one hand, each project at least mentions partially covers ethical and legal frameworks in its deliverable documents. However, this is rather a determinant to be included and examined in this report. The projects that put special emphasis on addressing the ethical and societal implications of big data systems are relatively limited, and their number can be increased.

3.6.2.7 User-Centric Data Protection Approach

---

[149] BPR4GDPR. (n.d.). Innovation Proposal. 48. Retrieved March 20, 2021, from https://www.bpr4gdpr.eu/about/research-description/.

[150] DEFeND. (n.d.). What is the Defend Project - Defend Project. Retrieved March 20, 2021, from https://www.defendproject.eu/.

One of the most frequently covered issues in notable projects is the user-centered data protection approach. This issue is usually covered as "giving back the control of the data to its owner." In many projects, we see platform proposals in which the data subject can edit the consent dynamically. The SPECIAL project utilizes sticky policies that are attached to data in metadata form so that the data receivers would always know to what extent they can use and share personal data.[151] The MOSAICrOWN project also provides deployable solutions that allow data owners to maintain control of the data-sharing process.[152]

3.6.2.8 Automation of GDPR Compliance, Auditing, and Risk Assessment Tools

There has been a discussion on whether it is healthy to automate regulatory actions or whether hardcoding laws is a dangerous practice because the legal field is argumentative and dynamic.[153] While this theoretical discussion continues, many projects implement some sort of automation in data privacy compliance, auditing, and risk assessment. While SPECIAL aims to achieve automated compliance by utilizing sticky policies[154], BPR4GDPR[155], SMOOTH[156] , and PDP4E[157] projects propose automated auditing with risk assessment tools and data governance frameworks.

3.6.2.9 Gaining Momentum on Explainable AI Solutions to Address Ethical and Societal Implications

As we approach the mass adoption of AI in sensitive fields, one of the hotly debated data privacy issues is the right to explanation.[158] Utilizing AI systems in sensitive fields may cause the violation of several ethical principles such as non-discrimination, transparency, and accountability.[159] Although explaining the decisions of AI systems is an important issue for the ethical and societal implications of big data technologies, the projects that aim to offer solutions in Explainable AI just started to experience a momentum. While none of the projects started before 2018 have an XAI component in their projects, we are seeing increasing popularity following the GDPR's enforcement

---

151 SPECIAL. (n.d.). Home. Retrieved March 21, 2021, from https://www.specialprivacy.eu/.

152 MOSAICrOWN. (n.d.). Homepage. Retrieved March 21, 2021, from https://mosaicrown.eu/.

153 Timan & Mann. *Data Protection in the AI Era*, 14.

154 SPECIAL. (n.d.). Home. Retrieved March 21, 2021, from https://www.specialprivacy.eu/

155 BPR4GDPR. (n.d.). Innovation Proposal. 48. Retrieved March 20, 2021, from https://www.bpr4gdpr.eu/about/research-description/.

156 SMOOTH. (n.d.). About Smooth Project. Retrieved March 20, 2021, from https://smooth-platform.eu/about-smooth-project/.

157 Yod, S. M. (2019). PDP4E - D 2.4 Overall system requirements. https://www.pdp4e-project.eu/deliverables/.

158 Sartor, G. (European U. I. of F. (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. Panel for the Future of Science and Technology (STOA), 1st, 76-79. https://doi.org/10.2861/293.

159 Yalcin, *Examination of Current AI systems*, 2-3.

year of 2018. Out of the 24 projects examined, 5 projects, GenoMed4ALL, XAI, NL4XAI, XMANAI, TAPAS propose and offer explainable AI solutions.[160]

3.7 Final Remarks

Designing and implementing GDPR-compliant algorithms for big data processing is a challenging task that requires expertise in legal, ethical, and technical domains. This report aimed to provide a comprehensive guideline for the data privacy and data protection ecosystem. To achieve this goal, the fundamentals of relevant technical topics, namely, data processing, big data, and AI, are explained. To understand the legal framework, the relevant GDPR provisions and the liabilities of the data processors and controllers are also detailed. To better understand the background of the GDPR, the most common eight ethical issues were listed and explained. After covering the fundamentals, we identified the main Big Data challenges that are relevant to data protection and privacy. After identifying the legal obligations of the data processors and controllers and the challenges, we covered the privacy-preserving technologies (PPTs) that might address these obligations and issues. To have a thorough review, we analyzed 24 notable projects that are almost exclusively funded as per the Horizon2020 program. After analyzing the most notable European projects, we conducted a quantitative analysis of the challenges and the solutions in which we also identified the issues that require additional attention. While the integration of existing tools is selected by most of the projects as the main goal, the development of the PPTs is usually disregarded. Therefore, a clear need for interdisciplinary research projects that focus on developing novel PPT solutions (e.g., explainable AI, sticky policies, homomorphic encryption) is observed.

4 Need for Explainability in AI and Decision Making to Enhance Data Privacy

In the previous chapter, we carefully analyzed the data processing ecosystem and its players. We covered the main data processing challenges and organizational measures and privacy preserving technologies to address these challenges. From this chapter onwards, the focus will be on Explainable AI. In this chapter, we will cover the significance of explainable AI for today's data society and justification for the second part of the thesis.

Artificially intelligent (AI) systems offer many benefits to individuals and public & private institutions. Thanks to AI systems and software automation, the services which require a high level of human involvement may be provided quickly with low to no human involvement using machine learning. With the help of applied statistics and affordable computing power, engineers can develop AI systems to complete difficult

[160] GenoMed4All. (n.d.). About. 2021. Retrieved March 20, 2021, from http://genomed4all.eu/about/.

tasks such as designing driverless cars, building machine translation software, or developing algorithmic profiling systems.[161]

Since the primary goal of AI systems is to increase efficiency and accuracy,[162] machine learning engineers often overlook the explainability of their systems. The assumption of an engineer tends to be that as long as the model accurately predicts the result of a future event, its outcomes will satisfy the relevant parties. Even though accurately predicting an event outcome is the most important task that the AI systems have, we cannot diminish the system performance to just accuracy since there will be erroneous predictions made by the same system. Whenever an automated decision causes damages to one of the key data privacy actors, such as data subjects or third parties, because of an incorrect prediction, there may be liabilities and obligations as well as violations of fundamental rights & freedoms. In these situations, the reasoning of the AI systems will be crucial to understanding the logic behind the incorrect prediction and conducting an accountability check. For instance, in a recent study in the U.S. Fintech sector, the researchers found that mortgage refinancing algorithms used in the U.S. -as well as the professionals in this field- discriminate against Latin and African American borrowers.[163] Even the legitimate-business-necessity interpretation is taken into account,[164] the research shows that at least 6% of the minority applications are rejected due to purely discriminatory practices.[165] The hidden discrimination and bias in credit applications is just a simple and less harmful example of what AI systems can cause. Soon, armed UAVs with AI systems, AI judges, and AI police bots will take over their respective jobs, where accountability and liability are a significant part of the process. They will make decisions in irreversible matters which involve fundamental rights and freedoms such as the right to live, the right to bodily integrity, and the right to freedom.[166] Therefore, discriminatory or incorrect decisions may cause significant material and moral damages.[167]

---

[161] Etzioni, A, Etzioni, O.. *Keeping AI Legal* (2007). Vand. J. Ent. & Tech. L. XIX, 1. 2, https://perma.cc/UQ7F-7VYX.

[162] Domingos, P. *A few useful things to know about machine learning* (2012). Communications of the ACM 55, 10. 3, ISSN: 00010782, DOI:10.1145/2347736.2347755

[163] Bartlett, R. P. et al., "Consumer Lending Discrimination in the FinTech Era," SSRN Electronic Journal, November 2017, 1, DOI:10 . 2139 / SSRN. 3063448, https://dx.doi.org/10.2139/ssrn.3063448.

[164] Bartlett, *Consumer Lending Discrimination*, 3-4.

[165] Bartlett, *Consumer Lending Discrimination*, 21.

[166] Samek W, Wiegand T, Müller K-R. *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models* (2017). ITU Journal: ICT Discoveries 1, no. Special Issue 1: 2-3, arXiv: 1708.08296, http://arxiv.org/abs/1708.08296.

[167] Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos, "Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making," Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019): 1-3, ISSN: 2159-5399, DOI:10 . 1609/aaai.v33i01.33011418, arXiv: 1903.10598, www.aaai.org.

4.1 The Contribution and the Unique Nature of the Research

There are two main categories observed when the previous studies are analyzed. Studies that are in the first group are done by legal scholars researching the possible adverse effects of the widespread use of AI systems that are not explainable in sensitive domains. Since the usage of AI systems in sensitive fields may raise questions regarding violation of fundamental rights & freedoms and may create significant damages due to incorrect decisions, determining the parties' accountability and liability are very substantial. Therefore, legal scholars focus on the accountability and liability of the parties when damages are suffered, and rights are violated.

On the other hand, the second group of studies are conducted by machine learning and data science expert and focuses on the statistical analysis of the AI systems, transferability of the trained models to new areas, and cause-and-effect relationships between explanatory and response variables. These groups of researchers focus on understanding the decision-making process of the trained AI system.[168] However, legal reasoning might have to contain more information regarding the event that a technical expert foresees.

Therefore, there must be a bridge between the legal scope of explainability and the field of explainable artificial intelligence. Only with this bridge study the expectations of the public and law community may be met by the technical researchers. Therefore, this research will act as a bridge between the expectations of the public, the law community, and the works of the technical experts in building meaningful explainable AI systems.

Utilizing the increased network connectivity (thanks to the Internet), robotics & software automation, and cheap computing power, humanity is entering into an era where the mainstreamed and repetitive tasks are fully automated with artificially intelligent systems. Large enterprises and governments have already utilized intelligent systems in many of their tasks. However, this is still the beginning of the AI era. With the advancements in machine learning, Intelligent systems will increasingly be used in sensitive tasks. Therefore, the decisions of the Intelligent systems will be subject to many civil and penal disputes. Therefore, explaining the decision-making mechanism of these systems (i.e., explainability) will be a very crucial component of securing justice in a healthy society. This research aims to satisfy this need by approaching it from a law-oriented perspective as well as bearing the technical side in mind. By reviewing the latest academic and business literature and by experimenting on the current explainable AI and AI models, legally acceptable XAI systems will be developed and presented.

---

[168] Barredo Arrieta, A., Díaz-Rodríguez, N., del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 8. https://doi.org/10.1016/J.INFFUS.2019.12.012.

4.2 The Explainability vs. Accuracy Problem

There are a number of different algorithms that may be used for machine learning which have different levels of success on the accuracy metrics. Although the traditional algorithms are highly interpretable, AI engineers are likely to prefer deep learning algorithms over traditional algorithms due to the high performance of deep learning algorithms on accuracy metrics (see Fig 2).[169]

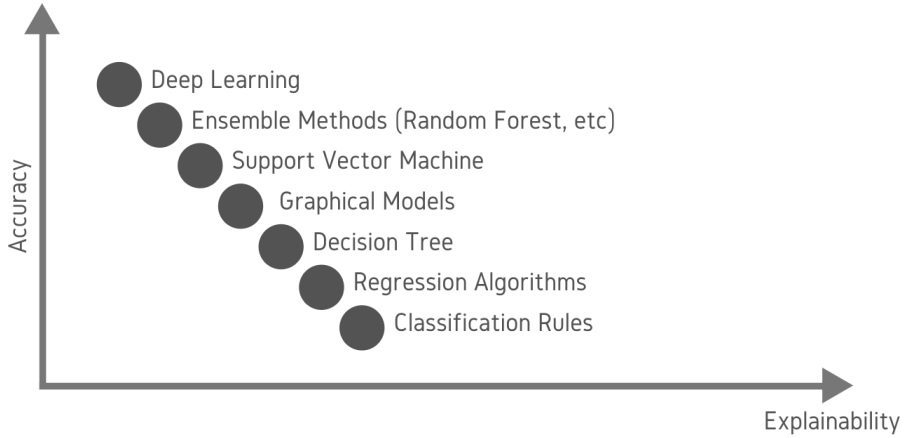


Fig 2. Accuracy-Explainability Plot of Various AI Algorithms[170]

In other words, machine learning algorithms that are highly explainable usually have low accuracy performance in a relative sense, especially when there is an abundance of data. Therefore, as long as there is no constraint on computational power and there is enough data, ML engineers tend to select models with higher accuracy while ignoring their low-level explainability. In addition, the popularity of predefined machine learning libraries (e.g. Keras, Tensorflow, PyTorch, Scikit Learn) also contributes to the widespread use of black-box models and to the negligence of the explainability in the AI systems.[171]

If a new wave of research does not solve the negative correlation observed between explainability and accuracy, in a near future, AI judges, soldiers, armed drones, police officers, and other sensitive AI systems must have to use the algorithms with high accuracy; therefore, with low explainability. The low level of AI system explainability will certainly be problematic in securing the right to explanation, particularly in areas

---

169 Preet Gandhi, Explainable Artificial Intelligence, 2019, accessed November 12, 2019, https://www.kdnuggets.com/2019/01/explainable-ai.html.

170 Turek, Matt, *Explainable Artificial Intelligence (XAI)* (2016). 1, accessed November 12, 2019, https://www.darpa.mil/program/explainable-artificial-intelligence

171 Bernhard Waltl and Roland Vogl, "Explainable Artificial Intelligence-the New Frontier in Legal Informatics," Jusletter IT, 2018, 3.

such as transportation, security, medicine, finance, legal, and military.[172] One may argue that the right to explanation is not a widely accepted and essential right today. However, it is not hard to foresee that with the new advancements in technology, the significance of this right will gain momentum, and it will soon become part of the fundamental rights & freedoms. For instance, in the U.S., credit scoring decisions must already be given with reasoning; therefore, algorithms used for credit scoring must be explainable.[173] On the other hand, the decisions made in these fields may constitute a violation of the traditional fundamental rights & freedoms as well. For instance, the decision of an AI judge without reasoning -regardless of its accuracy- will violate the right to a fair trial.[174]

The scope of explainability must also be examined from ethical and legal standpoints as well as a cognitive perspective to propose suggestions to develop truly explainable systems. Due to the rapid increase in the number of publications in the field, the cluster of the explainable AI literature created its own nomenclature with a variety of adjacent terms including, but not limited to, understandability, intelligibility, comprehensibility, transparency as well as interpretability and explainability. In the next section, we will first clarify the terminology confusion caused by using similar terms interchangeably in XAI.[175]

4.3 Term Clarification on Explainability

Since the field of Explainable Artificial Intelligence is relatively in its infancy, researchers have been using several terms similar to explainability. In fact, in the very beginning of the field, interpretability was a more prevalent term used instead of explainability.[176] Although the term explainability gained more popularity and has become the mainstream term used by the majority of the researchers as the field matures, there are still several adjacent and rival terms that should be defined here and possibly distinguished from explainability. In a non-exhaustive manner, these terms can be listed as follows:

- Understandability or Intelligibility
- Comprehensibility
- Interpretability
- Explainability
- Transparency
- Explicability
- Predictability
- Legibility

---

172 Turek, *Explainable Artificial Intelligence*.

173 Jiahao Chen, "Fair lending needs explainable models for responsible recommendation," FATREC 2018, 2018, 2, arXiv: 1809.04684v1.

174 European Court of Human Rights, Guide on Article 6: Right to a Fair Trial (Criminal Limb), technical report (2013), 32, https://perma.cc/C4XN-AE8N.

175 Barredo and Díaz-Rodríguez, *Explainable AI*, 5.

176 Barredo and Díaz-Rodríguez, *Explainable AI*, 3.

- Readability

In the upcoming subsections, we will briefly cover how these adjacent terms are used in the literature. Although this practice may not solve the term clarification issues entirely, it can reduce the vagueness we observe in the field of Explainable AI.

4.3.1 Understandability or Intelligibility

Understandability and intelligibility are two terms that are often used interchangeably. It refers to a model's characteristic to allow humans to understand its inner logic without any other explanation. For instance, the understandability of the black box models is very limited without the contribution of the explainability techniques. Understandability can be also considered as a goal that we want to achieve with the explainability of the AI systems. Understandability is the term usually used by cognitive scientists.[177]

4.3.2 Comprehensibility

Comprehensibility refers to the representability of the AI system's knowledge in a human-understandable manner. Therefore, it requires AI system outputs to contain symbolic descriptions of the entities that are similar to a human expert's output for the same prediction. This comparison can be made with the semantic and structural properties of the AI system's output.[178] Comprehensibility is an important property of AI systems, which makes them more simulatable, which is a property of model transparency. The importance of comprehensibility and simulatability will be covered in the dedicated Explainable AI section in more detail.

4.3.3 Interpretability

Interpretability usually refers to the capability of explaining a model's behavior to human understandable terms. Interpretability is usually used by the technical community referring to the technical interpretability of the model algorithm and decision-making process. Since the field of artificial intelligence often lead by the computer scientists with statistical expertise, interpretability of AI models, the Explainable AI studies are also started by this sub-community of researchers. Therefore, they often used the term interpretability in their studies and aimed to make the AI models technically more interpretable. Only in recent years the term explainability took over as the most popular Explainable AI term.[179]

---

[177] Lim, B. Y., Yang, Q., & Abdul, A. M. (2019). Why these Explanations? Selecting Intelligibility Types for Explanation Goals. *IUI Workshops*. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-20.pdf

[178] Fernandez, A., Herrera, F., Cordon, O., Jose Del Jesus, M., & Marcelloni, F. (2019). Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine*, *14*(1), 69–81. https://doi.org/10.1109/MCI.2018.2881645

[179] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 3.

4.3.4 Explainability

Compared to the term interpretability, explainability of an AI system refers to a broader scope. In addition to the interpretability of a model, to be deemed as explainable, the model should employ proper means to communicate the explanations generated by the model or the explainability technique.[180] While every explainable model is also interpretable, the reverse would not be a valid statement. After pioneered by the technical sub-community of Explainable AI researchers, the field of Explainable AI has gained popular among the data scientists with domain expertise who seek overall AI system explainability in their particular use cases. Therefore, the goal of Explainable AI was extended to provide adaptable explanations to end-users and the researchers, apart from developing interpretable models.

4.3.5 Transparency

Transparency is usually used as the opposite of the opaqueness of the model. A transparent model is understandable without any additional techniques. Fully transparent models are referred to as white-box models, while opaque models are referred to as black-box models. The level of transparency of the ML models can be -from less transparent to more transparent- defined by their algorithmic transparency, decomposability, and, finally, simulatability. [181] As mentioned earlier, the comprehensibility of the model can contribute to the model's transparency. Only comprehensible models can be simulatable by humans. On the other hand, the interpretability of a model will contribute to its algorithmic transparency. Finally, transparent models can provide explanations without external explainability techniques. However, the AI systems that relies on black-box (non-transparent) models can still be explainable with post-hoc explainability techniques.

4.3.6 Explicability

Explicability refers to an ethical principle that requires the AI system’s processes to be transparent, its capabilities and purpose to be openly communicated, and its decisions are explainable to those who are directly or indirectly affected by them.[182] Explicability is listed as one of the four ethical principles that compose the foundation of Trustworthy AI Framework with a number of fundamental rights. AI HLEG points out that explicability principle must be respected ensure that AI systems are developed, deployed and used in a trustworthy manner.

---

[180] Barredo and Díaz-Rodríguez, *Explainable AI*, 5.
[181] Barredo and Díaz-Rodríguez, Explainable AI, 5.
[182] AI HLEG, *Trustworthy AI*, 13.

4.3.7 Predictability

Predictability refers to the matching an observer's expectations with an AI system's actual behavior in each situation. Therefore, predictable AI systems tend not to surprise its observers with unexpected outcomes.[183] Although predictability of an AI system is not necessarily required for all applications, to achieve the reliability of the system, it might be an important property.

4.3.8 Legibility

Legibility refers to the quality of AI systems to expose their intention from their behaviors. In other words, by looking at its structure and past behavior, a legible AI system's objective can be detected with minimal effort. Therefore, by analyzing its processes, we can observe an AI system's goal.[184]

4.3.9 Readability

Readability refers to the quality of AI systems to have their notion of behavior to be human-readable. Therefore, readable AI systems can empower the users to quickly capture its processes and logic and easily predict how it will behave. [185]

In the literature, we observe several terms that are used with explainability, sometimes in a complimentary manner while synonymously at other times. In this section, we briefly covered nine of these terms; however, this list is far from being exhaustive. In the next section, we will identify the main Explainable AI goals pursued by different AI stakeholders.

4.4 Goals of XAI

The motivation of the previous studies varies depending on the interests of different Explainable AI stakeholders such as (i) data subjects, (ii) domain experts, (iii) data scientists & developers, (iv) company managers, and finally, (iv) regulatory entities. The literature shows that different stakeholders seek different explainability components in their AI systems. While domain experts look for trustworthiness, transferability, and confidence, regulatory entities and data subjects seek causality, policy awareness, and fairness.[186] On the other hand, some goals, such as informativeness, are desired by all the stakeholders. In this section, we will briefly cover these XAI goals and the stakeholders interested in these goals.

---

[183] Sado, F., & Loo, C. (2020). *Explainable goal-driven agents and robots-a comprehensive review and new framework*, *4*. https://www2.informatik.uni-hamburg.de/wtm/publications/2021/SLKW21/2004.09705v1.pdf

[184] Sado & Loo, *Explainable goal-driven agents*, 4.

[185] Sado & Loo, *Explainable goal-driven agents*, 4.

[186] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 8.

4.4.1 Trustworthiness

Trustworthiness is the confidence in a model to behave as intended when it was designed. Trustworthiness is a property of Explainable AI systems which are sought by domain experts and the users of these systems.[187] Trust in an AI system does not only require technical trust in model development and deployment, but also require socio-technical qualities such as trustworthiness of all actors and processes throughout the entire AI lifecycle. Especially the definition of the trustworthiness of AI-HLEG requires AI systems to be lawful, ethical, and robust, which refers to the socio-technical part.[188] Although trustworthiness is an important goal and component of overall explainability of AI systems, it does not automatically a model explainable. In other words, not every trustworthy AI system is explainable.[189]

4.4.2 Causality and Causability

One of the main goals of XAI techniques is to allow detecting causal relationship between variables. Standard versions of machine learning models are not capable of providing causal relationships. Instead, they can provide correlation information between variables. On the other hand, causal reasoning requires hypothesis testing with domain knowledge. There has been a long tradition of causal reasoning studies in both statistics and psychology. Domain experts can propose hypotheses containing causal relationships of variables based on the machine learning output and their domain expertise. These hypotheses can only be tested with the explanations provided by AI systems. Therefore, being able to provide explanations that can be used for causal reasoning is one of the goals of XAI. This capability of AI is referred to as causability[190] and is often sought by domain experts, managers, and regulatory entities.[191]

4.4.3 Transferability

Transferability refers to the ability to reuse the inferred knowledge that exists in an AI system in other cases, and it is one of the highly sought goals of XAI studies.[192] XAI techniques can help generating explanations which may help to understand the inner logic of ML models, which may also help with the knowledge transferability of these models. However, transferability does not directly make a model explainable, but instead, it is a goal that the domain experts and data scientists would like to achieve with Explainable AI techniques.[193]

---

187 Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 8.

188 AI HLEG, Trustworthy AI, 6-7.

189 Rawal, A., McCoy, J., Rawat, D., Sadler, B., & Amant, R. (2021). *Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives*. https://doi.org/10.36227/TECHRXIV.17054396.V1

190 Rawal & McCoy, *Recent Advances in Trustworthy XAI*, 3.

191 Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 8.

192 Barredo and Díaz-Rodríguez, *Explainable AI*, 8.

193 Rawal & McCoy, *Recent Advances in Trustworthy XAI*, 4.

4.4.4 Informativeness

Informativeness, as being one of the main XAI goals, refers to capability of providing information on a number of issues related to the AI system. These issues include inner logic of the model, information about individual decisions, dataset used for training and testing, and all the other relevant data. Apart from being targeted by data scientists, users, domain experts, and managers, it is a goal sought by regulatory agencies since informativeness is a mandatory components of AI systems, which are covered by relevant GDPR articles.[194] Informativeness can empower data subjects (i.e., users of AI systems) to review the decisions that affect them directly or indirectly. They can seek human oversight or can challenge the decision before relevant authorities. These safeguards can only be used when the AI systems are informative.

4.4.5 Confidence

Closely related to trustworthiness, confidence is goal of XAI that focuses on a model's robustness and stability. Explainable AI techniques can provide explanation about an AI system's robustness and stability. For example, the level of determinism and sensitivity analysis results can be helpful to establish confidence in AI systems.[195]

4.4.6 Fairness

Fairness covers the development, deployment, and use of AI systems and covers non-discrimination and elimination of bias. The data collected and used for model training contains our historical biases, stigmas, and discriminatory practices.[196] Therefore, AI systems can infer these discriminatory logical rules and cause unfair practices. Explainable AI can initially empower data scientists to remove existing biases from the dataset and its effect on model training. The, during the use of AI systems, Explainable AI can empower data subject's to properly use the safeguards defined in the relevant articles of GDPR to check if they are being subject to discriminatory practices. Finally, it can help regulatory entities to conduct audits and investigation against discriminatory practices.[197]

4.4.7 Accessibility

Accessibility is wide aimed goal by the XAI community. It refers to data subject's ability to access to relevant information about individual decisions and general logic of the AI systems. Understanding the relevant information about an AI system is particularly difficult for non-technical data subjects. In addition, the channels used to present the explanations should be accessible by the receivers. Finally, AI systems should employ

---

[194] Barredo and Díaz-Rodríguez, *Explainable AI*, 8.
[195] Barredo and Díaz-Rodríguez, *Explainable AI*, 8.
[196] AI HLEG, Trustworthy AI, 12-13.
[197] Rawal & McCoy, *Recent Advances in Trustworthy XAI*, 11.

adaptable ways to present information to the receivers so that they can properly access to the explanation that would be useful for them.[198]

4.4.8 Interactivity

Interactivity can closely be associated with accessibility. While accessibility refers to have access adaptable and user-centric explanations, interactivity refers to receivers' capability to interact with the AI system. This interaction can be with the explanation interface as well as with the model or dataset.[199] This minor goal of XAI is still significant for the protection of the right to explanation and for the realization of Trustworthy AI.

4.4.9 Privacy Awareness and Compliance

Employing AI systems in sensitive areas can lead to privacy issues. For example, profiling someone's sensitive personal data and using this to provide AI-enabled services can be against GDPR requirements. Privacy awareness is an important goal that is sought by the data subjects.[200] Explainable AI is essential to understand AI systems that comply with the mandatory data privacy laws and regulations for individual decisions and provide explanations to the data subjects who are directly or indirectly affected by the decision. In addition, Explainable AI techniques can be effective when regulatory entities conduct general privacy compliance investigation.

4.5 Current Developments in Explainable AI from Legal and Technical Perspectives

The state of the art in artificial intelligence may be evaluated from different perspectives such as legal, ethical, technical, and even cognitive psychology. From the legal and ethical perspectives, the significance of the explainability of AI systems has already caught the attention of European and American policymakers and scholars to some extent. As mentioned above, the right to explanation in credit scoring has been a long-standing right in the U.S. On the European side, with the EU General Data Protection Regulation (GDPR) enacted in 2016, the right to explanation was strengthened with Art. 13-15 of the GDPR -which all read "The data subject shall have... access to ... the existence of automated decision-making... " and "... meaningful information about the logic involved"-.[201] On most occasions, the party influenced by the AI system is not aware of the parameters used in the model and the sampling of the data (train and test

---

[198] Barredo and Díaz-Rodríguez, *Explainable AI*, 9.

[199] Barredo and Díaz-Rodríguez, *Explainable AI*, 9.

[200] Barredo and Díaz-Rodríguez, *Explainable AI*, 10.

[201] Bryce Goodman and Seth Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"," AI Magazine, 2017, 6, ISSN: 07384602, doi:10.1609/aimag.v38i3.2741, arXiv: 1606.08813.

data), as well as which parameter is given more weight for the prediction. The E.U. and the U.S. have already had a set of preliminary rules in place to mitigate this problem. As mentioned above, fundamental rights and freedoms, such as the right to a fair trial or the right to life, may be used as a shield against unfair practices in AI systems. From the technical perspective, designing explainable AI systems is also significant for recognizing cause and effect relationships to improve existing systems.

We have already covered some of the Horizon-funded European projects (e.g., XAI, NL4XAI, XMANAI, TAPAS) that propose explainability solutions. On the other hand, before the initial European attempts, The Defense Advanced Research Projects Agency (DARPA) had already initiated the first project aiming to improve the explainability of AI systems. In a competition held by DARPA, eleven U.S. universities, in partnership with industrial players and European universities, have proposed novel explainable AI systems. These participant universities have suggested explainable learners (a combination of explainable models and explanation interfaces) in addition to their research on the psychological model of explanation. Their systems may focus on one of these three subcategories: (i) Deep Explanation, (ii) Interpretable Models, and (iii) Model Induction. Briefly, Deep explanation teams aim to develop modified deep learning models in which explainable features may be extracted. Interpretable model teams focus on traditional & causal methods (e.g., And-Or grammars, Hierarchical Bayesian Networks, and Random Forests) and try to develop more explainable models (more structured, interpretable, and causal). Finally, Model induction teams try to induce novel models by testing the black box models.[202]

CP	Performer	Explainable Model	Performer
Both	UC Berkeley	Deep Learning	Reflexive and Rational
	Charles River	Causal Modeling	Narrative Generation
	UCLA	Pattern Theory+	3-level Explanation
Autonomy	Oregon State	Adaptive Programs	Acceptance Testing
	PARC	Cognitive Modeling	Interactive Training
	CMU	Explainable RL (XRL)	XRL Interaction
Analytics	SRI International	Deep Learning	Show and Tell Explanation
	Raytheon BBN	Deep Learning	Argumentation and Pedagogy
	UT Dallas	Probabilistic Logic	Decision Diagrams
	Texas A&M	Mimic Learning	Interactive Visualization
	Rutgers	Model Induction	Bayesian Teaching

Fig 3. The Ongoing XAI Research in the U.S.[203]

[202] Gunning, David, Explainable Artificial Intelligence (XAI), Technical Report (2017), 10-18, https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

[203] Gunning, *Explainable Artificial Intelligence*, 10-18.

4.6 Final Remarks

Ideally, when an AI system is used for matters which involve public services, justice, and other heavily regulated areas, the system must be transparent and explainable.[204] In addition to the explainable nature of the machine learning models, this also means that the feature selection process must be accessible publicly or upon request. Furthermore, the sampling principles of the training and test datasets must be very well explained. Finally, the design of the model must be explainable by nature which indicates both the use of explainable machine learning algorithms and also the availability of an interface with which the administrator of the AI system or the relevant authorities can analyze the results.

In the following Chapter, we will move on to the cognitive model for explanations. After defining how humans reason and how machine learning models reason, we will attempt to understand how we can use machine explanations to assist human reasoning and mitigate some of the human biases. Following the cognitive reasoning part, we will move on to analyzing the GDPR's right to explanation framework with designated safeguards. We will analyze the ideal explainability standards for legal compliance. We will also associate these safeguards with the Trustworthy AI principles and identify their contribution to the realization of Trustworthy AI.

After the cognitive, legal, and ethical analysis, we will conduct a technical analysis of Machine Learning and Deep Learning models. Additionally, we will briefly cover relevant issues such as the history of AI, machine learning problems and paradigms, the AI development cycle, and the black-box nature of neural networks.

Following the machine learning and deep learning overview chapter, explainability techniques in different stages of AI development cycle will be identified and presented in a structured format. While most XAI research focuses on model explainability, there are several stages of the AI development cycle where the overall explainability of the system can be strengthened. The relevant explainability techniques are applied throughout the whole AI development cycle starting from the data collection stage until after individual predictions are outputted by the system. In other words, the entirety of AI system components (the algorithm, selection of the features, sampling of the train data, test data, validation data, presentation logic of the explanations, and all the other aspects) must be transparent and interpretable for a comprehensively explainable system. On the other hand, where the model explainability is in question, these techniques can be grouped into two main categories: (i) Ante-hoc approaches and (ii) Post-hoc approaches.[205] Ante-hoc approaches aim to achieve explainability with the model design and other assisting methods, whereas the post-hoc approaches aim to extract explanations from existing models.[206] A third approach usually considered within the

---

[204] Turek, *Explainable AI*.

[205] Kacper Sokol and Peter Flach, "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, December 2019, 57- 58, doi:10.1145/3351095.3372870, arXiv: 1912.05100, http://arxiv.org/abs/1912.05100%20http://dx.doi.org/10.1145/3351095.3372870.

[206] Barredo and Díaz-Rodríguez, Explainable AI, 10-12.

post-hoc approaches is the global mimic approach, which aims to generate models that mimic the overall behavior of complex black box models.[207] However, for the sake of simplicity, global mimic approach will be covered under post-hoc explainability.

After these chapters, we will become capable of combining these inputs from a variety of domain fields including, cognitive psychology, law, ethics, and information technologies. Therefore, we will be informed about different stakeholders' pain points, interests, and expectations. Besides, we will also know the goals that these stakeholders try to achieve with Explainable AI. Finally, we will know the limitations of Explainable AI in these fields and the available resources to strengthen the explainability of the AI systems. The research will continue to examine these approaches and concepts to achieve the desired explainability goals as the designed framework requires. After identifying these boundaries for legally, ethically, and technically acceptable explainability, we will have an opportunity to propose a comprehensive checklist to guarantee a desired level of explainability. These findings will at least result in a shift in the explainability axis of the Accuracy vs. Explainability plot, shown in Fig 4.

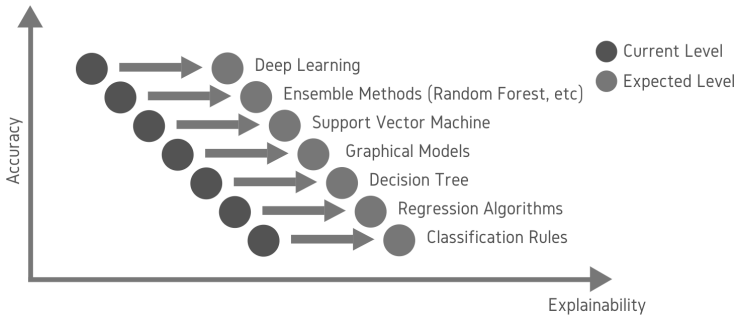


Fig 4. The Shift Aimed with the research on the Accuracy Explainability Plot

5 Cognitive Analysis of Explanations and Explainability

In the previous chapter, we identify the various goals that different stakeholders try to achieve with Explainable AI. From trustworthiness to causality, from accessibility to privacy awareness, the realization of these goals depends on the explanations generated by the AI systems whose decisions need to explain these goals. Therefore, explanations are one of the main tools that help these stakeholders achieve their goals regardless of the model and data they use.

From a philosophical and psychological standpoint, explainability refers to the degree to which humans can understand, comprehend, and reason with a decision or an action. Explainability can be achieved with the transparency of the decision-making process of AI systems. However, full transparency cannot always be achieved, especially in the existence of black box models. In addition, apart from the model explainability, the

[207] Sokol & Flach, *Explainability Fact Sheets*, 58.

overall AI system explainability also requires explainability at pre-modeling, post modeling stages as well as management-level policies towards sustainable explainability of the AI system. Therefore, researchers have been proposing and developing explainability techniques to mitigate these transparency issues. While these techniques can generate informative data that potentially contains explanations, it is important to understand what constitutes an explanation and how we can achieve explainability with this data. To be able to answer these questions, we need to understand how humans reason and explain their decisions and behaviors. Then, we need to analyze how Explainable AI techniques generate explanations and compare them with human explanation processes.

For this analysis, we will rely on the framework proposed by Wang, Yang, Abdul, and Lim[208] where they aim to generate user-centric explanations with Explainable AI techniques.[209] Their framework was inspired by multidisciplinary studies and relies on social sciences, as Miller suggests.[210] Lim's proposal builds on top of Hoffman's[211] and Klein's[212] work on how humans formulate and accept explanations. Therefore, the multidisciplinary nature of the framework is in line with the multidisciplinary approach of this thesis. In addition, the framework is designed for the developers to be used in software development processes, which is also in line with the practical approach we adopted in this thesis.

5.1 How People Reason and Explain Things

In this section, we will scrutinize how human beings reason and generate explanations. To understand the explanation process of humans, we first need to understand the goals that we -as human beings- aim to achieve with explanations. Then, we move on to reasoning approaches (e.g., deductive, inductive, and abductive reasoning) to understand how humans use these approaches to explain behaviors or decisions. After understanding the scope of these reasoning approaches, we move on to comparing two similar concepts: causal attribution and causal explanations and how they relate to each other.

5.1.1 Explanation Goals

Depending on the role of the receiver, the goal that was to be achieved with the explanation may vary. In a broad sense, explanations are often used for **filtering causes** to simplify the observation complexity and **generalize observations** into a conceptual model to **predict future outcomes**. Additionally, explanations are often used to

---

208 Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *Undefined*. https://doi.org/10.1145/3290605.3300831

209 Lim & Yang, *Selecting Intelligibility Types*, 4-5.

210 Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.48550/arxiv.1706.07269

211 Hoffman, R. R., & Klein, G. (2017). Explaining Explanation, Part 1: Theoretical Foundations. *32*(3), 68–73. https://doi.org/10.1109/MIS.2017.54

212 Klein, G. (2018). Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems*, *33*(2), 83–88. https://doi.org/10.1109/MIS.2018.022441353

strengthen the **transparency** of the decision-making process, which is also a legal requirement in sensitive use cases. In connection with the transparency of the system, explanations can also lead to **better decision-making**. If the system does not behave in a particular manner, explanations can help for **tracing and debugging** problems and issues. Since explanations tend to make AI systems more transparent, they also **increase the trustworthiness** of these systems.[213]

For instance, from a practical perspective, if the end user is the receiver of the explanation, the goal might be able to build trust of the user to use the system so that the engagement increases, which leads to higher profit. Providing explanations can be done with purely compliance purposes as well. While several legal norms (e.g., GDPR Art. 13-15, Art. 22, and Recital 71) dictates the AI system developers to provide explanations under certain circumstances. The AI system developers may also choose to comply with non-binding legal or ethical principles to develop and maintain more sustainable AI systems. Finally, explanations can contribute to the realization of Trustworthy AI by strengthening compliance with ethical principles, such as **human oversight**, **fairness, diversity, and non-discrimination,** as well as **accountability**.[214]

5.1.2 Inquiry and Reasoning

Humans mainly reason using three distinct approaches, namely deduction, induction, and abduction, and also their derivatives such as analogical reasoning and hypthetico-deductive reasoning. **Deduction** refers to the process of reasoning from premises to the conclusion. It is usually used synonymously with the top-down reasoning approach.[215] For instance, if our premise is that all fish can swim, if we use deductive reasoning, we can conclude that since sharks are a type of fish, they can swim. In contrast, **induction** refers to reasoning from a single observation to generally applicable conclusions. Inductive reasoning is the reverse of deductive reasoning and is often referred to as bottom-up reasoning.[216] For instance, we observe different fish species such as sharks, dolphins, whales, and salmon and that they can all swim. With inductive reasoning, we can conclude that all fish can swim. However, without observing the swimming capabilities, the conclusion of the inductive reasoning is highly questionable. **Abduction** is a reasoning approach similar to inductive reasoning. However, in inductive reasoning, the most likely explanation is prioritized and used for generalization. Inductive reasoning is often referred to as inference to the best explanation. Finally, **analogical reasoning** refers to reasoning from instance to instance, and it is referred to as case-based reasoning. It is regarded as a weak form of inductive reasoning, yet it is often used for legal reasoning to generate explanations based on precedence and analogy. Therefore,

---

213 Wang & Yang, *Designing Explainable AI*, 4.

214 Yalcin, O. G. (2022). The Role of the Right to Explanation and Its Safeguards in the Realization of Trustworthy AI. *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence*, 179-182.

