



Ca' Foscari  
University  
of Venice

Master Degree programme  
in Digital and Public Humanities

Final Thesis

Ethical reflections on the Principle of Explainability  
in the GDPR and AI Act

**Supervisor**

Chiar.mo Prof. Alessandro Bernes

**Assistant supervisor**

Chiar.ma Prof.ssa. Fabiana Zollo

**Graduand**

Lia Cimino

Matriculation Number: 898897

**Academic Year**

2023/2024

## **Abstract**

In recent years, the significant increase in the use of big data and artificial intelligence systems has triggered a profound transformation in our society.

To address this phenomenon, various new regulations have been introduced in Europe, including the GDPR and the AI Act, aimed at governing these changes.

This thesis seeks to explore how the principle of "explainability" is regulated and ensured within the framework of European data protection laws, particularly when personal data is processed through automated procedures, and on artificial intelligence systems.

The first chapter examines the right to explanation as regulated by the GDPR (Regulation (EU) No. 2016/679). In this context, the "right to explanation" refers to the user's right to understand the mechanisms through which their personal data is processed. This right must be upheld by companies handling such data and should be safeguarded when users' personal data is processed by artificial intelligence systems, requiring the algorithms involved to be explainable.

The second chapter analyses the right to explainability as governed by the AI Act (EU Artificial Intelligence Act). In this framework, the right to explainability pertains to ensuring that artificial intelligence systems, especially high-risk ones, are transparent and comprehensible to users.

In the third and final chapter of the thesis, I develop an ethical reflection on the principle of explainability by referencing the content of the European regulations previously examined, specifically the AI Act and the GDPR. In particular, I focus on the ethical justifications underlying the principle of explainability and analyses the ethical challenges stemming from the current regulatory framework established by the GDPR and the AI Act.

<b>Introduction</b>	3
<b>Chapter I <u>GDPR and Explainability</u></b>	7
1.1. From the first definition of privacy to the right to data protection	7
1.2. Privacy in Europe: the development of the GDPR	13
1.3. The importance of explainability: the problem of information asymmetry	20
1.4. Explainability	25
1.5. Explainability under GDPR	26
1.5.1. Article 22	26
1.5.2. Articles 13,14,15	33
1.6. The Absence of a Right to Explanations in the GDPR. Another perspective	38
1.7. Application of the "explainability" principle: practical cases	42
<b>Chapter II <u>AI Act and Explainability</u></b>	56
2.1. Artificial Intelligence and the challenge of harmonizing its definition	56
2.2. From Turing to the AI Act: the evolution of artificial intelligence and its regulation	63
2.3. The need for ethical reflection: a focus on explainability and transparency	68
2.4. The AI Act: an introduction to its complexity	73
2.4.1. Overview of the AI Act	73
2.4.2. Structure and purpose of the regulation	75
2.5. Explainability and Transparency	77
2.5.1. Article 13	81
2.5.2. Article 14	85
2.5.3. Article 86	89
2.5.4. Preliminary considerations on the right to explainability under AI Act	91
2.6. European Office for AI	96
2.7. AI Pact	98
<b>Chapter III <u>Ethical reflections on the principle of explainability in the GDPR and AI Act</u></b>	100
3.1. Ethics, society, scientific development, and law: the need to understand to avoid fear	100
3.2. Explainability: reflections on practical ethics	111
3.2.1. Ethical necessity of explainable artificial intelligence: importance of the human-AI relationship of trust	111
3.2.2. Explainability, a tool in service of the human being	112
3.2.3. Explainability and the specific needs of users: explainability as a relative concept	114
3.2.4. Ethical need of transparency of data	118
3.2.5. Conflicting needs: the ethics of explainability and technological and economic development	119
<b>Conclusion</b>	122
<b>References</b>	127

## **Introduction**

In modern times, the rising application of big data and artificial intelligence (AI) technologies has triggered a massive transformation in our society, which has fundamentally changed many aspects of our daily lives, economic systems, and social and political relationships. Today, data is an important asset underlying automated decision-making processes and advanced technologies. The increasing pervasiveness of these technologies presents complex questions related to the protection of individual rights, the demand for transparency, the determination of accountability, and the reliability of intelligent systems. More particularly, the widespread use of artificial intelligence and automation of decision-making presents significant challenges in protecting personal information ("personal data"), where individuals may be affected by algorithms and automated decisions without understanding fully what governs those outcomes.

In response to these concerns, European institutions have enacted concrete rules for the control of the use of such advanced technologies while protecting the fundamental rights of the people. Two of the most important regulations in this field are the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act (AI Act), both of which are aimed at protecting personal data and fostering the ethical use of artificial intelligence, with particular focus on ensuring transparency in decision-making processes and requiring human oversight in automated decisions.

One important notion that figures in both frameworks is that of "explainability," referring to the right of individuals to understand how their personal data is used and the logic followed by automated decision-making. Explainability plays a crucial role,

especially when the opaque decision-making processes and algorithms may influence the individual in ways that are not easy to fathom. The right to explainability goes further than just providing information; it should also include ensuring that people are provided with a clear and understandable description of the mechanisms behind automated processing and decision-making so that, in case of need, they can adequately defend their rights and freedoms.

The objective of this thesis is to conduct a thorough analysis of the regulatory mechanisms that govern and guarantee the principle of explainability within the context of European legislation with reference to the subject of data protection and artificial intelligence.

The first chapter of the thesis will examine the right to explainability under the GDPR, in an effort to describe how this right is manifested in the context of processing personal data, especially in situations where data is processed by algorithms or automated systems. It focuses on how the GDPR ensures the mechanisms behind data processing are accessible to individuals to understand how an automated decision could affect them. Organizations dealing with such data would be required to provide adequate explanations. The "transparency" obligation under the GDPR will be explored, examining its practical application and the challenges posed by the complexity of the technologies involved.

The second chapter will focus on the right to explainability in relation to the AI Act (Artificial Intelligence Act), which serves as the European legislation regulating the

artificial intelligence, especially in relation to AI systems identified as high-risk. In this context, the right to explainability pertains not merely to data protection but also to the comprehension of decisions rendered by AI systems. The chapter will analyze the specific requirements set forth by the AI Act to ensure that artificial intelligence systems are transparent and comprehensible to users, especially those used in high-risk contexts.

The third and final chapter will offer an ethical analysis regarding the principle of explainability, highlighting the moral and social justifications that underlie this right while scrutinizing the ethical challenges arising under the current European regulatory framework. This investigation in ethics will dwell on the moral consequences associated with keeping transparency in automated decision-making systems, assessing the challenges met when trying to realize these principles in a regulatory context often missing clear guidelines or totally inadequate to handle the rapid strides taken by the said technology.

The objective of this thesis is to examine the extent to which existing European regulations effectively respond to the increasing intricacy of novel technologies and to assess whether the right to explainability, as established by the GDPR and the AI Act, can truly facilitate meaningful user control over automated decision-making processes. Further, the research will delve into the ethical dilemmas attached to the artificial intelligence regulation and consider how such regulations might develop in response to impending issues relating to transparency, accountability, and personal data

protection in an increasingly

## Chapter I

### GDPR and explainability

#### 1.1. From the first definition of privacy to the right to data protection

Norberto Bobbio wrote that human rights are "*historical rights, that are, born under certain circumstances marked by struggles for the defense of new freedoms against old powers, gradually, not all at once, and not once and for all.*" <sup>1</sup>

As Bobbio states, specific historical contingencies lead people to question their condition and needs. These questions prompt individuals to become aware of particular necessities, encouraging the creation of appropriate frameworks to address them. Human rights, therefore, are not static principles fixed in history but elements that gain importance and become essential following specific historical events. This is why Bobbio refers to them as historical rights rather than purely human rights.

This perspective provided by Bobbio is useful for understanding the concept of privacy and its historical development. It shows how privacy evolved over time to become the subject of structured European regulations such as the GDPR.

The modern concept of privacy has ancient origins and has undergone significant <sup>2</sup> changes and adaptations over time, evolving according to specific historical periods, the communities in question, and the individuals addressing the issue.<sup>3</sup>

The concept of privacy, in its strict sense, developed from the Western desire to draw

---

<sup>1</sup> Norberto Bobbio, *L'età dei diritti*, Torino, Einaudi, 1997, p. XIII

<sup>2</sup> Janice Richardson, *Law and the Philosophy of Privacy*, New York, Routledge, 2017, pp. 1-214

<sup>3</sup> Gianmarco Cifaldi, *Evolution of Concepts of Privacy and Personal Data Protection under the Influence of Information Technology Development*, in *Sociology and Social Work Review*, Volume 7 (Issue 1), 2023, pp. 35-60: <https://ricerca.unich.it/retrieve/7798a6bc-f30c-46c0-8397-19cbb84ee841/Evolution-of-Concepts-of-Privacy-and-Personal-Data-Protection-under-the-Influence-of-Information-Technology-Development.pdf>

a clear boundary between the personal sphere and the surrounding world. This need to distinguish what belongs to an individual from what is public reflects a desire to protect and delineate a private space separate from the collective dimension. This sentiment became a necessity starting in the 19th century, coinciding with the mass migration of populations from rural areas to cities.<sup>4</sup>

This urbanization phenomenon led to a redefinition of public and private concepts: the overcrowding of urban areas made people more sensitive to and protective of their private sphere, thus enforcing a clear distinction between public and private spaces. This shift made private spaces more valuable and subject to more intense protection.<sup>5</sup>

The modern notion of privacy dates back to 1890, when the Harvard Law Review, a monthly publication in the United States, featured an article titled *The Right to Privacy*, written by jurists Samuel Warren and Louis Brandeis. In this article, the right to privacy was defined as "*the right to be let alone*,"<sup>6</sup> meaning the right to be left undisturbed to enjoy one's private life.

The authors introduced this definition in response to the behavior of newspapers of the time, which often published unfiltered depictions of guests attending the social events of the urban bourgeoisie. Samuel Warren, in particular, sought to claim the confidentiality of private life by establishing limits on external interference.