215 Wang & Yang, *Designing Explainable AI*, 4.

216 Wang & Yang, *Designing Explainable AI*, 4.

despite its weak strength, it has applications in a number of sensitive fields. [217] In addition to the conventional form of reasoning, it is possible to observe hybrid reasoning formats as well. **Hypthetico-deductive reasoning** is one of them. In hypthetico-deductive reasoning, the reasoning person would observe and identify a problem, form an induction-based hypothesis as induction from observations and deduce the predictions relevant to the hypothesis, and further test the hypothesis.[218]

5.1.3 Causal Attribution and Explanations

Causal attributions are the articulations of internal or external factors that have an influence on an outcome or observation. Causal attributions do not have to be explanations since they do not always contain information about the main cause that leads to a output. They provide -on the other hand- important information that can lead to potential causes, which can lead to the generation of causal chains.[219] Together with explanations, causal attributions are important components in creating causal chains.

Causal explanations refer to explanations that inform the receivers about the selected causes of the observations and predictions. Causal explanations are usually contrastive between a fact and a foil. Therefore, they clarify why an event occurred in contrast to another.[220] Due to their contrastive nature, they are easily comprehensible and simulatable by humans. Causal explanations can also answer the "why not" questions to infer why a foil did not occur. Therefore, the selected subset of the causes can provide counterfactual explanations for changing the outcome for given inputs.[221] In other words, counterfactual explanations are the explanations that can explain why an outcome would not have occurred if something has not had happened.[222]

5.1.4 Decision Theories

Decision theories try to explain how people make decisions under uncertainty based on their expected value, risk, and probabilistic distribution of the outcomes.[223] Based on people's level of risk aversion, the expected value of the alternative outputs can vary,

---

[217] Lim & Yang, *Selecting Intelligibility Types*, 4-5.

[218] Lawson, A. E. (2000). The Generality of Hypothetico-Deductive Reasoning: Making Scientific Thinking Explicit. *American Biology Teacher*, *62*(7), 482–494. https://doi.org/10.2307/4450956

[219] Wang & Yang, *Designing Explainable AI*, 5.

[220] Jacovi, A., Swayamdipta, S., Ravfogel, S., Elazar, Y., Choi, Y., & Goldberg, Y. (2021). Contrastive Explanations for Model Interpretability. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 1597–1611. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.120

[221] Lim & Yang, *Selecting Intelligibility Types*, 5.

[222] Molnar, C., & Dandl, S. (n.d.). Counterfactual Explanations. In *Interpretable Machine Learning*. Retrieved August 22, 2022, from https://christophm.github.io/interpretable-ml-book/counterfactual.html

[223] Wang & Yang, *Designing Explainable AI*, 5.

and cause variations of the decisions based on their personality.[224] Different perception of risk can usually be rooted back to external factors. For example, a person who is on the brink of bankruptcy can be more risk-averse when making a new investment decision. When reasoning is done on a relatively insignificant matter, the risk perception can be loosened such as choosing a cookie policy when visiting a social media platform. Therefore, decision theory is concerned with the underlying reasons behind these varying choices. Decision theory is both a theory of beliefs and a theory of choice. Therefore, it deals with both the agent's beliefs, desires, and attitudes (i.e., preference attitudes) and how these attitudes adhere together.[225]

5.2 The Concepts that Empower Explainable AI Techniques to Generate Explanations

Algorithmic explanations are created based on different paradigms, but they can be linked back to human reasoning. By generating explanations, these techniques can support human reasoning and specific methods of scientific inquiry (e.g., Bayesian probability, similarity modeling, and queries). To achieve these goals, they can generate explanations in different formats and data structures.

In this section, we will cover the important concepts that empower AI systems to generate inferences and explanations. We will first describe Bayesian probability, which dictates the decision mechanisms of many machine learning algorithms. Then, we cover how similarity modeling using historical observations to predict the labels of new observations powers most machine learning algorithms' inner logic. In this process, we scrutinize how intelligibility queries can be used to obtain meaningful information regarding an AI system's behavior and decisions. The responses received with intelligibility queries can be used to form explanation pieces (i.e., XAI elements) that can be stored in standard or special data structures. Finally, these pieces can be visualized with data visualization methods to be consumed and understandable by humans more easily.

5.2.1 Bayesian Probability

Bayes' Theorem is a mathematical equation to calculate the conditional probability of a given event. It is widely used in statistical calculations and for inductive reasoning operations.[226] Bayesian probability (i.e., conditional probability) refers to the probability of an event based on the known conditions related to that event. Therefore, according

---

[224] Fishburn, P. C. (1990). Utility Theory and Decision Theory. *Utility and Probability*, 303–312. https://doi.org/10.1007/978-1-349-20568-4_40

[225] Steele, K., & Stefánsson, H. O. (Winter 2020). *Decision Theory*. Stanford Encyclopedia of Philosophy. Retrieved October 29, 2022, from https://plato.stanford.edu/archives/win2020/entries/decision-theory/

[226] Joyce, J. (2021). *Bayes' Theorem*. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem/

to Bayes' theorem, the probability of an event depends on the prior conditions, probabilities, and likelihoods. For instance, according to Bayes' theorem, a model should consider that likelihood of an output is less likely to be a rare label.[227]

5.2.2 Similarity Modeling

Similarity modeling is the concept of associating objects and observations with other similar objects and distinguishing them from them based on their unique values. Clustering algorithms, classification models such as logistic regressions, dimensionality reduction algorithms such as PCA, autoencoders, collaborative filtering, and case-based reasoning rely on similarity modeling. Most of these machine learning models rely on training data to generate their models. Explanations generated from these models often rely on inductive and analogical reasoning. Identifying causal attributions with additional effort can help receivers to understand the underlying logic behind the relations and the rulesets.[228]

5.2.3 Intelligibility Queries and Types

Intelligibility queries can be described as requests that are responded with pieces of information that can be combined with other intelligibility queries to compose meaningful explanations. Lim identifies several intelligibility queries that can be outputted from an AI system. The non-exhaustive list of queries consists of inputs, output, certainty, why, why not, how to, what if, and when.[229] For instance, the query "when" can be used to obtain information regarding the relevant times about a system output. The query "when" can be used to learn the time of the decision, time of the observation, the time of the expiry of a certain event, decision, or opportunity. Therefore, different varieties can be created from one query type. The queries can be in question format such as what, when, or what if as well as in system state component format. One can request to obtain information about the list of inputs (i.e., dataset) used to train a model as well as the set of inputs to generate a certain output.

While providing information on input, output, and certainty (i.e. system state) has been historically easier to achieve, generation explanations for the other queries have become more difficult with the advancements in the black box machine learning models. Jiarpakdee and Tantithamthavorn provide query examples for intelligibility:[230]

- ***"What:** What is the logic behind the AI/ML models?*
- ***Why:** Why is an instance predicted as TRUE?*
- ***Why Not:** Why not is an instance predicted as FALSE?*

---

227 Wang & Yang, *Designing Explainable AI*, 5.

228 Wang & Yang, *Designing Explainable AI*, 5.

229 Wang & Yang, *Designing Explainable AI*, 6.

230 Jiarpakdee, J., Tantithamthavorn, C., Dam, H. K., & Grundy, J. (n.d.). *A Theory of Explanations*. Retrieved August 20, 2022, from https://xai4se.github.io/xai/theory-of-explanations.html

- ***How To:*** *How can we reverse the prediction of an instance (e.g., from TRUE to FALSE) generated by the system?*
- ***What If:*** *What would the system predict if the values of an instance are changed?"*

The explanations derived from these queries can be combined to generate comprehensive explanations. These explanations can help boost human understanding of the reasoning of AI systems.

5.2.4 Explainable AI Elements

While there are dozens of XAI techniques, the building blocks that compose the XAI explanations tend to group around several elements. The first element is the feature attribution or influence that demonstrates which feature has a negative or positive significant effect on an outcome. Another element is the similar or different instances of the training data, prototypes, criticism, and counterfactual examples. Names and values of input or outputs and definitions of rules are examples of other elements.[231]

5.2.5 Data Structures

Explanations can be in almost any type of data structure. While strings are the preferred options for textual explanations, integers and floats can be used for numerical explanations. In more complex explanations, lists and dictionaries can be useful to present explanations with multiple components. Logical clauses can be described with decision trees, and ontologies can be represented in the form of graphs that consists of nodes and edges. By using graphs, deductive reasoning operations can be conducted. With approximation and abstraction of patterns, extendable objects can be used to represent models, and using these objects, analogical and inductive reasoning can be employed.[232]

5.2.6 Data Visualization and Graphing

In addition to a wide variety of data structures, visuals, graphs, and plots are other forms that explanations can be generated. Line plots can be useful to represent sequential data that can contribute to the transparency of the datasets. Node-link diagrams can show model structures that would be useful to strengthen a model's transparency. Tornado diagrams and scatterplots can be useful for detecting correlations between features. Other alternative methods, such as text highlighting, can be used for case-based reasoning. Partial dependency plots can be used for visualizing how feature attribution varies across different feature values. Finally, tables or other tabular objects can be used for sensitivity analysis and simulation studies.[233]

---

[231] Wang & Yang, *Designing Explainable AI*, 6.
[232] Wang & Yang, *Designing Explainable AI*, 6.
[233] Wang & Yang, *Designing Explainable AI*, 6.

5.3 How XAI Can Help Mitigating Human Errors

Although ideal human reasoning relies on cognitive paradigms such as deductive, inductive, or analogical reasoning, human beings tend to reason based on the dual process model. The dual Process Model asserts that human decision-making processes can be grouped under two distinctive approaches: System 1 (i.e., fast thinking) and System 2 (i.e., slow thinking).[234]

System 1 thinking relies on heuristics, where there is a scarcity of time and resources. In System 1 thinking, during their cognitive process, humans use representativeness heuristics and inductive reasoning to make decisions based on previously observed similar events. Therefore, if a person is more experienced, he can make decisions faster and more easily. In System 2 thinking, the cognitive process is slower and requires analytical thinking where a person employs rational reasoning. The existence of domain knowledge and comprehending the knowledge with its semantic connections are important. In contrast with System 1 thinking, System 2 thinking relies more on deductive reasoning, the hypothetico-deductive model, and Bayesian reasoning. While System 2 thinking is more systemic and yields more reliable decisions, System 1 thinking can interfere with the System 2 thinking process as part of bounded rationality.[235] Bounded rationality is a decision-making process that prioritizes satisfice instead of optimization, which often leads to suboptimal decisions.[236] Both thinking paradigms can lead to erroneous decisions. System 1 Decision errors occur due to several heuristic biases. Although there are numerous heuristic biases,[237] for the sake of simplicity, we will cover the most relevant heuristic biases that are included in Wang & Yang's work. These biases are (i) representativeness bias, (ii) availability bias, (iii) anchoring bias, and (iv) confirmation. Regardless of the type of heuristic bias, oversimplification, overconfidence, fatigue, and time pressure are often the main sources of System 1 decision errors. On the other hand, misaligned trust in tools or experts can lead to errors in System 2 decision thinking. Making decisions based on the output of uncalibrated tools or incorrect information provided by domain experts can lead to errors.[238]

5.3.1 Mitigating Representativeness Bias

Explainable AI techniques can be useful to mitigate some of these biases and can increase the accuracy of decisions. Representativeness bias is a mental shortcut to pre-

---

[234] Wang & Yang, *Designing Explainable AI*, 7.

[235] Wang & Yang, *Designing Explainable AI*, 7.

[236] *Bounded Rationality*. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/bounded-rationality

[237] Humphreys, P., & Berkeley, D. (1983). Problem Structuring Calculi and Levels of Knowledge Representation in Decision Making. *Advances in Psychology*, *16*(C), 122-123. https://doi.org/10.1016/S0166-4115(08)62197-4

[238] Wang & Yang, *Designing Explainable AI*, 8.

dicting the likelihood of an event, and it relies on the similarity of the event to an existing mental prototype.[239] In other words, human beings tend to create a mental prototype and these mental prototypes can be used to estimate the likelihood of an event. For instance, thinking that a person with glasses is highly intellectual and smart is an example of representative bias. The reason for his glasses could be just genetic predisposition instead of reading too much to the point of eyesight deterioration. XAI explanations can provide new prototypes which represent different outcomes with their similarity index to challenge the validity of the current prototype generalization.

5.3.2 Mitigating Availability Bias

Availability bias is the process of decision making based on emotional cues, familiar facts, and vivid images left mental impressions on someone. Availability heuristics can reduce the decision-making time; however, they also challenge human ability to accurately judge the likelihood of probabilistic events.[240] Availability bias suggests that a single memorable moment have an outsized influence on decision making compared to steady observations with long term effect. For example, availability bias suggests that a catastrophic event we see in the news can cause us to think that everything is going bad. The XAI explanations that show the actual occurrence probability of certain outcomes can be used to mitigate availability bias.[241]

5.3.3 Mitigate Anchoring Bias

Anchoring bias is a cognitive bias that causes humans to rely primarily on a piece of irrelevant information that was provided in the beginning. For instance, in a court hearing, when judges are given examples of lower-level crimes where convicts are sentences to lower amount of jail time, they can rule for a lower jail time for the current case where the crime is not even related to the previous examples. Hearing lower numbers clouds the judgment of these judges for no logical reason. Another example could be how much someone is willing to pay for a product. When they search through web to buy a product for a certain budget, the judgment of the consumers can be clouded after being exposed to different products with higher prices. XAI explanations can provide evidence-based information about why the irrelevant information is not significant for an outcome.[242] Seeing input attributions for multiple outcomes and how attributions are contrasted for different hypotheses can help mitigate the anchoring bias.[243]

---

[239] *Representativeness Heuristic*. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/representativeness-heuristic

[240] *Availability Heuristic*. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/availability-heuristic

[241] Wang & Yang, *Designing Explainable AI*, 8-9.

[242] *Anchoring Bias*. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/anchoring-bias

[243] Wang & Yang, *Designing Explainable AI*, 8-9.

5.3.4 Mitigate Confirmation Bias

Confirmation bias refers to the underlying tendency to pay attention to the evidence that supports our current beliefs. In other words, confirmation bias is a cognitive strategy that suggests searching for evidence that best supports our hypotheses. Therefore, when we adopt confirmation bias strategy, instead of looking for independent observations to understand if a hypothesis is supported by evidence, we tend to only look for the observations that supports our hypothesis. Confirmation bias may cause creating causal relationships between unproven or irrelevant events.[244] To mitigate confirmation bias, hypothetico-deductive reasoning can be employed. In addition, backward-driven reasoning should also be discouraged as it can lead to spurious causality. XAI explanations that prioritize input attributions to generate hypotheses could help mitigate the confirmation bias.[245]

5.3.5 Moderating Trust

While the previous issues are relevant to System 1 errors, misaligned trust in tools, experts, or systems can cause System 2 errors. Explanations are powerful in establishing trust in AI systems. However, trust can only be established when the model is reliable. Therefore, uncovering the model's poor performance can cause distrust in the model. Therefore, explanations can also help humans whether to trust in the tools, experts, or systems in question. Sharing input, output, and certainty attributes can help the receiver whether to establish trust in an AI system.[246]

5.4 Final Remarks

While explanations have several legal and ethical benefits for compliance and accountability purposes, they can also contribute to the accuracy of the decision-making process. They can be used to improve model accuracies and aid people in making better decisions. Both automated systems and humans tend to make mistakes regardless of their decision-making process due to the underlying risks that come with their reasoning. In this section, we covered how human beings reason, how XAI systems generate explanations, and how these explanations can be useful to mitigate some of the shortcomings of human reasoning in both System 1 and System 2 thinking.

Another issue to cover for better trust in decision-making is the stages where these explanations should be generated. As we mentioned in this section, XAI elements can be both about system state and intelligibility queries. The system state contains information on the inputs of the decision-making process as well. Therefore, dataset or premodeling explainability is one of these stages. In addition, intelligibility queries and output attributes are relevant to the model explainability and post-hoc explainability.

---

[244] *Confirmation Bias*. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/confirmation-bias

[245] Wang & Yang, *Designing Explainable AI*, 8-9.

[246] Wang & Yang, *Designing Explainable AI*, 8-9.

Certainty of the model predictions is another issue which are analyzed under model testing, evaluation, and parameter tuning, which are subject to benchmark analysis. Finally, overall evaluation of the models and their explainability are subject to policy and management level explainability actions. In the next Chapter, we will first cover the legal framework defined by the GDPR, particularly Art. 13-15, 22 and Recital 71. Then, we will extract the safeguards defined under these articles and associate them with the Trustworthy AI principles, an ethical framework proposed by AI-HLEG of European Commission. Therefore, while this section covers the cognitive analysis of explainability, the next section's scope covers the legal and ethical norms around explainability. When combined together, these two chapters will clarify the social sciences aspects and the requirements of the explainability of the AI systems.

6 The Right to Explanation and Trustworthy AI

In today's data-driven society, previously unknown issues such as profiling and algorithmic decision-making have become an everyday reality. While some fields do not raise concerns, using these technologies in sensitive fields such as law, finance, military, law enforcement, and human resources causes human rights and privacy concerns.[247]

Amid growing concerns about automated decision-making systems and profiling of data subjects, in May 2018, European Union's new General Data Protection Regulation, *or Regulation 2016/679*, came into effect, replacing the Data Protection Directive of 1995, *or Directive 95/46/EC*.[248] Although Data Protection Directive was an essential step toward data privacy and protection, it did not provide EU-wide direct enforceability since it was enacted as a directive rather than a regulation. The GDPR - on the other hand- is enacted as a regulation and, therefore, does not require a secondary procedure to be implemented at the national level.[249] Thus, the GDPR embodies direct EU-wide enforceability.

One of the novelties that came with GDPR is much disputed, "the right to explanation." Although the Data Protection Directive created the preliminary version of a right to explanation, GDPR took it to a whole new level. In a narrow definition, the right to explanation refers to a data subject's right to receive information from the data controller in relation to automated decision-making or profiling.[250] However, the actual scope of the right to explanation is not limited to receiving basic-level information about an AI system. In contrast, it should be regarded as an umbrella right with several safeguards to enhance the trustworthiness of the entirety of the AI systems.

---

[247] Malgieri, *The Right to Explanation*, 2-3.

[248] Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242. https://doi.org/10.1007/s13347-017-0263-5.

[249] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*. https://doi.org/10.1609/aimag.v38i3.2741.

[250] Zingales, N. (2021). Right to Explanation. In *Glossary of Platform Law and Policy Terms* (pp. 279–281). https://hdl.handle.net/10438/31365.

In this paper, we will first review the GDPR articles relevant to the right to explanation and automated decision making -including profiling-, then analyze the suitable GDPR safeguards to ensure the realization of Trustworthy AI, and understand the relationship between these safeguards and the associated ethical principles within the GDPR's right to explanation framework.

6.1 Right to Explanation in the European Union

There has been a heated discussion in academia regarding the existence of an effective right to explanation in the EU GDPR. While the predominant stand is on the existence of a right to explanation in the text of GDPR, some scholars claim that the "restrictive, unclear, or even paradoxical" nature of the GDPR makes it unfeasible to trigger any explanation-related right [6.3].[251] While a right to explanation is not explicitly stated in the binding articles of GDPR,[252] the legal framework articulated by the GDPR embodies several adjacent rights and safeguards, which together may constitute a right to explanation.[253] While Art. 13-15 aims to regulate a right to explanation in case of automated decision making, Art. 22 limits the use cases of automated decisions and creates several safeguards in the event of their use.

GDPR Art. 13(2)(f), 14(2)(g), and 15(1)(h) are the provisions that define similar rights and obligations for different scenarios, which create the first part of the right to explanation. These provisions are also the main battlefield between two groups who claim and oppose the existence of a right to explanation in GDPR. While a healthy discussion on this issue is essential to find the best application of the legal framework, the discussion, which initially started between two papers, seems to pay little attention to the text of the relevant GDPR provisions.[254] While the first paper, by Goodman and Flaxman, is for the existence of a groundbreaking and powerful right to explanation, the second paper, by Wachter, Mittelstadt, and Floridi, opposes this idea and asserts that GDPR does not articulate a right to explanation and claims that GDPR sets out other safeguards and rights to protect data subjects.[255]

On the other hand, Art. 22(1) gives data subjects the right not to be subject to a decision based solely on automated processing except if the decision is (a) necessary for a contractual relationship, (b) is authorized by the Union or Member State law that lays down suitable safeguard measures to protect data subject's rights, freedoms, and legitimate interests, and finally, (c) based on the explicit consent of the data subject.

---

[251] Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a Right to Explanation is Probably Not the Remedy You are Looking for. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2972855.

[252] Winikoff, M., & Sardelic, J. (2021). Artificial Intelligence and the Right to Explanation as a Human Right. *IEEE Internet Computing*, *25*(2), 108–112. https://doi.org/10.1109/MIC.2020.3045821.

[253] Malgieri, *The Right to Explanation*, 5-8.

[254] 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*, 145. https://doi.org/10.15779/Z38M32N986.

[255] Selbst & Powles, *Meaningful information*, 238.

For the application of Art. 22(1)(a) and (c), the data subject must be provided with suitable safeguards to obtain human intervention, express his or her point of view, and finally, contest the decision.[256]

Recital 71 goes one step further and extends the safeguards with two additional rights: (i) the right to challenge the decision and (ii) the right to obtain an explanation of the decision reached after assessment. Although Recital 71 is not directly enforceable,[257] Article 15(1)(h) implicitly creates the right to obtain an ex-post explanation.[258]

When read together, according to Articles 13(2)(f), 14(2)(g), 15(1)(h), and Article 22, in the event where a data subject is subject to a "decision based solely on automated processing", "which produces legal effects concerning him or her or similarly significantly affects him or her", he or she has a right to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject",[259] as demonstrated in Fig 5:

GDPR Article	Scope
Art 13 (2) (f)	Information to be provided where personal data are collected from the data subject
Art 14(2)(f)	Information to be provided where personal data have not been obtained from the data subject
Art 15(1)(h)	Right of access by the data subject
Art 22(3)	Automated individual decision-making, including profiling
Recital 71	Profiling

Fig 5. The GDPR Articles Relevant to the Right to Explanation

There are two main requirements to trigger a right to receive meaningful information: (i) the decision should be made based solely on automated processing and (ii) this decision should produce a legal effect or affect the data subject significantly. When interpreting the first requirement, the "solely" component should not be interpreted too rigid to avoid causing the right to be ineffective. An insignificant level of human intervention should not deem this right ineffective. Additionally, the second requirement clearly states that a decision does not have to create a legal effect on the data subject, in the event where the decision creates a significant economic or social effect concerning the data subject, a right to explanation can be triggered. When these requirements are fulfilled, a data subject can exercise his or her right to receive meaningful information. But a right to explanation should also be meaningful and impactful.[260]

---

[256] Malgieri, *The Right to Explanation*, 5-8.

[257] Edwards, L., & Veale, M. (2018). Enslaving the Algorithm: From a "right to an Explanation" to a "right to Better Decisions"? *IEEE Security and Privacy*, *16*(3), 46–54. https://doi.org/10.1109/MSP.2018.2701152.

[258] Malgieri, *The Right to Explanation*, 11.

[259] Selbst & Powles, *Meaningful information*, 235.

[260] Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*. https://doi.org/10.1609/aimag.v38i3.2741.

According to Selbst and Powles, there are four components to having a meaningful and impactful right to explanation. First, the "meaningful information" should be interpreted subjectively based on the data subject since Art. 13-15 particularly aims at data subjects. Second, meaningful information should either have instrumental or intrinsic value or enable a possible action. Thirdly, meaningful information should provide enough functionality to guarantee the data subject's exercise of rights. Finally, meaningful information requirements should be interpreted based on the facts of the case to avoid hampering innovation and R&D efforts in this field.[261] While the first three components are in favor of data subjects ((a) requiring explanations with protective interpretation, (b) providing instrumental or intrinsic value or enabling a possible action, and (c) having enough functionality to guarantee the exercise of rights), the fourth component tries to balance the impact of the first three to ensure the continuation of innovation in this field which leads to long-term prosperity and wealth. Only by actively managing and respecting these components, we can truly achieve responsible competitiveness.[262]

6.2 Safeguards around Right to Explanation

As mentioned earlier, an explanation for an automated decision should either have instrumental or intrinsic value or enable a possible action. In case of enabling a possible action, suitable safeguards mentioned in Art. 13-15, Art. 22(4), and Recital 71 have become important references for interpretation.[263] While Art. 13, 14, and 15 define the right to obtain information about automated decision-related processing as a common safeguard, Art. 22 explicitly mentions -in a non-exhaustive wording[264] -three safeguards regarding automated decision making: (i) the right to obtain human intervention on the part of the controller, (ii) the right to express one's point of view, and (iii) to contest the automated decision. Additionally, Recital 71 further expands the list of suitable safeguards with (a) the right to challenge the automated decision and (b) the right to obtain an explanation of the automated decision. Therefore, we can create a non-exhaustive list of safeguards – as shown in Fig 6- to ensure fairness and transparency of data processing where there is automated decision-making.

---

[261] Selbst & Powles, *Meaningful information*, 236.

[262] AI HLEG, Trustworthy AI, 9.

[263] Hamon, R., Junklewitz, H., Malgieri, G., Hert, P. de, Beslay, L., & Sanchez, I. (2021). Impossible explanations?: Beyond explainable AI in the GDPR from a COVID-19 use case scenario. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 549–559. https://doi.org/10.1145/3442188.3445917.

[264] Kaminski, M. E., Bertolini, A., Brennan-Marquez, K., Comandé, G., Cushing, M., Helberger, N., van Drunen, M., van Eijk, N., Eskens, S., Malgieri, G., Price, N., Sax, M., & Selbst, A. (2019). The Right to Explanation, Explained. *BERKELEY TECHNOLOGY LAW JOURNAL*, *34*, 189. https://doi.org/10.15779/Z38TD9N83H.

Safeguard	Legal Basis
The right to obtain information about automated decisions	GDPR Art 13(2)(f), 14(2)(f), and 15(1)(h)
The right to contest/challenge the automated decision	GDPR Art 22(3), Recital 71
The right to express one's point of view	GDPR Art 22(3)
The right to obtain human intervention	GDPR Art 22(3)
The right to obtain an explanation of the decision after assessment	GDPR Recital 71

Fig 6. The Safeguards relevant to the Right to Explanation and Their Legal Basis in the GDPR

6.2.1 The right to obtain information about automated decisions

The first and most important safeguard for the right to explanation is the right to obtain information about automated decision-making. This safeguard requires data controllers to provide information on (i) the existence of automated decision-making, (ii) meaningful information about its logic, and finally (ii) the significance and the envisaged consequences of such automated decisions. This safeguard does not require data controllers to provide a specific explanation for a particular decision. Instead, it requires an explanation of the general mathematical logic used in the decision-making process.[265] In literature, scholars often limit the scope of the right to explanation merely to this safeguard.[266] However, in a broader sense, this safeguard does not cover the true boundaries of the right to explanation and secure the relevant ethical principles.

6.2.2 The right to contest/challenge the automated decision

While Article 22(3) mentions the right to "contest" a decision, Recital 71 takes a step further and mentions the right to "challenge" the automated decision. While contesting simply refers to adjusting or reviewing a decision, challenging a decision refers to requesting to identify the inadequateness of the decision to deem it ineffective. Despite the claims made by some scholars about the existence of the right to explanation, some argue that the mere existence of a right to contest a decision as a safeguard requires a right to explanation.[267] Without prejudice to the unbinding nature of Recital 71, we can assume that when there is an automated decision that creates a significant or legal effect, the data subject is provided with the right to contest/challenge the automated decision.[268]

---

[265] Winikoff & Sardelic*, The Right to Explanation as a Human Right*, 109.

[266] Kaminski & Bertolini, *The Right to Explanation, Explained*, 189.

[267] Bayamlıoğlu, E. (2021). The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation." *Regulation & Governance*. https://doi.org/10.1111/REGO.12391.

[268] Malgieri, *The Right to Explanation*, 6.

6.2.3 The right to express one's point of view

Expressing one's opinion is itself a safeguard as well as a component of another safeguard, the right to contest/challenge an automated decision. Expression of one's opinion has a different significance for contesting and challenging a decision. In contesting a decision, data subjects may express their opinion to change the outcome of the decision whereas expressing an opinion during the challenge of a decision can make the decision null and void. Therefore, the importance of expressing one's point of view can create different legal results.

6.2.4 The right to obtain human intervention

Obtaining human intervention is a safeguard explicitly stated in GDPR Art. 22(3) and many EU Member States laws.[269] This safeguard ensures that when a decision is successfully contested or challenged by the data subjects, they will not be subject to another almost identical automated decision. Besides, this safeguard should be enforced with the right to express one's point of view since, without the data subject's point of view, the chance to eliminate algorithmic bias may be difficult, which corresponds to damaging the fairness element of the automated decision-making process explicitly laid out in Articles 13(2) and 14(2). It is important to note here that Art. 22 applies to cases where a decision is based solely on automated processing. Although some scholars argue that using spurious human involvement limits the applicability of this safeguard and the other safeguards mentioned in Art. 22,[270] adding a layer of ineffective human oversight should not be regarded as an effective medium to bypass this safeguard.

6.2.5 The right to obtain an explanation of the decision after assessment

Apart from the right to explanation in a general sense, Recital 71 of the GDPR also mentions an extended right to explanation after other safeguards (e.g., requesting human intervention, contesting/challenging a decision, and expressing an opinion regarding a particular decision) are triggered. There are claims that since the term right to obtain explanation is used under non-binding Recital 71, some scholars claim that the GDPR does not create a binding right to explanation in Articles 13-15 and Article 22.[271] However, the right to obtain an explanation under Recital 71 is merely a subcategory of the right to explanation in question. This subcategory only covers the explanation about a decision after a safeguard triggers an assessment of this decision.

In addition, a valid assessment will require a review of the architecture and implementation of the algorithm. Therefore, this safeguard will require an explanation of the

---

[269] Malgieri, *The Right to Explanation*, 6.

[270] Edwards, & Veale. *Enslaving the Algorithm*, 6-7.

[271] 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*, 1–47. https://doi.org/10.2139/SSRN.2903469.

architecture and implementation of the algorithm, which is usually referred to as the "ex-ante" explanation in addition to the specific explanation about a particular decision.

6.3 The Ethical Principles on Explainability and Right to Explanation

Apart from the legal response to the issues brought about by the mass adoption of AI systems, scholars and policymakers often refer to digital ethics principles, which contribute to the protection of fundamental rights and freedoms. Almost every major institution, along with big tech companies, published their frameworks to address today and tomorrow's ethical problems.[272] While many of these frameworks contain common themes, this paper primarily considers the Trustworthy AI principles laid out by the European Commission's High-Level Expert Group. The High-Level Expert Group Report offers guidelines designed to guide the AI community towards lawful, ethical, and robust AI practices.[273]

One of the main arguments of the High-Level Expert Group for a system to be trustworthy is that we should be able to understand why it behaves in a certain manner and how it outputs the decisions. Explainable AI (i.e., XAI) is the up-and-coming AI subfield that tries to understand how AI systems behave in general or are related to individual decisions.[274] XAI can contribute to the realization of Trustworthy AI, particularly by helping the satisfaction of some of the seven key principles.[275] Although explainability is not one of these seven principles in the Trustworthy AI framework, some of these principles are directly associated with the explainability of automated decision-making systems and the relevant GDPR safeguards.

Some of these principles are explicitly mentioned under the relevant GDPR articles. Article 13(2) and 14(2) clearly states that data subjects should be provided with some information about the automated decision-making process to ensure "fairness" and "transparency." Therefore, we can safely assert that two of the ethical goals with the right to explanation are to guarantee fairness and transparency of the data processing related to automated decision-making.

Transparency is the most relevant Trustworthy AI principle that has a direct link to explainability. In the High-Expert Group report, explainability is placed under the transparency principle along with traceability and communication. Furthermore, the transparency principle requires transparency in all three pillars of AI systems (i.e., the data, the system, and the business model).[276] Therefore, ideally, a transparent automated decision-making system should provide explanations about the data it uses, its technical logic, and finally, its business model.

---

[272] Siau, K., & Wang, W. (2020). Artificial intelligence (AI) Ethics: Ethics of AI and ethical AI. *Journal of Database Management*, *31*(2), 74–87. https://doi.org/10.4018/JDM.2020040105.

[273] AI HLEG, Trustworthy AI, 2.

[274] AI HLEG, Trustworthy AI, 2.

[275] The seven requirements are (i) human agency and oversight, (ii) technical robustness and safety, (iii) privacy and data governance, (iv) transparency, (v) diversity, non-discrimination, and fairness, (vi) societal and environmental wellbeing, and (vii) accountability.

[276] AI HLEG, Trustworthy AI, 18.

Although the wording of GDPR Art. 13-15 uses the term 'fairness', the High-Level Expert Group expands this principle to 'Diversity, Non-Discrimination, and Fairness.' By expanding the inclusiveness and diversity throughout the entire AI system's life cycle, we can contribute to the realization of trustworthy AI. Discrimination and bias can occur at the algorithm level as well as the data and the business level. From bad data collection to inadvertent historic bias, incompleteness, and bad governance models, there is a variety of possibilities that unintentional bias is embedded into automated decision-making systems. Additionally, one may also encounter intentional discrimination.[277] At each stage of the life cycle, all five safeguards must be enabled and actionable as they are all suitable for the reinforcement of the fairness element of the automated decisions.

The accountability principle consists of audibility and minimization of negative impact. Therefore, to strengthen the accountability of automated decision-makers, we need to enable suitable safeguards that can be useful to detect negative issues such as algorithmic bias and discrimination. The assessment of algorithms, data, and design processes is particularly important when data subjects contest/challenge an automated decision.[278] Additionally, after a particular automated decision, the data subject's right to obtain an explanation requires this explanation to provide an adequate level of accountability feature. Apart from the assessment of the processes, if a potential issue is detected, the data controllers must enable measures to minimize its negative impact.

Finally, human agency is another important principle that interacts with the safeguards reviewed in this article. AI systems should enable mechanisms to receive external feedback about their performance on fundamental rights and human autonomy. In a trustworthy AI system, the data subject should be able to make informed decisions about the AI system. The right not to be subject to a decision based solely on automated processing in case of legal or significant consequences is a consequence of this principle. Apart from encouraging human autonomy, an AI system should also enable safeguards to allow for human oversights. Therefore, we can say that the right to obtain human intervention is a suitable tool to ensure human oversight of AI systems. On the other hand, all the safeguards support human agency; therefore, contribute the human autonomy in automated decisions and the consequences.

Fig 7 shows the trustworthy AI principles matched with the safeguards defined in the GDPR articles relevant to the right to explanation:

[277] AI HLEG, Trustworthy AI, 18.
[278] AI HLEG, Trustworthy AI, 18.

Transparency	Diversity, Non-Discrimination, and Fairness	Accountability	Human Agency and Oversight
• The right to obtain information about automated decisions • The right to obtain an explanation of the decision after	• The right to obtain information about automated decisions • The right to obtain human intervention • The right to express one's point of view • The right to contest/challenge the automated decision • Right to obtain an explanation of the decision after assessment	• The right to contest/challenge the automated decision • The right to obtain an explanation of the decision after assessment	• The right to obtain human intervention

Fig 7. The Right to Explanation Safeguards that directly affect Trustworthy AI Principles

While four of the seven principles are directly associated with the safeguards reviewed in this paper and therefore, with the right to explanation, the remaining three principles are not within the direct scope of the safeguards in question. However, other safeguards and technical & organizational measures are still effective mediums to ensure (i) technical robustness and safety, (ii) privacy and data governance, and (iii) societal and environmental wellbeing where there is an AI system and automated decision-making.

6.4 The Effect of the Right to Explanation on Trustworthy AI

In the light of the above explanation, we can finally create a relationship timeline diagram of the GDPR articles, the safeguards, and ethical principles that have an association with the right to explanation due to automated decision-making, as shown in Fig 8:

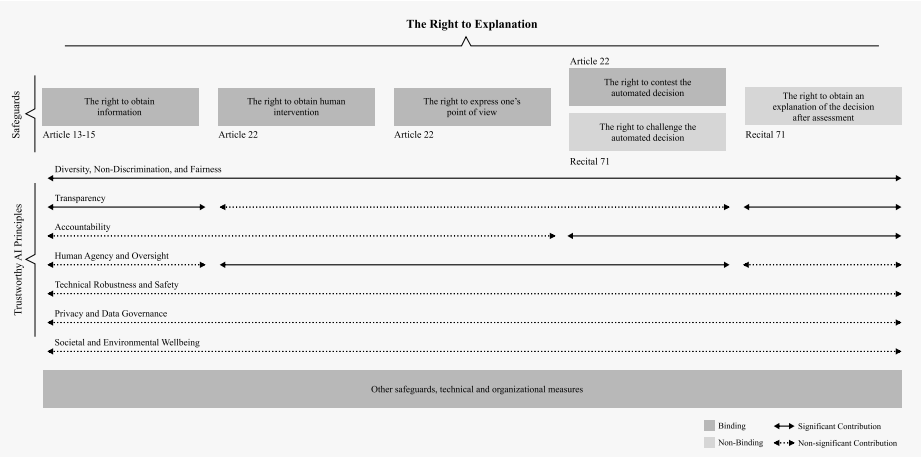


Fig 8. The relationship between GDPR articles, safeguards, and trustworthy AI principles

In Fig 8, dark gray boxes represent the safeguards that are mentioned in the binding articles of the GDPR whereas the light gray boxes represent the non-binding safeguards mentioned in Recital 71 of the GDPR. While dashed arrowed lines represent the non-significant contribution of the safeguards to the principle, solid arrowed lines represent a significant contribution. It is important to note that dashed lines may mean still mean

some contribution, but the level of the contribution of the safeguard on the ethical principle in question would be limited. Additionally, GDPR lays out several safeguards and technical & organizational measures that can strengthen all seven principles of realizing Trustworthy AI. However, they are not included in Fig 8 since they do not directly aim at automated decisions and the right to explanation.