The article begins with the following passage:

---

<sup>4</sup> Hannah Arendt, *The Human Condition*, London, The University of Chicago Press, 1958, pp. 68-73

<sup>5</sup> Samuel Warren and Louis Brandeis, *The right of Privacy*, Harvard Law Review, Vol.4, No.5 (Dec. 15, 1890) , pp. 1896-1898, [https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=2613&context=faculty\\_scholarship](https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=2613&context=faculty_scholarship)

<sup>6</sup> Samuel Warren and Louis Brandeis, *op.cit.*, pp.1896-1898

*"That the individual shall have full protection in person and in property is a principle as old as the common law; but it has been found necessary from time to time to define anew the exact nature and extent of such protection. Political, social, and economic changes entail the recognition of new rights, and the common law, in its eternal youth, grows to meet the demands of society. Thus, in very early times, the law gave a remedy only for physical interference with life and property, for trespasses vi et armis. Then the 'right to life' served only to protect the subject from battery in its various forms; liberty meant freedom from actual restraint; and the right to property secured to the individual his lands and his cattle. Later, there came a recognition of man's spiritual nature, of his feelings and his intellect. Gradually the scope of these legal rights broadened; and now the right to life has come to mean the right to enjoy life, the right to be let alone; the right to liberty secures the exercise of extensive civil privileges; and the term 'property' has grown to comprise every form of possession—intangible as well as tangible." <sup>7</sup>*

From this excerpt, it becomes evident that, as with Norberto Bobbio's "historical rights," Samuel Warren and Louis Brandeis also viewed individual rights as not fixed once and for all but as values that develop and assert themselves over time in response to political, social, and economic changes. According to the authors, the notion of privacy evolved alongside the concept of property, which transformed over time to encompass both tangible and intangible possessions.<sup>8</sup>

The right to privacy aims to protect the life of the individual, ensuring the ability to

---

<sup>7</sup> Samuel Warren and Louis Brandeis, *op.cit.* pp.1896-1898

<sup>8</sup> Samuel Warren and Louis Brandeis, *op.cit.* pp.1896-1898

fully enjoy one's existence and safeguard their personal sphere. This principle is designed not only to preserve the private lives of individuals and their private relationships but also to protect personal writings and any other creations of the intellect or emotions. In conclusion, the right described by Samuel Warren and Louis Brandeis seeks to protect the private lives of individuals, including their writings and private creations.

Like any other subjective position recognized by the social and legal system, this new right had to necessarily "coexist" with other rights. Balancing freedom of expression and information on one side, and the right to privacy and confidentiality of individuals on the other, thus began to emerge as a priority.

With the advent of the digital world and the massive processing of personal data of digital users, the scope of the right to privacy expanded significantly. This made it necessary to redefine this right in broader terms, encompassing the protection of personal data. In other words, the right to the protection of personal data stems from the recognition of the more general right to privacy, serving as a specific application and extension of it.<sup>9</sup>

For this reason, starting in the early 1990s, and with the widespread use of the Internet, the issue of protecting personal data began to be addressed concretely.<sup>10</sup> To better understand the structure and scope of this right, it is useful to analyse the definition of personal data. The definition referred to here is the modern one, elaborated within the framework of the GDPR, which provides a comprehensive and contemporary

---

<sup>9</sup> Paolo Guarda, Giorgia Bincoletto, *Diritto comparato della privacy e della protezione dei dati personali*, Ledizioni, Marzo 2023, pp. 55-57

<sup>10</sup> Paolo Guarda, Giorgia Bincoletto, *op.cit.* pp.55-97

understanding of what constitutes personal data in today's digital age. The definition is: *“Personal data means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person.”*<sup>11</sup>

Based on this broad definition of personal data, the right to data protection is understood as an individual's ability to control the information concerning them (which personal data is processed and how), which reflects their identity and outlines their persona in today's information society.

This right must be distinguished from the “negative freedom” of not being subjected to interference in private life and should instead be considered the foundation of the “positive freedom” to control the flow of one's own information.<sup>12</sup>

For this reason, the right to personal data protection is often interpreted as a person's right to informational self-determination, i.e., the free ability of each individual to define and determine themselves with regard to the information concerning them.<sup>13</sup>

The right to informational self-determination directly depends on individuals' ability to understand the fundamental mechanisms governing these processes related to the use of personal data.

---

<sup>11</sup> GDPR, Article 4

<sup>12</sup> Giusella Finocchiaro, *La protezione dei Dati personali in Italia*, Regolamento UE N.2016/679 e d.lgs. 10 Agosto 2018, n.101, Bologna Zanichelli, 2019

<sup>13</sup> Buttarelli, *Banche dati e tutela della riservatezza*, Milano, Giuffrè Editore, 1997, pp. 1-ss

A fundamental prerequisite for such understanding lies in the right to explainability.<sup>14</sup>

This concept will be explored further in the following chapters.

Today, due to the specificity and complexity of the right to privacy and the right to data protection, they are treated separately and are governed by different sources.

While they are interrelated, the right to privacy represents a broader concept within which the right to personal data protection is encompassed, following a general-to-specific relationship. The latter focuses specifically on the protection of personal information, whereas the former covers a wider range of rights linked to the confidentiality of individuals' private lives.

A key distinction between the two positions lies in the fact that, while the right to privacy has an essentially negative nature, consisting in the right not to disclose and to keep certain information private, the right to personal data protection aims at active control over such data. Furthermore, the right to privacy does not apply to all types of information but exclusively to those pertaining to strictly confidential matters.<sup>15</sup>

In this context, a significant example of a regulatory framework is Regulation 2016/679, known as the GDPR, which provides a comprehensive legal structure dedicated to the protection of personal data. Its purpose is to ensure control and security over individual information in the digital age.<sup>16</sup>

---

<sup>14</sup> Bryce Goodman, Seth Flaxman, *European Union regulation on algorithmic decision-making and a "right to explanation"*, Oxford, United Kingdom, Oxford Internet Institute and Department of Statistics, University of Oxford, 31 August 2016: <https://doi.org/10.48550/arXiv.1606.08813> pp.1-9

<sup>15</sup> Luigi Rendina, *Privacy vs protezione dati personali: attenti alla differenza, ne va della nostra identità*, Agenda Digitale, 30 October 2019: <https://www.agendadigitale.eu/sicurezza/privacy/privacy-e-protezione-dati-personali-cosa-sono-quali-differenze-cosa-e-cambiato-col-gdpr/>

<sup>16</sup> General Data Protection Regulation (GDPR)

Although the GDPR does not expressly enshrine the right to privacy as an autonomous and direct right of individuals, it acknowledges and emphasizes its significance in the Recitals, particularly Recital 4<sup>17</sup>. Here, it states that the processing of personal data should serve humanity and that the right to data protection is not absolute but must be balanced with other fundamental rights in accordance with the principle of proportionality. The Regulation also aligns itself with the Charter of Fundamental Rights of the European Union, ensuring respect for private and family life, home and communications, alongside the broader protection of personal data, thereby safeguarding confidentiality and human dignity through regulated data processing practices.

## **1.2. Privacy in Europe: the development of the GDPR**

The initial steps toward greater privacy protection date back to the years immediately following World War II. On December 10, 1948, the United Nations General Assembly adopted the Universal Declaration of Human Rights, the first document to recognize the right to privacy as a fundamental human right.

Article 12 of the Declaration states:

*"No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honor and reputation. Everyone has the right to the protection of the law against such interference or attacks."*<sup>18</sup>

This provision aims to ensure every individual's right to live their life without external

---

<sup>17</sup> GDPR, Recital 4

<sup>18</sup> United Nations, *Universal Declaration of Human Rights* <https://www.un.org/en/about-us/universal-declaration-of-human-rights>

intrusions interfering with their privacy, family life, private correspondence, and more.

It represents the first formal articulation of the right to privacy for all individuals.

Two years after the adoption of the Declaration, the European Convention on Human Rights (ECHR) was signed in Rome in 1950. This convention is an international treaty aimed at safeguarding human rights and fundamental freedoms within the legal systems of its member states.<sup>19</sup>

Article 8 of the Convention states:

➤ *Everyone has the right to respect for his private and family life, his home and his correspondence.*

➤ *There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety, or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.*

This provision also defines the essential terms of the right to privacy, while simultaneously establishing the need for its balance with other significant interests that may justify its limitation. For example, privacy rights can be restricted when national security or public safety is at risk, or when the health, freedom, or rights of others are threatened.

What differentiates Article 12 of the Declaration from Article 8 of the ECHR is the

---

<sup>19</sup> Council of Europe Portal, The European Convention on Human Rights, *The Convention in 1950* [https://www.coe.int/en/web/human-rights-convention/the-convention-in-1950?](https://www.coe.int/en/web/human-rights-convention/the-convention-in-1950?__hstc=312123171.1728202000.1728202000.1728202000.1728202000&__hssc=312123171.1728202000.1728202000.1728202000.1728202000&__hsid=312123171.1728202000.1728202000.1728202000.1728202000)

nature of the legal instrument that conveys these rights. While the Declaration laid the groundwork for the approval of the ECHR<sup>20</sup>, the latter is a legally binding international treaty that came into force following ratification by a number of states. In contrast, the Declaration is not legally binding within the domestic legal systems of individual states. However, it holds moral and interpretative significance, having been adopted by the international community as a reference text for human rights.

The Strasbourg Convention of 1981, also known as Convention 108 of the Council of Europe, is a key legal instrument for protecting individuals against the misuse of automated personal data processing. It remains the only legally binding international framework on this subject and is open to adherence by non-European Council member states, such as Uruguay, which joined in 2013.<sup>21</sup>

The Convention introduced personal data protection as a distinct concept, emphasizing that safeguarding fundamental rights and freedoms requires data processing to meet specific conditions. Article 5 outlines these principles, including fairness, legality, purpose limitation, and data quality. For example, personal data must be processed fairly, collected for legitimate purposes, and not retained beyond what is necessary. Sensitive data, such as race, political opinions, and health, can only be processed with

---

<sup>20</sup>The European Convention on Human Rights (ECHR) was the first binding international treaty inspired by the Declaration. It focuses solely on civil and political rights, excluding economic and social rights, which are considered difficult to enforce. The European Union (EU), although not explicitly referencing the Declaration, bases its laws and policies on its principles. The EU Charter of Fundamental Rights, adopted in 2009, incorporates many rights from the Declaration but goes further in certain areas, such as the explicit right to asylum (the Declaration only recognizes the right to seek asylum) and introduces new protections, such as the prohibition of the death penalty and the protection of personal data.  
[https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/757559/EPRS\\_ATA\(2023\)757559\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/757559/EPRS_ATA(2023)757559_EN.pdf)

<sup>21</sup> Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Strasbourg, 28 January 1981 <https://rm.coe.int/1680078b37?>

strict legal safeguards.<sup>22</sup>

The Convention applies to both public and private sectors, including law enforcement and judiciary activities. Its goal is to protect individuals from abuse and regulate transnational data flows, drawing inspiration from Article 8 of the European Convention on Human Rights, which enshrines the right to private and family life.<sup>23</sup>

Article 1 states:

*"The purpose of this Convention is to secure, in the territory of each Party, for every individual, whatever their nationality or residence, respect for their rights and fundamental freedoms, and in particular their right to privacy, with regard to the automatic processing of personal data relating to them ('data protection')."*<sup>24</sup>

All EU member states and the European Union itself have ratified the Convention. It regulates automated data processing while allowing states to extend protections to manual data processing if personal data is part of or intended for filing systems. The Convention also establishes the right of individuals to access their data, request corrections, and restricts data transfers to countries lacking equivalent legal protections.

As mentioned earlier, in the 1990s, with the widespread use of the Internet, the right to privacy evolved to include the right to personal data protection, reflecting the changing needs of the digital age.

The first European document entirely dedicated to regulating personal data protection

---

<sup>22</sup> Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, *op.cit.* <https://rm.coe.int/1680078b37?>

<sup>23</sup> Jody R. Westby, *International Guide to Privacy*, American Bar Association, 2004, pp. 87-89

<sup>24</sup> Art. 1, Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, *op.cit.* <https://rm.coe.int/1680078b37?>

was the Directive of the European Parliament and Council No. 95/46/EC.<sup>25</sup> This Directive, was adopted on October 24, 1995, and remained in effect until 2018, when it was replaced by the GDPR.<sup>26</sup>

This directive was introduced following the Maastricht Treaty and the establishment of the European Community, which created two new areas of competence: common foreign and security policy and cooperation in the fields of justice and home affairs. Additionally, the system of free movement of people and goods, known as the Schengen Area, was implemented.<sup>27</sup>

In this new institutional framework, the Directive 95/46/EC had a dual purpose: ensuring the protection of personal data on the one hand and fostering the development of the common market on the other. Its purpose was to strike a balance between the *protection of the fundamental rights and freedoms of natural persons, in particular their right to privacy with respect to the processing of personal data* (Article 1(1)), and *the free flow of personal data* (Article 1(2)).<sup>28</sup>

This need for harmonization emerged in response to the fragmented data protection regulations of every European member state. Each state followed its own rules, which hindered the functioning of the European single market. This situation was at odds with the European Union's primary goal of creating a common market for all member

---

<sup>25</sup> Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data: <https://eur-lex.europa.eu/eli/dir/1995/46/oj>

<sup>26</sup> European Data Protection Supervisor, *The History of the General Data Protection Regulation* [https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation\\_en?](https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en?)

<sup>27</sup> European Council, Council of the European Union, *How Maastricht changed Europe*, October 2024 <https://www.consilium.europa.eu/en/maastricht-treaty/>

<sup>28</sup> Art. 1, Directive 95/46/EC, *op.cit.* <https://eur-lex.europa.eu/eli/dir/1995/46/oj>

states.

What ultimately led to the adoption of the GDPR was the increasingly formal role, which the Directive 95/46/EC had assumed over the years. While it proved useful in harmonizing the European single market and establishing uniform rules for processing personal data for economic and commercial purposes, it fell short in effectively safeguarding the fundamental rights of individuals affected by data processing.<sup>29</sup>

Over time, this formalist approach reduced data protection to the fulfillment of bureaucratic procedures, such as providing information to data subjects and obtaining their consent for data processing. Consent became the central, and often sole, criterion for legitimizing the use of data, effectively enabling the *free flow of data*.<sup>30</sup>

As a result, member states interpreted this framework as a mere administrative formality, undermining its intended purpose of protecting individuals.<sup>31</sup>

The subsequent adoption of the GDPR was a response to both the shortcomings of Directive 95/46/EC and the rapid technological advancements of recent years. This new European regulation, together with Directive 2016/680 (regarding the protection of individuals in relation to the processing of personal data by competent authorities for the prevention, investigation, detection, or prosecution of criminal offenses or the

---

<sup>29</sup> Paul De Hert, Vagelis Papakonstantinou, *The proposed data protection Regulation replacing Directive 95/46/EC: A sound system for the protection of individuals*, Sciencedirect.com, Computer Law and Security Review, April 2012 [10.1016/j.clsr.2012.01.011](https://doi.org/10.1016/j.clsr.2012.01.011) pp.130-142

<sup>30</sup> Giusella Finocchiaro, *op.cit.*

The dual objective is achieved by ensuring that data flows in compliance with the rules established by the Regulation. The need for balance arises precisely from this, from reconciling two seemingly opposing needs: data protection and data circulation. These are realized through the “secure” circulation of data, meaning circulation that complies with the provisions of the Regulation.

<sup>31</sup> Paul De Hert, Vagelis Papakonstantinou, *op.cit.* [10.1016/j.clsr.2012.01.011](https://doi.org/10.1016/j.clsr.2012.01.011) pp.130-142

execution of criminal penalties, as well as the free movement of such data<sup>32</sup>), define the “EU Data Protection Package.”<sup>33</sup>

The GDPR’s philosophy can be summarized as the promotion of a data management approach centered on personal data protection. It requires a constant assessment of the risks to individuals’ rights arising from data processing.<sup>34</sup>

This regulation moved beyond the previous approach, which had focused on adhering to predefined procedures (standardized processes for data management) without concretely evaluating the potential risks associated with these processes.

To this end, the European legislator revisited the Directive 95/46/EC, making seemingly minor changes to many provisions but significantly altering the substantive approach. These changes reshaped the entire framework, giving a new meaning and content to all provisions, even though the formal structure of the legal framework remained largely unchanged.<sup>35</sup>

It should be noted that the GDPR, upon its drafting and adoption, became a benchmark for personal data protection not only within EU member states but also on a global scale.<sup>36</sup>

---

<sup>32</sup> *Direttiva (UE) 2016/680 del Parlamento Europeo e del Consiglio del 27 Aprile 2016* <https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32016L0680>

<sup>33</sup> European Commission, Legal framework of EU data protection [https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection\\_en?](https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection_en?)

<sup>34</sup> Yuhong Yan, The Risk-Based Approach to Personal Data Protection and the Response of the International Trade Law, *Beijing Law Review*, Vol.14. No.3, September 2023 [10.4236/blr.2023.143067](https://doi.org/10.4236/blr.2023.143067) pp.1250-1270

<sup>35</sup> Giusella Finocchiaro, *op cit.*

<sup>36</sup> The history of data protection regulation [https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation\\_en](https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en)

From a chronological and procedural standpoint, the process began in 2012 when the European Commission introduced the first draft of the European Regulation in order to update the legislation regarding data processing and free data flow.<sup>37</sup>

On May 4, 2016, Regulation (EU) 2016/679, known as the "GDPR," was published in the Official Journal of the European Union and entered into force on May 25, 2016<sup>38</sup>.

As a "self-executing" legal instrument, the GDPR is directly and immediately applicable across member states, becoming effective upon its approval.

### **1.3. The importance of explainability: the problem of information asymmetry**

As previously mentioned above, the increasing use of big data and artificial intelligence (AI) systems has triggered a profound societal shift, changing how society is perceived and defined: the transition has occurred from an Information Society<sup>39</sup> to a Black Box Society, where the internal working of many systems, particularly those driven by AI, are opaque and difficult to understand.<sup>40</sup>

This term originates from the concept developed by Frank Pasquale in his book *The Black Box Society: The Secret Algorithms that Control Money and Information*. In his work, the author illustrates the asymmetry characterizing the relationship between large technology companies and institutions that control AI systems and the individuals (or users) who use their services.

The asymmetry described by Pasquale refers to the disparity between the AI companies

---

<sup>37</sup> Giuseppe Cassano, Vincenzo Colarocco, Giovanni Battista Gallus, Francesco Paolo Micozzi, *Il Processo di Adeguamento al GDPR*, Milano, Giuffrè Editore, 2022, pp.1-ss

<sup>38</sup> GDPR, Article 99

<sup>39</sup> D.F Wallace, *The Pale King, An Unfinished Novel*, Little, Brown and Company, 2011 p.85

<sup>40</sup> Frank Pasquale, *The Black Box Society: The secret Algorithms that control Money and Information*, Cambridge-London, Harvard University Press, 2015, pp.1-38

and their users in access to information about how AI systems operate. On one side, there are these entities with access to vast amounts of data and advanced analytical tools. On the other, there are users with limited competence to understand how these systems operate, making it difficult for them to challenge decisions they believe to be unfair.

This asymmetry is especially evident in the field of personal data management, where AI is widely used to improve the efficiency of processes that handle large volumes of personal data. The challenge arises because personal data processing often relies on AI systems using algorithms that are not easily understandable or transparent, not only to end users but sometimes even to the data controllers themselves.

It is evident that this lack in understanding algorithmic functioning complicates efforts to ensure regulatory compliance in personal data management.<sup>41</sup>

The problem of information asymmetry has been addressed by the European Union through the introduction of the "explainability" requirement in the GDPR, which has become a crucial point to ensure the reliability, transparency, and accountability of automated decision-making systems.