The order of the safeguards is created in a timeline structure. Although data subjects may skip one or more of them, in an automated decision-making process, the safeguards are likely to be triggered from left to right. We enter the realm of the right to explanation with the right to obtain information laid out in Art. 13-15. After receiving the initial information from the data controller, if the data subjects are not satisfied with the information and the decision, they can trigger the human intervention, express opinions, and contest the decision safeguards mentioned in Art. 22. Art. 22 lists these three safeguards and does not limit the potential safeguards with these three by using the term 'at least'. Recital 71 further expands these safeguards with the right to challenge a decision and the right to obtain an explanation. Although Recitals are not binding, combined with the non-exhaustive listing of Art. 22 safeguards, we can safely assume that Recital 71 safeguards will be taken very seriously by the administrative authorities and judiciary when the binding Art. 13-15 and Art. 22 are interpreted.

We can also see to what extent safeguards contribute to the seven principles of Trustworthy AI. All five safeguards are effective measures to strengthen the 'diversity, non-discrimination, and fairness' principle as they require a considerable amount of reasoning, human interaction, and model assessment. The transparency principle is mainly strengthened by the right to obtain information defined in Art. 13-15 and the right to obtain an explanation defined in Recital 71. They require data controllers to provide information from the existence of automated decisions to the inner logic of the models, and by some interpretations, even to specific explanations about a particular decision, which enhance the transparency property of the AI systems. These safeguards also force data controllers to use explainable models from the beginning to comply with them. Although each step contributes to the accountability feature, the right to contest/challenge a decision and the right to obtain an explanation may open a direct channel to administrative or judicial bodies, which increases the accountability of the AI systems. Finally, obtaining human intervention, expressing one's point of view, and contesting/challenging a decision will certainly contribute to strengthening the human agency and overview principle. The other principles are not directly affected by these safeguards, but when we interpret the provisions of GDPR (primarily, Art. 13-15, Art. 22, and Recital 71), we should take them into consideration. For example, adopting a too rigid interpretation for explanation requirements can damage the technical robustness of the AI systems. In addition, when designing solutions for compliance with these safeguards, data controllers should also consider the 'privacy and data governance' and 'societal and environmental wellbeing' principles for a sustainable business model.

6.5 Final Remarks

In this Chapter, we covered the GDPR safeguards and Trustworthy AI principles that are relevant to and associated with the right to explanation and explainability of AI systems.

Understanding the legal and ethical norms around explainability is essential to have a sustainable AI ecosystem since automated decisions are becoming more widely used in every part of our lives and companies, institutions, and governments take advantage of AI systems to increase their revenue, profitability, and service quality. Although some of these systems do not pose threats to the individuals whose data have been processed and used for automated decision-making and profiling, some of these systems can output automated decisions that can produce significant social, economic, or legal outcomes. In such events, GDPR lays out several safeguards specifically aimed at addressing the issues that may surface due to automated decisions along with many other general-purpose safeguards that can also be used for the protection of fundamental rights and freedoms. These specific safeguards and the rights created around these safeguards create a comprehensive framework for the right to explanation. This ecosystem of safeguards also serves for the realization of Trustworthy AI by strengthening the ethical principles described by the High-Level Expert Group of the European Commission. Therefore, it was essential to analyze GDPR's legal framework, which contains the abovementioned rights and safeguards around the right to explanation, and the ethical principles that are highlighted by the High-Level Expert Group. As a result, we create a relationship timeline diagram to show the timeline of an automated decision process with the relevant safeguards suitable at each stage and their contribution to the seven principles of Trustworthy AI.

7 Technical Overview of Machine Learning and Deep Learning

Now that we covered the social aspects of AI explainability, in this Chapter, we aim to make an introduction to the field of machine learning and to clarify the scope of similar domains, particularly deep learning. This Chapter also aims to introduce popular machine learning models briefly, compare different machine learning approaches and concepts, and identify the steps of the machine learning development cycle.

Apart from the general overview of machine learning, we dive into the inner processes of neural networks and deep learning to understand the black-box model structure. In addition, since deep learning is a subsection of machine learning, the concepts covered throughout this Chapter are usually applicable to deep learning by extension.

7.1 What is Machine Learning?

Computers do not have cognitive abilities, and they cannot reason on their own. However, they are powerful machines for processing data, and they can complete difficult calculation tasks in a small amount of time. They can process anything so long as we

provide them with detailed, step-by-step logical and mathematical instructions. So, if we can represent the cognitive abilities of a human with logical operations, it seems possible that computers can develop cognitive skills.

Consciousness is a highly debated topic in artificial intelligence, focusing on the question of whether computers can become conscious. While the discussion centers on the possibility of machines fully mimicking human consciousness (known as *General AI*), the replication of specific human skills for specific tasks (*Narrow AI*) is also considered to be part of artificial intelligence.[279]

The term "Machine Learning" was first coined in 1959 by Arthur Samuel, an IBM scientist and pioneer in the field of computer gaming and artificial intelligence.[280] Throughout the 50s, 60s, and 70s, the early work on neural networks was conducted with the goal of mimicking the human brain. However, real-life applications of neural networks were unfeasible for a long time due to the limitations of computer technology. The fundamental machine learning research on other ML techniques (i.e., non-deep learning techniques which require fewer computer resources) was popularized in the 80s and 90s. The advancements in computer technology during this period partially allowed the adoption of machine learning applications in real life. As the years passed, the limitations due to immature computer technology were mostly eliminated, particularly in recent years.[281] Although we always strive for better and more efficient computing power and storage, today, we can at least quickly build models, test them, and even deploy ML models on cloud for the whole world to use. The field of machine learning has become a vibrant ecosystem thanks to the abundance of data, efficient data storage technologies, and faster & cheaper processing power. Fig 9 summarizes the timeline of the artificial intelligence development trends by decade:

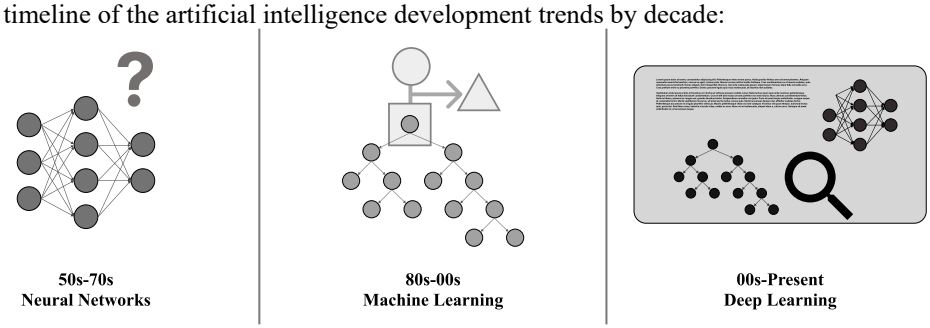


Fig 9. The Timeline of Artificial Intelligence Development Trends

[279] Wirth, N. (2018). Hello marketing, what can artificial intelligence help you with?: *International Journal of Market Research*, *60*(5), 435–438. https://doi.org/10.1177/1470785318776841

[280] Petrik, M. (n.d.). Introduction to Machine Learning. In *University of New Hampshire*. Retrieved August 28, 2022, from https://www.cs.unh.edu/~mpetrik/teaching/intro_ml_17_files/class1.pdf

[281] Foote, K. D. (2021, December 3). *A Brief History of Machine Learning*. Dataversity. https://www.dataversity.net/a-brief-history-of-machine-learning/

Machine learning is considered a sub-discipline in the field of artificial intelligence. Machine Learning (ML) studies aim to automatically improve the performance of the computer algorithms designed for particular tasks with experience. In a machine learning study, the experience is derived from the training data, which may be defined as the sample data collected on previously recorded observations. Through this experience, machine learning algorithms can learn and build mathematical models to make predictions and decisions. The learning process starts with feeding training data (e.g., examples, direct experience, basic instructions), which contains implicit patterns, into the model. Since computers have more processing power than humans, they can find these meaningful patterns in the dataset within a short amount of time. These patterns are -then- used to make predictions and decisions for new observations. The learning may continue even after deployment if the developer builds a suitable machine learning system that allows continuous training.[282]

> "*Previously, we might use machine learning in a few sub-components of a system. Now we actually use machine learning to replace entire sets of systems, rather than trying to make a better machine learning model for each of the pieces.*"
>
> --Jeff Dean

There is an ever-increasing use of machine learning applications in different fields. These real-life applications vary to a great extent. Some use cases in selected fields may be listed as follows:

- **Healthcare**: Medical diagnosis given the patient's symptoms,
- **E-commerce**: Predicting the expected demand,
- **Law**: Reviewing legal documents and alerting lawyers about problematic provisions,
- **Social Network**: Finding a good match given the user's preferences on a dating app, and
- **Finance**: Predicting the future price of a stock given the historical data.

These five real-life applications are a small set of potential use cases, and there are hundreds, if not thousands, of potential machine learning applications and dozens of algorithms for these applications. These methods are usually grouped into four main approaches: (i) Supervised Learning, (ii) Semi-supervised Learning, (iii) Unsupervised Learning, and (iv) Reinforcement Learning.

Each method contains distinct features in its design, but they all follow the same underlying principles and conform to the same theoretical background. In the upcoming

---

[282] Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., & Zhang, J. D. (2020). An Introduction to Machine Learning. *Clinical Pharmacology and Therapeutics*, *107*(4), 871–885. https://doi.org/10.1002/CPT.1796

sections, we will cover these different approaches in more detail. But first, we will conduct another term clarification between Machine Learning and its adjacent fields: (i) Artificial Intelligence, (ii) Deep Learning, (iii) Big Data, and (iv) Data Science.

7.2 Scope of Machine Learning and Its Relation to Adjacent Fields

In literature, we often see terms such as artificial intelligence, machine learning, deep learning, big data, and data science to address the same or similar concepts. There is a slight level of ambiguity about the differences between these terms. In this section, we clarify this ambiguity and identify the differences so that our statements can be clearer.

7.2.1 Artificial Intelligence

Artificial Intelligence (AI) is a broad umbrella term, and its definition varies across different textbooks. The term AI is often used to describe computers that simulate human intelligence and mimic "cognitive" abilities that humans associate with the human mind. Problem-solving and learning are examples of these cognitive abilities. The field of AI contains machine learning studies since AI systems are capable of learning from experiences. Machines with artificial intelligence are capable of:

- Understanding and interpreting data,
- Learning from data, and
- Making 'intelligent' decisions based on insights and patterns extracted from data.

These terms are highly associated with machine learning. Thanks to machine learning, AI systems can learn and excel at their level of consciousness. Machine learning is used to train AI systems and make them smarter.

7.2.2 Deep Learning

Deep learning (DL) is a sub-field of machine learning that exclusively uses multiple layers of neurons to extract patterns and features from raw data.[283] These multiple layers of interconnected neurons create artificial neural networks (ANNs). An ANN is a special machine learning algorithm designed to simulate the working mechanism of the human brain. There are many different types of artificial neural networks intended for several purposes. In summary, deep learning algorithms are a subset of machine learning algorithms.

---

[283] Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *Sn Computer Science*, *2*(5), 377. https://doi.org/10.1007/S42979-021-00765-8

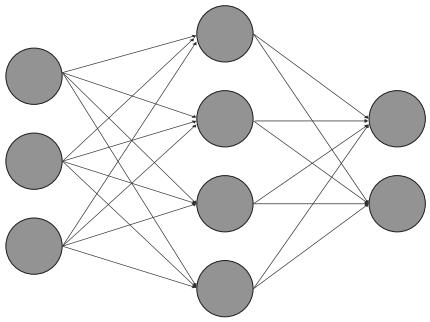


Fig 10. An Example of Multi-layer Artificial Neural Network Architecture

Just as in machine learning, all four approaches (supervised, semi-supervised, unsupervised, and reinforcement learning) can be utilized in deep learning. When there is an abundance of data and enough computing power, deep learning almost always outperforms the other machine learning algorithms. Deep learning algorithms are especially useful in image processing, voice recognition, and machine translation.

7.2.3 Data Science

Data science is an interdisciplinary field that sits at the intersection of artificial intelligence, particular domain knowledge, information science, and statistics. Data scientists use various scientific methods, processes, and algorithms to obtain knowledge and draw insights from observed data.[284] In contrast with machine learning, the goal of a data science study does not have to be automated decision making with trained models. Data science studies often aim to extract knowledge and insight to support the human decision-making process without creating an AI system. Therefore, although there is an intersection between data science and the other adjacent fields, the field of data science differs from them since it does not have to deliver an intelligent system or a trained model.

7.2.4 Big Data

Big data is a field that aims to efficiently analyze a large amount of data that cannot be processed with traditional data-processing methods and applications. Data with more observation usually brings more accuracy, while high complexity may increase false discovery rates. The field of big data studies on how to efficiently capture, store, analyze, search, share, visualize, and update data when the size of a dataset is very large. [285] Big data studies can be used both in artificial intelligence (and its sub-fields) and in

[284] Sarker, *Data Science and Analytics*, 376.
[285] Sarker, *Data Science and Analytics*, 377.

data science. Big data sits at the intersection of all the other fields mentioned above since its methods are crucial for all of them.

7.2.5 The Taxonomy Diagram

The relationship between these adjacent terms may be visualized in the following taxonomy diagram:

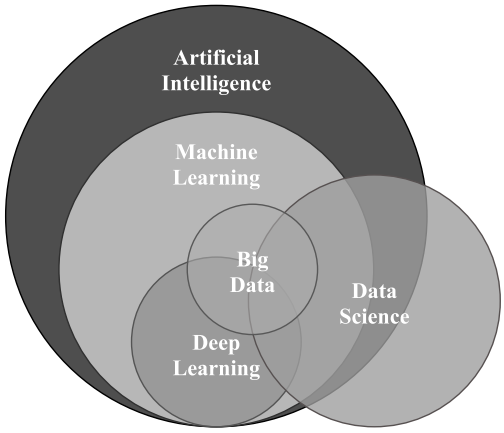


Fig 11. The Taxonomy of the Artificial Intelligence and Data Science Subfields

This taxonomy is almost clear evidence of the reasons behind the ambiguity. There is a complex hierarchical structure between these fields and the higher-level field is not always fully inclusive of the more specific field. Therefore, some of the concepts are applicable to other fields as well. For example, most of the explanations under this Chapter is, although not necessarily, also valid for Deep Learning as well. Therefore, the naming practices are not necessarily incorrect, but they are confusing. Therefore, it is vital to know the intersections and subtractions of these fields.

7.3 Machine Learning Approaches and Models

Top machine learning approaches are categorized depending on the nature of their feedback mechanism for learning. These different approaches may be listed as follows:[286]

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

---

[286] Rebala, G., Ravi, A., & Churiwala, S. (2019). *An Introduction to Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-15729-6

Most of the machine learning problems may be addressed by adopting one of these approaches. Yet, we may still encounter complex machine learning solutions that do not fit into one of these approaches. In this section, we will briefly cover the scope of these four main machine learning approaches, along with their application examples and relevant ML models. Defining and clarifying the scope of these approaches are essential to quickly uncover the nature of an AI problem, analyze the resources, and develop suitable solutions. In addition, it can also be useful to set cognitive, legal, and ethical goals when designing Explainable AI systems.

7.3.1 Supervised Learning

The supervised learning approach can be adopted when there is a dataset containing the records of the response variable values (or labels). Depending on the context, this data with labels is usually referred to as "labeled data" and "training data". [287] When we try to predict a person's height using his weight, age, and gender, we need the training data that contains people's weight, age, and gender info along with their real heights. This data allows the machine learning algorithm to discover the relationship between height and the other variables. Then, using this knowledge, the model can predict the height of a given person.

For example, we can mark e-mails as 'spam' or 'not-spam' based on the differentiating features of the previously seen spam and not-spam e-mails such as the lengths of the e-mails and the use of particular keywords in the e-mails. Learning from training data continues until the machine learning model achieves a high level of accuracy on the training data.

There are two main supervised learning problems: (i) Classification Problems, and (ii) Regression Problems. In classification problems, the models learn to classify an observation based on its variable values. During the learning process, the model is exposed to a lot of observations with their labels.[288] For example, after seeing thousands of customers with their shopping habits and gender information, a model may successfully predict the gender of a new customer based on his/her shopping habits. Binary classification is the term used for grouping under two labels such as male and female. Another binary classification example might be predicting whether the animal in a picture is a 'cat' or 'not cat', as shown in Fig 12.

[287] Rebala & Ravi, *An Introduction to Machine Learning*, 19-21.
[288] Rebala & Ravi, *An Introduction to Machine Learning*, 19-21.

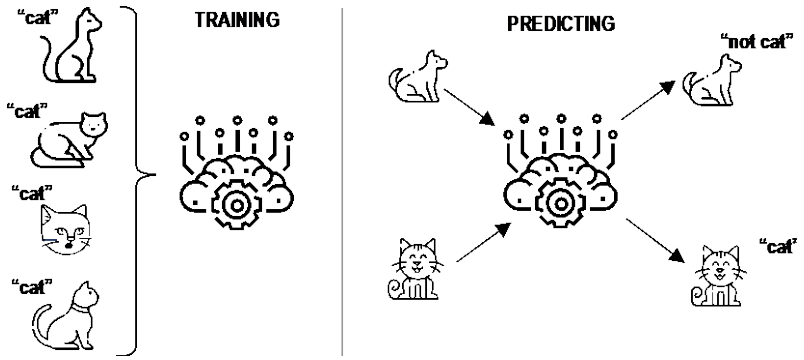


Fig 12. Classification Problem in Supervised Learning[289]

On the other hand, multilabel classification is used when there are more than two labels. Identifying and predicting handwritten letters and numbers on an image would be an example of multilabel classification.

In regression problems, the goal is to calculate a value by taking advantage of the relationship between the other variables (i.e., independent variables, explanatory variables, or features) and the target variable (i.e., dependent variable, response variable, or label). The strength of the relationship between our target variable and the other variables is a critical determinant of the prediction value, along with the values of the explanatory variables for the observation.[290] Predicting how much a customer would spend based on its historical data would be classified as a regression problem.

Linear and Logistic Regression: Linear regression is a linear approach to modeling the relationship between a numerical response variable (Y) and one or more explanatory variables (Xs). Logistic regression, on the other hand, is a slightly different method to model the probability of a particular class or event to exist, such as male/female for gender. Therefore, linear regression is used for regression problems whereas logistic regression is mostly used for classification problems.[291]

Decision Trees and Ensemble Methods: A decision tree is a flowchart-like structure and a decision support tool that connects the potential decisions and uncertain events with their probabilities to create a model that predicts possible outcomes. We can also ensemble multiple decision trees to create more advanced machine learning algorithms, such as random forest algorithms.[292]

---

[289] Icons made by Freepik, Those icons, Eucalyp, from www.flaticon.com.

[290] Rebala & Ravi, *An Introduction to Machine Learning*, 19-21.

[291] Worster, A., Fan, J., & Ismaila, A. (2007). Understanding linear and logistic regression analyses. *Canadian Journal of Emergency Medicine*, *9*(2), 111–113. https://doi.org/10.1017/S1481803500014883.

[292] Dietterich, T. G. (2000). Ensemble methods in machine learning. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1857 LNCS*, 1–15. https://doi.org/10.1007/3-540-45014-9_1/COVER.

Support Vector Machines: A support vector machine constructs a hyperplane to separate a space which can be used for classification, regression, or outlier detection. For example, a three-dimensional space (*e.g., a cube*) can be separated into smaller pieces with a two-dimensional hyperplane (*e.g., a square*). This will help to group observations into two different classes. The potential applications can be much more complicated than this example.[293] Support Vector Machine is a popular machine learning algorithm due to its high accuracy performance and relatively low-level computing source requirements.

K-Nearest Neighbors: The k-nearest neighbors algorithm is a machine learning algorithm that may be used for classification and regression problems. k is a user-defined constant, which represents the number of neighbor observations to be included in the algorithm. In classification problems, the neighbors of a new unlabeled observation are used to predict the label of this new observation based on the labels of the neighbors.[294]

Neural Networks (Multilayer perceptron): Feedforward only neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs) are often used in supervised learning problems, which will be covered in the upcoming section.

7.3.2 Unsupervised Learning

Unsupervised learning is an approach used in machine learning algorithms to draw inferences from datasets that do not contain labels. Unsupervised learning is mainly used in clustering analysis. Clustering analysis is a grouping effort in which the members of a group (i.e., a cluster) are more similar to each other than the members of the other clusters. There are several clustering methods available, and they usually utilize a type of similarity measure based on selected metrics such as Euclidean or probabilistic distance.[295] Bioinformatic sequence analysis, genetic clustering, pattern mining, and object recognition are some of the clustering problems which may be tackled with the unsupervised learning approach.

Another use case of unsupervised learning is dimensionality reduction. Dimensionality is equivalent to the number of features used in a dataset. In some datasets, you may find hundreds of potential features stored in individual columns. In most of these datasets, several of these columns are highly correlated. Therefore, we should either select the best ones (i.e., *feature selection*) or extract new features by combining the existing ones (i.e., *feature extraction*). This is where unsupervised learning comes into play. [296] Dimensionality reduction methods help us create neater and cleaner models that are free of noise and unnecessary features.

---

[293] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.

[294] Kramer, O. (2013). K-Nearest Neighbors. In *Dimensionality Reduction with Unsupervised Nearest Neighbors*, 51, 13-14. Springer. https://doi.org/10.1007/978-3-642-38652-7_2.

[295] Rebala & Ravi, *An Introduction to Machine Learning*, 22.

[296] Kramer, K-Nearest Neighbors, 13-14.

Unsupervised learning may also be used in anomaly detection problems and generative systems. I will briefly mention some of the popular unsupervised machine learning models below.

Hierarchical Clustering: Hierarchical clustering is an unsupervised machine learning algorithm used to group unlabeled observations having similar characteristics incrementally. Hierarchical clustering can be agglomerative (bottom-up approach) or divisive (top-down approach). The hierarchy of the clusters is represented as a tree or a dendrogram.[297]

K-Means Clustering: K-means clustering is a popular unsupervised machine learning algorithm. K is a user-assigned constant representing the number of clusters to be created. K-means clustering groups observations into k distinct clusters based on the distance to the center of a cluster. [298]

Principal Component Analysis (PCA): PCA is widely used for dimensionality reduction. PCA finds a linear combination of two or more variables, which are called principal components. This procedure reduces the dimensional complexity of the model so that the problem may be visualized and analyzed more quickly as the model is trained more easily as well. [299]

Neural Networks: Autoencoders, Deep Belief Nets, Hebbian Learning, Generative adversarial networks (GANs), and Self-organizing maps are some of the neural networks used for unsupervised learning. The details and the applications of some of these network structures will be covered in the upcoming chapters.

7.3.3 Semi-Supervised Learning

Semi-supervised learning is a machine learning approach that combines the characteristics of supervised learning and unsupervised learning. A semi-supervised learning approach is particularly useful when we have a small amount of labeled data with a large amount of unlabeled data available for training. Supervised learning characteristics help take advantage of the small amount of label data. In contrast, unsupervised learning characteristics are useful to take advantage of a large amount of unlabeled data. Although supervised learning is a powerful approach, labeling data -to be used in supervised learning- is a costly and time-consuming process. On the other hand, a large amount of data can also be beneficial even though they are not labeled. So, in real life, semi-supervised learning may shine out as the most suitable and the most fruitful machine learning approach if done correctly. In semi-supervised learning, the process usually starts with clustering the unlabeled data. Then, we use the labeled data to label the clustered unlabeled data. Finally, a significant amount of now-labeled data is used to train machine learning models. Semi-supervised learning models can be very powerful

---

[297] Nielsen, F. (2016). Hierarchical Clustering. In *Undergraduate Topics in Computer Science. Undergraduate Topics in Computer Science* (pp. 195–196). Springer, Cham. https://doi.org/10.1007/978-3-319-21903-5_8.

[298] Nielsen, *Clustering with k-Means*, 197.

[299] Kramer, *K-Nearest Neighbors*, 13-14.

since they can take advantage of a high volume of data. Semi-supervised learning models are usually a combination of transformed and adjusted versions of the existing machine learning algorithms used in supervised and unsupervised learning. [300] This approach is successfully used in areas like speech analysis, content classification, and protein sequence classification. The similarity of these fields is that they offer abundant unlabeled data and only a small amount of labeled data.

Some of the popular approaches and methods used in semi-supervised machine learning problems are as follows:[301]

- Diagnostic Techniques,
- Generative Techniques,
- Input-based Regularization Methods:
    - The Cluster Assumption Techniques,
    - The Fisher Kernel,
    - Co-training.

7.3.4 Reinforcement Learning

Reinforcement learning is one of the primary approaches to machine learning concerned with finding optimal agent actions that maximize the reward within a particular environment. The agent learns to perfect its actions to gain the highest possible cumulative reward. There are four main elements in reinforcement learning:[302]

- **Agent:** The trainable program which exercises the tasks assigned to it
- **Environment:** The real or virtual universe where the agent completes its tasks.
- **Action:** A move of the agent which results in a change of status in the environment
- **Reward:** A negative or positive remuneration based on the action.

Reinforcement learning may be used in both the real world as well as in the virtual world. For instance, one may create an evolving ad placement system deciding how many ads to place to a website based on the ad revenue generated in different setups. The ad placement system would be an excellent example of real-world applications. On the other hand, we can train an agent in a video game with reinforcement learning to compete against other players, which are usually referred to as bots. Finally, virtual and real training of robots in terms of their movements are done with the reinforcement

---

[300] Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. *Https://Doi.Org/10.2200/S00196ED1V01Y200906AIM006*, *6*, 1–116. https://doi.org/10.2200/S00196ED1V01Y200906AIM006.

[301] Seeger, M. (2006). A Taxonomy for Semi-Supervised Learning Methods. *Semi-Supervised Learning*, 7-12. https://infoscience.epfl.ch/record/161326

[302] van Otterlo, M., & Wiering, M. (2012). Reinforcement Learning and Markov Decision Processes. In *Reinforcement Learning. Adaptation, Learning, and Optimization* (Vol. 12, pp. 3–42). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27645-3_1/COVER.

learning approach. Some of the popular reinforcement learning models may be listed as follows:[303]

- Q-Learning,
- State-Action-Reward-State-Action (SARSA),
- Deep Q Network (DQN),
- Deep Deterministic Policy Gradient (DDPG),

One of the disadvantages of the existing deep learning frameworks is that they lack comprehensive module support for reinforcement learning, and TensorFlow is no exception. Deep reinforcement learning can only be done with extension libraries built on top of existing deep learning libraries such as Keras-RL, TF.Agents, and Tensorforce or dedicated reinforcement learning libraries such as Open AI Baselines and Stable Baselines.

7.3.5 Evaluation of Different Machine Learning Approaches

We briefly covered the four main machine learning approaches: (i) Supervised learning, (ii) unsupervised learning, (iii) semi-supervised learning, and (v) reinforcement learning. We also mentioned or briefly explained the most popular machine learning models for each approach. These models can be placed in a taxonomy model:

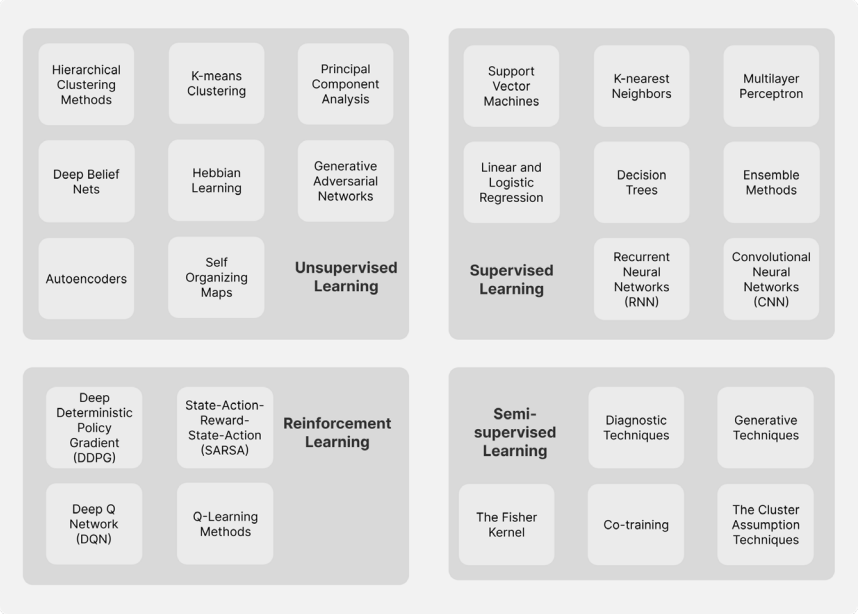


Fig 13. A Taxonomy of the Popular Machine Learning Models

[303] Janetzky, P. (2021, August 29). *Deep Reinforcement Learning: From SARSA to DDPG and beyond*. Towards Data Science. https://towardsdatascience.com/deep-reinforcement-learning-from-sarsa-to-ddpg-and-beyond-458100c2fda8.

These approaches are applied to machine learning problems with several potential algorithms to solve various sub-problem sets. While supervised learning solves classification and regression problems, unsupervised learning deals with dimensionality reduction and clustering. Semi-supervised learning combines supervised learning and unsupervised learning approaches to take advantage of unlabeled data for classification tasks, whereas reinforcement learning is used to find the perfect set of actions for the highest reward. A summary of the characteristics of these approaches may be found in Fig 14:

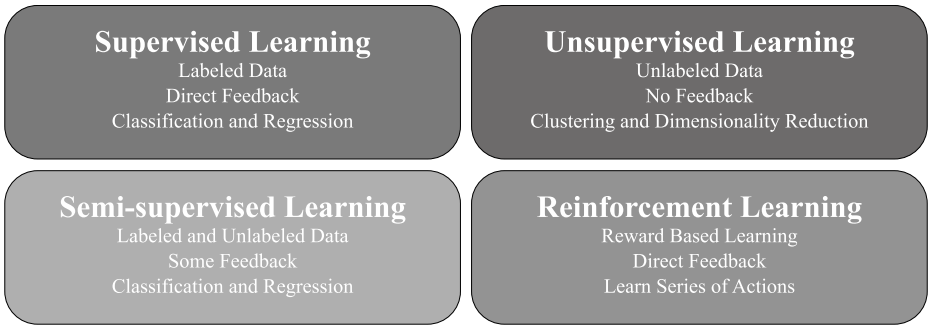


Fig 14. A Summary of the Characteristics of the Machine Learning Approaches

In this paper, the focus will be on supervised learning algorithms. The stages of machine learning are structured by taking the supervised learning flow into account. Therefore, the most applicable problem sets will be regression and classification problems. On the other hand, clustering and dimensionality reduction problems can also benefit from explainable AI techniques as well as semi-supervised learning problems.[304] Finally, although we will not pay much attention in this paper, number of studies focus on explainability of reinforcement learning algorithms.[305]

7.4 Stages of Machine Learning Development Cycle

Thanks to years of machine learning studies, we have a standardized machine learning development process flow where we can accurately build and train models. Although

[304] Montavon, G., Kauffmann, J., Samek, W., & Müller, K. R. (2022). Explaining the Predictions of Unsupervised Learning Models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13200 LNAI*, 117–138. https://doi.org/10.1007/978-3-031-04083-2_7/FIGURES/8.

[305] Wells, L., & Bednarz, T. (2021). Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. *Frontiers in Artificial Intelligence*, *4*, 48. https://doi.org/10.3389/FRAI.2021.550030/BIBTEX.

you might see slightly altered process flows in other sources, the fundamentals remain the same. The steps of a machine learning process may be listed as follows:[306]

- Data Collection
- Data Processing
- Model Selection
- Training
- Evaluation
- Hyperparameter Tuning
- Prediction

7.4.1 Data Collection

Data is the fuel of machine learning models. Without proper data, we cannot reach our expected destination: high accuracy. This data must be of high quality as well as in large volumes. Therefore, both the quality and quantity of the gathered data are significant for a successful machine learning project. In fact, gathering data is one of the most challenging parts of machine learning projects.[307] This stage is also one of the most delicate stages of the development cycle since a poorly planned data collection practice can damage several Trustworthy AI principles such as fairness and accountability. Data collection is a stage suitable for explainability techniques. The output of this stage is a representation of data in a tabular (e.g., SQL tables, CSV files) or non-tabular format (NoSQL databases, JSON files, dictionaries).

7.4.2 Data Pre-Processing and Cleaning

After collecting the data, this data should be transformed into a format that the machine learning algorithms can accept. Therefore, first, initial cleaning and transformations to the dataset is applied to remove the initial noise. This part may include several tasks, including, but not limited to, dealing with missing values, removing duplicates, correcting errors, converting data structures (e.g., from string to float), normalizing the data, and generating dummy variables. [308] After the initial cleaning and transformation, the data is usually randomized to eliminate any unwanted correlation due to the timing of data gathering. After cleaning and randomizing our data, data visualization tools are employed to discover relationships between variables that may help during the model building process. It is important to use explainability techniques to detect discriminatory patterns at this stage. In addition to the biases and discriminatory patterns, issues such as class imbalances and outliers can also be identified with data visualization.

---

[306] Bhattacharya, S. (2021). *A Primer on Machine Learning in Subsurface Geosciences* (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-030-71768-1

[307] Lehr, D., & Ohm, P. (2017). Playing with the Data: What Legal Scholars Should Learn about Machine Learning. *U.C. Davis Law Review*, *51*. https://heinonline.org/HOL/Page?handle=hein.journals/davlr51&id=667&div=&collection=

[308] Lehr & Ohm, *Playing with the Data*, 681-683.

After fully processing its data, the prepared dataset is split into training and evaluation (*i.e., test*) datasets. [309]

7.4.3 Model Selection

After collecting, processing, and splitting the data into prepared testing and evaluation sets, depending on the problem, a number of alternative machine learning models are trained to find the best performing model. Since writing the code from scratch is a redundant and cumbersome task, a number of machine learning libraries has become popular, which reduces the development time. However, this also damages the developer's understanding of the model's inner logic, which makes the overall system less explainable.

7.4.4 Training

The training step is usually combined with model selection since, for model selection, these models should be trained, and their performance should be evaluated. In the training stage, processed training data is fed into the model(s) for the optimization of the variable coefficients and loss minimization. The goal of training is to make the highest number of correct predictions or the lowest amount of error. For example, if we are using linear regression with a single explanatory variable, the linearly optimized regression line equation would be the following: [310]

$$y = m*x + b$$

Equation 1. Linear Regression Equation

Notation:
y: response variable
x: explanatory variable
m: slope
b: intercept

Linear regression model tries to find the perfect slope (*m*) and intercept (*b*) values to minimize aggregated measures of the difference between the actual y values and y predictions. The process for perfecting the model is done iteratively over several training steps until no further performance increase on the selected performance metric is observed.

7.4.5 Evaluation

Immediately after training the model with training dataset, evaluation dataset that the model has never been fed before is used to measure the performance of the model out-

[309] Bhattacharya, *A Primer on Machine Learning*, 62.
[310] Rebala & Ravi, *An Introduction to Machine Learning*, 64-65.

side of its known universe. This previously unseen data provides with an objective performance score that is more reliable. The ideal training/test split ratios for datasets are usually 80/20, 90/10, or 70/30, depending on the domain. In some cases, data scientist also set aside a validation dataset.[311]

Especially when the available observations for training are limited, one of the useful evaluation techniques used by data scientists is cross-validation.[312] Evaluation is a particularly important step to check for overfitting. Machine learning models are overly eager when it comes to optimization. They tend to create a very complex set of variable values to capture all the variance in our data. However, this may lead to overfitting problems[313] when we deploy the model in real life since perfecting a model using a limited amount of training data creates a short sighting effect. The model should be highly accurate but also flexible. In machine learning development, developers should find a good balance between the bias & variance. There has to be a balance between the level of statistical bias introduced to the system and the level of the variance observed so that the model provides meaningful and reliable predictions in real life.[314]

These are some of the properties to look out for when evaluating a machine learning model. Let's say we were careful about bias & variance trade-off and overfitting, and we used cross-validation for training our model. But *how are we going to measure the success of our model?* This is where we choose performance terms, depending on our problem.

Performance Terms for Classification Problems: Developers and data scientists usually rely on **Confusion Matrix** to understand how our model performed. Confusion matrix does not only allow us to calculate the **accuracy** of the model but also **recall**, **precision**, and **F1-score** of the model performance.[315]

---

[311] Lehr & Ohm, *Playing with the Data*, 696-700.

[312] See. ***Cross-Validation*** *is an alternative resampling technique used for evaluation. In k-fold cross-validation, the dataset is split into k number of groups. One group is kept as testing data, and this group is switched k times. So, each group is used for testing once. In the end, we have a much more reliable performance evaluation.*

[313] See. ***Overfitting*** *is a machine learning problem that occurs when the model is too closely fits the observations. When the model has an overfitting problem, it tends to perform well for training data but performs poorly for testing data and in the real world.*

[314] See. ***Bias & Variance Trade-off*** *is a property of machine learning models. Bias is the assumptions made by the model to simplify the optimization process. Variance is the amount of chance that the estimate of the target function with different data. While bias brings simplicity to the model, you may be way off to have reliable predictions, whereas the variance damages the ability to obtain meaningful results.*

[315] Klein, B. (2022, July 5). *Confusion Matrix in Machine Learning*. Python-Course.Eu. https://python-course.eu/machine-learning/confusion-matrix-in-machine-learning.php

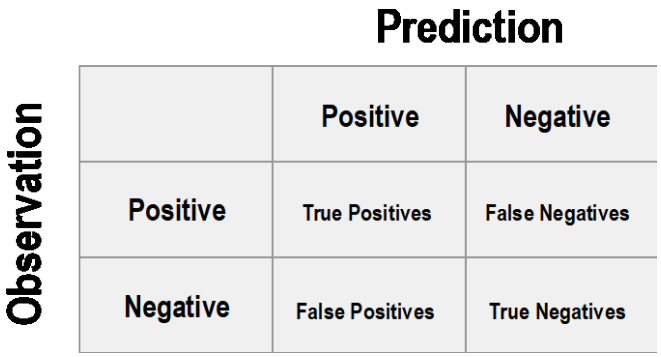


Fig 15. Confusion Matrix for Classification Problems

Performance Terms for Regression Problems: Developers and data scientists usually use **error-based metrics** to measure model performance. The difference between real observation and prediction is called an error. With an aggregative calculation, we might find metrics such as **Root Mean Squared Error (RMSE), Mean Absolute Error (MSE)**, and other metrics.[316] These metric values are useful to measure the model's success for a particular regression problem.