As decision-making algorithms are increasingly used in critical areas like healthcare, finance, justice, and employment, there is a growing need to understand how these systems work, especially when their decisions impact individuals and society. Discussing explainability means recognizing that, while artificial intelligence can improve efficiency and automation, it must also be applied with principles of

---

<sup>41</sup> Giuseppe Mobilio, *L'intelligenza artificiale e le regole giuridiche alla prova il caso paradigmatico del GDPR*, federalismi.it, n. 16/2020 pp.266-298

transparency and accessibility. Explainability is not just a technical matter, it is an ethical and regulatory issue that ensures these technologies are used fairly and responsibly. Primarily, the concept of explainability is central to obtain and maintain user trust in automated systems. When people use a service that involves decision-making algorithms, they want to understand how and why certain decisions were made, especially if these decisions impact their daily lives. A system that acts like a “black box,” providing results without a clear explanation of the reasoning behind them, generates uncertainty and mistrust. People often feel manipulated or vulnerable when they don’t know how their personal data is being used or why certain decisions are made. Without clear explanations, it becomes even harder for users to make informed decisions about how their data is handled and how those decisions may affect their lives. In situations where transparency is essential for accountability, explainability helps uncover the decision-making processes, making the system’s reliability more evident.<sup>42</sup>

Explainability also plays a vital role in fighting AI discrimination and bias. Complex algorithms, such as those based on machine learning techniques, can harbor implicit biases that reflect or amplify social and cultural inequalities. For example, if a recruitment algorithm is trained on data that includes gender or racial biases, it runs the risk of perpetuating those biases in its results.<sup>43</sup> Explainability allows developers

---

<sup>42</sup> Andrea Ferrario and Michele Loi. *How Explainability Contributes to Trust in AI*. In 2022 ACM Conference on Fairness, Accountability, and Transparency, June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, pp.1457-1466 <https://doi.org/10.1145/3531146.3533202>

<sup>43</sup> Chen, Z. *Ethics and discrimination in artificial intelligence-enabled recruitment practices*. *Humanit Soc Sci Commun* 10, 2023 pp.1-12 <https://doi.org/10.1057/s41599-023-02079-x>

and users to identify and correct any inaccuracy in decision-making models increasing fairness and inclusivity. When an algorithm works in a transparent and easy way to understand, it becomes possible to identify and fix issues related to bias or discrimination. This helps ensure that decisions are fair and based on appropriate standards.<sup>44</sup>

The need for explainability also stems from the requirement to comply with data protection and human rights regulations. Regulations like the GDPR require transparency, meaning people must be informed about how their data is processed and how decisions affecting them are made. In cases such as automated decisions about access to financial services or healthcare, the law often requires the ability to challenge or request a review of decisions made without sufficient explanation.<sup>45</sup> Moreover, explainability is a key factor in ensuring that automated decisions are not just fair but also easy to understand. Explaining how a decision-making system works means unpacking complex processes into transparent, available, and easily understandable language for non-experts. This approach not only fosters trust among users but also enables faster and more efficient measures by decision-makers to address issues. A system that provides answers without a clear rationale not only causes confusion, but it may also hide errors, malfunctions, or even manipulations either by the developers or external actors.<sup>46</sup> And this is the because explainability is fundamental to ensuring democratic and collective control over the use of technologies; it enables users,

---

<sup>44</sup> Simon Chandler, *How Explainable AI is Helping Avoid Bias*, Forbes, 2020 <https://www.forbes.com/sites/simonchandler/2020/02/18/how-explainable-ai-is-helping-algorithms-avoid-bias/>

<sup>45</sup> Chen, Z. *op.cit.* pp. 1-4

<sup>46</sup> Chen, Z. *op.cit.* pp.1-4

regulators, and developers to continuously review the effectiveness and ethical impacts of decision-making systems.

Finally, explainability is a principle that supports responsible innovation. With the increasing integration of advanced technologies, such as artificial intelligence, into society, there is a need to ensure that these technologies are developed and used in ways that maintain key principles like fairness, privacy, and justice. Integrating explainability into an AI framework means creating models that are not only effective but also ethically responsible. Technological progress must go hand in hand with a clear understanding of the ethical and legal consequences of these technologies. In this way, explainability goes beyond technical performance—it becomes a fundamental principle for developing innovations that uphold human rights and operate transparently and fairly.<sup>47</sup>

In short, explainability is crucial because it ensures transparency, fairness, reliability, and accountability in automated systems. It also supports the ethical, fair, and respectful implementation of these technologies. Above all, explanations increase understanding, simultaneously supporting governance, rectification, and monitoring, hence making decision-making fairer and more accountable.

---

<sup>47</sup> Nicol Turner Lee, *Making AI more explainable to protect the public from individual and community harms*, Brookings, November 29, 2023, pp.1-11 <https://www.brookings.edu/articles/making-ai-more-explainable-to-protect-the-public-from-individual-and-community-harms/>

## 1.4. Explainability

Broadly, the right to explanation refers to the user's right to understand the mechanisms by which their personal data is processed. This right must be upheld by entities operating in this specific field, ensuring that personal data is processed by AI systems that use explainable algorithms.<sup>48</sup>

AI systems, at their core, depend on datasets for training and carrying out their tasks. These algorithms create new representations of reality based on "datafications" (i.e., interpretations of data) to achieve their goals. When the data used is correct and reliable, the system operates effectively.<sup>49</sup>

However, when anomalies or unexpected results arise, it becomes necessary to question the validity and reliability of the decisions made. In such circumstances, the issue of explainable algorithm becomes crucial when it is necessary to scrutinize or validate the outcomes of an algorithmic system.<sup>50</sup>

In these cases, the lack of explainability would make it challenging, if not impossible, to assess whether the system operates correctly or whether there are biases, errors, or distortions in the decision-making process.

To address this, the right to explanation must be applied in two stages: *ex-ante* explanation and *ex-post* explanation. In practice, information about how algorithms function can be provided to users either before (*ex-ante*) or after (*ex-post*) a particular decision is made.

---

<sup>48</sup> Erica Palmerini, *Decisioni algoritmiche e diritto dei dati*, giudicedonna.it, Numeri 1-2/2023, pp.1-20

<sup>49</sup> Mortaji, S. T. H., & Sadeghi, M. E. *Assessing the reliability of artificial intelligence systems: Challenges, metrics, and future directions*. International Journal of Innovation in Management Economics and Social Sciences, 4(2), 2024, pp.1-13 [www.ijimes.ir](http://www.ijimes.ir)

<sup>50</sup> Mortaji, S. T. H., & Sadeghi, *op.cit.* pp.1-3

When information is provided in advance, the ex-ante explanation usually gives a general overview of how the algorithms work and the types of decisions they may make. In these cases, there is a risk that the explanation will be vague or unclear, not providing the information people need to understand clearly.<sup>51</sup>

On the other hand, informations provided to users after a particular decision has been made (*ex-post*) may be more specific and detailed. This second approach is generally seen as more effective in protecting individuals' rights, as it provides an explanation directly related to a specific decision.<sup>52</sup>

## **1.5. Explainability under GDPR**

In the GDPR, the right to explainability is addressed, even though not explicitly, in several provisions, particularly Articles 13, 14, 15, and 22, and in Recital 71.

In my opinion, as I will explain in the rest of the essay, the issue should be understood by considering the connection between the right to automated decision-making and the right to challenge or contest decisions that affect one's rights or interests.

### **1.5.1. Article 22**

Article 22 of the GDPR establishes a general right for individuals not to be subjected to decisions based solely on automated processing, including profiling, when such decisions have legal consequences or significantly impact them. This right is especially

---

<sup>51</sup> Gianclaudio Malgieri , Giovanni Comandé, *Why a Right to Legibility of Automated Decision- Making Exists in the General Data Protection Regulation, 2017*, in *International Data Privacy Law*, Volume 7, Issue 4, November 2017, pp. 243-265, <https://doi.org/10.1093/idpl/ix019>

<sup>52</sup> Gianclaudio Malgieri , Giovanni Comandé, *op.cit.* pp.243-265

important, for example, in cases involving the processing of sensitive data (as outlined in paragraph 4).

However, this right is excluded in certain specific cases, as explicitly outlined in the same provision, i.e., when the decision is necessary for the execution or conclusion of a contract between the data subject and the data controller, when the decision is authorized by Union or Member State law, or when the decision is based on the explicit consent of the data subject.

In all these cases, where an individual may be subjected to a decision based solely on automated processing of data, the rights, freedoms, and legitimate interests of the individual must still be protected. Specifically, in the first and third cases, the individual must be guaranteed, as a minimum protective measure, the right to request human intervention, express their opinion, and challenge the decision.

In essence, Article 22 regulates Automated Decision-Making (ADM), establishing that such processing is generally prohibited, except in cases explicitly identified, where specific safeguards must be in place for individuals subject to these automated decisions.<sup>53</sup>

---

<sup>53</sup> GDPR, Article 22:

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests;  
or

(c) is based on the data subject's explicit consent.

The guarantees may concern the right of the data subject to "intervene" in the decision-making process by requesting human intervention from the data controller or expressing their opinion about the decision.

An additional safeguard that must be ensured in such cases concerns the right of the individual affected by ADM to contest the adopted decision.

It is clear that the exercise of these rights, which must still be guaranteed, requires individuals to have adequate information about the automated processing involved. This information must include, at a minimum, the existence of the automated processing and its key characteristics.

Without this information, the rights mentioned above risk remaining mere formalities without real effectiveness.

For this reason, in my opinion the recognition of this right is particularly relevant for interpretative purposes and should lead us to believe that the right to an explanation is indeed contemplated by the GDPR.<sup>54</sup>

---

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

<sup>54</sup> Anticipating what will be more fully illustrated in the par. 1.6, highlight right now that on this point, there are differing opinions in the literature, Giuseppe Mobilio, *op.cit.*, pp.291 *"In seeking to ground this principle in a currently applicable legal basis, it should be noted that the GDPR does not explicitly enshrine the right in question, as this provision is only mentioned in Recital 71 and was not incorporated into the main body of the Regulation. Based on a restrictive interpretation of the GDPR, supported also by the preparatory works, it has been concluded that a true right to obtain an ex post explanation of a specific decision cannot be established. At most, a right to be informed ex ante about the existence of an automated decision-making process and the logic employed could be identified.*

In this case, it should be understood that the information provided to the data subject must include the right to receive an explanation of the decision made, at least in a way that ensures the protection of the rights of the individual affected by the automated processing.<sup>55</sup>

This interpretation is supported not only by Recital 71 but also by the nature of the right to contest recognized in Article 22.<sup>56</sup>

---

*Conversely, according to others, a right to comprehensibility can be derived from the combined provisions of Articles 13 and 14 (notification obligations), Article 22 (prohibition of being subject to automated decision-making), as well as the right of access (Article 15) and the right to obtain information regarding the processing (Article 12). Similarly, the same conclusion could be reached by favouring not a strictly literal interpretation of the GDPR, but rather one aimed at protecting and exercising the rights of the data subject.*

*Adopting a systematic interpretation, it has also been suggested that the GDPR guarantees the 'legibility' of data and analytic algorithms, understood as the comprehensibility of their architecture, that is, how the algorithms function and transparency in their implementation, meaning their commercial use."*

<sup>55</sup> Carlo Colapietro, *Algorithms between transparency and protection of personal data* in *Federalismi.it - Observatory on transparency*, 22.02.2023, pp.151-174: according to which there is not a general obligation of "explainability" but this obligation would be imposed in relation to the concrete case if the mere "understandability" "is not sufficient to guarantee a balance between the interest underlying the use of an algorithm and the protection of the interested party. Not, therefore, a general obligation on the part of the owner, but instead an obligation that arises after the evaluation of the concrete case, if therefore the principles already established to protect the interested party are not sufficient".

<sup>56</sup> Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (GDPR), Recital 71:

*"The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.*

*Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyze or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behavior, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her.*

*However, decision-making based on such processing, including profiling, should be allowed where expressly authorized by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or*

To further clarify, it is important to note that the doctrine have also stated that *“the disclosure is not an end in itself but is only a means of knowing the algorithmic act in order to allow a reasoned challenge before a judge. Transparency is therefore established as the measure of effectiveness of the equation: knowability of the acts/its challengeability in court”*.<sup>57</sup>

---

*national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent.*

*In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.*

*Such measure should not concern a child.*

*In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organizational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimized, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject, and prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect.*

*Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.”*

<sup>57</sup> *Giovanna De Minico, Justice and artificial intelligence: a changing balance, in Rivista AIC Associazione italiana dei costituzionalisti, n. 2/2024, del 29.04.2024, p.93, who writes thus: “We have now reached the third and final link in this chain prodromal to the automated decision: the contestability of the act in court. The judge must be able to view it, without encountering factual or legal obstacles in his control of lawfulness. The judge cannot accept the infallibility of the intelligent outcome, but first he will have to retrace the predictive statistical reasoning that led to a certain output and, only then, will he be able to decide whether or not to share the intelligent reasons. Therefore, his judicial review is not focused on the extrinsic correctness of the act; it is not forced to stay within the boundaries of the consequentality between premises and conclusions and/or the reliability between outcomes and premises, but can move towards the legality of the input data, verifying the absence of bias and other discriminatory ingredients. In our opinion, the judge will be able to re-examine even the reasonableness of the evaluation criteria assigned to the mechanical mind; in an expression, he will be able to retrace ex post the logic of deterministic reasoning according to a method of investigation that closely resembles the ab intriseco syndicate on technical discretion, separated by an evanescent border line from the scrutiny on the merits with which he has no difficulty in confusing himself. This penetrating scrutiny must make at least two*

This right, which is essentially an expression of the right to defense<sup>58</sup>, necessarily presupposes a correct and complete understanding of the decision being contested.

Otherwise, the right to contest would be rendered useless.

It is well established that the right to defend oneself is effectively upheld when the reasons for the contested act are clearly communicated.<sup>59</sup>

---

*factors available to the judge: one positive, the other negative. Precisely, an adequately motivated algorithmic act and the absence of industrial secrecy that can be opposed to his request for access to the source code, otherwise the union will only be able to remain on the surface. These two guarantees are necessary because they ensure a technology that is understandable, examinable and challengeable in the courts for the incidence of errors and biases, not unlike what happens with the human mind. If this were not the case, could we still say that technology is at the service of the person? Anthropocentrism focuses on a human being, free and responsible for choosing whether or not to use mechanical intelligence because he is aware of the reality he faces. The opacity of the arcana technologicum is nothing other than a re-edition of the ancient legal dogma, assisted by an atypical normality, which makes today's inscrutability less contestable than yesterday's due to that veil of objectivity and scientific certainty which gives it a semi-judicial immunity".*  
<https://www.rivistaaic.it/it/rivista/ultimi-contributi-pubblicati/giovanna-de-minico/giustizia-e-intelligenza-artificiale-un-equilibrio-mutevole>

<sup>58</sup> The right to contest the established decision entails the right to exercise a form of protection against that decision. This should materialize in the right to have access to a remedy that allows the individual affected by the "automated" decision to, for example, file a complaint or an appeal to assert their rights and interests they believe have been infringed by the decision.

This conclusion can also be drawn from the fact that this "safeguard" follows and complements the other guarantees provided by the regulation in favor of the individual affected by the decision namely, the right to obtain human intervention and the right to express their opinion.

The first constitutes a procedural safeguard (there must be human intervention), the second provides a safeguard of participation in the decision-making process (the ability to express one's view), and the third offers a safeguard of protection against the decision taken (the ability to contest the decision).

<sup>59</sup> It is a rule that is more logical than legal that knowledge of the decision to be contested, and its content, is the fundamental prerequisite for exercising the right to contest it. For example, in administrative matters, it is well-established that the reasoning behind an act, understood as the indication of the factual elements and legal evaluations on which the decision is based, constitutes a means of defense for the recipient of the decision. See Court of Justice of the EU, Grand Chamber, December 2, 2009, No. 89; Court of Justice of the EU, Grand Chamber, December 22, 2008, No. 333: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62008CJ0089>

For similar reasons, the Italian Administrative Court, in cases involving "algorithmic administration," affirms the citizen's right to know the algorithm used by the administration

This interpretation appears consistent with, and thus supported by, the EU interpretative principle of the so-called "principle of effectiveness," according to which a European provision should preferably be interpreted in a way that enables the achievement of its intended objective<sup>60</sup>. This provision should be understood as guaranteeing that the individual has the right to truly understand how the algorithm behind the contested decision works. Only then the individual can effectively exercise their right to challenge the decision. The hermeneutic hypothesis presented here aligns with what is stated in Recital 71 of the GDPR, which, after mentioning the data subject's right "*not to be subjected to a decision, which may include a measure, assessing personal aspects concerning them, based solely on automated processing, and producing legal effects concerning them or similarly significantly affecting them,*" clarifies that, in certain cases, decisions based on such automated processing may be allowed. It specifies that "*in any case, such processing should be subject to appropriate safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express their opinion, to obtain an explanation of the decision reached after such an assessment, and to contest the decision.*"

Thus, the right to receive an explanation of the decision is an essential prerequisite for the right to contest that same decision.

---

(so-called knowability) and to obtain an explanation of how the algorithm works in an understandable manner (so-called comprehensibility). Otherwise, the recipient's right to defence would be undermined. See Regional Administrative Court (TAR) of Campania (Naples), Section III, November 14, 2022, No. 7003.

<sup>60</sup> Pursuant to the so-called "principle of effectiveness," EU laws must be interpreted in a way that "ensures the effectiveness of the aforementioned regulations and directive, as well as the effective judicial protection required by Union law," Court of Justice of the EU, First Chamber, July 13, 2023 (Cases 363/21-364/21), par. 100. This is a corollary of the teleological interpretative criterion; see T. Tridimas, *The Court of Justice and Judicial Activism*, in *European Law Review*, 1996, p. 208

The wording of the Recital therefore aligns with the interpretation outlined above and it provides a reasonable understanding of the rule, in line with the protections that the GDPR aims to provide for individuals subject to the automatic decisions under by Article 22. In this sense, Recital 71 supports the interpretation of the provision, as, although it is not legally binding, it provides valuable guidance for protecting the rights of data subjects by reinforcing the interpretation of EU regulations.

### **1.5.2. Articles 13, 14, 15**

An essential condition for the exercise of the *minimum* rights guaranteed by Article 22 for the data subject affected by decisions based solely on automated processing is being informed of the existence of such automated decisions and their content. Without this information, the rights provided by Article 22 cannot be effectively exercised.

This is addressed by the provisions contained in Articles 13, 14, and 15 of Chapter III (Rights of the Data Subject), Section 2 (Information and Access to Personal Data) of the GDPR.

Specifically, Articles 13(2)(f) and 14(2)(g) require data controllers to (i) inform data subjects about the existence of automated decision-making processes and to (ii) provide meaningful information about the logic of automated processing and the significance and consequences of such processing for the data subject.

The difference between these two provisions is in how the information is collected: Article 13 applies when personal data is collected directly from the data subject; Article 14 refers to cases where data is collected by the data controller from a source other than the data subject. For example, this might include online tracking (via

cookies or other tracking technologies), purchasing contact lists (a company buys contact lists from another company for marketing purposes), or public sources (where data is extracted from public records or publicly accessible databases, such as the registry office).

In addition to the right to be informed about the existence and characteristics of automated decision-making processes, the European legislator aims to grant data subjects an autonomous right of access to this information. Thus, Article 15 guarantees data subjects the right to access information about the processing of their data, allowing them to obtain (i) specific details about the functioning of automated decision-making processes concerning them, (ii) meaningful information about the logic of the algorithm, (iii) the significance of the processing, and its possible consequences.

This article introduces the right of access to information, which can be exercised by the user at any time, even after the conclusion of an automated decision-making process.

A substantial difference can be observed among these provisions, which is particularly relevant for protecting data subjects in the context of personal data processing. Articles 13 and 14 provide for *ex ante* notifications about the expected data processing, focusing only on access to general information about how the system operates. Article 15, however, concerns the user's right to obtain specific information about the processing of their personal data.<sup>61</sup>

This latter provision, unlike the previous ones, recognizes the right of data subjects to access *ex post* information regarding the existence of an automated decision-making

---

<sup>61</sup> Gianclaudio Malgieri , Giovanni Comandé, *op.cit.* pp.243-265

process, even (but not only) after the decision-making process has been initiated or concluded.<sup>62</sup>

From this, it can be deduced that the information related to the logic behind the processing should have different characteristics: in the first case (right to notification), such information necessarily concerns the logical criteria generally applicable to the processing; in the second case (right to access), the information may refer (if the decision about the specific processing has already been made) to the logical criteria applied concretely to the decision that affected the user.