7.4.6 Hyperparameter Tuning

With the evaluation step, developers can generate performance metrics for each variation of the model for both training and test datasets; they can tune model hyperparameters to increase our performance even further. Learning rate, number of training steps, initialization values, epoch size, batch size, and distribution type are some of the hyperparameters that can be used to maximize the mode performance.[317] Hyperparameter tuning is usually referred to as an artwork rather than a science. Data scientists use their intuition to try different combinations of hyperparameters to achieve the highest performance.

7.4.7 Prediction

Following the training, evaluation, and hyperparameter tuning, the model development is completed. Therefore, the training model can be used to make predictions for a previously unseen observation. The prediction step should not be seen as the end of the learning process. After receiving real-world feedback, developers go back and train, evaluate, and tune our model further to address the ever-changing nature of data science problems.

---

[316] Dickson, M. C., Bosman, A. S., & Malan, K. M. (2022). Hybridised Loss Functions for Improved Neural Network Generalisation. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST*, *405 LNICST*, 169–181. https://doi.org/10.1007/978-3-030-93314-2_11

[317] Lehr & Ohm, *Playing with the Data*, 696-700.

In this section, we covered the stages of the machine learning development cycle; the stages are sometimes combined or presented separately. However, the tasks completed in the cycle are standardized. Since deep learning is a sub-field of machine learning, these stages are mostly applicable to deep learning problems except for the subtle differences that goes beyond the scope of this thesis.

7.5 Deep Learning Overview and Black Box Problem

The popularity of deep learning algorithms and neural networks has gained momentum in recent years. There is a very good reason for deep learning's increasing popularity: **its uncanny accuracy performance**. Especially when there are abundant data and available processing power, deep learning is the choice of machine learning experts.[318] The performance comparison between deep learning and traditional machine learning algorithms is shown below in Fig 16.

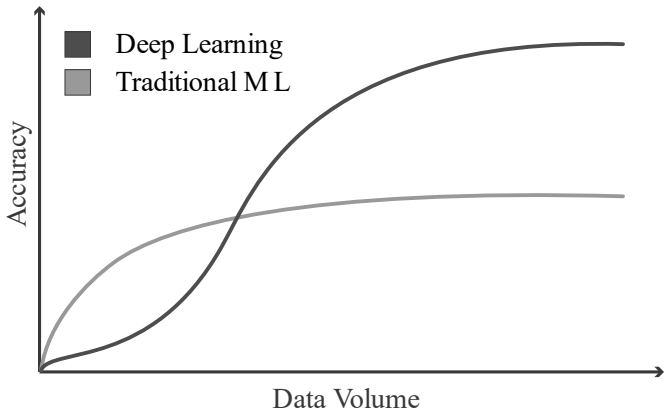


Fig 16. Deep Learning vs. Traditional ML Comparison on Accuracy

Deep learning is a subfield of machine learning which imitates data processing and pattern generation capabilities of the human brain for automated decision making. The distinct accuracy curve of deep learning compared to the other machine learning algorithms contributed to its widespread use and adoption by machine learning experts. Deep learning is made possible thanks to artificial neural networks. Artificial Neural Networks are the network structure that simulates the neurons in human brains so that deep learning can take place. Below you may find an example of an artificial neural network (ANN) with deep learning capability.

---

[318] Pu, Y., Apel, D. B., Liu, V., & Mitri, H. (2019). Machine learning methods for rockburst prediction-state-of-the-art review. *International Journal of Mining Science and Technology*, *29*(4), 565–570. https://doi.org/10.1016/J.IJMST.2019.06.009

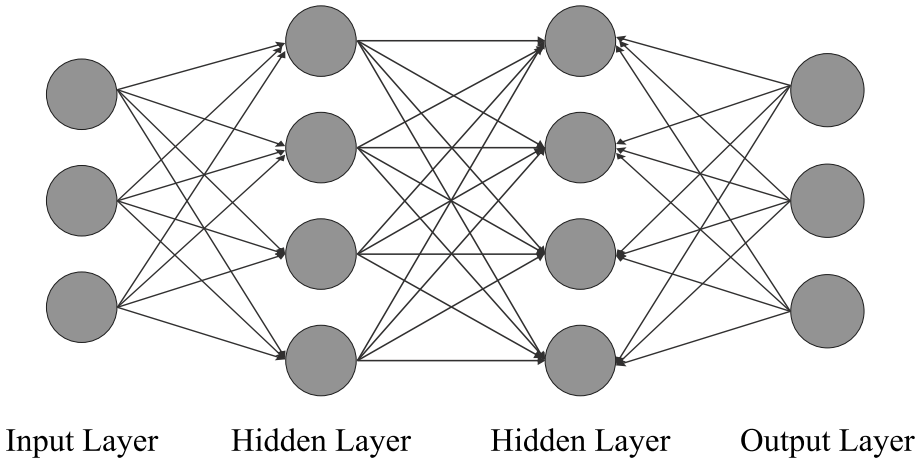


Fig 17. A Depiction of Artificial Neural Networks with Two Hidden Layers

One might think that deep learning is a newly invented field that has recently overthrown other machine learning algorithms. However, the field of artificial neural networks and deep learning dates to the 1940s. The recent rise of deep learning is mainly due to a high amount of available data and -more importantly- due to cheap and abundant processing power.
In this section, we will identify and define at the critical concepts that we often use in deep learning, including (i) activation functions, (ii) loss functions, (iii) optimizers and backpropagation, (iv) regularization, and (v) feature scaling. However, before starting the concept definitions, we will cover the history of artificial neural networks and deep learning.

7.5.1 Timeline of Neural Networks and Deep Learning Studies

The timeline of neural networks and deep learning studies does not consist of a series of uninterrupted advancements. In fact, the field of artificial intelligence experienced a few downfalls, which are referred to as AI winters. The history of neural networks and deep learning starts in 1943 and although experiences downfalls, it continues uninterruptedly until today.

Development of Artificial Neurons – In 1943, the pioneer academics Walter Pitts and Warren McCulloch published the paper "*A Logical Calculus of the Ideas Immanent in Nervous Activity,"* where they presented a mathematical model of a biological neuron called ***McCulloch Pitts Neuron***. The capabilities of McCulloch Pitts Neuron are minimal, and it does not have a learning mechanism. The importance of McCulloch Pitts Neuron is that it lays the foundation for deep learning. In 1957, Frank Rosenblatt published another paper, titled "The Perceptron: A Perceiving and Recognizing Automa-

ton", where he introduced the **Perceptron** with learning and binary classification capabilities.[319] The revolutionary Perceptron model -risen to its place after Mcculloch Pitts Neuron- has inspired many researchers working on artificial neural networks.

Backpropagation – In 1960, Henry J. Kelley published a paper titled "Gradient Theory of Optimal Flight Paths," where he demonstrates an example of continuous backpropagation. In 1962, Stuart Dreyfus improved backpropagation with chain rule in his paper, "The Numerical Solution of Variational Problems." Paul Werbos was first in the U.S. to propose that backpropagation could be used for neural nets after analyzing it in depth in his Ph.D. Thesis in 1974. [320] The term backpropagation was coined in 1986 by Rumelhart, Hinton & Williams, and these researchers have popularized its use in artificial neural networks.[321]

Training and Computerization – In 1965, Alexey Ivakhnenko, usually referred to as the "Father of Deep Learning," built a hierarchical representation of neural networks and successfully trained this model by using a polynomial activation function. In 1970, Seppo Linnainmaa found automatic differentiation for backpropagation and was able to write the first backpropagation program. This development may be marked as the beginning of the computerization of deep learning. In 1971, Ivakhnenko created an 8-layer neural network, which is considered a deep learning network due to its multilayer structure.

AI Winter – In 1969, Marvin Minsky and Seymour Papert wrote the book Perceptrons, in which he fiercely attacks the work of Frank Rosenblatt, the Perceptron. This book caused devastating damage to AI project funds, which triggered **an AI winter** that lasted until the 1980s. [322]

Convolutional Neural Networks – In 1980, Kunihiko Fukushima introduced the Neocognitron, the first convolutional neural networks (CNNs), which can recognize visual patterns. In 1982, Paul Werbos proposed the use of backpropagation in neural networks for error minimization, and the AI community has adopted this proposal widely. In 1989, Yann LeCun used backpropagation to train CNNs to recognize handwritten digits in the MNIST dataset. [323]

---

[319] Kurenkov, A. (2020, September 27). *A Brief History of Neural Nets and Deep Learning*. Skynet Today. https://www.skynettoday.com/overviews/neural-net-history

[320] Kurenkov, *A Brief History of Neural Nets*.

[321] Chauvin, Y., & Rumelhart, D. E. (1995). *Backpropagation: Theory, Architectures, and Applications*. Lawrence Erlbaum Associates. https://www.routledge.com/Backpropagation-Theory-Architectures-and-Applications/Chauvin-Rumelhart/p/book/9780805812596

[322] Kurenkov, *A Brief History of Neural Nets*.

[323] Caceres, P. (n.d.). The Convolutional Neural Network. In *Introduction to Neural Network Models of Cognition*. Retrieved August 29, 2022, from https://com-cog-book.github.io/com-cog-book/features/recurrent-net.html

Recurrent Neural Networks – In 1982, John Hopfield introduced the Hopfield Network, which is an early implementation of recurrent neural networks (RNNs). Recurrent neural networks are revolutionary algorithms that work best for sequential data. In 1985, Geoffrey Hinton, David H. Ackley, and Terrence Sejnowski proposed Boltzmann Machine, which is a stochastic RNN without an output layer. In 1986, Paul Smolensky developed a new variation of the Boltzmann Machine, which does not have intra-layer connections in input and hidden layers, which is called a Restricted Boltzmann Machine. Restricted Boltzmann Machines are particularly successful in recommender systems. In 1997, Sepp Hochreiter and Jürgen Schmidhuber published a paper on an improved RNN model, Long Short-Term Memory (LSTM), which we will also cover under the RNN Chapter. In 2006, Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh combined several Restricted Boltzmann Machines (RBMs) and created Deep Belief Networks, which improved the capabilities of RBMs.[324]

Capabilities of Deep Learning – In 1986, Terry Sejnowski developed NETtalk, a neural network-based text-to-speech system that can pronounce English text. In 1989, George Cybenko showed in his paper "Approximation by Superpositions of a Sigmoidal Function" that a feed-forward neural network with a single hidden layer ***can solve any continuous function***.[325]

Vanishing Gradient Problem - In 1991, Sepp Hochreiter discovered and proved the vanishing gradient problem, which slows down the deep learning process and makes it impractical. After 20 years, In 2011, Yoshua Bengio, Antoine Bordes, and Xavier Glorot showed that using Rectified Linear Unit (ReLU) as the activation function can prevent vanishing gradient problem.[326]

GPU for Deep Learning – In 2009, Andrew Ng, Rajat Raina, and Anand Madhavan, with their paper "Large-scale Deep Unsupervised Learning using Graphics Processors", recommended the use of GPUs for deep learning since the number of cores found in GPUs is a lot more than the ones in CPUs. This switch reduces the training time of neural networks and makes their applications more feasible.[327] Increasing use of GPUs for deep learning has led to the development of specialized ASICS for deep learning

---

[324] Caceres, The Recurrent Neural Network. In *Introduction to Neural Network Models of Cognition*. Retrieved August 29, 2022, from https://com-cog-book.github.io/com-cog-book/features/recurrent-net.html

[325] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 1989 2:4*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

[326] Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 315–323. https://proceedings.mlr.press/v15/glorot11a.html

[327] Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. *ACM International Conference Proceeding Series*, *382*. https://doi.org/10.1145/1553374.1553486

(e.g., Google's TPU)[328] along with official parallel computing platforms introduced by GPU manufacturers (e.g., Nvidia's CUDA and AMD's ROCm).

ImageNet and AlexNet – In 2009, Fei-Fei Li launched a database with 14 million labeled images, called ImageNet. The creation of the ImageNet database has contributed to the development of neural networks for image processing since one of the essential components of deep learning is abundant data. Ever since the creation of the ImageNet database, yearly competitions were held to improve the image processing studies. In 2012, Alex Krizhevsky designed a GPU trained CNN, AlexNet, which increased the model accuracy by 75% compared to earlier models. [329]

Generative Adversarial Networks – In 2014, Ian Goodfellow came up with the idea of a new neural network model while he was talking with his friends at a local bar This revolutionary model, which was designed overnight, is now known as Generative Adversarial Neural Networks (GANs), which is capable of generating art, text, poems, and it can complete many other creative tasks.[330]

Power of Reinforcement Learning – In 2016, Deepmind trained a deep reinforcement learning model, AlphaGo, which can play the game of Go, which is considered a much more complicated game compared to Chess. AlphaGo beat the World Champion Ke Jie in Go in 2017.[331]

Turing Award to the Pioneers of Deep Learning – In 2019, the three pioneers in AI, Yann LeCun, Geoffrey Hinton, and Yoshua Bengio shared the Turin Award. This award is proof that shows the significance of deep learning for the computer science community. [332]

7.5.2 Structure of Artificial Neural Networks

Before diving into essential deep learning concepts, let's take a look at the journey of the development of today's modern deep neural networks. Today, we can easily find examples of neural networks with hundreds of layers and thousands of neurons, but before the mid-20th century, the term artificial neural network did not even exist. It all

---

[328] *Cloud Tensor Processing Units (TPUs)*. (n.d.). Google Cloud. Retrieved August 29, 2022, from https://cloud.google.com/tpu/docs/tpus

[329] *Brief History of Deep Learning from 1943-2019*. (2019, November 4). Machine Learning Knowledge. https://machinelearningknowledge.ai/brief-history-of-deep-learning/

[330] Giles, M. (2018, April). The GANfather: The man who's given machines the gift of imagination | MIT Technology Review. *MIT Technology Review*. https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/

[331] *AlphaGo*. (n.d.). Deepmind. Retrieved August 29, 2022, from https://www.deepmind.com/research/highlighted-research/alphago

[332] *History of Deep Learning*, Machine Learning Knowledge.

started in 1943 with a simple artificial neuron -McCulloch Pitts Neuron- which can only do simple mathematical calculations with no learning capability. [333]

McCulloch Pitts Neuron. The McCulloch Pitts Neuron was introduced in 1943, and it is capable of doing only basic mathematical operations. Each event is given a Boolean value (0 or 1), and if the sum of the event outcomes (0s and 1s) surpasses a threshold, then the artificial neuron fires.[334] A visual example for OR and AND operations with McCulloch Pitts Neuron is shown in Fig 18:

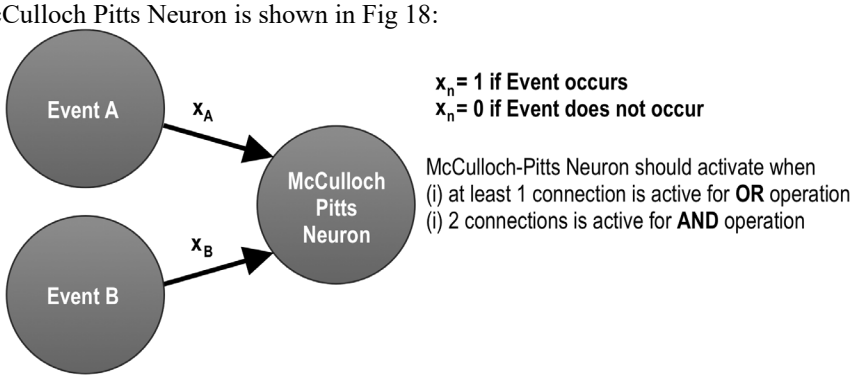


Fig 18. McCulloch Pitts Neuron for OR and AND operations

Since the inputs from the events in McCulloch Pitts Neuron can only be Boolean values (0 or 1), its capabilities were minimal. This limitation was addressed with the development of the Linear Threshold Unit (LTU).

Linear Threshold Unit (LTU). In a McCulloch Pitts Neuron, the significance of each event is equal, which is problematic since most real-world events do not conform to this simplistic setting. To address this issue, Linear Threshold Unit (LTU) was introduced in 1957. In an LTU, weights are assigned to each event, and these weights can be negative or positive. The outcome of each even is still given a Boolean value (0 or 1), but then is multiplied by the assigned weight. The LTU is only activated if the sum of these weighted event outcomes is positive.[335] In Fig 18, you may find a visualization of LTU, which is the basis for today's artificial neural networks.

---

[333] *History of Deep Learning*, Machine Learning Knowledge.

[334] Kurenkov, *A Brief History of Neural Nets*.

[335] Gurney, K. (1997). *An Introduction to Neural Networks* (1st ed.). CRC Press. https://doi.org/10.1201/9781315273570

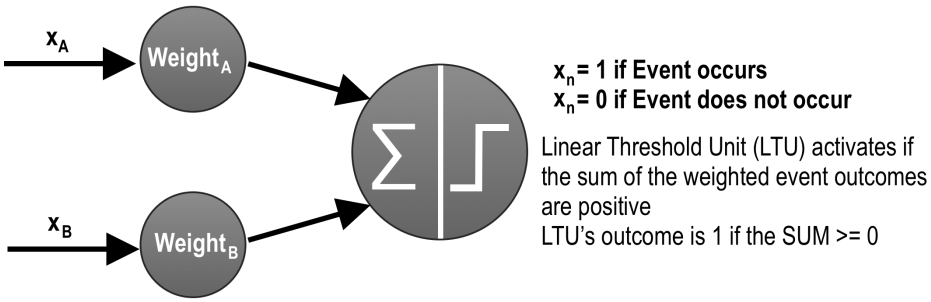


Fig 19. Linear Threshold Unit (LTU) Visualization

Perceptron. Perceptron is a binary classification algorithm for supervised learning and consists of a layer of LTUs. In a Perceptron, LTUs use the same event outputs as input. The perceptron algorithm can adjust the weights to correct the behavior of the trained neural network. In addition, a bias term may be added to increase the accuracy performance of the network. When there is only one layer of Perceptron, it is called a single-layer perceptron. There is one layer for outputs, along with a single input layer that receives the inputs. When hidden layers are added to a single-layer perceptron, we end up with a multilayer perceptron (MLP). An MLP is considered a type of deep neural network, and the artificial neural networks we build for everyday problems are examples of MLP. [336] Below in Fig 20, you may find an example visualization of a single-layer Perceptron:

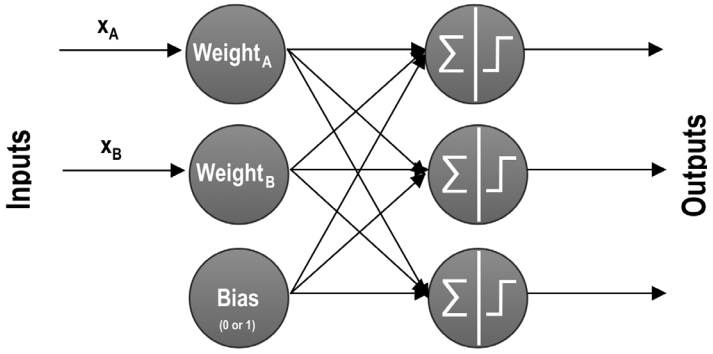


Fig 20. An Example of a Single Layer Perceptron Diagram

A Modern Deep Neural Network. The deep neural networks we come across today are improved versions of multilayer perceptron (MLP). We often use a more complex activation function than a step function (0 or 1) such as ReLU, Sigmoid, Tanh, and SoftMax. Modern deep neural networks usually take advantage of one of the gradient descent methods for optimization.[337] An example modern deep neural network is shown in Fig 21 below:

---

[336] Gurney, *An Introduction to Neural Networks*.
[337] Kurenkov, *A Brief History of Neural Nets*.

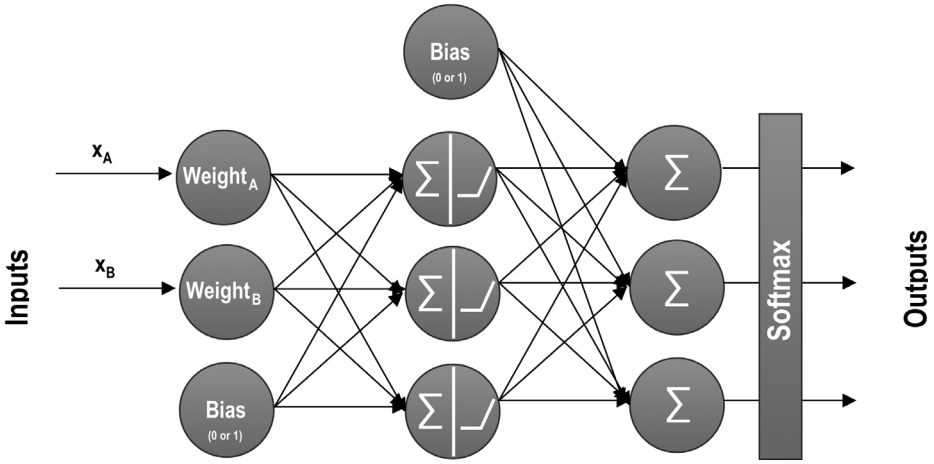
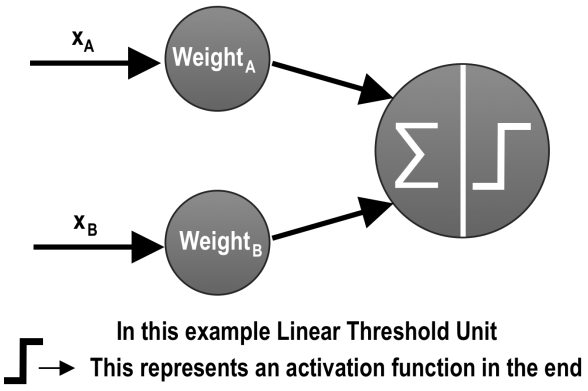


Fig 21. A Modern Deep Neural Network Example

Now that we are more informed about the journey to develop today's modern deep neural networks, which started with the McCulloch Pitts Neurons, we can move on to defining and understanding the essential deep learning concepts used in deep learning applications.

7.5.3 Activation Functions

An activation function is a function used to help artificial neural networks to learn complex patterns from the data. An activation function is usually added to the end of each neuron, which affects what to fire to the next neuron.[338] In other words, as shown in Fig 22, the activation function of a neuron gives the output of that neuron after being given an input or set of inputs.



[338] Chauhan, N. S. (2022, April 26). *An Overview of Activation Functions in Deep Learning*. https://www.theaidream.com/post/an-overview-of-activation-functions-in-deep-learning

Fig 22. An Example LTU Diagram with Activation Function in the End

Activation functions introduce a final calculation step that adds additional complexity to artificial neural networks. Therefore, they increase the required training time and processing power. Despite this negative side effect, activation functions increase the capabilities of the neural networks to use relevant information and suppress irrelevant data points. Without activation functions, neural networks would only be performing a linear transformation. Although avoiding activation functions makes a neural network model simpler, the model will be less powerful and will not be able to converge on complex pattern structures. A neural network without an activation function is essentially just a linear regression model.

There are a number of different activation functions we can use in our neural networks. A non-exhaustive list of activation functions may be found below: [339]

- Binary Step
- Linear
- Sigmoid (Logistic Activation Function)
- Tanh (Hyperbolic Tangent)
- ReLU (Rectified Linear Unit)
- SoftMax
- Leaky ReLU
- Parameterized ReLU
- Exponential Linear Unit
- Swish

Among these activation functions, Tanh, ReLU, and Sigmoid activation functions are widely used for single neuron activation. Also, the SoftMax function is widely used after layers. You may find the X-Y plots for Tanh, ReLU, and Sigmoid functions in Fig 23.

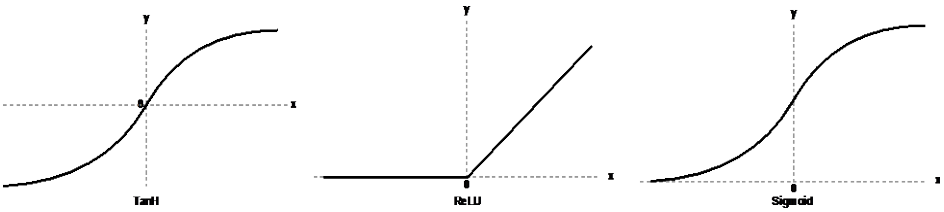


Fig 23. Plots for Tanh, ReLU, and Sigmoid Functions

Depending on the nature of the problem, one activation function may perform better than the other. Even though ReLU, Tanh, and Sigmoid functions usually converge well in deep learning, we should try all possible functions and optimize our training to

[339] Chauhan, *An Overview of Activation Functions in Deep Learning*.

achieve the highest accuracy performance possible. A straightforward comparison between ReLU, Tanh, and Sigmoid can be made with the following bullet points: [340]

- ReLU function is a widely used general-purpose activation function. It should be used in hidden layers. In case there are dead neurons, Leaky ReLU may fix potential problems.
- The sigmoid function works best in classification tasks.
- Sigmoid and Tanh functions may cause the vanishing gradient problem.

The best strategy for an optimized training practice is to start with ReLU and try the other activation functions to see if the performance improves.

7.5.4 Loss (Cost or Error) Functions

Loss functions are functions that are used to measure the performance of a deep learning model for given data. It is usually based on error terms, which is calculated as the distance between the real (measured) value and the prediction of the trained model.

$$e_i = y_i - \hat{y}_i$$

Error = Measured Value - Predicted Value

Equation 2. The Error Term Equation

Therefore, we can calculate an error term for each prediction we make. When working with millions of data points, to be able to derive insights from these individual error terms, we need an aggregative function so that we can come up with a single value for performance evaluation. This function is referred to as the **loss function, cost function, or error function, depending on the context.**[341]

Several loss functions are used for performance evaluation, and choosing the right function is an integral part of model building. This selection must be based on the nature of the problem. While Root Mean Squared Error (RMSE) function is the right loss function for regression problems in which we would like to penalize large errors, multi-class cross-entropy should be selected for multi-class classification problems.

In addition, to be used to generate a single value for aggregated error terms, the loss function may also be used for rewards in reinforcement learning. In most AI systems, loss functions are used with error terms, but it is possible to use loss functions as a reward measure.

Several loss functions are used in deep learning tasks. Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) are some of the appropriate loss functions for regression

---

[340] Chauhan, *An Overview of Activation Functions in Deep Learning*.
[341] Dickson & Bosman, *Hybridised Loss Functions*, 2-3.

problems. For binary and multi-class classification problems, we can use variations of Cross-Entropy (i.e., *Logarithmic*) function.

7.5.5 Optimization in Deep Learning

Now that we covered activation and loss functions, it is time to move on to weight and bias optimization. Activation functions used in neurons and layers make final adjustments on the linear results derived from weights and bias terms. We can make predictions using these parameters (weights and biases). The distances between the actual values and the predicted values are recorded as error terms. These error terms are aggregated into a single value with loss functions. In addition to this process, optimization functions make small changes to the weights and biases and measure the effects of these changes with loss functions. This process helps to find the optimal weight and bias values to minimize errors and maximize the accuracy of the model. This training cycle is shown in Fig 24 below:

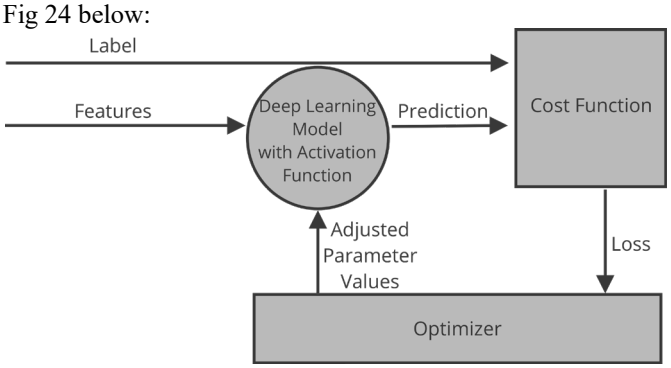


Fig 24. Deep Learning Model Training with Cost Function, Activation Function, and Optimizer

There are several optimization algorithms and challenges encountered during the optimization process. In this section, we will briefly introduce these functions and challenges. But first, we will cover an essential optimization concept: *Backpropagation*.

7.5.6 Backpropagation

The backpropagation algorithm is an essential component in neural network architecture used for iteration in parallel with the optimizer. It serves as a central mechanism by which neural networks learn. The name explains itself since the word propagates means is to transmit something. Therefore, the word backpropagation means "transmitting information back". This is what the backpropagation algorithm precisely does: It takes the calculated loss back to the system, which is used by the optimizer to adjust the weights and biases.[342] This process may be explained step by step, as shown below:

---

[342] Chauvin & Rumelhart, *Backpropagation: Theory, Architectures, and Applications*.

- **Step 1** – The trained neural network makes a prediction with the current weights and biases,
- **Step 2** – The performance of the neural network is measured with a loss function as a single error measure,
- **Step 3** – This error measure is backpropagated to the optimizer so that it can re-adjust the weights and biases, and
- **Repeat**

By using the information provided by the backpropagation algorithm, optimization algorithms can perfect the weights and biases used in the neural network. Let's look at the optimization algorithms (i.e., *optimizers*), which are used in parallel with the backpropagation mechanism.

7.5.6.1 Optimization Algorithms

An optimization algorithm may be defined as an algorithm helping another algorithm to maximize its performance without delay. Deep learning is one field where optimization algorithms are widely used. The most common optimization algorithms used in deep learning tasks are listed as follows:[343]

- Adam
- Stochastic Gradient Descent (SGD)
- Adadelta
- Rmsprop
- Adamax
- Adagrad
- Nadam

Note that all these optimizers are readily available in the existing deep learning libraries such as TensorFlow and PyTorch as well as the loss and activation functions. They are easily used by the data scientists without having to know a deep understanding of their inner logic. Although this standardization reduces the development time, it can also pose issues regarding overall explainability of the AI systems. The most common ones used in real applications are Adam Optimizer and Stochastic Gradient Descent (SGD) Optimizer. To have a general idea about how the optimization functions work, we will cover the Gradient Descent & SGD algorithm.

7.5.6.1.1 Gradient Descent and Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent is a variation of gradient descent methods. SGD is widely used as an iterative optimization method in deep learning. The roots of SGD date back to the 1950s, and it is one of the oldest -yet most successful- optimization algorithms. Gradient Descent methods are a family of optimization algorithms used to minimize

---

[343] Murugan, P., & Durairaj, S. (2017). *Regularization and Optimization strategies in Deep Convolutional Neural Network*. https://doi.org/10.48550/arxiv.1712.04711

the total loss (or cost) in neural networks. There are several gradient descent implementations: The original Gradient Descent -or Batch Gradient Descent- algorithm uses the whole training data per epoch.[344] Stochastic (Random) Gradient Descent (SGD) selects a random observation to measure the changes in total loss (or cost) because of the changes in weights and biases. Finally, mini-batch Gradient Descent uses a small batch so that training may still be fast as well as reliable. [345]

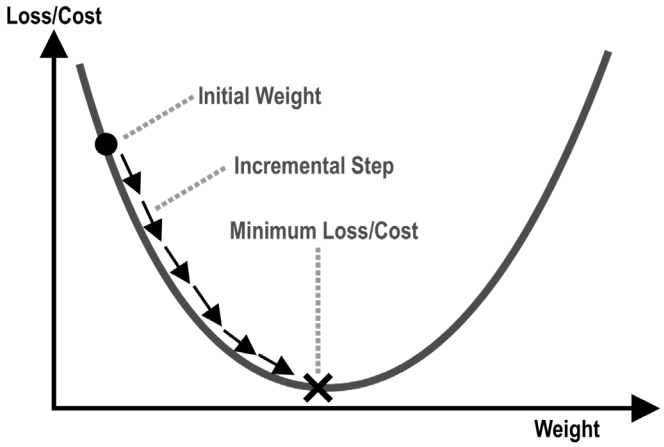


Fig 25. A Weight-Loss Plot Showing Gradient Descent

Fig 25 shows how Gradient Descent algorithms work. Larger incremental steps are taken when the machine learning expert selects a faster learning rate.

Learning Rate can be described as the parameter in optimization algorithms that regulates the step size taken at each iteration while moving forward a minimum of a loss/cost function. With a fast-learning rate, the model converges around the minimum faster, yet it may overshoot the actual minimum point. With a slow learning rate, optimization may take too much time. Therefore, a machine learning expert must choose the optimal learning rate, which allows the model to find the desired minimum point in a reasonable time.[346]

We will not cover the other optimization algorithms since they are mostly altered or improved implementations of gradient descent methods. Therefore, understanding the gradient descent algorithm will be enough for the purpose of this thesis.

In the next section, we see the optimization challenges which negatively affect the optimization process during training. Some of the optimization algorithms, as mentioned earlier, were developed to mitigate these challenges.

---

[344] See. ***Epoch*** *is the hyperparameter that represents the number of times that the values of a neural network are to be adjusted using the training dataset.*

[345] Murugan & Durairaj, *Regularization and Optimization Strategies*, 2.

[346] Bhattacharya, *A Primer on Machine Learning*, 87-88.

7.5.6.2 Optimization Challenges

There are three optimization challenges we often encounter in deep learning. These challenges are (i) Local Minima, (ii) Saddle Points, and (iii) Vanishing Gradients. Let's briefly discuss what they are. [347]

7.5.6.2.1 Local Minima

In neural network training, a simple loss-weight plot with a single minimum might be useful to visualize the relationship between the weight and the calculated loss for educational purposes. However, in real-world problems, this plot might contain many local minima, and our optimization algorithm may converge on one a local minimum rather than the global minimum point.[348] Fig 26 shows how our model can be stuck at a local minimum.

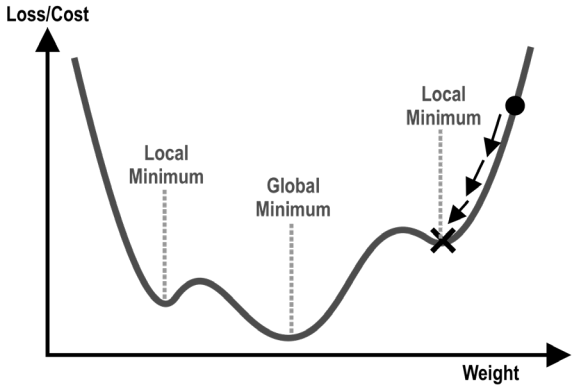


Fig 26. A Weight-Loss Plot with Two Local Minima and a Global Minimum

7.5.6.2.2 Saddle Points

Saddle points are stable points in the graphs that the algorithm cannot figure out whether it is a local minimum or a local maximum. Both sides of a saddle point have zero slopes. Optimizers using more than one observation for loss calculation may be stuck in a saddle point. Therefore, Stochastic Gradient Descent is a suitable solution for saddle points.[349] A simplified graph with saddle point is shown in Fig 27:

---

[347] Murugan & Durairaj, *Regularization and Optimization Strategies*, 2-6.

[348] Goodfellow, I. J., Vinyals, O., & Saxe, A. M. (2014). Qualitatively characterizing neural network optimization problems. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1412.6544

[349] Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in Neural Information Processing Systems*, *27*.

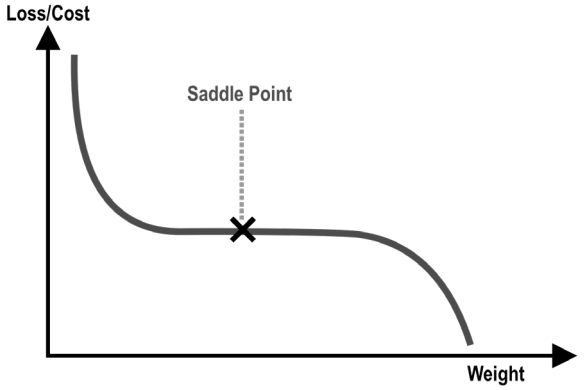


Fig 27. A Weight-Loss Plot with Two Local Minima and a Global Minimum

7.5.6.2.3 Vanishing Gradients

Excessive use of certain activation functions (e.g., sigmoid) may negatively affect the optimization algorithm. It becomes difficult to reduce the output of the loss function since the gradient of the loss function approaches zero. An effective solution to the vanishing gradient problem is to use ReLU as the activation function in hidden layers. Sigmoid activation function -the main reason for the vanishing gradient problem- and its derivative is shown in Fig 28:

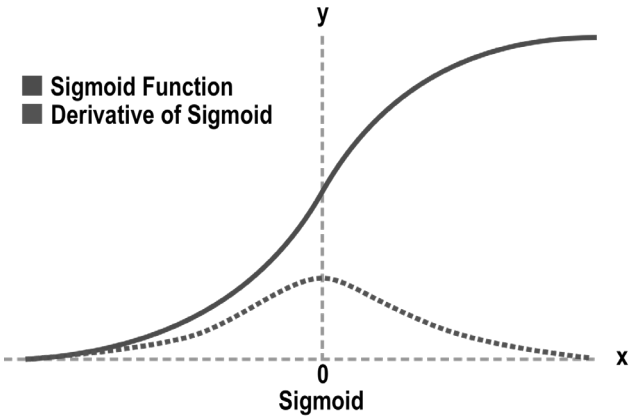


Fig 28. The Sigmoid Function and Its Derivative

To be able to solve these common optimization challenges, we should try and find the best combination of activation functions and optimization functions so that our model correctly converges and finds an ideal minimum point.

7.5.6.3 Overfitting and Regularization

Another important concept in deep learning and machine learning is overfitting. In this section, we cover the overfitting problem and how to address overfitting with regularization methods.

7.5.6.3.1 Overfitting

In the beginning of this Chapter, we already briefly covered the concept of overfitting briefly for machine learning. Overfitting is also a challenge in deep learning. When neural networks fit a limited set of data points too tightly, its performance is usually jeopardized in the real world. Underfitting is also not a desired situation since it would not achieve a good accuracy level.[350] Underfitting and overfitting problems are shown in Fig 29.

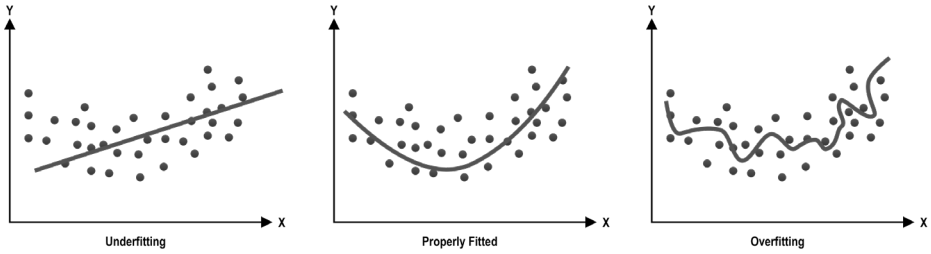


Fig 29. Underfitting and Overfitting in X-Y Plot

The solution to the underfitting problem is building a good model with meaningful features, feeding enough data, and training enough. On the other hand, more data, removing excessive features, and cross-validation are proper methods to fight the overfitting problem. In addition, we have a group of sophisticated methods to overcome overfitting problems, namely, regularization methods.

7.5.6.3.2 Regularization

Regularization is a technique to fight overfitting. There are several possible methods used for regularization, which may be listed as follows:[351]

- Early Stopping
- Dropout
- L1 & L2 Regularization
- Data Augmentation

Early Stopping – Early stopping is a very simple -yet effective- strategy to prevent overfitting. Setting enough epochs (training steps) is crucial to achieving a good level of accuracy. However, you may easily go overboard and train your model to fit too

[350] Jabbar, H. K., & Khan, R. Z. (2014). Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study). *Computer Science, Communication and Instrumentation Devices*, 163–172. https://doi.org/10.3850/978-981-09-5247-1_017

[351] Murugan & Durairaj, *Regularization and Optimization Strategies*, 9.

tightly to your training data. With early stopping, the learning algorithm is stopped if the model does not show a significant performance improvement for a certain number of epochs. [352]

Dropout – Dropout is another simple -yet effective- regularization method. With dropout enabled, our model temporarily removes some of the neurons or layers from the network, which adds additional noise to the neural network. This noise prevents the model from fitting to the training data too closely and makes the model more flexible. [353]

L1 & L2 Regularization – These two methods add an additional penalty term to the loss function, which penalizes the errors even more. For L1 regularization, this term is a lasso regression, whereas it is ridge regression for L2 regularization. L1 & L2 Regularizations are particularly helpful when dealing with a large set of features. [354]

Data Augmentation – Data augmentation is a method to increase the amount of training data. By making small transformations on the existing data, we can generate more observations and add them to the original dataset. Data augmentation increases the total amount of training data, which helps prevent the overfitting problem. [355]

7.5.6.4 Feature Scaling

Another crucial concept in deep learning is feature scaling. Feature scaling is a method to normalize the range of features so that neural networks perform more accurately. When the range of the values of a feature varies considerably, some objective functions may not work correctly in machine learning models. For instance, classifiers usually calculate the distance between two data points. When the variance of the values of a feature is large, this feature dictates this calculated distance, which means an inflated influence of this particular feature on the outcome. Scaling the value ranges of each feature helps to eliminate this problem. There are several feature scaling methods which are listed below:[356]

- **Standardization**: It adjusts the values of each feature to have zero-mean and unit variance.
- **Min-Max Normalization (Rescaling)**: It scales the values of each feature between [0, 1] or [-1, 1].
- **Mean Normalization**: It deducts the mean from each data point and divides the result to the max-min differential. It is a slightly altered and less popular version of min-max normalization.

---

[352] Murugan & Durairaj, *Regularization and Optimization Strategies*, 8.

[353] Murugan & Durairaj, *Regularization and Optimization Strategies*, 7.

[354] Murugan & Durairaj, *Regularization and Optimization Strategies*, 7-8.

[355] Murugan & Durairaj, *Regularization and Optimization Strategies*, 7.

[356] Vashisht, R. (2021, January 6). *When to perform a Feature Scaling?* Atoti. https://www.atoti.io/articles/when-to-perform-a-feature-scaling/

- **Scaling to Unit Length**: It divides each component of a feature by the Euclidian length of the vector of this feature.

Using feature scaling has two benefits in deep learning:

- It ensures that each feature contributes to the prediction algorithm proportionately.
- It speeds up the convergence of the gradient descent, therefore reducing the model training time.

7.6 Black-box Models

Black box systems can be described as systems whose inputs and outputs can be observed, but their inner mechanism is not known to outsiders. There are more than one reasons that a system can have a black box nature. For example, when a hedge fund develops a state-of-the-art trading algorithm, they may want to keep the inner logic of the system as a trade secret. This trading system is regarded as a black box system to the people who are outside of the hedge fund. This is an example of intentional black boxing.[357] Some systems can be inherently black boxes. For example, tracing the decision process of complex multi-layer neural networks are extremely difficult, which makes them black box models. Therefore, the black box characteristic of a system is a relative classification. In the world of XAI, black box systems usually refer to the systems whose inner logic cannot be explained by its nature, not because of hidden algorithmic process.

7.6.1 White box or Glass Box Models

In contrast with the black box models, white box models are transparent models whose inner mechanisms can be observed along with their inputs and outputs. Therefore, a true white box model would have algorithmic transparency, simulatability, and decomposability features.[358] White box models are also referred to as glass box or transparent models. Traditional machine learning algorithms such as linear regressions and decision trees are considered as white box models. On the other hand, black box models can be referred to as opaque models.

7.6.2 What Constitutes a Black Box Model

The main distinguishing feature between black box and white box models is the observability of their inner mechanisms. In a decision tree algorithm, we can create a tree structure with proper coefficients, which leads to a particular output when a set of input values are given. On the other hand, in a linear or logistic regression, we would have

---

[357] Pasquale, F. (2016). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674970847

[358] Barredo and Díaz-Rodríguez, *Explainable AI*, 10-12

an equation, which is optimized with a least squares algorithm, that provides coefficients for explanatory variables and constant values, which can easily be interpreted and traced. These algorithms are regarded as white box models. However, observability of the inner mechanism should reach an acceptable level. In other words, observability should not be limited to algorithmic transparency, which can be a complex equation that does not provide enough meaningful information to achieve Explainable AI goals.

On the other hand, in a multilayer perceptron (i.e., ANN), the level of transparency is very limited. Multilayer Perceptrons usually consist of several hidden layers with -often- a large number of neurons. The connection between the input layer and the output layer becomes quite abstract and hard to trace back. The approximation provided by a multilayer perceptron does not provide an insightful function.[359] While the coefficients in regression are what make it a white box, the weights (i.e., coefficients) in a neural network are not directly linked to the approximated function. An input characteristic can be very significant for one observation, whereas insignificant for another observation.

Therefore, when we cannot trace and observe the process from input values to the output, where the coefficients do not automatically associate with approximated function, or when the model does not have any of the three characteristics (i.e., algorithmic transparency, decomposability, and simulatability), the model is regarded as a black box model.

7.6.3 How to Open Black Boxes

The black box – white box issue is usually associated with the model explainability, and Explainable AI researchers have been working on methods to open the black box using post-hoc explanation methods. These post-hoc methods usually rely on external techniques to understand the inner mechanisms of black box model.[360] These methods will be explained in more details in the next Chapter's model explainability and post-hoc explanation techniques sections.

7.7 Final Remarks

In this chapter, we made an introductory analysis on machine learning, which also includes its subfield, deep learning. We compared the fields of artificial intelligence, machine learning, deep learning, data science, and big data. Then, we covered the main machine learning approaches: (i) supervised learning, (ii) unsupervised learning, (iii) semi-supervised learning, and (iv) reinforcement learning, and introduced some of the popular machine learning models used with these approaches. These concept definitions were followed by the steps of machine learning development cycles. This section

---

[359] Sarle, W. S. (2000, June 23). *How to measure importance of inputs?* SAS. ftp://ftp.sas.com/pub/neural/importance.html.

[360] Azodi, C. B., Tang, J., & Shiu, S. H. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends in Genetics*, *36*(6), 442–455. https://doi.org/10.1016/J.TIG.2020.03.005.

explained the necessary steps to successfully build & train a machine learning model using the collected and processed data.

After the general introduction to machine learning, we narrow down the scope to deep learning. We covered the timeline of artificial neural networks and deep learning, which provided valuable insights on the advancements in deep learning research. Following the deep learning timeline, we analyzed the structure of neural networks and the artificial neurons in detail. Also, we covered the fundamental deep learning concepts, including, (i) optimization functions, (ii) activation functions, (iii) loss functions, (iv) overfitting & regularization, and (v) feature scaling. Then, we finalized this chapter with a term clarification on black box and white box models and explanations on what constitutes a black box model.

In the next chapter, we will focus on the explainability techniques applicable in different stages of machine learning development cycle.

8 Strengthening Explainability at Different Stages of ML Lifecycle

As we already discussed in the previous chapters, with the increasing complexity of a model, it becomes more difficult to understand how the model behaves in a production environment.[361] Traditional machine learning models such as linear regressions and decision trees are inherently interpretable, and their algorithms allow for generating explanatory information. For instance, when an AI system using a linear regression makes a prediction after training, it also provides valuable interpretable information about the significance and coefficients of its explanatory variables. However, their accuracies have been empirically lower compared to more modern models. Considering that the creation of linear regression dates back to the 19th century, the accuracy difference between this algorithm and the modern algorithms that has been developed in the 20th and 21st centuries is not surprising.[362] The most successful models were mostly developed in the 20th and the 21st centuries to solve multi-dimensional and multi-variate problems, which encapsulate high complexity and great abstraction in their nature.[363, 364] Only with models like support vector machines and neural networks have we been able to achieve 90+% accuracies consistently and tackle complex problems in the natural language processing and computer vision spaces. However, this high level of accuracy comes at a cost, which can lead to previously unforeseen hazards: the block-box problem. Trusting a model to be fair and respect fundamental rights solely based on its past

---

[361] Hu, B., Tunison, P., Vasu, B., Menon, N., Collins, R., & Hoogs, A. (2021). XAITK: The explainable AI toolkit. *Applied AI Letters*, *2*(4), e40. https://doi.org/10.1002/AIL2.40

[362] Foote, *History of Machine Learning*.

[363] *Support Vector Machine*. (n.d.). Retrieved August 15, 2022, from https://cml.rhul.ac.uk/svm.html

[364] Hardesty, L. (2017). *Explained: Neural networks | MIT News | Massachusetts Institute of Technology*. April 14,. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

performance is not a secure approach if the values that we hold at the highest status are at the risk of being undermined.

Model-explainability is the stage where researchers and developers can tackle the black-box problem and can remedy the current negative correlation between accuracy and explainability. While model explainability takes the lion's share of the attention and efforts in the Explainable AI field, the true explainability of AI systems requires adopting and implementing methods at the pre-modeling and post-modeling stages as well. In the pre-modeling stage, understanding how the dataset was formed and processed, along with how the data was collected, may contribute to the overall explainability of an AI system. Explainable feature engineering, dataset description standardization, dataset summarization, and exploratory data analysis are some of the techniques that can be applied during data processing. In addition to these techniques, more ambitious efforts such as Linked Open Data can further strengthen the explainability of the dataset used for model training.[365]

In the post-modeling stage, during the training, evaluation, and hyperparameter tuning stages, a developer can conduct a model benchmark analysis with a focus on the explainability property to select models with a healthy balance of accuracy and explainability properties. When the trained model is used for predictions, and the model starts interacting with the data subjects. These data subjects should have access to an interface to receive meaningful information about a particular decision or the general logic of the system if they are affected by the model's predictions. In addition to the user interface, a presentation logic should transform plain data into the meaningful information specific to the case and the data subject so that the right to explanation of the data subjects can be protected and the trustworthiness property of the AI system can be preserved.[366]

Finally, in a more general sense, there are a number of management and policy-level measures that can help achieve more advancements and facilitation in the field of Explainable AI. Co-development of policies among local, national, and international level public and private institutions can bring people from different backgrounds to contest their opinions to create a suitable environment where explainability can support sustainable advancements in the field of artificial intelligence. Following these efforts, cooperated R&D efforts can help fruitful results to strengthen the explainability property of the AI systems in a faster manner. Finally, self-imposed or legally required explainability audits can guarantee an appropriate level of explainability for a given use case. For sensitive use cases, such as the cases in the medical or legal domain, the threshold can be set at a higher level than for the less sensitive use cases. In this section, we will cover pre-modeling, model, post-modeling, and policy-level explainability methods and techniques in more detail.

---

[365] *Linked Open Data*. (n.d.). Europeana Pro. Retrieved September 1, 2022, from https://pro.europeana.eu/page/linked-open-data

[366] Sovrano, F., Vitali, F., & Palmirani, M. (2020). Modelling GDPR-Compliant Explanations for Trustworthy AI. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12394 LNCS*, 10-11. https://doi.org/10.1007/978-3-030-58957-8_16.

8.1 Pre-Modeling Explainability

The research on Explainable AI mainly focuses on model explainability. However, the methods used by data analysts can also be powerful tools to secure fundamental rights and Trustworthy AI principles. These exploratory analyses and pre-processing methods can generate explanations and insights, which can be used to eliminate bias and discrimination and to determine accountability. The right to explanation does not only require providing any explanation, but it also requires the explanations to be justified. Therefore, the explanations generated by the AI systems must be in compliance with the right to explanation and Trustworthy AI principles. Standardization efforts and data linkage can be extremely powerful secondary tools that can break the negative correlation that we see between accuracy and explainability. In terms of pre-modeling explainability, we identify four types of activities that can contribute to the overall explainability of AI systems: (i) exploratory data analysis and data summarization, (ii) feature engineering, (iii) standardization activities, and (iv) linking data. In the following sections, we will cover these activities in more detail.

8.1.1 Exploratory Data Analysis and Data Summarization

Exploratory data analysis and data summarization contain various techniques that have been used by data analysts for a long time. Graphing, descriptive statistics, inferential statistics, and summary tables are some examples of these methods. The main purpose of these methods is to discover the correlations and associations between explanatory and response variables to build to generate insights to be used during decision making. Additionally, another important purpose of these groups of techniques can be eliminating bias and discriminatory practices. Graphing or data visualization methods are the methods of presenting data graphically that allows the viewers to identify trends and relationships visually and in a more intuitive manner.[367] Line plots, pie charts, bar plots, and heatmaps are examples of graphs among several dozens of variations.[368] For example, an analyst can easily spot severe cases of age, race, or gender biases using a grouped or stacked bar chart and work on eliminating the bias. In addition to graphs, summary tables can serve a similar purpose. By using summary tables, the raw information can be summarized in multiple ways.[369] Descriptive statistics summarize a set of observations such as mean, median, mode, sum, minimum, maximum, variance, and standard deviation.[370] A summary table or set of descriptive statistics can provide information about the median age, gender counts, or ethnic distribution of a dataset. This information can help analysts to detect unbalanced datasets, which can cause discriminatory model training. In addition to bias detection, these methods can even be used to provide

---

[367] Myatt, G. J. (2006). *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. John Wiley and Sons, 60-62. https://doi.org/10.1002/0470101024

[368] Sekar, M. (2022). Types of Visualizations. In *Machine Learning for Auditors* (pp. 155–167). Apress. https://doi.org/10.1007/978-1-4842-8051-5_16.

[369] Myatt, *Making Sense of Data*, 63-64.

[370] Myatt, *Making Sense of Data*, 63.

explanations at the post-modeling stage as part of the human oversight and accountability principles.

8.1.2 Feature Engineering

Feature engineering is an important data pre-processing step that helps transform, combine, and clean raw data and improve the quality and accuracy of the trained models.[371] Especially when building more traditional models, feature engineering plays an important role. Properly creating hybrid variables or simply cleaning the noisy data are some of the feature engineering techniques that are crucial to data preprocessing. In addition, feature engineering is a manual dimensionality reduction method that can simplify model complexity,[372] which makes them more explainable. We can repurpose some of these methods to eliminate biases to ensure that our data does not contain any inherent discriminatory features.

8.1.3 Standardization Activities

Feature extraction, explanatory data analysis, and visualization efforts are crucial to eliminate bias and enable human oversight. However, as the AI systems become more complex so are the datasets that they are trained on. Therefore, data standardization activities are essential to fully understand and explain the data in these datasets. Standardized column descriptions attached to the actual dataset are important to make sense of the features. Using a universal file format can be important to access the dataset and make explanations accessible to others. Finally, formatting the dataset in an understandable format such as in relational databases in key-value pairs contributes to the overall accessibility and understandability of the dataset. Finally, information about the data collection and preparation procedures are crucial to detect bias and ensure fairness.[373]

8.1.4 Linking Data

Another important tool to increase the overall explainability of the AI systems is the linked data. The linked data is any form of interconnected and structured data that can be accessed and traced with semantic queries. It builds on top of the extended versions of standard Web technologies such as HTTP, RDS, and URIs. The Linked Data concept proposes to use HTTP URIs not only to identify Web documents but also arbitrary real-

---

[371] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, *8*, 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942.

[372] Reddy & Reddy, *Dimensionality Reduction Techniques*, 2.

[373] Adhikari, A., Wenink, E., van der Waa, J., Bouter, C., Tolios, I., & Raaijmakers, S. (2022). Towards FAIR Explainable AI: a standardized ontology for mapping XAI solutions to use cases, explanations, and AI systems. *The15th International Conference on PErvasive Technologies Related to Assistive Environments*, 562–568. https://doi.org/10.1145/3529190.3535693.

world entities.[374] While the current standard data of the Web usually contains information only for human readers; the linked data contains additional information that can be read by computers. This concept allows users to trace data from one source to another, creating an endless web of data.[375] The current internet protocols do not require the use of linked data. However, the increasing popularity of the linked data concept can create opportunities for the strengthening of the explainability property of AI systems. With the linked data, data analysts can take advantage of metadata to trace the source of a data point back to its roots. They can also see a more detailed view of the data in question with the attached metadata. Tracing back data to its source can be extremely important for the Trustworthy AI principles, particularly for accountability. By an extension of this principle, Linked Data can help determine the liability in case of breach of fundamental rights and freedoms.[376]

8.2 Model Explainability

Model explainability is on its way to becoming an important concern for AI systems, especially for sensitive fields such as medicine, law, finance, and recommendation systems. With regards to explainability property, there are two main types of models: (i) transparent models and (ii) black box models. In black box models, the mapping from input to output is invisible to the user, whereas in transparent models, users can mathematically analyze the mappings.[377] With the introduction of the right to explanation under GDPR and the Trustworthy AI principles, explainable and transparent models become more preferred over black-box models when they have similar accuracy levels. However, as mentioned in the previous chapters, the current state of the art does not offer highly accurate and explainable models.[378] Especially when there is enough data, computing power, and technical talent, black box models continuously outperform the transparent models. Therefore, in recent years, research activities focusing on increasing the explainability property of the black box models have seen exponential growth.[379] Following these research efforts, today, we have a wide range of explainability techniques, which should be categorized to be able to understand in its entirety. In this chapter, we will cover both transparent models and black box models with their distinctive features and selected techniques. There are five criteria in the literature that

---

[374] Bizer, C. (2009). The emerging web of linked data. *IEEE Intelligent Systems*, *24*(5), 87–92. https://doi.org/10.1109/MIS.2009.102.

[375] Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). Linked data on the web (LDOW2008). *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, 1265–1266. https://doi.org/10.1145/1367497.1367760.

[376] Rodríguez-Doncel, V., Santos, C., Casanovas, P., & Gómez-Pérez, A. (2016). Legal aspects of linked data – The European framework. *Computer Law and Security Review*, *32*(6), 799–813. https://doi.org/10.1016/J.CLSR.2016.07.005.

[377] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 10.

[378] Gholizadeh, S., & Zhou, N. (2021). *Model Explainability in Deep Learning Based Natural Language Processing*. https://doi.org/10.48550/arxiv.2106.07410

[379] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 2.

has been used to categorize the model explainability techniques, namely (i) stage, (ii) scope, (iii) problem type, (iv) input data, and (v) output data.[380] Although there are other categorizations criteria such as application field and construction approach, within this thesis, the techniques will not be analyzed based on them.[381]

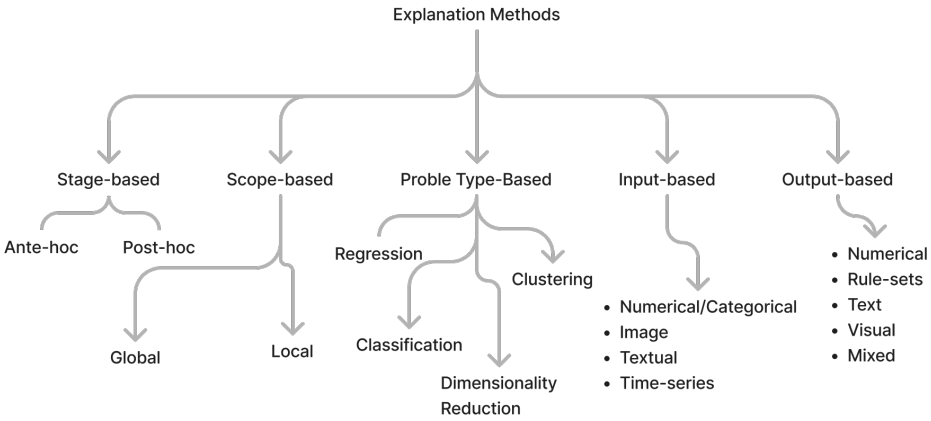


Fig 30. Classification of XAI Methods into Hierarchical System[382]

8.2.1 The Stage-based Categorization

The stage criterion refers to the stage where the model generates explanations. While the transparent models can create explanations in an ante-hoc manner, black box models can output explanations using post-hoc explanation methods. In other words, transparent models can provide explanations from the beginning, and during the training, black box models have to rely on external techniques to provide explanations for already trained models during testing.[383]

8.2.1.1 Ante-hoc Explainability and Transparency

Ante-hoc explainability requires for the underlying ML model to be inherently transparent. Therefore, ante-hoc explainability is a property of the transparent models that already provide a good level of interpretability. The transparency of a transparent model can be at different levels. In literature, we observe three levels of transparency that are ordered from less transparent to more transparent respectively: (i) Algorithmic transparency, (ii) decomposability, and (iii) simulatability. Algorithmic transparency refers to the user's ability to understand a model's process to produce an output from its given

---

[380] Vilone, G., & Longo, L. (2021). Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction 2021, Vol. 3, Pages 615-661*, *3*(3), 615–661. https://doi.org/10.3390/MAKE3030032.

[381] Vilone & Longo, *Explainable Artificial Intelligence*, 5.

[382] Vilone & Longo, *Explainable Artificial Intelligence*, 6.

[383] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 10-12.

input. For instance, a linear regression model is algorithmically transparent because the correlation between its variables and its error surface can easily be analyzed and reasoned. On top of algorithmic transparency, decomposability of a ML model would mean a higher level of overall model transparency property. Since decomposability refers to the user's ability to understand and explain the parts of a model, including its parameters, inputs, and calculations, the models with complex features may not be decomposable. Finally, the third layer would be simulatability, which refers to the user's ability to simulate a model in a human comprehensible manner. [384] Highly complex ML models such as the complex and large linear regressions or decision trees are difficult to simulate, whereas simple linear regression or decision tree variations are usually more simulatable.[385]

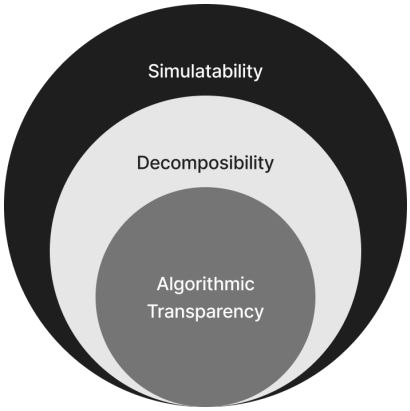


Fig 31. The Level of Transparencies of Transparent ML Models

Depending on its algorithm and case-specific complexity, machine learning models may have different levels of transparency, and we can explain their decision-making logic to a certain degree. Machine learning models such as linear/logistic regressions, decision trees, k-nearest neighbors, rule-based learners, general additive models, and Bayesian models are considered transparent models. On the other hand, tree ensemble models, support vector machines, multi-layer neural networks, and other complex neural network types are considered black box models.[386] In the next sections, we will analyze the transparent models' level of transparency under different circumstances.

Linear and Logistic Regression: Linear regression is one of the oldest machine learning models that is widely used by social scientists and econometrics studies.[387] Logistic

[384] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 10-11.

[385] Vilone & Longo, *Explainable Artificial Intelligence*, 11.

[386] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 10-11.

[387] Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine*, *3*(1), 2. https://doi.org/10.1186/1751-0473-3-17/TABLES/6.

regression adds another layer to a generic linear regression model, and instead of fitting the values to a line, it fits them to a sigmoid curve.[388] The practical difference for this additional operation is to use the model for classification tasks. While linear regression models are used for regression tasks, logistic regression models are used for classification tasks. [389] Simple linear and logistic regressions with small sample sizes and limited number of variables can be simulatable and decomposable. However, as their size and variable count increases, they lose these properties and are only left with algorithmic transparency. Additionally, another determinant for the level of transparency is the characteristics of the variables. If the variables used in a model are synthetic variables generated with feature engineering methods that include more than one raw feature, then the model's transparency properties will diminish.[390]

Decision Tree: Decision trees are inherently transparent models that can easily satisfy the three layers of transparency. Decision trees are graphical models that consist of nodes, edges, probabilities, and value information.[391] They can be built with a minimum amount of technical knowledge, and therefore, it is a popular machine learning model among social scientists that has been around for a very long time.[392] Although a decision tree can have all three layers of transparency (e.g., simulatability, decomposability, and algorithmic transparency), not all decision trees have stimulability and decomposability properties. The main determinant of a decision tree's transparency level is its size. A simple decision tree with a small number of edges and nodes can be recreated by humans with simple mathematical knowledge. Such a decision tree is regarded as simulatable. As the number of nodes and edges increases, it becomes non-simulatable; however, as long as the node structure is not very complex, it can still be decomposable to its components. When the size reaches a certain point with a complex node-edge structure, the only property of transparency that can be observed in the said decision tree would be algorithmic transparency. Since the inherent algorithmic structure of decision trees are transparent, we can always understand the decision process of a decision tree, which deems it algorithmically transparent in all circumstances. Although decision trees are highly explainable and very intuitive, their inference properties are relatively weaker, and therefore, derivative methods combining multiple decision trees (e.g., random forest, gradient boosting, tree ensemble models) are developed. Although these models perform much better in terms of accuracy, their explainability is reduced to a level where only post-hoc explainability techniques can be used.[393]

---

[388] Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2010). An Introduction to Logistic Regression Analysis and Reporting, *96*(1), 4. https://doi.org/10.1080/00220670209598786.

[389] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 13.

[390] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 14.

[391] Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*, *27*(3), 223. https://doi.org/10.1016/S0020-7373(87)80053-6.

[392] Utgoff, P. E. (1989). Incremental Induction of Decision Trees. *Machine Learning 1989 4:2*, *4*(2), 162. https://doi.org/10.1023/A:1022699900025.

[393] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, *14-15*.

K-Nearest Neighbors: K-Nearest Neighbors (KNN) model relies on a simple yet effective inference logic to make predictions. A given observation is placed in a multi-dimensional space and compared with the previously recorded observations. To make a prediction on this observation, it takes the averages of a number of closest historical observations and the predictions about them. This number is defined by the developer, and the letter K is used to represent it. The developer can change the K value to create the best-performing version of the model. While the most common label is selected for classification problems, the average of the true values is calculated for the regression problems. The reliability of KNN predictions depends on the distance and the similarity between the observation in question and the selected neighbors. A smaller distance between the observation and its neighbors means a closer relationship and lower error margins. When the number of K is small, the model is highly simulatable by humans. On the other hand, as the K grows, its simulatability and decomposability properties disappear, leaving it only algorithmically transparent.[394] However, under any level of complexity, K-nearest neighbors can provide at least some level of transparency and thus, it has been widely accepted in the fields where model interpretability is sought.[395, 396]

Rule-based Systems: Rules based systems usually consist of a set of conditional rules that can be expressed in the form of IF … THEN … statements and their more complex combinations. The rule-based systems are perhaps the most transparent systems among other transparent models since the algorithm contains well-defined rulesets. In fact, many post-hoc explainability techniques' main strategy is creating rule-based systems to generate explanations for the predictions made by black-box systems.[397, 398] However, just as with the other transparent models, there is a trade-off between the complexity and the explainability properties of rule-based systems. As the coverage and the specificity of the rules increase, the rule-based systems start to lose their simulatability and decomposability, respectively. [399]

---

[394] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, *16*.

[395] Li, L., Umbach, D. M., Terry, P., & Taylor, J. A. (2004). Application of the GA/KNN method to SELDI proteomics data. *Bioinformatics (Oxford, England)*, *20*(10), 1638–1640. https://doi.org/10.1093/BIOINFORMATICS/BTH098.

[396] Maciej Serda, Becker, F. G., Cleary, M., Team, R. M., Holtermann, H., The, D., Agenda, N., Science, P., Sk, S. K., Hinnebusch, R., Hinnebusch A, R., Rabinovich, I., Olmert, Y., Uld, D. Q. G. L. Q., Ri, W. K. H. U., Lq, V., Frxqwu, W. K. H., Zklfk, E., Edvhg, L. v, …)2004(فاطمی, ح. An kNN Model-based Approach and its Application in Text Categorization. *Uniwersytet Śląski*, *7*(1), 559–570. https://doi.org/10.2/JQUERY.MIN.JS.

[397] Wang, Q., Shen, Y. P., & Chen, Y. W. (2002). Rule extraction from support vector machines. *ESANN*, *28*(2), 106–110. https://doi.org/10.1007/3-540-28803-1_10.

[398] Johansson, U., König, R., & Niklasson, L. (2004). The Truth is In There - Rule Extraction from Opaque Models Using Genetic Programming. *FLAIRS*.

[399] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 16.

Generalized Additive Models: Generalized additive models are generalized linear models where the smooth functions of significant predictor variables are linearly correlated with the response variable. They contain properties of generalized linear models as well as additive models. Generalized additive models are popular methods used in finance[400] and risk management[401] fields. Due to their understandability and explainability properties, they are useful in explaining the relationship between explanatory and response variables. Depending on their level of complexity, they can be simulatable as long as the smooth functions are limited within human cognitive abilities. As the smooth functions and the explanatory variable count increases, the models become less simulatable and decomposable. Finally, the most complex variations of the models can only have algorithmic transparency.[402]

Bayesian Networks or Directed Graphical Models: Bayesian Networks or Directed Graphical Models are probabilistic machine learning models. In a Bayesian network, while nodes represent random variables, the edges encode conditional independent relations between the nodes they are connecting.[403] Bayesian networks can demonstrate the relationship between the explanatory and response variables clearly with the help of edges. Bayesian networks are inherently transparent models whose algorithm is transparent. Additionally, when the model's complexity does not exceed a certain level, they are also decomposable and simulatable. Bayesian networks have been around for a very long time, and it has been a popular choice for applications in numerous fields, such as econometrics, finance, robotics, gaming, and cognitive modeling.[404]

8.2.1.2 Post-hoc Explainability

When an ML model does not meet any of the criteria mentioned above, an external method should be implemented to generate explanations for its decisions. These methods can only be used after the model makes a prediction. While there are several techniques to generate post-hoc explanations, the first criterion to categorize them is their applicability to all models. While some of these methods are applicable to all the ML models, called *model-agnostic techniques*, some of them can only be used for specific models, called *model-specific techniques*. Model agnostic techniques for post hoc explainability are designed to be used for any ML model, and they usually rely on model simplification, feature relevance estimation, and visualization techniques. On the other hand, model-specific post hoc explainability techniques are often grouped as the ones

---

[400] Taylan, P., & Weber, G. (2007). New Approaches to Regression in Financial Mathematics by Additive Models. *System Research and Information Technologies*. http://www.ict.nsc.ru/jct/content/t12n2/Taylan.pdf.

[401] Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, *23*(2), 129. https://doi.org/10.1002/ASMB.658

[402] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, *17*.

[403] Lafferty John, Liu Han, Wasserman Larry. Directed Graphical Models, https://www.stat.cmu.edu/~larry/=sml/DAGs.pdf.

[404] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 16.

that aim to extract information from shallow ML models and the techniques devised for deep learning models. [405]

8.2.1.2.1 Model Agnostic Post-hoc Explainability Methods

Model specific post-hoc explainability techniques are the explainability techniques that can be plugged into any model regardless of their nature. Since they are model agnostic, they are not usually affected by the inner logic of the ML models. Therefore, the complexity of the ML model would not directly be a determinant in using model-agnostic explainability methods. However, when the model architecture is complex, it may follow difficult-to-capture methods, which may influence the performance of the explainability techniques.
These methods often rely on one of several explanation strategies. The most common strategy, among others, is mimicking the patterns and simplifying the decision-making processes of the underlying ML models to generate explanations with these simplified but transparent models. Another strategy is to test the significance of the features by making small changes on a selected feature. Another strategy is to only highlight the decision-making process to provide more information about the features that caused a particular output.[406] In the following sections, we will provide more details on these strategies and briefly cover the explainability techniques that uses one of these strategies.

Explanation by Simplification and Local Explanations: Explanation by simplification is a powerful strategy which aims to mimic the behavior of ML models and generate transparent explainer models that can be used to generate explanations. In the literature, we observe three main branches under this strategy. The first branch of techniques relies on rule extraction techniques. Local Interpretable Model-Agnostic Explanations (LIME) technique is one of the most popular techniques in the literature and it builds locally linear models around the predictions of a black box model to generate explanations.[407] Another popular method G-REX[408] is also based on rule extraction to generate explanations. Finally, CNF (Conjunctive Normal Form) or DNF (Disjunctive Normal Form) techniques are used to extract features for a human-interpretable

---

405 Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 21-23.

406 Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 18.

407 Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 99. https://doi.org/10.48550/arxiv.1602.04938.

408 Konig, R., Johansson, U., & Niklasson, L. (2008). G-REX: A versatile framework for evolutionary data mining. *IEEE International Conference on Data Mining Workshops, ICDM Workshops 2008*, 971–974. https://doi.org/10.1109/ICDMW.2008.117.

model.[409] Another branch of techniques tries to create a transparent model, such as a decision tree with model simplification as a model extraction process. Finally, the last branch of techniques tries to use a simplification approach to audit black-box models. This approach includes comparing black-box risk scoring models and (ii) applying statistical tests to check if the auditing data lacks certain key features.[410]

Feature Relevance Explanations: Feature relevance explainability techniques are developed to describe the inner logic of a black box model by measuring the influence and significance of each feature in a trained model. Along with the simplification approach, the feature relevance approach is one of the most popular approaches in the literature. The techniques in this family usually rely on one of the following approaches: (i) influence functions, (ii) sensitivity-based, (iii) game theory inspired, (iv) saliency-based, (v) interaction-based, and (vi) others. [411] Among all the feature relevance techniques, SHapley Additive exPlanations (SHAP) [412] is perhaps the most popular technique, and it calculates an additive feature importance score for each prediction with significant properties such as local accuracy, missingness, and consistency. Game Theory Inspired methods such as coalitional Game Theory[413] and local gradients models are used to measure the contribution of each feature to the model predictions are part of this approach. These methods try to observe the changes to be made in a feature to observe changes in the outputs. Similar to this approach, several sensitivity analysis techniques try to measure the importance of each feature for decision-making at local and global levels. While sensitivity methods try to create a general overview on the significance of all features, saliency methods (e.g., Automatic STRucture IDentification method (ASTRID)) [414] focus on the attributes that made a significant contribution to a particular output. As a result, they find the most significant features to create a new model whose accuracy cannot be distinguished from the original black box model. On the other hand, influence function-based techniques try to trace the attributes of a particular prediction back to the training data by using the non-trained model architecture, gradients, and Hessian-vector products.[415]

---

[409] Su, G., Wei, D., Varshney, K. R., & Malioutov, D. M. (2016). Interpretable Two-level Boolean Rule Learning for Classification. *CoRR*, *abs/1606.05798*. http://arxiv.org/abs/1606.05798.

[410] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 18.

[411] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 20.

[412] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-December*, 4766–4775. https://doi.org/10.48550/arxiv.1705.07874.

[413] Saad, W., Han, Z., Debbah, M., Hjørungnes, A., & Başar, T. (2009). Coalitional game theory for communication networks. *IEEE Signal Processing Magazine*, *26*(5), 77–97. https://doi.org/10.1109/MSP.2009.000000.

[414] Henelius, A., Puolamäki, K., & Ukkonen, A. (2017). *Interpreting Classifiers through Attribute Interactions in Datasets*. https://doi.org/10.48550/arxiv.1707.07576.

[415] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 20.

Visual Explanations: Model agnostic visual explanation techniques help the end-users with more intuitive and comprehensible means. Since model-agnosticism requires the ability to use any ML model, developing model-agnostic visual explanations pluggable directly into a model is not very easy. Visual explainability techniques often use a layer of feature relevance techniques, and the visualizations are built on top of the output of these techniques. These visualization techniques are usually grouped as (i) Conditional, Dependence, and Shapley Plots, (ii) Sensitivity-Saliency plots, and (iii) Other techniques.[416]

8.2.1.2.2 **Model Specific Post-hoc Explainability Methods**

While model-agnostic explainability techniques can be plugged into different black-box models, they might be limited in their capabilities to generate useful explanations in certain cases. Therefore, there are also a growing number of model-specific explainability techniques that are tailored for particular black-box models. While some of these models focus on shallow ML models that are regarded as black-box, such as tree ensembles, random forests, and multiple classifier systems, as well as support vector machines, the others focus on deep learning models, such as Multilayer Perceptron, Convolutional Neural Networks, and Recurrent Neural Networks.