The right of access guaranteed by Article 15 therefore appears to be essential for exercising the right to contest the decisions adopted under Article 22.<sup>63</sup>

Just like in Italian law, where access to administrative acts is governed by Law 241/1990, access to information about automated data processing involved in a contested decision is crucial for challenging that decision. Infact, it would seem be illogical to provide information that, due to its generality and vagueness, would make it impossible to exercise the right to contest provided by the regulation.

---

<sup>62</sup> "Incidentally, it has been highlighted that Article 15, unlike Articles 13 and 14, has the advantage of providing a right that can be exercised by the data subject rather than merely imposing an obligation on the data controller. Furthermore, it allows for overcoming the temporal limitations set by Articles 13 and 14, enabling the individual to obtain information even if the processing has already commenced, is currently being carried out, or has even resulted in a decision. This further underscores the importance of transparency for individuals affected by automated administrative activities in both investigatory and decision-making contexts." cfr. Cons. Stato, sez. VI, 4 febbraio 2020, n. 881

<sup>63</sup> The guarantee of transparency, aimed at ensuring, on a procedural level, the corresponding right to take action and defend oneself in court, is thus fulfilled in this case by the GSE's own acts (specifically, the notice initiating the procedure and the final decision). These acts not only outlined the reasons for the algorithm's modification but also, by explicitly providing the mathematical formula used, accompanied by a legend explaining the parameters employed, clearly detailed the computerized decision-making mechanism (the so-called "knowability"). Additionally, they explained its functioning in terms that are understandable even to users without technical expertise (the so-called "comprehensibility"). TAR Lazio, sez. IIIter, 20 febbraio 2024, n. 3402.

From the analysis of the articles just described, some conclusions can be drawn about the right to explanation enshrined in the GDPR.

First of all, the right to be informed should be distinguished from the right to explanation.<sup>64</sup> The substantial difference between these two rights mainly emerges when considering the "moment" when the user obtains information regarding the processing of their personal data. Receiving information usually refers to receiving *ex ante* notifications, through which individuals are informed from the outset, in general terms, about data processing or automated profiling.

In contrast, receiving an explanation implies obtaining *ex post*, specific, and objective information regarding how the data was processed and the decisions made concerning a particular subject.<sup>65 66</sup>

For the reasons mentioned above, concerning the strong connection between the right

---

<sup>64</sup> Gianclaudio Malgieri, Giovanni Comandé, *op.cit.* pp.243-265

<sup>65</sup> Carlo Colapietro, *Algorithms between transparency and protection of personal data* in *Federalismi.it - Observatory on transparency*, 22.02.2023, pp.151-174: "In this regard, note how the scope of application of the art. 15 compared to that of the articles. 13 and 14 are different: while the latter refer to a moment prior to the start of processing, and therefore justify the request for information relating exclusively to the data and processes in progress at that moment, art. 15 refers to a phase subsequent to the start of the processing, therefore legitimizing the request for further information, which cannot be the same as provided for by the articles. 13 and 14, but which may instead coincide, for example, with information relating to the new data generated by the algorithm, as well as information relating to the functioning, at the time of the request, of the algorithm itself, limited to "what is legally necessary to protect the person in the process of technological heterodimerization of his identity". There he refers to R. MESSINETTI, *The protection of the human person versus Artificial Intelligence. Decision-making power of the technological apparatus and right to explanation of the automated decision*, in *Contract and Business*, n. 2, 2019, pg. 885.

<sup>66</sup> Regarding the content and effects of Article 15, it is useful to refer to Recital 60 of the GDPR, which states that "the principles of fair and transparent processing imply that [...] the controller [provides] the data subject with any further information necessary to ensure fair and transparent processing, taking into account the specific circumstances and context in which the personal data are processed." Furthermore, it should be noted that, pursuant to Article 12 of the GDPR, the data controller has an obligation to facilitate "the exercise of the data subject's rights under Articles 15 to 22."

to challenge an automated decision and the right to access the necessary information to do so, I believe that the right to an explanation is a fundamental part of the GDPR framework.

However, the following hermeneutical clarification must be added.

Regardless of the greater or lesser effectiveness of the information provided under Articles 13, 14, and 15, it is evident that in all these cases, data subjects are not actively involved. They remain mere recipients of information, with their level of understanding or interaction not being taken into account, and thus the effectiveness of this information is not ensured.<sup>67</sup>

It is undeniable that the right to receive information or explanations increases transparency (providing details about the purpose, commercial goals, socio-legal implications, and concrete effects) but this does not automatically guarantee full understanding by individuals. Consider, for example, the numerical and mathematical complexity of algorithms, which can sometimes be difficult to interpret even for the data controllers themselves.<sup>68</sup>

Articles 13 and 14 require data controllers to provide *ex-ante* explanations to ensure that users understand the relevant algorithm. However, in this specific case, due to the type of information provided, often data subjects do not have the possibility to fully comprehend how the system in question works, as the approach used is not entirely understandable and, therefore, is not fully transparent or exhaustive. A similar reasoning can be applied to the right to access under Article 15 and the right to

---

<sup>67</sup> Francesco Sovrano, Fabio Vitali, Monica Palmirani, *Modeling GDPR-Compliant Explanations for Trustworthy AI*, September 2021, pp. 3-5 <https://arxiv.org/pdf/2109.04165>

<sup>68</sup> Francesco Sovrano, Fabio Vitali, Monica Palmirani, *op.cit.* pp. 3-5

explanation under Article 22, as read in conjunction with Recital 71.

The essential point, therefore, is that, in addition to the “what”, the information and explanations mentioned above, there is also the issue of the “how”: how this data is communicated to the data subjects.<sup>69</sup>

The problem seems to be not so much whether there is a right to an explanation, but rather how this right is implemented according to GDPR.

In this regard, the doctrine argues that the GDPR ensures both the legibility of data and algorithms, meaning the comprehensibility of how algorithms work, and the transparency of their use, referring to the commercial application of the algorithm. Legibility, in this context, means providing a clear and personalized explanation of how specific data is processed. According to the authors, this approach would resolve the debate between ex-ante and ex-post explanations (whether information is provided before or after the automated processing) and the issue of whether the right is to receive a specific explanation or just general information.<sup>70</sup>

#### **1.6. The absence of a right to explanations in the GDPR. Another perspective**

In the section above we have discussed the GDPR, provides for a right of individuals to receive explanations about automatic decision-making affecting them. For completeness of the analysis, it is appropriate to account for the opposing theses that argues that such right is not clearly provided for in the legislation, in any case not in a legally enforceable way. In this opinion, while the GDPR places a lot emphasis on

---

<sup>69</sup> Castets-Renard, C. *Accountability of algorithms in the GDPR and beyond: A European legal framework on automated decision-making*. Fordham Intellectual Property, Media and Entertainment Law Journal, 2019, Article 3, pp. 94-110 <https://ir.lawnet.fordham.edu>

<sup>70</sup> Castets-Renard, C, *op.cit.* pp.94-110

transparency and the rights of users, it makes no explicit provision for the right of an explanation about automatic decision-making.

This view is based on an analysis of the regulatory text, the legislative history, and the technical feasibility of explainability. It suggests that, while the regulation promotes transparency, it does not guarantee a legally binding right to receive explanations.<sup>71</sup>

The arguments proposed are as follow: there is a lack of explicit wording within the legally binding provisions of the GDPR.

Articles 13 to 15 provide that when automated decisions are taken data controllers “provide meaningful information about the logic involved”. It does not require data controllers to provide individual detailed explanation to each affected person.

Article 22, which regulates automated decision-making, provides for the right of individuals not to have decisions concerning them be taken exclusively by automated assessment but not the right to receive an explanation about them. This may be an important distinction: rather than assuming the GDPR creates a substantive right to receive explanations, a more limited type of transparency could be at stake.<sup>72 73</sup>

---

<sup>71</sup> Bryan Casey, Ashkon Farhangi, Roland Vogl, *Rethinking Explainable Machine: The GDPR’s “Right to Explanation” Debate and the rise of algorithmic audit enterprise*, 2019, Berkeley Technology Law Journal, pp.145-155, <https://doi.org/10.15779/Z38M32N986>

<sup>72</sup> Wachter, S., Mittelstadt, B., & Floridi, L. *Why A Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*. International Data Privacy Law, 2017, pp. 76–99

<sup>73</sup> A right to explanation derived from Article 22(3) is subject to several significant limitations. Article 22(2) outlines three exceptions where the general prohibition on automated decision-making does not apply: (a) when it is necessary for entering into or performing a contract; (b) when authorized by Union or Member State law that includes safeguards for data subjects; or (c) when based on the data subject’s explicit consent. However, the safeguards provided in Article 22(3)—including the right to human intervention, expression, and contesting a decision—only apply to cases meeting exceptions (a) or (c), but not (b). This means that when automated decision-making is authorized by law, no explicit safeguards are mandated, leaving their definition to national legislation.

Another significant discussion concerns the role of Recital 71, which is frequently cited in support of a Right to Explanation. However, under EU law, recitals may be used as interpretative guidelines do not contain binding provisions, which makes them unenforceable. While Recital 71 includes language suggesting an entitlement to explanations, the absence of specific wording in the legally binding Articles of the GDPR indicates that the legislator did not intend to establish a formal Right to Explanation. This interpretation is further supported by the legislative history, as earlier drafts of the GDPR contained stronger transparency requirements, including an more explicit Right to Explanation, which was later removed. This omission reflects both pragmatic concerns about the feasibility of enforcing such a right and a broader reluctance by regulators to impose stringent obligations on data controllers.<sup>74</sup>

In practice, and from a technological perspective, a genuine right to explanation presents significant challenges. Most modern machine learning models, particularly those based on deep learning, are inherently "black boxes" and even their developers often have little idea how specific outputs are produced. There are, of course, some explainability techniques, but even those provide approximations rather than proper explanations for concrete decisions. Yet, a legally required right to an explanation would, due to the lack of standardized methods to produce meaningful and consistent

---

Additionally, the exemption for contractual necessity presents further limitations. The GDPR does not define what qualifies as "necessary" for contract performance, creating a risk that data controllers may unilaterally determine necessity without requiring user consent. Unlike the explicit consent exception in Article 22(2)(c), the contractual necessity exception allows automated decisions without prior consent, meaning data subjects might not be aware of such decisions beyond general notification and access rights under Articles 13-15. While data subjects retain the right to contest, express their views, or seek human intervention for decisions covered under Article 22(3), they cannot object to the use of automated decision-making itself in these cases. Wachter, S., Mittelstadt, B., & Floridi *op.cit.* pp.76-99

<sup>74</sup> Wachter, S., Mittelstadt, B., & Floridi *op.cit.* pp.76-99

explanations, face considerable challenges regarding its enforcement. Besides, existing provisions of the GDPR already establish a framework of accountability without requiring a right to explanation. Impact assessments, transparency obligations, and the right to contest automated decisions are part of the provisions contained in the regulation.