Tree Ensembles, Gradient Boosting, Random Forests, and Multiple Classifier Systems: Tree ensemble methods are part of traditional machine learning algorithms that are built on top of decision trees. Decision tree algorithms are popular ML algorithms that have been used for decades, and they provide a good level of accuracy performance. In addition, decision tree algorithms are simulatable, decomposable, and algorithmically transparent, which makes them ideal transparent models. On the other hand, decision trees tend to overfit, and they still have room for accuracy improvements. When multiple decision trees are stacked or combined, they constitute a powerful model with a high level of accuracy, but also their explainability property reduces with the aggregation operations used to combine multiple decision trees. Therefore, to be able to explain the outputs of tree ensembles, we need to rely on post-hoc explainability techniques. While model-agnostic explainability techniques are still pluggable to tree ensemble models, there are several model-specific explainability techniques tailored to tree ensemble models. These models can be grouped under explanation by simplification and feature relevance techniques. [417]

In the explanation by simplification branch, an approach used by Simplified Tree Ensemble Learner (STEL)[418] is to simply train a less complex and more explainable model with the randomly selected observations from the original training data that follow the same probability distribution. Another approach is to train one simple and one complex model and use the simple model to interpret the outputs of the more complex and accurate model. In the feature relevance family, one approach is to measure the

---

[416] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 20.

[417] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 21.

[418] Deng, H. (2019). Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, *7*(4), 277–287. https://doi.org/10.1007/s41060-018-0144-8.

Mean Decrease Accuracy (MDA) and Mean Increase Error (MIE)[419] scores of the random forest performance when a particular feature is changed. Another approach is to develop a framework that can provide with the information to change an observation's label from one class to another. Both these approaches try to measure the significance of features for the labels, which provides an opportunity for explainability. Finally, a number of other approaches are developed to address explainability in tree ensemble methods. Stacking With Auxiliary Features (SWAF)[420] aims to generate and integrate explanations in ensembles, while DeepSHAP stack tree ensembles with deep learning models to create explanation maps and improve the overall explainability of the AI systems.[421]

Support Vector Machines: Support Vector Machines (SVM) are perhaps the highest-performing traditional machine learning models. Regardless of the nature of the problem (e.g., classification, regression, and anomaly detection), SVMs generate hyperplane(s) in a high-dimensional space. The higher distance between the training-data point translates to lower error margins. Prior to deep learning algorithms, SVMs were regarded as the top-performing ML models, and they are still used in several use cases, such as when the computing power is limited. However, this high-performing ML model also has the lowest explainability property among the other traditional ML models. Therefore, several explainability techniques are tailored specifically for SVMs to increase their explainability property. These methods usually follow one of the following approaches: (i) simplification, (ii) local explanations, (iii) visualizations, and (iv) explanations by example.

SVM specific post-hoc explanations can be generated in a number of methods. One of these methods is to create rule-based models or extract fuzzy or eclectic rules designed specifically for SVMs by using support vectors. Another method is to create hyper-rectangles from the intersection between support vectors and the hyperplane(s) to create rules. The third method is using training data as a component to create rules. This method is used in several proposed techniques (e.g., Hyper-rectangle Rule Extraction)[422] to find specific prototype vectors for each class and define hyper-rectangles around these classes, which then can be used for generating explanations. Furthermore, Bayesian systems can be used in combination with SVMs where Bayesian systems are used to generate explanations for the decisions made by SVMs. Finally, a number of

---

[419] Tran, T. T., & Draheim, S. (2020). *Explainability vs . Interpretability and Methods for Models' Improvement*.

[420] Rajani, N. F., & Mooney, R. J. (2016). Stacking With Auxiliary Features. *IJCAI International Joint Conference on Artificial Intelligence*, *0*, 2634–2640. https://doi.org/10.48550/arxiv.1605.08764.

[421] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 21.

[422] Barakat, N. H., & Bradley, A. P. (2007). Rule extraction from support vector machines: A sequential covering approach. *IEEE Transactions on Knowledge and Data Engineering*, *19*(6), 729–741. https://doi.org/10.1109/TKDE.2007.190610.

visualization techniques tailored for SVMs can be used to generate explanations for the outputs of these SVMs. [423]

Multi-layer Perceptron: Multilayer Perceptrons, or multi-layer neural networks, had a groundbreaking effect in the world of artificial intelligence. With the abundance of data, high computing power, and proper technical expertise, they tend to outperform all traditional machine learning models in predictive analytics cases. However, especially as the number of hidden layers increases, a deep learning model's explainability comes under scrutiny. Therefore, in addition to model agnostic post-hoc explainability techniques, researchers proposed several model-specific explainability techniques tailored for multi-layer perceptron following different approaches such as model simplification, feature relevance, text or visual explanations, and local explanations. While interpreting a single perceptron is quite easy, the multi-layer architecture makes the relationship more abstract. [424] The DeepRED[425] algorithm decomposes a multi-layer perceptron into single Perceptrons to extract rules using multiple decision trees and rules. Interpretable Mimic Learning uses gradient boosting trees to generate explainable models from Multi-layer Perceptrons. DeepLIFT[426] method relies on feature relevance and computes importance scores in a multi-layer perceptron. Finally, several model-specific techniques, such as PatternNet and PatternAttribution[427], are proposed to generate theoretically sound explanations from Multi-layer Perceptrons.[428]

Convolutional Neural Networks: Convolutional neural networks are powerful network that dominates the world of computer vision today. In most computer vision problems such as image recognition & classification, image generation, video processing, and other similar tasks, convolutional layers are always present, and researchers proposed several model-specific explainability techniques to generate explanations for the computer vision problems. Although there are several techniques, in terms of their purpose, they can be grouped under two categories. The first group consists of the techniques that trace the output back to its input space to discover the parts of the input image that was significant for the output such as Gradient-weighted Class Activation

---

[423] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 22-23.

[424] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 23-24.

[425] Zilke, J. R., Mencía, E. L., & Janssen, F. (2016). DeepRED - Rule Extraction from Deep Neural Networks. *Discovery Science*, *9956 LNAI*, 457–473. https://doi.org/10.1007/978-3-319-46307-0_29.

[426] Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. *Proceedings of the 34th International Conference on Machine Learning*, 3145–3153. https://proceedings.mlr.press/v70/shrikumar17a.html.

[427] Kindermans, P. J., Schütt, K. T., Alber, M., Müller, K. R., Erhan, D., Kim, B., & Dähne, S. (2017, October 24). Learning how to explain neural networks: PatternNet and PatternAttribution. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1705.05598.

[428] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 23-24.

Mapping (Grad-CAM).[429] The second group consists of the techniques that are developed to understand the inner logic of the network and the inferences taking place in the intermediate layers. These methods can adopt several approaches such as explanation by simplification, feature relevance, visual explanations, and architecture modification. In addition to visual explanations, some of them generate text explanations.

Recurrent Neural Networks: Recurrent Neural Networks often used on time-series and sequential data for especially predictive analytics applications. In addition, they are often used for NLP tasks since natural language sentences work in sequences as well. In terms of their methods, the model-specific explainability techniques for Recurrent Neural Networks can be grouped under two categories: (1) by understanding the inferences of the RNN model, (ii) by modifying RNN architecture to provide insights about individual decisions. RETAIN (REverse Time AttentIoN)[430] model uses a two-level neural attention model to detect past patterns. An RNN equipped with SISTA (Sequential Iterative Soft-Thresholding Algorithm)[431] can have a higher explainability thanks to sequence of correlated observations with a sequence of sparse latent vectors. On the other hand, an RNN combined with Hidden Markov Model (HMM)[432] can have accuracy of the RNN while explainability of HMM.[433]

8.2.2 The Scope-based Categorization

The scope-based categorization is another criterion for grouping explainability techniques.[434] In an AI system, we might need to obtain two types of explanations. The first one would be about the entire decision-making process and the overall rules applicable to every prediction. These explanations are global explanations, which are also covered under GDPR Art. 13-15. In addition to the global explanations, we might also generate explanations for a particular decision. This explanation would be specific to a particular

---

[429] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7.

[430] Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2016). RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. *Advances in Neural Information Processing Systems*, 3512–3520. https://doi.org/10.48550/arxiv.1608.05745.

[431] Wisdom, S., Powers, T., Pitton, J., & Atlas, L. (2016, November 22). Interpretable Recurrent Neural Networks Using Sequential Sparse Recovery. *Neural Information Processing Systems*. https://doi.org/10.48550/arxiv.1611.07252.

[432] Krakovna, V., & Doshi-Velez, F. (2016, June 16). Increasing the Interpretability of Recurrent Neural Networks Using Hidden Markov Models. *International Conference on Machine Learning*. https://doi.org/10.48550/arxiv.1606.05320.

[433] Barredo Arrieta and Díaz-Rodríguez, Explainable AI, 27.

[434] Das, A., & Rad, P. (2020). *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*. https://doi.org/10.48550/arxiv.2006.11371.

use case and a particular individual. Therefore, it may not contain generally applicable rules. GDPR Recital 71 regulates a safeguard that requires data controllers to provide local explanations.[435] Therefore, the explanations in both scopes can be crucial under certain circumstances and, furthermore, required for legal compliance.

8.2.3 The Problem Type-based Categorization

Depending on the problem that the AI system is tackling, the explanations can vary. In classification problems, the model tries to predict an output among a number of categories accurately. Therefore, the explanations are likely to focus on distinguishing observation from the members of the different categories. On the other hand, in a regression problem, the explanation focuses on the contributor of numerical output and tries to measure the effect of the features on a particular output.[436]

8.2.4 The Input-based Categorization

The input type is an important criterion for the Explainability methods since in many cases, the applicable methods will be determined based on the input type. Although there can be other input types, the most common input types for the machine learning algorithms are (i) numerical/categorical inputs, (ii) image inputs, (iii) textual inputs, and (iv) time-series inputs.[437]

Numerical/Categorical Inputs: Numerical/categorical inputs are most common inputs of the machine learning algorithms. For the development of expert systems and decision support systems, numerical/categorical inputs are the main source of input observations.[438]

Image Inputs: Image inputs are usually used in computer vision problems such as object detection or image generation. In some cases, in hybrid problems, image inputs can be used to generate textual outputs.[439]

Textual Inputs: Textual inputs are usually used in Natural Language Processing (NLP) and Natural Language Understanding (NLU) problems. Models solving NLP and NLU problems such as algorithmic translation or sentiment analysis require textual inputs.[440]

---

[435] Dam, H. K., Tran, T., & Ghose, A. (2018). Explainable Software Analytics. *Proceedings - International Conference on Software Engineering*, 53–56. https://doi.org/10.48550/arxiv.1802.00603.

[436] Vilone & Longo, *Classification of XAI Methods*, 6.

[437] Vilone & Longo, *Classification of XAI Methods*, 3.

[438] Vilone & Longo, *Classification of XAI Methods*, 3.

[439] Vilone & Longo, *Classification of XAI Methods*, 3.

[440] Vilone & Longo, *Classification of XAI Methods*, 3.

Time-Series Inputs: Time series are sequential set of numerical or categorical data that are measured in a particular time interval in a predetermined time periods. Data scientists often use the relationship between each observation in the sequences to predict the trends and the future movement of the sequences.

8.2.5 The Output-based Categorization

One of the most important criteria for categorizing explanations is the output format. In literature, for taxonomy creation purposes, especially post-hoc explainability methods are usually categorized based on the output format, then further categorized by the utilized methods (e.g., sensitivity analysis, and finally by the input type (e.g., textual inputs).[441]

Numerical Outputs: Several explainability techniques aims to measure the contribution of an input variable with quantitative metrics. Shapley Additive Explanations (SHAP) combines input features in a linear manner to create an approximation of the underlying model. This method is called additive feature attribution and a variation of SHAP, namely TreeExplainer, also take advantage of additive feature attribution for model approximation. While SHAP assumes independence of the features, TreeExplainer may also be effective of interacting features.[442]

Additionally, various techniques rely on input perturbation and change the observed input values to cause a change in the prediction. Global Sensitivity Analysis (GSA), Feature Importance, and Feature Perturbation are some of the methods that relies on individual or group-level input perturbation to generate numerical explanations.[443]

Rule Outputs: When using black box models, generating human understandable rulesets are usually impossible. However, these rulesets can be extremely important to comprehending how AI systems make decisions. Therefore, some post-hoc explainability techniques aim to generate rulesets or simplified models so that users can understand the underlying, more complex black box models' inner logic. For instance, Genetic Rule Extraction (G-REX)[444] uses genetic algorithms to generate conditional rules (IF-THEN) with mathematical operators (e.g., AND/OR). The rules that G-REX extracts contain fact-based reasoning for the predictions of the underlying model and suggest a set of counterfactual events with a list of changes to the variables that lead to a different outcome. Finally, the general reasoning combined with counterfactual events creates a hierarchical ruleset covering the entire space of possibilities, which is regarded as a global explanation of the underlying model. [445]

While G-REX creates global explanations using genetic algorithms, GLocalX employs genetic algorithms to generate local explanations aiming for specific predictions

---

[441] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 12.
[442] Vilone & Longo, *Classification of XAI Methods*, 4.
[443] Vilone & Longo, *Classification of XAI Methods*, 6.
[444] Konig & Johansson, *G-REX*, 972.
[445] Vilone & Longo, *Classification of XAI Methods*, 11.

outputted by a model. While both G-REX and GLocalX[446] are model agnostic and applicable to all models, there are several model-specific explanation techniques tailored for specific models. We can group these techniques into three classes: (i) decompositional techniques, (ii) pedagogical techniques, and (iii) eclectic techniques.

Decompositional techniques -as the name suggests- decompose the components of a specific model algorithm and generate a set of rules or sub-models, which can successfully mimic the behavior of the underlying model. Neural Network Knowledge Extraction (NNKX)[447] generates binary decision trees from multi-layered feed-forward neural networks by grouping the activation values of the last layer and propagating them back to the input to generate clusters. Discretizing Hidden Unit Activation Values by Clustering, Validity Interval Analysis (VIA), and Discretized Interpretable Multi-Layer Perceptrons (DIMLPs) are examples of model-specific decompositional explanation techniques outputting rulesets.[448]

Pedagogical techniques rely on creating sets of rules and test them with empirical observations. When they receive positive feedback, they keep the rules whereas negative feedbacks results with the replacement of the rule. By testing the potential rulesets on many observations, the techniques cover the entire input space and creates a comprehensive ruleset. For instance, Rule Extraction from Neural Network Ensemble (REFNE)[449] extracts symbolic rules from neural network ensemble instances. Then, algorithm randomly creates rules and test them with categorical attributes to see if these rules are applicable to all the instances. Other examples of pedagogical techniques are C4.5 Rule-PANE, DecText, TREPAN, and Tree Regularization.[450]

Finally, eclectic models combine the components from both decompositional and pedagogical approach. [451]

Textual Outputs: Textual outputs can be useful when a competent explanation can only be presented in the form of word sequences. For example, denial of a loan application requires proper communication with the applicant that should contain a combination of numerical outputs with their reasoning. Textual explanations may provide information about meaningful information about the algorithmic logic involved by means of semantic mapping from models to symbols.[452]

---

[446] Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX - From Local to Global Explanations of Black Box AI Models. *Artificial Intelligence*, *294*, 103457. https://doi.org/10.1016/J.ARTINT.2021.103457.

[447] Bondarenko, A., Aleksejeva, L., Jumutc, V., & Borisov, A. (2017). Classification Tree Extraction from Trained Artificial Neural Networks. *Procedia Computer Science*, *104*, 556–563. https://doi.org/10.1016/J.PROCS.2017.01.172.

[448] Vilone & Longo, *Classification of XAI Methods*, 12.

[449] Chakraborty, M., Biswas, S. K., & Purkayastha, B. (2022). Rule extraction using ensemble of neural network ensembles. *Cognitive Systems Research*, *75*, 36–52. https://doi.org/10.1016/J.COGSYS.2022.07.004.

[450] Vilone & Longo, *Classification of XAI Methods*, 13.

[451] Vilone & Longo, *Classification of XAI Methods*, 12.

[452] Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 12.

Most-Weighted-Path is an example of explanation techniques that outputs textual explanations. By tracing the neurons in a neural network, starting from the output neuron to the input neurons, Most-Weighted-Path generates textual explanations indicating the most relevant features for predicting an output category. [453] InterpNET[454] relies on activation values of a DNN to generate textual explanations of the classifications done by underlying models and generate sentences containing causal relationships. Most-Weighted-Combination, Maximum-Frequency-Difference, and Mycin are some of the model-specific examples of explanation techniques that create textual explanations.[455]

Visual Outputs: Visual explanations can be powerful tools when generating explanations for computer vision problems and provide valuable information about the model's behavior.[456] Pixel marking after certain decomposition operations can be powerful tools to understand which parts of an image attained most significance for the output. Salient masks, Layer-Wise Relevance Propagation (LRP), Spectral Relevance Analysis (SpRAy), and Middle- Level Feature Relevance (MLFR) are some examples of these techniques.[457]

The underlying logic behind many of these techniques is associated with the dimensionality reduction techniques for human interpretability.[458] In addition to pixel marking on images, there is another group of visual explanation techniques that use plots and graphs. Sensitivity Analysis outputs plots demonstrating local gradients, which can be used to detect the modifications required for changing label predictions. Individual Conditional Expectation (ICE), Partial Importance (PI), and Individual Conditional Importance (ICI) plots can demonstrate how labels react to a value change in features.[459]

Mixed Outputs: Some explainability techniques can combine more than a single output format to offer added value to their users. Many of them combine visuals or textual explanations with numerical explanations to help users truly comprehend the explanations. Functional ANOVA Decomposition can measure the influence of non-additive interactions and present findings in numerical and visual formats. Justification Narratives is a model agnostic technique that can output graphs and textual explanations by mapping crucial values underlying a classification to a semantic space.

---

453 Vilone & Longo, *Classification of XAI Methods*, 18-19.

454 Barratt, S. (2017). InterpNET: Neural Introspection for Interpretable Deep Learning. *KDD Undergraduate Consortium*. https://doi.org/10.48550/arxiv.1710.09511.

455 Vilone & Longo, *Classification of XAI Methods*, 19-20.

456 Knaflic, C. N. (2015). *Storytelling with Data: A Data Visualization Guide for Business Professionals*. Wiley, 151-152. https://www.wiley.com/en-us/Storytelling+with+Data%3A+A+Data+Visualization+Guide+for+Business+Professionals-p-9781119002253.

457 Vilone, G., & Longo, L. (2021). Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction 2021, Vol. 3, Pages 615-661*, *3*(3), 615–661. https://doi.org/10.3390/MAKE3030032.

458 Barredo Arrieta and Díaz-Rodríguez, *Explainable AI*, 12.

459 Vilone & Longo, *Classification of XAI Methods*, 21.

Apart from usual formats, prototype lists, positive and iconic prediction lists, adversarial examples, and misrepresented samples to generate contrastive explanations are some of the less common mixed explanation outputs covered in the literature.[460]

8.3 Explainability Benchmarking during Training, Evaluation, and Hyperparameter Tuning

During the development process of AI systems, developers tend to conduct benchmark analyses on alternative models. Given the time and computational limitations, they try to achieve the maximum level of accuracy. Accuracy, precision, recall, and F1 scores are some of the common performance metrics for classification problems, whereas MSE, RMSE, and MAE are some of the performance metrics applicable to regression problems. With the introduction of the mandatory right to explanation safeguards under GDPR and ethical principles published by international institutions, developers must include explainability measures in their benchmarking analysis.

While there are many explainability techniques today, there is not enough study on how to systematically benchmark these techniques. Several researchers propose the OpenXAI framework for evaluating and benchmarking post-hoc explanation techniques. OpenXAI framework uses three criteria to compare the performance of explainability techniques: (i) faithfulness, (ii) stability, and (iii) fairness.[461]

Faithfulness is a measure of understanding how faithfully a given explanation can mimic the underlying model. There are two categories of faithfulness: (a) ground-truth faithfulness and (b) predictive faithfulness. OpenXAI uses several metrics to calculate Ground-truth faithfulness on the agreement between ground-truth explanations and explanations generated by the state-of-the-art methods. For predictive faithfulness, OpenXAI relies on the Prediction Gap on Important feature perturbation (PGI) and the Prediction Gap on Unimportant feature perturbation (PGU) metrics.[462]

Stability metrics measure how much the explanations change with small perturbations to the input. OpenXAI uses Relative Input Stability (RIS), Relative Representation Stability (RRS), and Relative Output Stability (ROS) to measure the maximum change in explanation due to small perturbation in the inputs, model parameters, and output prediction probabilities.[463]

The fairness criterion measures how the performance metrics vary across different minority and majority groups. Fairness can be measured by averaging the other explainability metrics across different groups. A wider gap among the different groups for faithfulness and stability performances means the existence of more unfairness among these groups. [464]

---

[460] Vilone & Longo, *Classification of XAI Methods*, 30.

[461] Agarwal, C., Saxena, E., Krishna, S., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). *OpenXAI: Towards a Transparent Evaluation of Model Explanations*. https://doi.org/10.48550/arxiv.2206.11104.

[462] Agarwal & Saxena, *OpenXAI*, 6.

[463] Agarwal & Saxena, *OpenXAI*, 6.

[464] Agarwal & Saxena, *OpenXAI*, 6.

8.4 Explanation Interface and Post Deployment Presentation Logic

Achieving pre-modeling and modeling explainability of the AI systems are crucial steps to protect fundamental rights and freedoms; however, what makes these efforts meaningful is the ability to present them to the persons affected by these systems. According to European Commission's AI-HLEG, AI systems should be accessible, have a universal design, and emphasize user-centricity.[465]

In a general sense, accessibility refers to allowing all data subjects from different backgrounds and groups have the access to the same AI capabilities. This access also includes access to explanations. When an AI system can provide proper explanations for the majority while ignoring the ethnic or religious minorities, the accessibility to these AI systems will be limited and unfair. Connected to accessibility, universal design refers to the accessibility of AI systems by the widest possible range of users. User centricity refers to the specialized accessibility for every user. Therefore, providing one-size-fits-all explanation techniques which cannot tailor their outputs for the users cannot be deemed in line with the Trustworthy AI principles.[466]

Trustworthy AI mentions explicability as a necessary principle for the foundation of Trustworthy AI. Although there might be a small difference between explicability and explainability, they are often used interchangeably. According to AI-HLEG, explicability is crucial to building and maintaining users' trust in an AI system. It requires processes to be transparent, important details about the AI systems to be properly communicated, and decisions to be explainable to the extent possible. For block-box systems, it requires other measures of explicability (e.g., traceability, auditability and transparent communication on system capabilities) to be enabled to protect fundamental rights and freedoms.[467] However, considering the advancements in the field of Explainable AI, even black box systems can and should provide explanations.

In addition to the foundation of Trustworthy AI, for its realization, there are four Trustworthy AI that are affected by the explainability component of AI systems, namely, (i) diversity, non-discrimination, and fairness, (ii) accountability, (iii) human oversight, and (iv) transparency.[468] Therefore, explanations should serve one or more of the Trustworthy AI principles, and their communication should be made accordingly. GDPR introduces several safeguards under Art. 13-15, 22, and Recital 71 to protect the data subject's right to explanation and for the realization of Trustworthy AI. These safeguards can be listed as follows: [469]

- The right to obtain information about automated decisions (SG1),
- The right to contest/challenge the automated decision (SG2),
- The right to express one's point of view (SG3),
- The right to obtain human intervention (SG4),
- The right to obtain an explanation of the decision after assessment (SG5).

---

[465] AI HLEG, Trustworthy AI, 18-19

[466] AI HLEG, Trustworthy AI, 19.

[467] AI HLEG, Trustworthy AI, 13.

[468] Yalcin, The *Right to Explanation and Trustworthy AI, 185*.

[469] Yalcin, *The Right to Explanation and Trustworthy AI*, 181.

Based on the Trustworthy AI principles and GDPR requirements, here is the list of players that needs to have access to explanations:

Safe-guard	Data Subject	Processor / Controller Representative	Administrative Bodies	Judiciary
SG1	Always	No	Sometimes	Sometimes
SG2	Always	No	Sometimes	Always
SG3	Always	Sometimes	Sometimes	Sometimes
SG4	Always	Always	Sometimes	Sometimes
SG5	Always	No	Sometimes	Sometimes

Table 4. The Players Who Can Request Explanations for RtE Safeguards

Therefore, data subjects, data controllers and their representatives, and judicial and administrative bodies may require explanations depending on the use case of the explanations. While explanations provided to the data subject should be simplified and easily consumable, explanations provided to the judiciary should be detailed since it will be examined by the matter experts. On the other hand, the explanations provided to the data subjects can contain sensitive information, but the explanations for the same prediction about a data subject should not reveal the same information to an administrative body. Therefore, there are a number of dimensions of the explanations that require special attention. These points support the AI-HLEG's recommendation on avoiding One-Size-Fits-All solutions as different players need different explanations.[470]

8.4.1 Presentation Logic

When explainability techniques generate an explanation, the AI systems should enable an outlet (i.e., user interface) to communicate this explanation. The explanations provided in a user interface should not be just plain information. Instead, they should provide understandable, content-specific, and adaptable explanations depending on the user interacting with the interface. To achieve these goals, a presentation logic should be employed to customize the raw explanations flowing from the model. Therefore, a presentation logic to convert plain explanation data into meaningful information for the specific person for a specific purpose is necessary. Only with this approach can a judge obtain the explanation they need to proceed with the judicial procedure, a data subject can understand why they have been subject to an automated decision, or a representative of the data controller can review the automated decision affecting a data subject.[471]

---

[470] Sovrano, F., Vitali, F., & Palmirani, M. (2021). Making Things Explainable vs Explaining: Requirements and Challenges Under the GDPR. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13048 LNAI*, 5-6. https://doi.org/10.1007/978-3-030-89811-3_12/FIGURES/2.

[471] Sovrano & Vitali, *Making Things Explainable*, 7-8.

8.4.2 User Interface

The outputted explanation from the presentation logic would be sent to a user interface so that the authorized person can view the customized explanations. The format of the user interface can be in different formats, from an online or mobile platform to an offline outlet. In fact, explanations can even be provided in a physical copy. However, the procedure to obtain an explanation should not be cumbersome since such procedures will harm the accessibility of the explanation. In addition, a user interface allows users to interact with the system, which would contribute to the overall explainability score of the AI systems.

8.5 Management Level Contribution to Explainability

Apart from the development of an individual AI system, there are various occasions, necessities, and opportunities that can be used to strengthen the explainability ecosystem. They go beyond the development stages of individual AI systems and often create effects at the sector, national, or international level. From lower to higher levels, explainability audits collaborative R&D efforts, and cooperated development of policies are some of these contributors to the overall ecosystem.

8.5.1 Explainability Audits

Explainability requires a multi-disciplinary approach for competency. It protects several ethical principles. The details of the explanations to be provided when there is an automated decision maker are specified under specific laws (GDPR in the European Union). There are several metrics introduced by the XAI researchers to measure the explainability performance of the models. Therefore, there are various issues to be checked for compliance and performance, and conducting systematic audits is an important step to ensure the protection of fundamental rights and freedoms along with Trustworthy AI principles.

From a technical perspective, explainability auditing can be used to measure the status of the system explainability and the benefit of the explainability techniques for a system. From the technical perspective, there are several components that adds up to the overall explainability of the system. Functional explainability measures the transparency of the model and whether it can provide global and local explanations. Faithfulness -as mentioned above- refers to the system's reliability and trustworthiness. In other words, it refers to whether a system can provide reliable explanations for the causal chain of the decision process. Interactivity refers to the adaptability and user-centricity of the system. Explainability trade-off refers to the system's loss of accuracy for the sake of higher explainability and whether this is feasible.[472]

---

[472] Langer, M., Baum, K., Hartmann, K., Hessel, S., Speith, T., & Wahl, J. (2021). *Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives*. *Proceedings of the IEEE International Conference on Requirements Engineering*, *2021-September*, 165. https://doi.org/10.1109/REW53955.2021.00030.

From the psychological perspective, several auditing dimensions measure if the explanations provided by the system can fulfil the receiver's needs. Understandability refers to the contribution of the explanations to the receiver's overall perception about the system. Context-dependency refers to the customizability of the explanations for the goal and needs relevant to the context. Usability refers to the ease-of-use for the receiver, which usually aims to measure the usability of the user interface. A complex user interface can make the overall explainability ineffective. Honesty refers to whether the system provides non-deceptive explanations. With the increasing significance of the dark patterns, honesty of the system has become an important dimension to be measured.[473]

From the legal perspective, there are several applicable laws with several articles that should be taken into account for compliance. The first and foremost important law in the EU is GDPR, and apart from the compliance with the right to explanation related articles, compliance with the general data processing principles laid out in Art. 5, 6, 32, and others are an important dimension for legal compliance. In addition, compliance with the Cybersecurity Act and the draft AI Regulation are the other important dimensions of legal compliance. Finally, compliance with the general rules of law and fundamental rights and freedoms are important contributors to legal compliance. [474]

From the ethical perspective, there are various ethical principles that may be subject to auditing. First of all, diversity, non-discrimination, and fairness, accountability, human oversight, and transparency principles are directly related to the explainability of the AI systems. In addition to these principles, auditing for the other Trustworthy AI principles would be important to achieve the trustworthiness of the AI systems. [475]

8.5.2 Collaborative R&D Efforts

In addition to firm-level explainability auditing, national and international level collaborative R&D efforts can create important opportunities for XAI innovation. Explainable AI is an interdisciplinary field that requires expertise from a number of fields to obtain meaningful results. In addition, explainability property is often perceived as an obstacle rather than a facilitator, and in such situations, collaborative R&D efforts supported by governmental organizations can help firms and research institutions to take the leap towards a more advanced state-of-the-art.

For XAI, DARPA has started an initiative with this goal. In 2015, DARPA introduced the Explainable Artificial Intelligence (XAI) Program with the goal to enable end users to better understand, trust, and manage AI systems. Between 2017-2021, 12 teams consisting of US universities collaborating with European counterparts and institutions proposed their approaches and application. While 11 teams focused on the Explainable Learners category, one team was selected to develop the Psychological Models of Explanation. Several machine learning models such as traceable probabilistic models, causal models, and explanation techniques such as state machines derived

---

[473] Langer & Baum, *Explainability Auditing*, 166.
[474] Langer & Baum, *Explainability Auditing*, 166-167.
[475] Langer & Baum, *Explainability Auditing*, 166-167.

from reinforcement learning models, Bayesian teaching, and GAN dissection were explored and developed. Apart from the explainable models and post-hoc explainability techniques, researchers tested the psychological effectiveness of explanations generated by machine learning algorithms.[476]

In the European Union, Horizon 2020 and its successor, Horizon Europe, provides a research program schema for universities and institutional partners in the European Union or associated countries to create consortiums to conduct concerted research in a particular field. In recent years, with projects such as. XAI,[477] NL4XAI,[478] and XMANAI[479] brought several universities and institutional partners to develop explainability solutions that can contribute to the compliance efforts in the European Union.

8.5.3 Cooperated Development of Policies

Cooperated development of policies is an important concept for the realization of the Trustworthy AI in a global scale. In today's world where cloud computing infrastructures are the norm to serve the services of the AI systems, data is a fluid property traveling from jurisdiction to another. A service provider in one country can have servers across the globe in multiple jurisdictions providing its services to people in all jurisdictions. For example, the U.S. based social media services such as Facebook, Instagram, Twitter, YouTube collects data from all over the world, process and store this data in multiple jurisdictions and provide AI-enabled services to the data subjects. It is very likely for them to disregard or overcome the efforts of a single jurisdiction. This reality shows the significance of the cooperated policy development at the international level.

There is already a cooperated policy development efforts in the European Union. Powerful regulations such as General Data Protection Regulation (GDPR) and Cybersecurity Act are some of the examples of successful cooperated policy development efforts. Especially, as a result of the concerted efforts of the European Countries for the enactment of the GDPR, not only European citizens enjoyed a higher protection with the introduction of digital rights such as the right to explanation, but the global standards on data privacy increased and other jurisdictions started to follow the trend. With the enactment of the draft Artificial intelligence Act, the standards will be taken to the next level and will contribute to the global understanding of Trustworthy AI.[480]

In a field where advancements can lead to existential threats, not only at the European level, but at the United Nations level, cooperated AI policies should be developed

---

[476] Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, *2*(4), e61. https://doi.org/10.1002/AIL2.61.

[477] *Xai - Website*. (n.d.). Retrieved August 20, 2022, from https://xai-project.eu/.

[478] *About – NL4XAI – Interactive Natural Language Technology for Explainable Artificial Intelligence*. (n.d.). Retrieved August 20, 2022, from https://nl4xai.eu/about/.

[479] *Explainable Manufacturing Artificial Intelligence | XMANAI Project | Fact Sheet | H2020 | CORDIS | European Commission*. (n.d.). Retrieved August 20, 2022, from https://cordis.europa.eu/project/id/957362.

[480] *Artificial intelligence act*. (2022, January 14). Think Tank - European Parliament. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792.

and explainability of the AI systems can be prioritized since it will contribute to the understandability of these systems' inner logic, which may prevent major issues for individuals, communities, and even civilizations.[481]

8.6 Final Remarks

In this chapter, we have dived into the explainability techniques that are applicable in the different stages of the AI system development, deployment, use. As pointed out several times, the main focus of the Explainable AI research has been leaning towards ante-hoc model explainability and to opening the black-box with post-hoc explainability techniques. However, the overall AI explainability cannot be achieved with model explainability since model explainability is rather limited with the interpretability property of the AI models.

The overall explainability of an AI system include components from the pre-modeling explainability where the explanations for the data used to train the model are generated. Additionally, after the model selection, adopting standards to compare the explainability level of different models in standardized benchmark analysis is another important step to strengthen the overall AI system explainability. Creating a presentation logic that is adaptable and customizable for the need of the explanation receiver and creating an interactive and user-centric explanation interface are two other important components that make explainability practically meaningful for the society at both individual and collective levels. Finally, the management level explainability policies and measures can ensure the sustainable development of Explainable AI systems.

9 Designing GDPR-Compliant and Trustworthy XAI Systems

9.1 Introduction

In this Chapter, we will combine our findings on the right to explanation, explainability, and Trustworthy AI that are scattered in several fields including law, ethics, psychology, and data science. By considering these fields that are distilled from this wide background of relevant fields, we will propose recommendations based on what we covered in the previous chapters that can serve the needs of the stakeholders in the field of Explainable AI. Only by adopting this interdisciplinary, pragmatic, and comparative approach, we can present our recommendations that are applicable to develop GDPR-compliant and Trustworthy XAI systems.

In today's world, the field of data processing and artificial intelligence are two closely associated fields. While the roots of the today's data processing activities date back to the 19th century, in case where we consider that any type of data recording can be considered data processing, these roots go back to several thousand years. On the

[481] Urban, T. (2015, January 27). *The Artificial Intelligence Revolution: Part 2*. Wait But Why. https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html.

other hand, the field of artificial intelligence is relatively new. We see the initial regression studies dating back to the beginning of the 19th century.[482] In the beginning, the fields of data processing and artificial intelligence did have limited interactions since artificial intelligence was only a theoretical field. However, everything changed with the rise of computers. The advancements in computer technologies made high processing power, cheap data storage, and ease of information exchange available to the masses. With these developments, researchers started to notice the potential of artificial intelligence when a large count of observations is recorded and properly processed. This awareness led to the development of a strong interaction between the fields of data processing and artificial intelligence.

While in the beginning, the outputs of the data processing activities were used for simple tasks for the identification or simple data analysis tasks, Artificial Intelligence applications have become a major use case of these recordings. These applications started to fulfill a wide range of tasks, from simple automation tasks, such as consumer loan applications, to complex systems, such as automated driving. While some of these applications operate in minimal-risk areas, such as recommender systems for video streaming, some of them operate in high-risk areas, such as robot-assisted surgery. Depending on the risk level of these AI applications, the data processing activities should be conducted in a GDPR-compliant manner. While GDPR sets special norms for the processing of special categories of personal data, the general data processing principles are defined under GDPR Art. 5.

With the increasingly wider use of AI applications and close association of these two fields, namely, artificial intelligence and data processing, we see the social and ethical implications of these technologies more prevalent in our everyday lives. These changes vary from labor market movements to discriminatory practices for ethnic and religious minorities. According to a study conducted in 2013 by the Department of Science Engineering of the University of Oxford, 47% of current jobs will be affected by automation and artificial intelligence in a period of 20 years.[483] On the other hand, AI-enabled loan application applications in the U.S. are responsible for around 6% of discriminatory practices against ethnic minorities. Therefore, there are serious ethical and societal consequences of these AI applications, which require technical and organizational measures to maintain the legality of these applications. GDPR Art. 32 is the main article regulating these technical and organizational measures along with a number of other safeguards defined under their relevant articles, which defines a non-exhausting list of measures in line with the principles set out in GDPR Art. 5. These organizational measures certainly require the availability of information on input, output, and certainty of data processing activities in the existence of AI systems as well as other intelligibility

---

[482] Stanton, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, *9*(3), 2. https://doi.org/10.1080/10691898.2001.11910537.

[483] Au-Yong-Oliveira, M., Canastro, D., Oliveira, J., Tomás, J., Amorim, S., & Moreira, F. (2019). The role of AI and automation on the future of jobs and the opportunity to change society. *Advances in Intelligent Systems and Computing*, *932*, 348–357. https://doi.org/10.1007/978-3-030-16187-3_34.

query results such as why, when, what if, and so on.[484] Therefore, data processors and controllers are obliged to keep records of their data processing activities and maintain the ongoing confidentiality, integrity, availability, and resilience of processing systems and services.

In addition, GDPR Art. 13-15, Art. 22, and Recital 71 create a legal framework around the right to explanation and define several safeguards, which require the employment of Explainable AI techniques. These safeguards give the data subjects to the right to receive information about the inner logic of the model and specific explanations about a decision that affects them directly or indirectly. They also provide the right to seek human oversight and the right to contest or challenge a decision. Providing information to the data subject, having a domain expert in the loop, or an administrative or judicial body to review the decision require different levels of intricacies and sophistication. Data subjects may require simplified explanations since they are usually less informed about the domain in question. In addition, these explanations can contain any type of personal data without privacy concerns since the personal data is about them. On the other hand, the representative of the data processor or controller may need to see the inner logic of the decision-making process to understand how the system made a particular decision. Furthermore, special categories of the data may be kept hidden from the representative if this data does not play a significant role in the decision in question. The explanation provided must be adaptable for the receiver as well as it should be meaningful. Therefore, an explanation should provide meaningful information about the data and the system to the receiver, not just any information.

Another set of goals aimed at explainability is to provide interactivity, accessibility, and availability of the explanations to these actors. First, the explanations generated by the system should be presented in an interactive and accessible environment. This requirement can be achieved via a user interface that lives on the cloud, but alternative ways can still be GDPR-compliant. The user interface should enable the users to understand the underlying logic and the input behind a decision that affects them directly or indirectly. This interface should also give them the capability to interact with the inputs, variables, and explanations to generate more insights into how they can positively improve themselves. The interface should also provide enough explanations to enable the user to use the safeguards under the right to explanation framework.