Safeguards are implied for the data owner, without those safeguards becoming impracticably burdensome for the data controller. Instead of formalizing an individual Right to Explanation, the GDPR aims at procedural fairness and provides for oversight mechanisms, which better fits its overall objectives of data protection and accountability in the wider sense.<sup>75</sup>

Thus -according to this interpretation- the belief that a right to explanation may be assumed from provisions contained in the GDPR may be unfounded. The European Commission's 2012 proposal aimed to modernize data protection laws and enhance transparency, particularly through Article 22, which introduced safeguards against fully automated decision-making. However, while the regulation emphasizes transparency, it does not guarantee an enforceable right to explanation. While there are some transparency requirements, they do not amount to an enforceable right to know the exact reasons for an automated decision. The GDPR instead provide for the right of individuals to access general information on the data processing system and the potential consequences of its outcomes, but does not impose an obligation, for data

---

<sup>75</sup> Castets-Renard, C. *Accountability of algorithms in the GDPR and beyond: A European legal framework on automated decision-making*. *Fordham Intellectual Property, Media and Entertainment Law Journal*, 2019, Article 3, pp.94-11, 119-121

controllers to keep records of exhaustive individualized explanations as to how a single decision has been made.

In conclusion therefore, according to the opinion of some of the doctrine, the lack of explicit legal wording, the non-binding character of Recital 71, and the technical impossibility of explainability all confirm that a right to explanation does not exist under the GDPR.<sup>76</sup>

### **1.7. Application of the "explainability" principle: practical cases**

The shift from examining European regulations and related principles to verifying their practical application allows not only to identify their limits and natural application difficulties (which is normal for such an innovative discipline, applied to new phenomena in full and rapid evolution), but also to better define their scope and effectiveness.

For this reason, in completing this analysis, it seems useful to delve into the factual reality of applying the explainability principle, considering that the AI Act's entry into

---

<sup>76</sup> Critics argue that the scope of the GDPR's safeguards against algorithmic discrimination is limited, as Article 22 applies only to fully automated decisions, excluding cases involving minimal human intervention. Additionally, the requirement that automated decisions must produce "legal or similarly significant effects" is seen as restrictive, excluding cases like online behavioral advertising or price discrimination. Moreover, the "right not to be subject to automated decision-making" remains ambiguous, as it is unclear whether it prevents such processing outright or merely allows data subjects to block it. Another critique concerns the effectiveness of safeguards under Article 22(4), particularly the so-called "right to explanation," which is mentioned only in Recital 71, making it legally non-binding. Articles 13 to 15 primarily ensure notification of automated decision-making rather than detailed justifications for individual cases. Even Article 15, which allows for access to personal data, appears to focus on ex-ante rather than ex-post transparency. Furthermore, concerns exist over the ability of data controllers to withhold algorithmic details based on trade secret protections, as explicitly acknowledged in Recital 63. *Gianclaudio Malgieri, Giovanni Comandé, op.cit. pp. 6-7*

force is very recent.

As mentioned, the "explainability" requirement for automated decisions, prescribed by the regulation just examined, the GDPR, addresses the common need to ensure transparency in automated decision-making processes; transparency is a key element for an AI system that respects fundamental rights and freedoms, with humans as both the users and recipients. As has already been said, interpreters have immediately highlighted that one of the main limitations of the GDPR lies in its ability to guarantee the explainability of automated decisions.<sup>77</sup>

Its general formulation leaves considerable room for interpretation, creating uncertainty about its practical application. The clearest reference is contained in Recital 71, which suggests that individuals have the right to receive an explanation when their data are processed through automated decisions; however, this guidance, due to its normative location, does not find a binding application. Moreover, the right to an explanation, indirectly established by Article 15, paragraph 1, letter h), which provides the right of access to "meaningful information about the logic involved," does not appear to be supported by sufficient operational details to make it effectively enforceable. Therefore, the lack of a clear and unambiguous definition of what is meant by "explanation" under the GDPR is the first problem that risks leading to the development of divergent interpretations, resulting in inconsistent application among EU Member States.

In this regard, interesting, and very current, are the Conclusions of Advocate General

---

<sup>77</sup> Sandra Wachter, Brent Mittelstadt, Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, in *International Data Privacy Law*, Volume 7, Issue 2, May 2017, Pages 4-7, <https://dx.doi.org/10.2139/ssrn.2903469>

Jean Richard De La Tour, presented on September 12, 2024, in case C-203/22 pending before the Court of Justice of the European Union "*CK with the intervention of Dun & Bradstreet Austria GmbH, Magistrat der Stadt Wien.*" These conclusions will be examined by the Court for the purpose of deciding the case (not yet published), and they shed light on the delicacy and complexity of the application of this principle, further confirming the limits of the regulations already noted by interpreters.

The dispute concerns an issue in the credit scoring sector, where algorithms are used to determine consumers' creditworthiness: these companies, active in selling evaluations based on the automated collection and processing of personal data, did not comply with the information obligations set out by the GDPR, arguing that they were merely data providers. However, with the recent ruling in the Shufa Holding AG case (CJEU, 1st Chamber, 7.12.2023 - case C-634/21), the Court of Justice of the European Union ruled that even the mere processing of a score constitutes a "decision" under Article 22, as it produces significant effects on the legal situation of the individuals concerned.<sup>78 79</sup>

---

<sup>78</sup> Francesca Mattassoglio, *La Corte di giustizia europea, algoritmi e credit scoring. L'apertura del vaso di Pandora delle società che si "limitano" a elaborare gli scoring*, in *DB non solo diritto bancario.it, Dialoghi di diritto dell'Economia, Note*, 10 gennaio 2025, pp.1-14

<sup>79</sup> Judgment of the EU Court of Justice of December 7, 2023 (preliminary ruling request submitted by the Verwaltungsgericht Wiesbaden — Germany) — OQ / Land Hessen, Case C-634/21, SCHUFA Holding: "*Article 22(1) of Regulation (EU) 2016/679 of the European Parliament and of the Council of April 27, 2016, on the protection of natural persons with regard to the processing of personal data, as well as on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), must be interpreted as meaning that: The automated calculation, by a company providing commercial information, of a probability score based on personal data concerning an individual and relating to their ability to meet future payment obligations constitutes 'automated decision-making concerning natural persons' within the meaning of this provision, if such a probability score decisively determines the conclusion, execution, or termination of a contractual relationship with that individual by a third party to whom the probability score is communicated.*"

In the case C-203/22 examined by the Advocate General, an Austrian citizen was denied the conclusion or extension of a mobile phone contract, which would have involved a monthly payment of €10.00, due to insufficient financial capacity. The alleged insufficient solvency of the customer was justified based on an automated creditworthiness evaluation carried out by a scoring company specialized in providing such assessments.

At that point, the customer requested the relevant authorities in Austria, responsible for data protection, to acquire the information regarding the logic used in the automated decision-making process followed by the scoring company. This led to a legal dispute over what information the customer should have access to and what the scoring company was obliged to provide. This prompted Austrian judges to refer to the Court of Justice for an interpretative question regarding the right of access granted to the customer under Article 15, paragraph 1, letter h), of the GDPR, specifically concerning significant information on the logic used in the automated decision-making process related to their personal data.

This document seems extremely relevant for the present study, as it represents a factual application of the right to "explainability" established by the GDPR, addressing the definitional issue of the so-called significant information on the logic used in the context of the automated decision-making process concerning personal data processing.

First of all, the Advocate General's statement is noteworthy, emphasizing that the concept of "significant information about the logic used" in an automated decision-making process should be understood in a functional sense. That is, it must be

interpreted in light of the fact that "*the additional information obligations of the data controller and the related additional access rights of the data subject are explained by the purpose pursued by Article 22 of the regulation, which aims to protect individuals against specific risks to their rights and freedoms arising from automated processing of personal data, including profiling*" (par. 50 and 51 of the Conclusions).

This confirms the *ratio* of this right, which constitutes the essential prerequisite for ensuring the protection of fundamental rights and freedoms of individuals in the face of AI decisions.

Therefore, following this functional approach, the preferred interpretation is one that ensures the data subject to exercise their GDPR rights based on this information, including the right to express their opinion on an automated decision and, if necessary, challenge it. In simple terms, the accessible information should allow the data subject to defend themselves. To this end, the Advocate General emphasizes that the "explanation" must allow the data subject both to verify the accuracy of the personal data concerning them and the information related to the logic used in the corresponding automated decision, and to determine the existence of consistency and an objectively verifiable causal link between, on one hand, the method and criteria used and, on the other, the result reached by the automated decision.

In practice, the communicated information must allow the data subject to check "*whether it is truthful and, therefore, whether the automated decision in question is actually based on accurate information*" (par. 68 of the Conclusions).

On the basis of these premises, the Advocate General concludes by stating that "*the 'significant information on the logic used' in an automated decision-making process*

*must allow the data subject to exercise the rights granted to them under the GDPR and, in particular, under Article 22 of that regulation. This requires, first, that the person can obtain concise, easily accessible, and easily understandable information, formulated in simple and clear language, about the method and criteria used for that decision. Second, this information must be sufficiently complete and contextualized to enable the person to verify its accuracy and to determine whether there is consistency and an objectively verifiable causal link between, on the one hand, the method and criteria used and, on the other, the result reached by the automated decision"* (par. 71 of the Conclusions).