Apart from the chronological stages of Explainable AI development, namely, (i) data collection and processing, (ii) AI development, and (iii) development of explanation interface, the managerial level explainability activities can be regarded as the fourth pillar of Explainable AI system development. The goal at this stage is to comply with the industry standards and, possibly, improve the state-of-the-art of Explainable AI. In the European Union, GDPR's right to explanation framework defines the legal standards for the explainability of AI systems. However, "The Contradiction between Big Data innovation and Data Protection" -as one of the identified challenges in the third Chapter- can be observed in this field as well. Therefore, one of the technical goals of Explainable AI is to maintain the same level of model accuracy while increasing the explainability performance of the model. At the managerial level, AI system developers

[484] Wang & Yang, *Designing Explainable AI*, 6.

should contribute to the collaborative R&D and policy-development efforts for sustainable GDPR-compliant innovation in the field of Explainable AI. Finally, organizational measures in the scope of "Ethical and Technical Standards, Guidelines, Laws, and Codes of Conduct" should be closely followed and implemented for better compliance and higher accuracy performances.

In the next sections, we will cover these four areas (see Fig 31) in more detail and list the recommendations to develop truly Explainable Systems that respect GDPR's Right to Explanation Framework, which is crucial in realizing Trustworthy AI.

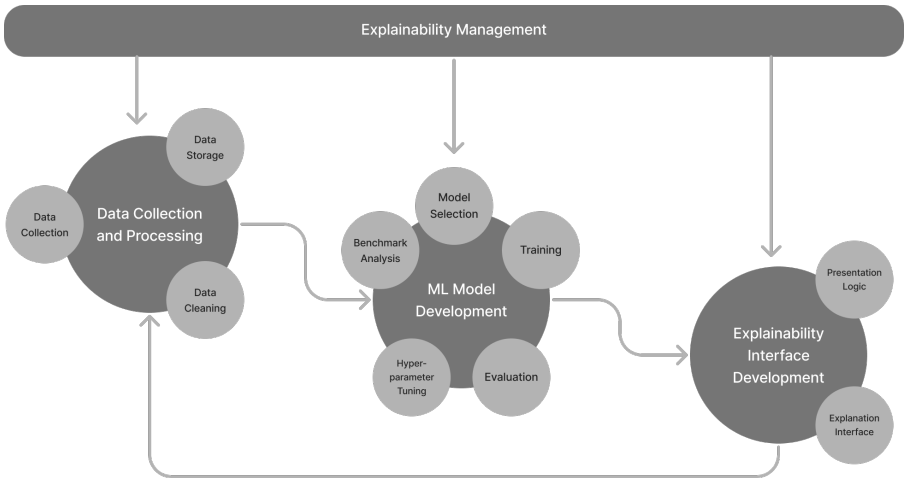


Fig 32. The Explainable AI Development Cycle

9.2 GDPR-Compliant and Trustworthy Data Collection and Processing Activities for XAI Systems

In machine learning studies, data processing is usually observed in the AI system development lifecycle. In its most simple form, data processing is the collection and cleaning of items of data to produce meaningful information. Therefore, data collection, data storage, and data cleaning are the main activities under data processing. In machine learning classifications, data storage activities are usually omitted and handled by big data technologies. Therefore, we often see data collection and data cleaning as the first two stages of the machine learning development cycle. However, to design Explainable AI systems, we will not follow this categorization. Data processing activities -including data storage- have their own legal and ethical principles and should be distinguished when conducting a thorough explainability analysis. Therefore, these activities are grouped into data processing, separately from the machine learning model development.

When data processing activities do not include any personal data, the legal requirements will be less burdensome. In addition, training models without personal data can be surprisingly effective in some top-performing AI systems. For instance, the manufacturing industry is a promising area for AI adoption, and most of the AI systems developed in this field rely only on the sensor data collected by industrial machines. Processing such data would have fewer legal compliance requirements. On the other hand, today's -especially consumer-facing- AI systems mainly rely on processing personal data. The AI systems such as loan application reviewers, CV-sorting software for recruitment procedures, and verification of the authenticity of travel documents operate in the realm of high-risk AI applications. They mainly rely on personal data processing. When processing personal data to be used in these high-risk applications, the explanations generated for such decisions must fulfill a higher level of compliance. Especially when the data collected contain special categories of personal data, the explanations should guarantee that the principles such as "Diversity, Non-discrimination, and Fairness" are not violated.[485] Therefore, the first question we must ask at this point would be the following:

- *Does the processed data contain any personal data? If yes, is any of this processed data part of a special personal data category?*

Since this thesis prioritizes designing Explainable AI systems that are within the scope of the right to explanation, the AI systems that do not rely on personal data and do not focus on the relevant XAI goals will not be prioritized in the recommendations. Therefore, the rest of this Section will focus on personal data processing activities.

When personal data is processed to develop AI systems, the processing should be conducted in line with the principles defined under GDPR Art. 5. The six principles that are explicitly mentioned in GDPR Art. 5 are as follows:

- **Lawfulness, fairness, and transparency:** The personal data should be processed lawfully, fairly, and in a transparent manner
- **Purpose limitation:** The personal data should be collected with a specific, explicit, and legitimate purpose.
- **Data minimization:** The collected personal data should be limited to what is necessary.
- **Accuracy:** The collected personal data should be kept up-to-date and removed or rectified without delay
- **Storage limitation:** The collected personal data should not be kept longer than necessary.
- **Integrity and Confidentiality:** The collected personal data should be kept under appropriate security measures

[485] *Regulatory Framework Proposal on Artificial Intelligence*. (n.d.). European Commission. Retrieved August 31, 2022, from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

Additionally, the second paragraph of Art. 5 requires the controller to be able to demonstrate compliance as an accountability requirement. Therefore, when designing Explainable AI systems, the developers should keep these principles in mind for each main data processing activity we will cover below.

9.2.1 Data Collection

Data collection is the first step of every data processing activity, which can be a physical or automatic collection of data. The data can be provided by the data subject as well as collected by the processor. In either case, the appropriate consent of the data subject should be obtained and kept throughout the continuation of the data processing activities. Therefore, in case one of the right to explanation related safeguards is triggered, the lawfulness of the data collection must be proven by the data controller or the data processor. Additionally, in line with one of the main Trustworthy AI principles, fairness should be also at the heart of the data collection activities. Especially in the event where the data is collected by the data processor with a data collection plan, the processor must ensure that the collected data does not inherently contain bias, which may cause training a discriminatory model. Therefore, the applicability of the transparency principle starts from the data collection step. Data minimization is also one of the principles that fit the best specifically for data collection activities. From the explainability perspective, the collected data should be minimized to the level to provide proper explanations as required by one of the right to explanation safeguards unless another lawful data activity requires more details. Another issue that may damage the explainability at the data collection stage is the integrity of the data. Data collection can be done by multiple persons or systems, and the divergences in collection practices can compromise the integrity of the dataset and can cause accuracy issues. To resolve this problem, there must be standardization policies in effect as an organizational measure.

For compliance at the data collection stage, standardization methods should be applied. Therefore, at the data collection stage, some of the important questions that might be helpful in designing better Explainable AI systems are:

- *Unless more is required for other lawful purposes, are the data collection practices limited to the extent that it is sufficient to generate explanations for the AI system that uses it?*
- *Was the data collection process conducted with fairness in mind? Were there policies to remove the biases in later stages if there were no checks at the data collection stage?*
- *Did the developer adopt standardized data collection policies and best practices to maintain the accuracy and integrity of the dataset that contains* personal data?

9.2.2 Data Storage

Data storage is the next step of the data lifecycle. When the data is collected, it is stored in storage devices that are either owned or leased on the cloud by the controller. While physical storage is relatively secure in terms of data privacy regulations, data is often

kept in the cloud due to ease of access, lower costs, and connectivity capabilities. On the other hand, keeping the collected data in the cloud means that the personal data is transferred to third parties, which are subject to data privacy compliance requirements, which are not within the scope of this thesis. However, we can still strengthen the explainability of AI systems at the data storage level. When data subjects use one of the safeguards and demand explanations, the explanation should include information about the input as well. These inputs will be saved in the data storage stage and should be accessible for a certain period of time. In addition, the integrity of the personal data and the dataset must be always ensured, which may be necessary to generate explanations using explainability techniques. The format that the data is stored (e.g., Tabular data, NoSQL database), whether to use cloud services, and whether to decentralize storage (i.e., centralized storage vs. decentralized storage) to maintain a scalable and robust data storage infrastructure are important decisions that should be made with the explainability considerations of the overall AI system in mind. To achieve these goals, we need to rely on standardization efforts (i.e., Ethical and Technical Standards, Guidelines, Laws, and Codes of Conduct). Some of the questions that may be helpful in designing Explainable AI Systems that comply with GDPR's right to explanation framework are as follows:

- *Does the developer rely on technical standards to ensure that the data storage activities align with the data processing principles, such as fairness, transparency, integrity, and accuracy?*
- *Is the stored data accessible by the explanation interface to provide information about the inputs directly and enable other integrated systems to generate explanations?*

9.2.3 Data Cleaning

Data cleaning is the step that has an inner data lifecycle with data storage. When the data is collected, it will be stored in its raw form. This raw form may contain several inefficiencies, and data cleaning is the stage where the raw data has been put under several data cleaning and manipulation techniques to generate a more valuable version of the collected data. These data cleaning and manipulation techniques can be applied for different purposes. Data cleaning activities can be conducted to remove the noise in the dataset. The noise refers to the collection of unnecessary and irrelevant information, and the raw data tends to contain a high amount of noise that may be observed differently for different data formats. For image data, noisy data usually refers to fuzzy images. For audio data, background sounds, white noise, or blips can be the source of the noise. For numerical and textual data, missing values, badly formatted data, and the use of incorrect data structure are some examples of noise. Cleaning noise requires relatively simple and harmless tasks, which tend not to distort the integrity of the dataset.

After the initial cleaning of the data, a new version of the data usually replaces the raw version since it contains more value. With this version, the data can be utilized for several applications, such as data analytics applications, IT systems, dashboards, and machine learning models. However, this data may still require additional cleaning and

processing operations. One of the major issues that can cause the violation of the "diversity, non-discrimination, and fairness" principle at the data cleaning stage is the imbalances in the dataset. Imbalanced data refers to the datasets where the target class has an imbalanced distribution of observations. For example, if an imbalanced dataset consisting of employees belonging mainly to one ethnic group is used to train an employee recommender system, this recommender system may continuously recommend candidates in this ethnic group. "Exploratory Data Analysis and Data Summarization" techniques can be effective techniques for detecting imbalances in the hands of talented and informed data scientists and domain experts. For example, plots, graphs, and statistical measures can be effective tools for detecting these imbalances. Once the imbalances are detected, developers can balance the dataset by using equal numbers of observations for each target class, which is one of the traditional and safest methods to deal with imbalances. On the other hand, in some use cases, removing the imbalances may mean that most of the data will be removed, leaving a small amount of data, which is not suitable for training a model. In such cases, developers can rely on "feature engineering" techniques at the data cleaning stage or "discrimination aware learning" at the machine learning model development stage.[486]

Apart from the positive obligation of the data controller to remove any type of discriminatory biases from the dataset, it also has a negative obligation not to manipulate the data to have an adverse effect on the data subjects. For example, data poisoning, a known attack format, can distort the integrity of the overall dataset, which can cause discriminatory decision-making affecting minority groups.[487] Maintaining the reliability of the data is related to two principles, namely, "Lawfulness, fairness, and transparency" and "Accuracy" of the data. To strengthen the explainability at the data cleaning stage, the following questions can be helpful:

- *When was the raw data cleaned, and what methods were used to deal with missing values? If imputation methods were used, were they tested using sensitivity analysis?*
- *Are there imbalances in the dataset? Which methods are used to remedy the imbalances? Are they in line with the data processing principles, particularly with "lawfulness, fairness, and transparency"*
- *Are the data cleaning activities properly documented for accountability analysis?*

---

[486] Ristanoski, G., Liu, W., & Bailey, J. (2013). Discrimination Aware Classification for Imbalanced Datasets. *International Conference on Information and Knowledge Management, Proceedings*, 1529–1532. https://doi.org/10.1145/2505515.2507836

[487] Nelson, B., Barreno, M., Jack, F., Anthony, C., Joseph, D., Rubinstein, B. I. P., Saini, U., Sutton, C., Tygar, J. D., & Xia, K. (2008, April). Exploiting Machine Learning to Subvert Your Spam Filter. *In Proceedings of First USENIX Workshop on Large Scale Exploits and Emergent Threats*.

9.2.4 Final Remarks

Data collection, storage, and cleaning are important steps toward the ML development cycle. Developing standard formats to conduct these operations are important to deal with the accountability issues and maintain the integrity of the dataset. While most research in Explainable AI focuses on model explainability, data explainability is still a very important property of overall AI system explainability.

The methods like (i) exploratory data analysis and data summarization, (ii) feature engineering, (iii) standardization activities (i.e., Ethical and Technical Standards and Guidelines), (iv) linking data can be effective techniques to increase the overall data explainability of the AI systems. Finally, apart from the operation-specific questions, the following questions can be useful to improve the data explainability of the AI systems:

- *Are the data collection, storage, and cleaning operations documented properly for providing explanations, especially to the domain experts and administrative and judicial bodies?*
- *Are the developers followed a standardized method in conducting the data operations?*
- *Are there records of metadata that may be helpful for some Explainable AI goals, such as trustworthiness and causability? Is the processed data in the linked data format?*

9.3 GDPR-Compliant and Trustworthy ML Model Development for XAI Systems

Following GDPR-compliant data collection, storage, and cleaning practices, developers have properly structured data that can be used in several data applications. Data visualization, dashboard development, data analysis, and other traditional IT applications are examples of traditional data applications. Another use case, which is within the scope of this thesis, is developing machine learning models to be used for automated decision-making.

If we list the stages of the ML development lifecycle chronologically, the list starts with the data collection and cleaning stages, which we have already covered in detail. After clearly defining the purpose of the AI application and creating the cleaned data, the following steps would be as follows:

- **Model Selection:** At this stage, developers select a model or multiple models based on the nature of the problem.
- **Training:** At the training stage, the selected model(s) is trained using the processed data.
- **Evaluation and Benchmark Analysis:** At this stage, the variations of a model or multiple models are evaluated and compared with a benchmark analysis.
- **Hyper-parameter Tuning:** In parallel with the evaluation stage, the model parameters are tuned to perform at their peak performance with parameter updates.
- **Making Predictions:** At this stage, the trained model is used to make predictions.

These stages should be regarded as parts of a cyclical process. These stages may have to be done over and over to create the best-performing model. For example, when developing a classifier for a MarTech platform to predict the churn probability of a customer, a developer can rely on several ML algorithms such as linear regression, decision tree, random forest, or multilayer perceptron (i.e., ANN). Developers often train and evaluate multiple models and conduct benchmark analysis to select the best-performing model. Several performance metrics based on aggregated error terms such as RMSE, MAE, and MAPE are used to select the best model for a regression task. However, these metrics only measure the accuracy of the model. While in the past, the focus of the developers was on creating the best-performing model in terms of accuracy, explainability is set to become an important component when comparing the overall performance of a model. In fact, there have been ongoing studies to develop metrics to quantify and measure the explainability of the models.[488] Therefore, one question that might be useful when developing Explainable ML models:

- *Are there any metrics used to measure the explainability of the models during the benchmark analysis, or did the model development entirely rely on the accuracy metrics?*

While the data collection, storage, and cleaning activities are automatically seen within the scope of the field of data processing, machine learning activities may not be considered data processing activities. However, machine learning activities are also data processing activities, and they should comply with the general principles of data processing norms. Therefore, the six principles defined under GDPR Art. 5 also apply to the ML development cycle. One general question we might ask at this point would be:

- *When developing ML models, do the developers take the GDPR Art. 5 principles into account, particularly "lawfulness, fairness, and transparency," "purpose limitation," and "integrity and confidentiality" principles?*

In addition to the GDPR Art. 5 principles, perhaps the more important norms are defined under GDPR Art. 13-15, Art. 22, and Recital 71 creating the right to explanation framework in the European Union. When we combine the wordings of these articles and recital provision, we can identify five distinct safeguards, which are:

- The right to obtain information about automated decisions,
- The right to contest/challenge the automated decision,
- The right to express one's point of view,
- The right to obtain human intervention, and
- The right to obtain an explanation of the decision after assessment.

When developing ML models, we should consider that the data subject who was affected by the model's predictions can trigger one of these safeguards, and therefore, the appropriate explanations must be provided to remain GDPR-compliant. In addition, for the quality of the explanations, we should consider that these safeguards are used to protect the following principles for the realization of Trustworthy AI:

---

[488] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for Explainable AI: Challenges and Prospects*, 11. https://doi.org/10.48550/arxiv.1812.04608.

- Human Agency and Oversight
- Transparency
- Accountability
- Diversity, Non-Discrimination, and Fairness

Therefore, when we analyze the stages of the ML model development, we need to take the above mentioned GDPR safeguards and Trustworthy AI principles into account. In the next sections, we will cover each stage of the ML development cycle and present our recommendations to make them GDPR-compliant, trustworthy, and more explainable.

9.3.1 Model Selection

The model selection stage is the stage where the developers select single or multiple machine learning models to solve a particular AI problem. Depending on the nature of the problem (e.g., regression, classification, clustering, dimensionality reduction, reinforcement learning), there are several potential models that developers can take advantage of. Therefore, usually, a subset of available ML models is selected to be trained. At this stage, the initial model selection is made based on the nature of the problem. Other factors, such as the amount of available data, might be an important factor since training neural networks requires a large amount of data. Additionally, available processing power is also a determinant in selecting neural networks as a potential model for the AI system.

In addition to these standard selection methods, another criterion should be the transparency of the models. While traditional ML models such as rule-based models, linear regressions, decision trees, and graphical models are highly interpretable, models such as ensemble trees, support vector machines, and neural networks are black-box systems whose decisions cannot be explained without external explainability techniques. On the other hand, according to the Draft AI Regulations, the AI applications are set to be split into groups based on their risk levels. Especially when designing high-risk AI applications, the explainability plan for the overall AI system becomes essential. The following questions might be useful to make the model selection stage more explainable:

- *Are the models used in the model selection phase inherently transparent and explainable, allowing ante-hoc explainability?*
- *If the models are not transparent, what types of post-hoc explainability techniques are used to provide explanations?*
- *According to the Draft AI Regulation, is the use case for the AI application part of high-risk or moderate-risk groups? If so, what are the explainability criteria used for selecting the ML model?*

9.3.2 Model Training

After the model(s) are selected, the next stage is to train the model using the processed data. The data is already cleaned earlier at the data cleaning stage. However, the developers still need to conduct the final operations (e.g., the train-test-validation split, data shuffling, feature-label split, standardization & normalization of the dataset) to feed the

data into the model as input variables. After these operations, the data is transformed into a format that the ML models can accept and use for learning. At this stage, one question that might be helpful in designing explainable AI systems are as follows:

- *At the final data operations before the model training, were the data operations, such as train-test-validation splits and data shuffling, conducted with a standardized methodology to mitigate potential bias issues?*

After the data preparation operations, the model is usually trained in a distributed fashion on the cloud, which is more in the realm of general data privacy issues and not in this thesis's direct scope.

9.3.3 Model Evaluation, Hyper-parameter Tuning, and Benchmarking

After the model training stage, developers train several models ready for testing. Assuming the train-test-validation split was done properly, at this stage, a testing dataset is used to evaluate the models' performances. Traditionally, numerical metrics such as MSE, RMSE, MAE, and MAPE are used for regression problems, whereas confusion matrices and metrics such as accuracy, precision, recall, and F1-score are used for classification problems. Developers often create tables to conduct benchmark analysis on these models. For unsupervised learning models, often very well-known problems and datasets are used to test the benchmark these models to the state-of-the-art public models. However, none of these metrics measure the explainability of the models, and therefore, the explainability scores of the models are usually not part of benchmark analysis. To design truly explainable AI systems, the model evaluation, tuning, and benchmarking activities should be also evaluated with explainability metrics. A few questions that may be helpful at this stage are as follows:

- *Is there any metric used during the model evaluation that measures the explainability of the model?*
- *If the model has explainability metrics, how effective are they in satisfying the triggered right-to-explanation safeguards?*
- *How can the explainability metric contribute to realizing Trustworthy AI with its relevant principles?*
- *How much weight is given to the explainability metrics when conducting benchmark analysis? Is this weight in line with the risk level of the AI application?*

9.3.4 Final Remarks

Explainability at the ML development cycle, including the model explainability, is perhaps the most challenging part of designing Explainable AI systems. While at the data level, proper data exploration and standardization techniques can increase the overall explainability of the AI system, at the ML development cycle stages, the explainability tasks might be more difficult due to the negative relationship between model accuracy and model explainability. Developers are put in a position to give up either part of the

accuracy or the explainability of their ML model. However, with the advancements in Explainable AI, the researchers are proposing various explainability techniques to open the black-box ML models without sacrificing their accuracy performances. Therefore, one important question to ask at this point is as follows:

- *Are developers consistently searching for better explainability techniques to catch up with the state-of-the-art in Explainable AI?*

In the next section, we will list our recommendations to develop better explainability interfaces that comply with the GDPR's right to explanation framework.

9.4 Post-Deployment Explainability with Explanation Interface and Presentation Logic

After the data and ML model development stages, the trained model is usually integrated into an interface to provide a particular service that it was designed for. This interface can be in several formats. Apart from Web-based, mobile, and desktop applications, command-line interfaces or interactive notebooks can also be suitable interface examples to provide AI services. These interfaces allow users to interact with the AI systems and handle critical operations such as user authentication and Input/Output (I/O) operations. As providing explanations was set to become an important requirement for AI systems, an explanation interface should also be integrated into these user interfaces to present the generated explanations. One of the fundamental components of Explainable AI systems is an explanation interface that the user can interact with to obtain appropriate explanations. This explanation interface should also take advantage of the other existing capabilities of the system, such as user identification as well as the Input/Output mechanism.

Since different data privacy actors will require different explanations, the user interface should be capable of adapting the explanations. There are several factors that may necessitate the adaptation of an explanation. Perhaps the most important one is the role of the user (e.g., the data subject, controller, or third party). The explanations should be simpler to understand without heavy jargon if they are created for the data subjects, whereas they should be detailed with a long line of justification, which may be traced back to its roots if they are used in a judicial proceeding. Connected with the roles, the level of complexity should also be adaptable based on the user's existing knowledge of the domain in question. One data subject may be an expert in the field and should be able to request more detailed explanations. The explanation interface should be able to address such requests.

The explanations provided by the explanation interface should serve the Explainable AI goals (e.g., trustworthiness, causability, transferability, informativeness, confidence, fairness, accessibility, interactivity, privacy awareness, and compliance). Besides, it should also empower data subjects to use the safeguards defined under GDPR Art. 13-15, Art. 22, and Recital 71. These safeguards also require these explanations to be

adaptable for different data privacy actors for human oversight and judicial and administrative proceedings. To achieve these goals, explanation interfaces should consist of two main components: (i) presentation logic, (ii) user interface.

9.4.1 Presentation Logic

Presentation logic refers to an algorithm that accepts raw explanation material and transforms it into explanations that can easily be consumed by the target user. Although there are several competent explainability techniques today, the information they provide cannot be comprehended by non-expert users. Providing them with a bulk of noisy raw data will not fulfill the GDPR requirements. Therefore, this data should be transformed based on the user profile and sent to the user interface for visualization and proper explanation components. Therefore, presentation logic should be designed in a way to deliver customized explanation insights, and while designing the presentation logic, answering these questions might be helpful:

- *How is the Explainable AI system transform the raw data into meaningful explanations for the user?*
- *Can the presentation logic adapt its explanations based on the user profile or status?*
- *Is the presentation logic designed to fulfill the right-to-explanation safeguards defined under GDPR?*
- *Can the presentation logic generate explanations containing the system state (input, output, and certainty) and reasoning queries (e.g., why, what, why not questions)?*

9.4.2 User Interface for Explanations

After the presentation logic algorithm receives raw explanation information on the data and model reasoning and generates customized explanations, the final outputs are presented in a user interface. The user interface for explanation is the point of interaction where users visually see and consume the explanation presented to them. There are several principles that should be considered to maximize the user experience. First, accessibility should be at the heart of designing user interfaces. In a broader sense, accessibility refers to several issues. The user interface should be reachable by the users. Hiding the explanation interface in the depths of the application interface will violate the accessibility principle. When the users access the interface to generate the explanations, they should be able to easily interact with the system without having to go through complex operations. Therefore, ease of use should be adopted as one of the main design principles.

The user interface should also allow interactivity. The user should be able to interact with the interface by changing the assumptions and conditions to get a more customized experience, which will increase its overall reasoning capabilities, an important measure for the explainability of the AI system.

- *Is the explanation interface easily accessible by the users (accessibility principle)?*

- *Is the interface designed in a manner that the user can easily navigate through to generate explanations?*
- *Does the interface empower users to interact with the interface to customize their explanations to have better reasoning capabilities (interactivity)?*

9.5 Explainability Management

Explainability management can be described as organized management activities to keep the explainability of the AI system at a certain level with audits, R&D efforts, and policy improvements. AI systems are living beings that require updates in a constant manner with periodic iterations. From minimal viable product status, they are improved and transformed into more complex and mature systems over time. During this transformation, the explainability of the AI system may be overlooked or disregarded by the system developers. Therefore, at the managerial level, there must be continuous efforts to reach and maintain a certain level of explainability. Several organizational and technical measures may be helpful to ensure that the AI Systems remain explainable.

First, standardized explainability audits can be useful to measure the explainability of the AI system at the data, model development, and post-deployment stages. These explainability audits may also be automated to a certain extent where the explainability level of the AI system is quantified. The higher the level of automation is achieved, the more up-to-date and frequent audits will be possible. The relevant questions about the explainability audits are as follows:

- *Are there standardized explainability audits conducted on the AI system? If so, to what extent are these audits automated, how often are they performed, and what is their scope?*
- *Do these audits apply controls for GDPR's right to explanation safeguards? Do they use any control mechanism for Trustworthy AI principles??*

Secondly, at the managerial level, the latest developments and standards in multiple domains must be researched and followed closely. To remain compliant with the GDPR requirements, the latest case law, ethical frameworks, and new explainability techniques should be researched, documented, and implemented to remain compliant. Therefore, some helpful questions here would be:

- *Do the managers have access to multidisciplinary (e.g., legal, ethical, and technical) experts knowledgeable about the latest developments in the field of explainable AI?*
- *Does the management closely follow and implement the newest legal, ethical, and technical standards to remain compliant with the right to explanation requirements defined under GDPR?*

9.6 Final Remarks

In this Chapter, we went over the stages of AI system development to present our assessment checklist to make the AI systems (i) GDPR-compliant, (ii) Trustworthy, and

(iii) Explainable. It is important to specify that the GDPR compliance and Trustworthiness components are within the scope of this thesis to the extent that they contribute to the explainability property of the AI systems. From the data operations at the pre-modeling stage to the ML model training and model benchmarking stages, from the explanation interface development stage to the implementation of explainability policies at the managerial level, 33 questions were listed to improve these three properties of the AI systems. The results of this interdisciplinary analysis (i.e., legal, ethical, psychological, and technical) conducted on the explainability of the AI systems is a practical assessment list. However, this assessment list should neither be regarded as an exhaustive list nor as a single instrument to ensure legal and ethical compliance. Instead, this list should complement the other assessment lists, such as Trustworthy AI Assessment List, and guidelines, and be further improved to obtain a more comprehensive list applicable to specific use cases.

10 Conclusion

The main goal of this thesis was to contribute to the multidisciplinary state of the art in Explainable AI to design GDPR-compliant and Trustworthy Explainable AI systems with the help of organizational and technical measures containing privacy-preserving technologies and explainability techniques. To have a more comprehensive understanding of the development of AI systems, this interdisciplinary analysis started in the field of data processing.

In the first Chapter of the thesis, we examined the general data privacy principles and the actors defined under GDPR. Then, we briefly covered the Trustworthy AI principles dictating the field of data privacy. Then, we analyzed the top European research project aiming at solving today's data privacy challenges with novel technical and organizational measures.

After the analysis of the data processing, we moved on to the field of Explainable AI and shared the justification for the research with a problem definition. In this section, we briefly covered the goals of Explainable AI and had a terminology clarification section. Since we already covered the projects in the EU, we briefly discussed the research projects that have taken place in the United States.

After the introductory Explainable AI section, we conducted a cognitive analysis of machine-generated explanations and discussed how they could be effective tools to help with imperfect human reasoning. Then, we analyzed GDPR's right to explanation framework and identified the safeguards defined under GDPR Art. 13-15, Art. 22, and Recital 71 and their contribution to the Trustworthy AI principles.

Then, to have a better understanding of the machine learning development process, we conducted a light technical analysis of machine learning and deep learning. Following this analysis, we explained why complex models have a black-box nature.

In the next Chapter, we covered the explainability techniques available to strengthen the explainability property of the AI systems starting from the data collection to the post-deployment stages. This analysis also included managerial-level techniques to strengthen the overall explainability of the AI systems.

In the final chapter of the thesis, we combine all the knowledge that came from numerous fields, including cognitive psychology, law, ethics, and data science, to summarize the findings in a multidisciplinary setting. This summary followed a chronological order, which started with the data processing stages, then followed by the ML model development stages, and post-deployment explainability stages. Finally, a managerial-level summary of the findings was included. In other words, this summary section followed the technical stages to develop Explainable AI systems in chronological order and identified the associated GDPR safeguards or principles along with the Trustworthy AI principles to strengthen the legal compliance and trustworthiness of the AI system. Finally, this Chapter also contained several useful questions to form a non-exhaustive assessment list that might be beneficial when designing GDPR-compliant and Trustworthy XAI systems.

11 Bibliography

1. *About – NL4XAI – Interactive Natural Language Technology for Explainable Artificial Intelligence. (n.d.). Retrieved August 20, 2022, from https://nl4xai.eu/about/*
2. *About A4Cloud | Cloud Accountability Project. (n.d.). Retrieved March 19, 2021, from http://a4cloud.eu/about.html*
3. *Adhikari, A., Wenink, E., van der Waa, J., Bouter, C., Tolios, I., & Raaijmakers, S. (2022). Towards FAIR Explainable AI: a standardized ontology for mapping XAI solutions to use cases, explanations, and AI systems. The15th International Conference on PErvasive Technologies Related to Assistive Environments, 562–568. https://doi.org/10.1145/3529190.3535693*
4. *AEGIS. (n.d.). Approach – AEGIS Big Data. Retrieved March 20, 2021, from https://www.aegis-bigdata.eu/approach/*
5. *Agarwal, C., Saxena, E., Krishna, S., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). OpenXAI: Towards a Transparent Evaluation of Model Explanations. https://doi.org/10.48550/arxiv.2206.11104*
6. *Allen, C. (2016). The Path to Self-Sovereign Identity. Life With Alacrity. http://www.lifewithalacrity.com/2016/04/the-path-to-self-sovereign-identity.html*
7. *AlphaGo. (n.d.). Deepmind. Retrieved August 29, 2022, from https://www.deepmind.com/research/highlighted-research/alphago*
8. *Analytics, B. I. G. D. (2019). Big Data Analytics. https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf*
9. *Anchoring Bias. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/anchoring-bias*
10. *Artificial intelligence act. (2022, January 14). Think Tank - European Parliament. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792*
11. *Au-Yong-Oliveira, M., Canastro, D., Oliveira, J., Tomás, J., Amorim, S., & Moreira, F. (2019). The role of AI and automation on the future of jobs and the opportunity to change society. Advances in Intelligent Systems and Computing, 932, 348–357. https://doi.org/10.1007/978-3-030-16187-3_34*
12. *Availability Heuristic. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/availability-heuristic*
13. *Azodi, C. B., Tang, J., & Shiu, S. H. (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends in Genetics, 36(6), 442–455. https://doi.org/10.1016/J.TIG.2020.03.005*
14. *Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., Siebourg-Polster, J., Steiert, B., & Zhang, J. D. (2020). An Introduction to Machine Learning. Clinical Pharmacology and Therapeutics, 107(4), 871–885. https://doi.org/10.1002/CPT.1796*
15. *Ball, A. (2012). Review of Data Management Lifecycle Models. University of Bath, 2. http://opus.bath.ac.uk/Thisversionismadeavailableinaccordancewithpublisherpolicies.*
16. *Barakat, N. H., & Bradley, A. P. (2007). Rule extraction from support vector machines: A sequential covering approach. IEEE Transactions on Knowledge and Data Engineering, 19(6), 729–741. https://doi.org/10.1109/TKDE.2007.190610*
17. *Barratt, S. (2017). InterpNET: Neural Introspection for Interpretable Deep Learning. KDD Undergraduate Consortium. https://doi.org/10.48550/arxiv.1710.09511*
18. *Barredo Arrieta, A., Díaz-Rodríguez, N., del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/J.INFFUS.2019.12.012.*

19. *Bayamlıoğlu, E. (2021). The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called "right to explanation." Regulation & Governance. https://doi.org/10.1111/REGO.12391*
20. *Berg, D. (2007). Bankruptcy prediction by generalized additive models. Applied Stochastic Models in Business and Industry, 23(2), 129–143. https://doi.org/10.1002/ASMB.658*
21. *Bernhard Waltl and Roland Vogl, "Explainable Artificial Intelligence-the New*
22. *Bhattacharya, S. (2021). A Primer on Machine Learning in Subsurface Geosciences (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-030-71768-1*
23. *Bizer, C. (2009). The emerging web of linked data. IEEE Intelligent Systems, 24(5), 87–92. https://doi.org/10.1109/MIS.2009.102*
24. *Bizer, C., Heath, T., Idehen, K., & Berners-Lee, T. (2008). Linked data on the web (LDOW2008). Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08, 1265–1266. https://doi.org/10.1145/1367497.1367760*
25. *Bondarenko, A., Aleksejeva, L., Jumutc, V., & Borisov, A. (2017). Classification Tree Extraction from Trained Artificial Neural Networks. Procedia Computer Science, 104, 556–563. https://doi.org/10.1016/J.PROCS.2017.01.172*
26. *BOOST4.0 CONSORTIUM. (2018). Boost 4.0 | Big Data for Factories. https://boost40.eu/*
27. *Boston Dynamics. (2019). Spot® | Boston Dynamics. In Boston Dynamics. https://www.bostondynamics.com/spot*
28. *Bounded Rationality. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/bounded-rationality*
29. *BPR4GDPR. (n.d.). Innovation Proposal. 48. Retrieved March 20, 2021, from https://www.bpr4gdpr.eu/about/research-description/*
30. *Brief History of Deep Learning from 1943-2019. (2019, November 4). Machine Learning Knowledge. https://machinelearningknowledge.ai/brief-history-of-deep-learning/*
31. *Bryce Goodman and Seth Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"," AI Magazine, 2017, 6, issn: 07384602, doi:10.1609/aimag.v38i3.2741, arXiv: 1606.08813.*
32. *Budig, T., Herrmann, S., Dietz, A., Pandl Supervisor, K., & Sunyaev, A. (n.d.). Trade-offs between Privacy-Preserving and Explainable Machine Learning in Healthcare. Retrieved February 1, 2021, from www.kit.edu*
33. *Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. Source Code for Biology and Medicine, 3(1), 1–8. https://doi.org/10.1186/1751-0473-3-17/TABLES/6*
34. *Caceres, P. (n.d.). The Convolutional Neural Network. In Introduction to Neural Network Models of Cognition. Retrieved August 29, 2022, from https://com-cog-book.github.io/com-cog-book/features/recurrent-net.html*
35. *Caceres, P. (n.d.). The Recurrent Neural Network. In Introduction to Neural Network Models of Cognition. Retrieved August 29, 2022, from https://com-cog-book.github.io/com-cog-book/features/recurrent-net.html*
36. *Cartledge, C. How Many Vs are there in Big Data. http://www.clc-ent.com/TBDE/Docs/vs.pdf*
37. *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, 145. https://doi.org/10.15779/Z38M32N986.*
38. *Chakraborty, M., Biswas, S. K., & Purkayastha, B. (2022). Rule extraction using ensemble of neural network ensembles. Cognitive Systems Research, 75, 36–52. https://doi.org/10.1016/J.COGSYS.2022.07.004*
39. *Chauhan, N. S. (2022, April 26). An Overview of Activation Functions in Deep Learning. https://www.theaidream.com/post/an-overview-of-activation-functions-in-deep-learning*

40. *Chauvin, Y., & Rumelhart, D. E. (1995). Backpropagation: Theory, Architectures, and Applications. Lawrence Erlbaum Associates. https://www.routledge.com/Backpropagation-Theory-Architectures-and-Applications/Chauvin-Rumelhart/p/book/9780805812596*
41. *Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., & Sun, J. (2016). RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. Advances in Neural Information Processing Systems, 3512–3520. https://doi.org/10.48550/arxiv.1608.05745*
42. *Choi, J. I., & Butler, K. R. B. (2019). Secure Multiparty Computation and Trusted Hardware: Examining Adoption Challenges and Opportunities. Security and Communication Networks. https://doi.org/10.1155/2019/1368905*
43. *Cloud Tensor Processing Units (TPUs). (n.d.). Google Cloud. Retrieved August 29, 2022, from https://cloud.google.com/tpu/docs/tpus*
44. *Confirmation Bias. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/confirmation-bias*
45. *Conocimiento, I. de I. del. (2016). Las 7 V del Big data: Características más importantes - IIC. Instituto de Ingeniería Del Conocimiento. https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/*
46. *CONSORTIUM, B. . (n.d.). About Boost 4.0. In Boost 4.0. Retrieved March 19, 2021, from https://boost40.eu/wp-content/uploads/2018/02/boost_leaflet.pdf*
47. *Copeland, B. J. (n.d.). Artificial Intelligence | Definition, Examples, and Applications. Britannica. Retrieved January 5, 2021, from https://www.britannica.com/technology/artificial-intelligence*
48. *Custers, B., La Fors, K., Jozwiak, M., Esther, K., Bachlechner, D., Friedewald, M., & Aguzzi, S. (2018). Lists of Ethical, Legal, Societal and Economic Issues of Big Data Technologies. SSRN Electronic Journal, 19. https://doi.org/10.2139/ssrn.3091018*
49. *Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals and Systems 1989 2:4, 2(4), 303–314. https://doi.org/10.1007/BF02551274*
50. *Da Bormida, M. (2019). Tackling ethical issues in a H2020 Project in the Big Data domain - AEGIS Ethics White Paper, 7-15.*
51. *Dam, H. K., Tran, T., & Ghose, A. (2018). Explainable Software Analytics. Proceedings - International Conference on Software Engineering, 53–56. https://doi.org/10.48550/arxiv.1802.00603*
52. *Miorandi, D., & Rizzardi A. Sticky Policies: A Survey, 3-4. Retrieved February 1, 2021, from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8807248.*
53. *Das, A., & Rad, P. (2020). Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. https://doi.org/10.48550/arxiv.2006.11371*
54. *Data Lifecycle | NNLM. (n.d.). National Library of Medicine. Retrieved January 5, 2021, from https://nnlm.gov/data/thesaurus/data-lifecycle*
55. *Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., & Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. Advances in Neural Information Processing Systems, 27.*
56. *Etzioni, A, Etzioni, O.. Keeping AI Legal (2007). Vand. J. Ent. & Tech. L. XIX, 1. 2, https://perma.cc/UQ7F-7VYX.*
57. *Gunning, David. Explainable Artificial Intelligence (XAI), Technical Report (2017), 10-18, https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.*
58. *Deborah Raji, I., Smart, A., White Google Margaret Mitchell Google Timnit Gebru Google Ben Hutchinson Google Jamila Smith-Loud Google Daniel Theron Google Parker Barnes Google, R. N., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for*

Internal Algorithmic Auditing ACM Reference Format, 1. https://doi.org/10.1145/3351095.3372873

59. *DECODE Tools | cryptography, authentication, anonymisation and data visualization. (n.d.). Retrieved March 20, 2021, from https://tools.decodeproject.eu/*
60. *DEFeND. (n.d.). What is the Defend Project - Defend Project. Retrieved March 20, 2021, from https://www.defendproject.eu/*
61. *Dellas, N. (2019). Initial Specification of BPR4GDPR architecture, 34-36.*
62. *Dickson, M. C., Bosman, A. S., & Malan, K. M. (2022). Hybridised Loss Functions for Improved Neural Network Generalisation. Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST, 405 LNICST, 169–181. https://doi.org/10.1007/978-3-030-93314-2_11*
63. *Dietterich, T. G. (2000). Ensemble methods in machine learning. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1857 LNCS, 1–15. https://doi.org/10.1007/3-540-45014-9_1/COVER*
64. *Domingo-Ferrer, J., & Blanco-Justicia, A. (2020). Privacy-Preserving Technologies. In International Library of Ethics, Law and Technology (Vol. 21, pp. 279–297). Springer Science and Business Media B.V. https://doi.org/10.1007/978-3-030-29053-5_14*
65. *Domingos, P. A few useful things to know about machine learning (2012). Communications of the ACM 55, 10. 3, ISSN: 00010782, DOI:10.1145/2347736.2347755*
66. *Domo. (2018). Data never sleeps 6.0: how much data is generated every minute?, 2. https://web-assets.domo.com/blog/wp-content/uploads/2018/05/18_domo_data-never-sleeps-6verticals.pdf*
67. *Deng, H. (2019). Interpreting tree ensembles with inTrees. International Journal of Data Science and Analytics, 7(4), 277–287. https://doi.org/10.1007/s41060-018-0144-8*
68. *e-SIDES / e-Sides. (n.d.). Retrieved February 11, 2021, from https://e-sides.eu/e-sides-project*
69. *E-SIDES. (n.d.). E-SIDES Project. Retrieved March 20, 2021, from https://e-sides.eu/e-sides-project*
70. *Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a Right to Explanation is Probably Not the Remedy You are Looking for. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2972855.*
71. *Edwards, L., & Veale, M. (2018). Enslaving the Algorithm: From a "right to an Explanation" to a "right to Better Decisions"? IEEE Security and Privacy, 16(3), 46–54. https://doi.org/10.1109/MSP.2018.2701152.*
72. *European Big Data Value Association. (2017). Strategic Research and Innovation Agenda. European Big Data Value, 4(October), 1–106. https://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf*
73. *European Court of Human Rights, Guide on Article 6: Right to a Fair Trial (Criminal Limb), technical report (2013), 32, https://perma.cc/C4XN-AE8N.*
74. *Explainable Manufacturing Artificial Intelligence | XMANAI Project | Fact Sheet | H2020 | CORDIS | European Commission. (n.d.). Retrieved August 20, 2022, from https://cordis.europa.eu/project/id/957362*
75. *FAQ Archive - RestAssured. (n.d.). Retrieved March 19, 2021, from https://restassuredh2020.eu/faq/*
76. *Fernandez, A., Herrera, F., Cordon, O., Jose Del Jesus, M., & Marcelloni, F. (2019). Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? IEEE Computational Intelligence Magazine, 14(1), 69–81. https://doi.org/10.1109/MCI.2018.2881645*
77. *Fishburn, P. C. (1990). Utility Theory and Decision Theory. Utility and Probability, 303–312. https://doi.org/10.1007/978-1-349-20568-4_40*

78. *Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. https://ssrn.com/abstract=3518482*
79. *Floridi, L. (2018). Soft ethics, the governance of the digital and the General Data Protection Regulation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379-380 (2133). https://doi.org/10.1098/rsta.2018.0081*
80. *Foote, K. D. (2021, December 3). A Brief History of Machine Learning. Dataversity. https://www.dataversity.net/a-brief-history-of-machine-learning/*
81. *Frontier in Legal Informatics," Jusletter IT, 2018, 3.*
82. *GDPR Article 15.1.h*
83. *GenoMed4All. (n.d.). About. 2021. Retrieved March 20, 2021, from http://genomed4all.eu/about/*
84. *Gholizadeh, S., & Zhou, N. (2021). Model Explainability in Deep Learning Based Natural Language Processing. https://doi.org/10.48550/arxiv.2106.07410*
85. *Giakoumopoulos, C., Buttarelli, G., & O'Flaherty, M. (2018). Handbook on European data protection law 2018. In Luxembourg: Publications Office of the European Union. https://doi.org/10.2811/58814*
86. *Giles, M. (2018, April). The GANfather: The man who's given machines the gift of imagination | MIT Technology Review. MIT Technology Review. https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/*
87. *Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 315–323. https://proceedings.mlr.press/v15/glorot11a.html*
88. *Goodfellow, I. J., Vinyals, O., & Saxe, A. M. (2014). Qualitatively characterizing neural network optimization problems. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. https://doi.org/10.48550/arxiv.1412.6544*
89. *Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision making and a "right to explanation." AI Magazine. https://doi.org/10.1609/aimag.v38i3.2741.*
90. *Google Books Team, T. (2020). Google Ngram Viewer. https://bit.ly/2LnR4pn*
91. *Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's explainable AI (XAI) program: A retrospective. Applied AI Letters, 2(4), e61. https://doi.org/10.1002/AIL2.61*
92. *Gurney, K. (1997). An Introduction to Neural Networks (1st ed.). CRC Press. https://doi.org/10.1201/9781315273570*
93. *Hamon, R., Junklewitz, H., Malgieri, G., Hert, P. de, Beslay, L., & Sanchez, I. (2021). Impossible explanations?: Beyond explainable AI in the GDPR from a COVID-19 use case scenario. FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 549–559. https://doi.org/10.1145/3442188.3445917.*
94. *Hardesty, L. (2017). Explained: Neural networks | MIT News | Massachusetts Institute of Technology. April 14,. https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414*
95. *Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & B. Scholkopf, "Support vector machines," in IEEE Intelligent Systems and their Applications, vol. 13, no. 4, pp. 18-28, July-Aug. 1998, doi: 10.1109/5254.708428.*
96. *Henelius, A., Puolamäki, K., & Ukkonen, A. (2017). Interpreting Classifiers through Attribute Interactions in Datasets. https://doi.org/10.48550/arxiv.1707.07576*
97. *High-Level Expert Group on Artificial Intelligence (European Commission). (2019). Ethics Guidelines for Trustworthy AI. In European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai*
98. *Hoffman, R. R., & Klein, G. (2017). Explaining Explanation, Part 1: Theoretical Foundations. Undefined, 32(3), 68–73. https://doi.org/10.1109/MIS.2017.54*

99. *Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. https://doi.org/10.48550/arxiv.1812.04608.*
100. *Home - SPECIAL. (n.d.). Retrieved January 30, 2021, from https://www.specialprivacy.eu/*
101. *Hu, B., Tunison, P., Vasu, B., Menon, N., Collins, R., & Hoogs, A. (2021). XAITK: The explainable AI toolkit. Applied AI Letters, 2(4), e40. https://doi.org/10.1002/AIL2.40*
102. *Humphreys, P., & Berkeley, D. (1983). Problem Structuring Calculi and Levels of Knowledge Representation in Decision Making. Advances in Psychology, 16(C), 121–157. https://doi.org/10.1016/S0166-4115(08)62197-4*
103. *Jabbar, H. K., & Khan, R. Z. (2014). Methods to Avoid Over-Fitting and Under-Fitting in Supervised Machine Learning (Comparative Study). Computer Science, Communication and Instrumentation Devices, 163–172. https://doi.org/10.3850/978-981-09-5247-1_017*
104. *Jacovi, A., Swayamdipta, S., Ravfogel, S., Elazar, Y., Choi, Y., & Goldberg, Y. (2021). Contrastive Explanations for Model Interpretability. EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings, 1597–1611. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.120*
105. *Janetzky, P. (2021, August 29). Deep Reinforcement Learning: From SARSA to DDPG and beyond. Towards Data Science. https://towardsdatascience.com/deep-reinforcement-learning-from-sarsa-to-ddpg-and-beyond-458100c2fda8*
106. *Jiahao Chen, "Fair lending needs explainable models for responsible recommen- dation," FATREC 2018, 2018, 2, arXiv: 1809.04684v1.*
107. *Jiarpakdee, J., Tantithamthavorn, C., Dam, H. K., & Grundy, J. (n.d.). A Theory of Explanations. Retrieved August 20, 2022, from https://xai4se.github.io/xai/theory-of-explanations.html*
108. *Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: the global landscape of ethics guidelines, 7. In arXiv.*
109. *Johansson, U., König, R., & Niklasson, L. (2004). The Truth is In There - Rule Extraction from Opaque Models Using Genetic Programming. FLAIRS.*
110. *Joyce, J. (2021). Bayes' Theorem. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2021/entries/bayes-theorem/*
111. *Kacper Sokol and Peter Flach, "Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches," FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, December 2019, 57- 58, doi:10.1145/3351095.3372870, arXiv: 1912.05100, http://arxiv.org/abs/1912.05100%20http://dx.doi.org/10.1145/3351095.3372870.*
112. *Kaminski, M. E., Bertolini, A., Brennan-Marquez, K., Comandé, G., Cushing, M., Helberger, N., van Drunen, M., van Eijk, N., Eskens, S., Malgieri, G., Price, N., Sax, M., & Selbst, A. (2019). The Right to Explanation, Explained. BERKELEY TECHNOLOGY LAW JOURNAL, 34, 189. https://doi.org/10.15779/Z38TD9N83H.*
113. *Kassir, S. (n.d.). Algorithmic Auditing: The Key to Making Machine Learning in the Public Interest. The Business of Government, 1–4. http://www.businessofgovernment.com/sites/default/files/Algorithmic Auditing.pdf*
114. *Kindermans, P. J., Schütt, K. T., Alber, M., Müller, K. R., Erhan, D., Kim, B., & Dähne, S. (2017, October 24). Learning how to explain neural networks: PatternNet and PatternAttribution. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. https://doi.org/10.48550/arxiv.1705.05598*
115. *Klein, B. (2022, July 5). Confusion Matrix in Machine Learning. Python-Course.Eu. https://python-course.eu/machine-learning/confusion-matrix-in-machine-learning.php*
116. *Klein, G. (2018). Explaining explanation, part 3: The causal landscape. IEEE Intelligent Systems, 33(2), 83–88. https://doi.org/10.1109/MIS.2018.022441353*

117. *Konig, R., Johansson, U., & Niklasson, L. (2008). G-REX: A versatile framework for evolutionary data mining. IEEE International Conference on Data Mining Workshops, ICDM Workshops 2008, 971–974. https://doi.org/10.1109/ICDMW.2008.117*
118. *Knaflic, C. N. (2015). Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley. https://www.wiley.com/en-us/Storytelling+with+Data%3A+A+Data+Visualization+Guide+for+Business+Professionals-p-9781119002253*
119. *Krakovna, V., & Doshi-Velez, F. (2016, June 16). Increasing the Interpretability of Recurrent Neural Networks Using Hidden Markov Models. International Conference on Machine Learning. https://doi.org/10.48550/arxiv.1606.05320*
120. *Kramer, O. (2013). K-Nearest Neighbors. In Dimensionality Reduction with Unsuper-vised Nearest Neighbors, 51, 13-14. Springer. https://doi.org/10.1007/978-3-642-38652-7_2*
121. *Kurenkov, A. (2020, September 27). A Brief History of Neural Nets and Deep Learning. Skynet Today. https://www.skynettoday.com/overviews/neural-net-history*
122. *La Fors, K. (n.d.). Human-centric big data governance: responsible ways to innovate privacy-preserving technologies. Retrieved March 20, 2021, from https://e-sides.eu/resources/e-sides-lessons-for-the-responsible-innovation-of-privacy-preserving-technologies-in-the-era-of-ai-karolina-la-fors-e-sides-beyond-privacy-learning-data-ethics-14-nov-2019-brussels*
123. *Langer, M., Baum, K., Hartmann, K., Hessel, S., Speith, T., & Wahl, J. (2021). Explainability Auditing for Intelligent Systems: A Rationale for Multi-Disciplinary Perspectives. Proceedings of the IEEE International Conference on Requirements Engineering, 2021-September, 164–168. https://doi.org/10.1109/REW53955.2021.00030*
124. *Lawson, A. E. (2000). The Generality of Hypothetico-Deductive Reasoning: Making Scientific Thinking Explicit. American Biology Teacher, 62(7), 482–494. https://doi.org/10.2307/4450956*
125. *Lehr, D., & Ohm, P. (2017). Playing with the Data: What Legal Scholars Should Learn about Machine Learning. U.C. Davis Law Review, 51. https://heinonline.org/HOL/Page?handle=hein.journals/davlr51&id=667&div=&collection=*
126. *Li, L., Umbach, D. M., Terry, P., & Taylor, J. A. (2004). Application of the GA/KNN method to SELDI proteomics data. Bioinformatics (Oxford, England), 20(10), 1638–1640. https://doi.org/10.1093/BIOINFORMATICS/BTH098*
127. *Lim, B. Y., Yang, Q., & Abdul, A. M. (2019). Why these Explanations? Selecting Intelligibility Types for Explanation Goals. IUI Workshops. http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-20.pdf*
128. *Lafferty John, Liu Han, Wasserman Larry. Directed Graphical Models, https://www.stat.cmu.edu/~larry/=sml/DAGs.pdf*
129. *Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 2017-December, 4766–4775. https://doi.org/10.48550/arxiv.1705.07874*
130. *Maciej Serda, Becker, F. G., Cleary, M., Team, R. M., Holtermann, H., The, D., Agenda, N., Science, P., Sk, S. K., Hinnebusch, R., Hinnebusch A, R., Rabinovich, I., Olmert, Y., Uld, D. Q. G. L. Q., Ri, W. K. H. U., Lq, V., Frxqwu, W. K. H., Zklfk, E., Edvhg, L. v, … □ ,□□□□□. (2004). An kNN Model-based Approach and its Application in Text Categorization. Uniwersytet Śląski, 7(1), 559–570. https://doi.org/10.2/JQUERY.MIN.JS*
131. *Malgieri, G. (2019). Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. Computer Law & Security Review, 35(5), 105327. https://doi.org/10.1016/J.CLSR.2019.05.002.*
132. *Markopoulos, I. (2020). Industry specific requirements analysis, definition of the vertical E2E data marketplace functionality and use cases definition I, 11. https://trusts-data.eu/*

133. *Martin, Y. S., & Kung, A. (2018). Methods and Tools for GDPR Compliance Through Privacy and Data Protection Engineering. Proceedings - 3rd IEEE European Symposium on Security and Privacy Workshops, EURO S and PW 2018, 108–111. https://doi.org/10.1109/EuroSPW.2018.00021*
134. *MHMD. (2019). My Health My Data. In My Health My Data. http://www.myhealthmydata.eu/*
135. *Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artificial Intelligence, 267, 1–38. https://doi.org/10.48550/arxiv.1706.07269*
136. *Molnar, C., & Dandl, S. (n.d.). Counterfactual Explanations. In Interpretable Machine Learning. Retrieved August 22, 2022, from https://christophm.github.io/interpretable-ml-book/counterfactual.html*
137. Montavon, G., Kauffmann, J., Samek, W., & Müller, K. R. (2022). Explaining the Predictions of Unsupervised Learning Models. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13200 LNAI*, 117–138. https://doi.org/10.1007/978-3-031-04083-2_7/FIGURES/8
138. *MOSAICrOWN. (n.d.). Homepage. Retrieved March 21, 2021, from https://mosaicrown.eu/*
139. *MOSAICrOWN. (n.d.). Research work. MOSAICrOWN. Retrieved March 20, 2021, from https://mosaicrown.eu/the-project/research-work/*
140. *Motivation & Objectives - TRUSTS. (n.d.). Retrieved March 19, 2021, from https://www.trusts-data.eu/motivation-objectives/*
141. *Mühle, A., Grüner, A., Gayvoronskaya, T., & Meinel, C. (2018). A survey on essential components of a self-sovereign identity. In Computer Science Review (Vol. 30, pp. 80–86). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cosrev.2018.10.002*
142. *Murugan, P., & Durairaj, S. (2017). Regularization and Optimization strategies in Deep Convolutional Neural Network. https://doi.org/10.48550/arxiv.1712.04711*
143. *Musketeer. (2020). ABOUT – MUSKETEER. https://musketeer.eu/project/*
144. *Myatt, G. J. (2006). Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining. John Wiley and Sons. https://doi.org/10.1002/0470101024*
145. *Myatt, G. J., & Johnson, W. P. (2014). Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining.*
146. *Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. Proceedings - IEEE Symposium on Security and Privacy, 111–125. https://doi.org/10.1109/SP.2008.33*
147. *Nelson, B., Barreno, M., Jack, F., Anthony, C., Joseph, D., Rubinstein, B. I. P., Saini, U., Sutton, C., Tygar, J. D., & Xia, K. (2008, April). Exploiting Machine Learning to Subvert Your Spam Filter. In Proceedings of First USENIX Workshop on Large Scale Exploits and Emergent Threats.*
148. *Nielsen, F. (2016). Hierarchical Clustering. In Undergraduate Topics in Computer Science. Undergraduate Topics in Computer Science (pp. 195–211). Springer, Cham. https://doi.org/10.1007/978-3-319-21903-5_8*
149. *Nielsen, F. (2016). Partition-Based Clustering with k-Means. In Undergraduate Topics in Computer Science. Undergraduate Topics in Computer Science (pp. 195–211). Springer, Cham. https://doi.org/10.1007/978-3-319-21903-5_8*
150. *Pasquale, F. (2016). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press. https://www.hup.harvard.edu/catalog.php?isbn=9780674970847*
151. *Pearson, S., & Casassa-Mont, M. (2011). Sticky policies: An approach for managing privacy across multiple parties. Computer, 44(9), 60–68, 60. https://doi.org/10.1109/MC.2011.225*
152. *Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2010). An Introduction to Logistic Regression Analysis and Reporting, 96(1), 3–14. https://doi.org/10.1080/00220670209598786*

153. *Petrik, M. (n.d.). Introduction to Machine Learning. In University of New Hampshire. Retrieved August 28, 2022, from https://www.cs.unh.edu/~mpetrik/teaching/intro_ml_17_files/class1.pdf*
154. *Preet Gandhi, Explainable Artificial Intelligence, 2019, accessed November 12, 2019, https://www.kdnuggets.com/2019/01/explainable-ai.html.*
155. *Project – XMANAI. (n.d.). Retrieved August 26, 2022, from https://ai4manufacturing.eu/project/*
156. *Pu, Y., Apel, D. B., Liu, V., & Mitri, H. (2019). Machine learning methods for rockburst prediction-state-of-the-art review. International Journal of Mining Science and Technology, 29(4), 565–570. https://doi.org/10.1016/J.IJMST.2019.06.009*
157. *Quinlan, J. R. (1987). Simplifying decision trees. International Journal of Man-Machine Studies, 27(3), 221–234. https://doi.org/10.1016/S0020-7373(87)80053-6*
158. *Raggett, Dave. (2019). Big Data Value Spaces for Competitiveness of European Connected Smart Factories 4.0.*
159. *Rajani, N. F., & Mooney, R. J. (2016). Stacking With Auxiliary Features. IJCAI International Joint Conference on Artificial Intelligence, 0, 2634–2640. https://doi.org/10.48550/arxiv.1605.08764*
160. *Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. ACM International Conference Proceeding Series, 382. https://doi.org/10.1145/1553374.1553486*
161. *Rawal, A., McCoy, J., Rawat, D., Sadler, B., & Amant, R. (2021). Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives. https://doi.org/10.36227/TECHRXIV.17054396.V1*
162. *Rebala, G., Ravi, A., & Churiwala, S. (2019). An Introduction to Machine Learning. Springer International Publishing. https://doi.org/10.1007/978-3-030-15729-6*
163. *Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. IEEE Access, 8, 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942*
164. *Regulatory Framework Proposal on Artificial Intelligence. (n.d.). European Commission. Retrieved August 31, 2022, from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai*
165. *Representativeness Heuristic. (n.d.). The Decision Lab. Retrieved August 23, 2022, from https://thedecisionlab.com/biases/representativeness-heuristic*
166. *Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session, 97–101. https://doi.org/10.48550/arxiv.1602.04938*
167. *Riccio, G. M., Peduto, A., Iraci, F., Briguglio, L., Sartin, E., Occhipinti, C., Gutiérrez, I., & Natale, D. (2021). The POSEID-ON Blockchain-Based Platform Meets the "Right to Be Forgotten", 14. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3745516*
168. *Ristanoski, G., Liu, W., & Bailey, J. (2013). Discrimination Aware Classification for Imbalanced Datasets. International Conference on Information and Knowledge Management, Proceedings, 1529–1532. https://doi.org/10.1145/2505515.2507836*
169. *Rizzo, A. (2017). MHMD Project Presentation. In My Health My Data, 4. http://www.myhealthmydata.eu/deliverables/D11.2-MHMD-Project-Presentation.pdf.*
170. *Robert P. Bartlett et al., "Consumer Lending Discrimination in the FinTech Era," SSRN Electronic Journal, November 2017, 1, DOI:10 . 2139 / SSRN. 3063448, https://dx.doi.org/10.2139/ssrn.3063448.*
171. *Rodríguez-Doncel, V., Santos, C., Casanovas, P., & Gómez-Pérez, A. (2016). Legal aspects of linked data – The European framework. Computer Law and Security Review, 32(6), 799–813. https://doi.org/10.1016/J.CLSR.2016.07.005*

172. *Rosenthal, M. D. (2000). Striving for Perfection: A Brief History of Advances and Undercounts in the U.S. Census. Government Information Quarterly, 17(2), 193–208, 202. https://doi.org/10.1016/S0740-624X(00)00027-7*
173. *Saad, W., Han, Z., Debbah, M., Hjørungnes, A., & Başar, T. (2009). Coalitional game theory for communication networks. IEEE Signal Processing Magazine, 26(5), 77–97. https://doi.org/10.1109/MSP.2009.000000.*
174. *Sado, F., & Loo, C. (2020). Explainable goal-driven agents and robots-a comprehensive review and new framework. ArXiv. https://www2.informatik.uni-hamburg.de/wtm/publications/2021/SLKW21/2004.09705v1.pdf*
175. *SAE International: Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems: J3016_202104 (2021)*
176. *Samek W, Wiegand T, Müller K-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models (2017). ITU Journal: ICT Discoveries 1, no. Special Issue 1: 2-3, arXiv: 1708.08296, http://arxiv.org/abs/1708.08296.*
177. *Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. Sn Computer Science, 2(5), 377. https://doi.org/10.1007/S42979-021-00765-8*
178. *Sarle, W. S. (2000, June 23). How to measure importance of inputs? SAS. ftp://ftp.sas.com/pub/neural/importance.html*
179. *Sartor, G. (European U. I. of F. (2020). The impact of the General Data Protection Regulation (GDPR) on artificial intelligence. Panel for the Future of Science and Technology (STOA), 1st, 76-79. https://doi.org/10.2861/293*
180. Seeger, M. (2006). A Taxonomy for Semi-Supervised Learning Methods. *Semi-Supervised Learning*, 15–32. https://infoscience.epfl.ch/record/161326
181. *Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. International Data Privacy Law, 7(4), 233–242. https://doi.org/10.1007/s13347-017-0263-5.*
182. *Sekar, M. (2022). Types of Visualizations. In Machine Learning for Auditors (pp. 155–167). Apress. https://doi.org/10.1007/978-1-4842-8051-5_16*
183. *Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision, 128(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7*
184. *Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX - From Local to Global Explanations of Black Box AI Models. Artificial Intelligence, 294, 103457. https://doi.org/10.1016/J.ARTINT.2021.103457*
185. *Sharma, S. (2019). Data Privacy and GDPR Handbook. In Data Privacy and GDPR Handbook. Wiley. https://doi.org/10.1002/9781119594307*
186. *Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. Proceedings of the 34th International Conference on Machine Learning, 3145–3153. https://proceedings.mlr.press/v70/shrikumar17a.html*
187. *Siau, K., & Wang, W. (2020). Artificial intelligence (AI) Ethics: Ethics of AI and ethical AI. Journal of Database Management, 31(2), 74–87. https://doi.org/10.4018/JDM.2020040105.*
188. *Sina Aghaei, Mohammad Javad Azizi, and Phebe Vayanos, "Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making," Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019): 1-3, ISSN: 2159-5399, DOI:10 . 1609/aaai.v33i01.33011418, arXiv: 1903.10598, www.aaai.org.*
189. *SMOOTH. (n.d.). About Smooth Project. Retrieved March 20, 2021, from https://smoothplatform.eu/about-smooth-project/*
190. *Sovrano, F., Vitali, F., & Palmirani, M. (2021). Making Things Explainable vs Explaining: Requirements and Challenges Under the GDPR. Lecture Notes in Computer Science (Including*

Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13048 LNAI, 169–182. https://doi.org/10.1007/978-3-030-89811-3_12/FIGURES/2

191. *Sovrano, F., Vitali, F., & Palmirani, M. (2020). Modelling GDPR-Compliant Explanations for Trustworthy AI. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12394 LNCS, 219–233. https://doi.org/10.1007/978-3-030-58957-8_16*
192. *SPECIAL. (n.d.). Home. Retrieved March 20, 2021, from https://www.specialprivacy.eu/*
193. *Stanton, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. Journal of Statistics Education, 9(3). https://doi.org/10.1080/10691898.2001.11910537*
194. *Steele, K., & Stefánsson, H. O. (Winter 2020). Decision Theory. Stanford Encyclopedia of Philosophy. Retrieved October 29, 2022, from https://plato.stanford.edu/archives/win2020/entries/decision-theory/*
195. *Su, G., Wei, D., Varshney, K. R., & Malioutov, D. M. (2016). Interpretable Two-level Boolean Rule Learning for Classification. CoRR, abs/1606.05798. http://arxiv.org/abs/1606.05798*
196. *Support Vector Machine. (n.d.). Retrieved August 15, 2022, from https://cml.rhul.ac.uk/svm.html*
197. *TAPAS – SESAR ER – Descripción corta del proyecto. (n.d.). Retrieved August 26, 2022, from https://tapas-atm.eu/*
198. *Taylan, P., & Weber, G. (2007). New Approaches to Regression in Financial Mathematics by Additive Models. System Research and Information Technologies. http://www.ict.nsc.ru/jct/content/t12n2/Taylan.pdf*
199. *The five Vs of big data | BBVA. (n.d.). BBVA. Retrieved January 5, 2021, from https://www.bbva.com/en/five-vs-big-data/*
200. *The seven requirements are (i) human agency and oversight, (ii) technical robustness and safety, (iii) privacy and data governance, (iv) transparency, (v) diversity, non-discrimination, and fairness, (vi) societal and environmental wellbeing, and (vii) accountability.*
201. *Timan, T., & Mann, Z. (2021). Data Protection in the Era of Artificial Intelligence: Trends, Existing Solutions and Recommendations for Privacy-Preserving Technologies. The Elements of Big Data Value. https://doi.org/10.1007/978-3-030-68176-0_7*
202. *Tran, T. T., & Draheim, S. (2020). Explainability vs . Interpretability and Methods for Models' Improvement.*
203. *Truex, S., Steinke, T., Baracaldo, N., Ludwig, H., Zhou, Y., Anwar, A., & Zhang, R. (2019). A hybrid approach to privacy-preserving federated learning. Proceedings of the ACM Conference on Computer and Communications Security, 1–11, 1. https://doi.org/10.1145/3338501.3357370*
204. *Turek, Matt, Explainable Artificial Intelligence (XAI) (2016). 1, accessed November 12, 2019, https://www.darpa.mil/program/explainable-artificial-intelligence.*
205. *Understanding the 7 Vs of Big Data. (2016, April 7). Impact. https://impact.com/marketing-intelligence/7-vs-big-data/*
206. *Utgoff, P. E. (1989). Incremental Induction of Decision Trees. Machine Learning 1989 4:2, 4(2), 161–186. https://doi.org/10.1023/A:1022699900025*
207. *Urban, T. (2015, January 27). The Artificial Intelligence Revolution: Part 2. Wait But Why. https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html*
208. *van Otterlo, M., & Wiering, M. (2012). Reinforcement Learning and Markov Decision Processes. In Reinforcement Learning. Adaptation, Learning, and Optimization (Vol. 12, pp. 3–42). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27645-3_1/COVER*
209. *Vashisht, R. (2021, January 6). When to perform a Feature Scaling? Atoti. https://www.atoti.io/articles/when-to-perform-a-feature-scaling/*
210. *Veeningen, M. (2020). SODA - Scalable Oblivious Data Analytics. SODA Project. https://soda-project.eu/*

211. *Vilone, G., & Longo, L. (2020). Explainable Artificial Intelligence: a Systematic Review. https://doi.org/10.48550/arxiv.2006.00093*
212. *Vilone, G., & Longo, L. (2021). Classification of Explainable Artificial Intelligence Methods through Their Output Formats. Machine Learning and Knowledge Extraction 2021, Vol. 3, Pages 615-661, 3(3), 615–661. https://doi.org/10.3390/MAKE3030032*
213. *Voigt, P., Wessing, T., & Bussche, A. (2018). The EU General Data Protection Regulation (GDPR) - A Practical Guide. In Irish Medical Journal (Vol. 111, Issue 5). https://www.springer.com/gp/book/9783319579580*
214. *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, 1–47. https://doi.org/10.2139/SSRN.2903469.*
215. *Walport, M. (2015). Distributed ledger technology: Beyond block chain. Government Office for Science, 1–88, 21. https://youtu.be/4sm5LNqL5j0*
216. *Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. Undefined. https://doi.org/10.1145/3290605.3300831*
217. *Wang, Q., Shen, Y. P., & Chen, Y. W. (2002). Rule extraction from support vector machines. ESANN, 28(2), 106–110. https://doi.org/10.1007/3-540-28803-1_10*
218. Wells, L., & Bednarz, T. (2021). Explainable AI and Reinforcement Learning—A Systematic Review of Current Approaches and Trends. *Frontiers in Artificial Intelligence*, *4*, 48. https://doi.org/10.3389/FRAI.2021.550030/BIBTEX
219. *Wing, J. M. (2019). The Data Life Cycle. Harvard Data Science Review. https://doi.org/10.1162/99608f92.e26845b4*
220. *Winikoff, M., & Sardelic, J. (2021). Artificial Intelligence and the Right to Explanation as a Human Right. IEEE Internet Computing, 25(2), 108–112. https://doi.org/10.1109/MIC.2020.3045821.*
221. *Wirth, N. (2018). Hello marketing, what can artificial intelligence help you with?: International Journal of Market Research, 60(5), 435–438. https://doi.org/10.1177/1470785318776841*
222. *Wisdom, S., Powers, T., Pitton, J., & Atlas, L. (2016, November 22). Interpretable Recurrent Neural Networks Using Sequential Sparse Recovery. Neural Information Processing Systems. https://doi.org/10.48550/arxiv.1611.07252*
223. *Worster, A., Fan, J., & Ismaila, A. (2007). Understanding linear and logistic regression analyses. Canadian Journal of Emergency Medicine, 9(2), 111–113. https://doi.org/10.1017/S1481803500014883*
224. *Wuyts, K. (2014). LINDDUN : a privacy threat analysis framework. https://distrinet.cs.kuleuven.be/software/linddun*
225. *Xai - Website. (n.d.). Retrieved August 20, 2022, from https://xai-project.eu/*
226. *Xai - Website. (n.d.). Retrieved August 26, 2022, from https://xai-project.eu/research-lines.html*
227. *Yalcin, O. G. (2020). Examination of current AI systems within the scope of right to explanation and designing explainable AI systems. CEUR Workshop Proceedings, 2598. https://ceur-ws.org/Vol-2598/paper-11.pdf.*
228. *Yalcin, O. G. (2022). The Role of the Right to Explanation and Its Safeguards in the Realization of Trustworthy AI. New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence, 178–187.*
229. *Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated Machine Learning: Concept and Applications. In ACM Trans. Intell. Syst. Technol (Vol. 10), 12:4. https://doi.org/*
230. *Yod, S. M. (2019). PDP4E - D 2.4 Overall system requirements. https://www.pdp4e-project.eu/deliverables/*

231. *Zilke, J. R., Mencía, E. L., & Janssen, F. (2016). DeepRED - Rule Extraction from Deep Neural Networks. Discovery Science, 9956 LNAI, 457–473. https://doi.org/10.1007/978-3-319-46307-0_29*
232. *Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. Https://Doi.Org/10.2200/S00196ED1V01Y200906AIM006, 6, 1–116. https://doi.org/10.2200/S00196ED1V01Y200906AIM006*
233. *Zingales, N. (2021). Right to Explanation. In Glossary of Platform Law and Policy Terms (pp. 279–281). https://hdl.handle.net/10438/31365.*

APPENDIX I: Checklist for Designing GDPR-Compliant Trustworthy XAI Systems

GDPR-Compliant and Trustworthy Data Collection and Processing Activities for XAI Systems

General Issues

- *Are the data collection, storage, and cleaning operations documented properly for providing explanations, especially to the domain experts and administrative and judicial bodies?*
- *Are the developers followed a standardized method in conducting the data operations?*
- *Are there records of metadata that may be helpful for some Explainable AI goals, such as trustworthiness and causability? Is the processed data in the linked data format?*

Data Collection

- *Does the processed data contain any personal data? If yes, is any of this processed data part of a special personal data category?*
- *Unless more is required for other lawful purposes, are the data collection practices limited to the extent that it is sufficient to generate explanations for the AI system that uses it?*
- *Was the data collection process conducted with fairness in mind? Were there policies to remove the biases in later stages if there were no checks at the data collection stage?*
- *Did the developer adopt standardized data collection policies and best practices to maintain the accuracy and integrity of the dataset that contains personal data?*

Data Storage

- *Does the developer rely on technical standards to ensure that the data storage activities align with the data processing principles, such as fairness, transparency, integrity, and accuracy?*
- *Is the stored data accessible by the explanation interface to provide information about the inputs directly and enable other integrated systems to generate explanations?*

Data Cleaning

- *When was the raw data cleaned, and what methods were used to deal with missing values? If imputation methods were used, were they tested using sensitivity analysis?*
- *Are there imbalances in the dataset? Which methods are used to remedy the imbalances? Are they in line with the data processing principles, particularly with "lawfulness, fairness, and transparency"*
- *Are the data cleaning activities properly documented for accountability analysis?*

GDPR-Compliant and Trustworthy ML Model Development for XAI Systems

General Issues

- *Are there any metrics used to measure the explainability of the models during the benchmark analysis, or did the model development entirely rely on the accuracy metrics?*
- *When developing ML models, do the developers take the GDPR Art. 5 principles into account, particularly "lawfulness, fairness, and transparency," "purpose limitation," and "integrity and confidentiality" principles?*

Model Selection

- *Are the models used in the model selection phase inherently transparent and explainable, allowing ante-hoc explainability?*
- *If the models are not transparent, what types of post-hoc explainability techniques are used to provide explanations?*
- *According to the Draft AI Regulation, is the use case for the AI application part of high-risk or moderate-risk groups? If so, what are the explainability criteria used for selecting the ML model?*

Model Training

- *At the final data operations before the model training, were the data operations, such as train-test-validation splits and data shuffling, conducted with a standardized methodology to mitigate potential bias issues?*

Model Evaluation, Hyper-parameter Tuning, and Benchmarking

- *Is there any metric used during the model evaluation that measures the explainability of the model?*

- *If the model has explainability metrics, how effective are they in satisfying the triggered right-to-explanation safeguards?*
- *How can the explainability metric contribute to realizing Trustworthy AI with its relevant principles?*
- *How much weight is given to the explainability metrics when conducting benchmark analysis? Is this weight in line with the risk level of the AI application?*

After Development

- *Are developers consistently searching for better explainability techniques to catch up with the state-of-the-art in Explainable AI?*

Post-Deployment Explainability with Explanation Interface and Presentation Logic

Presentation Logic

- *How is the Explainable AI system transform the raw data into meaningful explanations for the user?*
- *Can the presentation logic adapt its explanations based on the user profile or status?*
- *Is the presentation logic designed to fulfill the right-to-explanation safeguards defined under GDPR?*
- *Can the presentation logic generate explanations containing the system state (input, output, and certainty) and reasoning explanations (e.g., why, what, why not questions)?*

User Interface

- *Is the explanation interface easily accessible by the users (accessibility principle)?*
- *Is the interface designed in a manner that the user can easily navigate through to generate explanations?*
- *Does the interface empower users to interact with the interface to customize their explanations to have better reasoning capabilities (interactivity)?*

Explainability Management

Explainability Audits

- *Are there standardized explainability audits conducted on the AI system? If so, to what extent are these audits automated, how often are they performed, and what is their scope?*
- *Do these audits apply controls for GDPR's right to explanation safeguards? Do they use any control mechanism for Trustworthy AI principles?*

Co-operated Dev of Policies and R&D Efforts

- *Do the managers have access to multidisciplinary (e.g., legal, ethical, and technical) experts knowledgeable about the latest developments in the field of explainable AI?*
- *Does the management closely follow and implement the newest legal, ethical, and technical standards to remain compliant with the right to explanation requirements defined under GDPR?*