However, in his interpretation the Advocate General also addresses the issue of "how" this right can be concretely satisfied, taking into account the peculiarity of AI.

It should be added that, in the same document, the Advocate General also concludes that such an explanation does not necessarily include the obligation to provide the data subject with "*technical information that the latter would not be able to understand, such as the details of the algorithms used*" (par. 76 of the Conclusions). Therefore, the data controller would not be required to provide the algorithm used for processing.<sup>80</sup>

The arguments supporting this statement are mainly based on the purpose of the provision in question (Article 15, par. 1, letter h) of the GDPR, which requires

---

<sup>80</sup> It is useful to recall the Conclusions of the Advocate General expressed in the aforementioned judgment of the EU Court of Justice of December 7, 2023, OQ / Land Hessen, Case C-634/21, SCHUFA Holding, it is stated that: 58. "*For the reasons explained, I consider that the obligation to provide 'meaningful information about the logic involved' must be understood to include sufficiently detailed explanations of the meth used to calculate the score and the reasons for a certain result. In general, the controller should provide the data subject with general information, notably on factors taken into account for the decision-making process and on their respective weight on an aggregate level, which is also useful for him or her to challenge any 'decision' within the meaning of Article 22(1) of the GDPR.*"

providing significant information about the logic behind automated decision-making involving personal data. The goal is to ensure that the data subject can understand the information provided to them, enabling them to exercise their rights under the GDPR. Therefore, the Advocate General argues that understandable explanations are those that *"do not require particular technical expertise and are certainly more 'significant' than a complex mathematical formula"* (par. 73 of the Conclusions).

In simpler terms and to summarize, since the algorithm is often difficult to understand, even by experts, it would not fall under the "significant information" that the data controller is required to communicate to the individuals affected (and potentially harmed) by the algorithmic decision.

The interpretation offered by the Advocate General -which will need to be awaited to see whether it will be adopted by the Court- is that *"the regulation requires the data controller to provide significant information about the logic used, but not necessarily a complex explanation of the algorithms used or the disclosure of the complete algorithm (...). However, the information provided should be sufficiently complete for the data subject to understand the reasons behind the decision"* (par. 74 of the Conclusions).

With the further clarification that it is the responsibility of the data controller *"to ensure that the information provided is both accessible and complete so that the data subject can understand the process that led to the adoption of the automated decision they were subject to."*

The Advocate General further expresses that in case of a conflict between the exercise of the right of access provided in Article 15 of the GDPR and the rights of freedoms

of others, there should be a possibility to balance the rights and freedoms in question. In summary, he states that "*whenever possible, methods of personal data communication should be chosen that do not infringe the rights or freedoms of others, taking into account that such considerations should not lead to a refusal to provide the data subject with all the information*" (par. 90, 91).

The above interpretation, which is of significant relevance because it comes from an authoritative source and directly addresses the issue of explainability in relation to a factual case, still leaves open the issue of safeguarding the right to an effective remedy (Article 47 of the Charter of Fundamental Rights of the European Union). This is because excluding access to the algorithm, due to its difficulty in comprehension, there is a risk of limiting the data subject's right to defense.

Indeed, to make critical assessments on the automated decision, the data subject would largely have to rely on a "communication" from the data controller outlining the process and rationale behind the decision. This would likely be sufficient for the data subject's defense in most cases. However, this may not always be the case. It seems reasonable to assume that there could be situations where such communication is insufficient, and some understanding of the algorithm used would also be necessary.

In this context, the 'investigative' requests made by the court-appointed expert <sup>81</sup> in the

---

<sup>81</sup> Conclusions of Advocate General Jean Richard De La Tour, presented on September 12, 2024, in Case C-203/22 pending before the Court of Justice of the European Union "*CK v. with the intervention of Dun & Bradstreet Austria GmbH, Magistrat der Stadt Wien*":

*16. According to the appointed expert, in order to have a level of detail capable of supporting the issue of an enforcement order, in order to ensure the automated decision-making is intelligible and in order to verify the accuracy and consistency of the information provided, the data subject should receive, by virtue of the right of access guaranteed to him or her by*

underlying case that led to the European court proceedings are noteworthy, at least as an example. From reading the expert's instructions, it is clear that the complexity and extent of the information required are necessary for the individual to understand the reasons for the automated decision and, if necessary, challenge it. As outlined in the Conclusions, one might question whether the explanation obligation is sufficient to meet such detailed, complex "clarifications," as specified by the expert in the example considered.

---

*Article 15(1)(h) of the GDPR, in a sufficiently detailed and in-depth manner, the following minimum information:*

*– first, the personal data of the data subject which have been processed in order to formulate [credit assessment] factors and the basis on which those factors were formulated, specifying whether they have been weighted;*

*– secondly, the essential parts of the algorithm on which the automated decision-making is based, which in any event includes: the mathematical formula into which may be introduced, in the form of numerical values, all the information relevant to the calculation of the credit rating so that that formula produces that rating, and the intelligible explanation of all the values used in that formula, in particular those which are not directly derived from the information stored in respect of the data subject, and*

*– thirdly, the relevant information for establishing the link between the information processed and the valuation carried out, which includes, inter alia, a statement and adequate description of the valuation functions of all the values used in that formula; clarification of the information necessary to establish the link between the information and the valuation in the case of interval evaluations, and clarification of the land register or index functions used.*

*17. It is clear from the expert's report that only the disclosure of the mathematical formula and the valuation functions of all the values used in that formula would enable CK to understand the profile of her which was generated, meaning that it would be only on the basis of that information that she would be able to assert the rights conferred on her by Article 22(3) of the GDPR to express her point of view and challenge the decision based on automated processing.*

*18. According to that report, in order to enable the accuracy of the minimum information disclosed to be verified, D & B must also draw up and submit, in a relatively complete and detailed manner, and to serve as a basis for comparison, a list of all the information on at least 25 cases of comparable non-anonymised profiling which are contemporaneous with the profile generated in respect of CK of which were established using the same calculation rule."*

Moving to another area of application, which does not necessarily fall under the scope of the GDPR but rather concerns the protections afforded to citizens in the face of an automated decision, the opinion expressed by national administrative courts is particularly significant. These courts have, on several occasions, been called upon to adjudicate cases involving so-called "algorithmic decisions." This issue is particularly delicate, as it involves scenarios where the exercise of public authority is, at least in part, entrusted to artificial intelligence. The implications of these "algorithmic decisions" are evident in terms of safeguarding citizens' freedoms and respecting the limits of public authority.

Thus, in a legal framework different from the decisions of the European Court but motivated by similar reasons, the national administrative courts have also provided their interpretation of the explainability of automated decisions.

Although these are early rulings in a rapidly evolving field, a review of these decisions indicates a developing protective approach for individuals affected by automated decisions.

In the absence of specific regulation on the subject, national courts have had to address the issues raised by these new scenarios, relying on the principles and rules of the domestic legal system.<sup>82</sup>

---

<sup>82</sup> *"In the face of rapid progress, legislation may take time, but protection cannot, and the judiciary often becomes the frontline institution that must address the issue, having to provide a legal response, a solution, to the request for protection. This request arises unexpectedly, without prior announcements or definitions, through the challenge of a fully digital administrative act, which must still be examined, along with all other challenges to human decisions. As life evolves, so too does the administration, and digitalization permeates both. Progressively, irreversibly. All of this happens so quickly that there is no time to legislate in advance. And the first to address it is the judiciary. In Italy, this judiciary is the administrative court."* Luigi Carbone, *L'algoritmo e il suo giudice*, presentation at the conference "Digital

The fundamental rule adopted by administrative courts is that the use of algorithms in administrative activities does not exempt them from compliance with the principles that shape the legal system and govern such activities. Specifically, it has been explicitly stated that a "*fundamental need for protection in the use of so-called algorithmic tools is transparency in terms of the reasoning and/or justification of the decision*".<sup>83</sup>

It has been affirmed that "*the technical rule governing each algorithm remains, after all, a general administrative rule, constructed by humans and not by machines, and is then (only) applied by the latter, even if exclusively so*".<sup>84</sup>

Therefore, it has been deemed that, to verify the legitimacy of public administration's actions, there is a need to understand the programming language that translated legal rules into computer rules. As the doctrine has argued, this raises the issue of "*the proper formulation of the (computer) rules applied by the algorithm and, thus, the ability to verify the procedural model followed by the software to arrive at a specific decision. This, from an administrative law perspective, translates into a question concerning the correctness of the logical-legal reasoning that led to the creation of the act and its related justification*".<sup>85</sup>

In conclusion, according to the opinion of administrative case law, whenever an

---

*Administration – Daily Efficiency and Smart Choices,*" University of Naples Federico II, May 9–10, 2022, [www.giustizia-amministrativa.it](http://www.giustizia-amministrativa.it).

<sup>83</sup> Cons. Stato, sez. VI, 4.2.2020, n. 881, clarified that "*the use of algorithms must be correctly framed as an organizational tool, a procedural and investigative instrument, subject to the typical verifications of any administrative procedure, which remains the modus operandi of authoritative decision-making, to be carried out based on the legislation granting the power and the purposes assigned by that legislation to the public body holding the power.*"

<sup>84</sup> Cons. Stato, sez. VI, 8.4.2019, n. 2270

<sup>85</sup> Gherardo Carullo, "*Decisione amministrativa e intelligenza artificiale*", *Diritto dell'informazione e dell'informatica*, fasc. 3, 2021, pp. 431-461

automated administrative activity is involved using algorithms, verifying the legitimacy of the decision requires knowledge of the so-called "source code" of the algorithm, i.e., those computer rules written during the programming phase that govern the algorithm's functioning.<sup>86</sup>

The concept of understanding an automated decision presupposes knowing the algorithm "*according to an enhanced interpretation of the principle of transparency,*" whereby "*this understanding of the algorithm must be ensured in all its aspects: from its authors to the process used to develop it, to the decision-making mechanism, including the priorities assigned in the evaluation and decision-making process and the data selected as relevant. This is to verify that the criteria, assumptions, and outcomes of the automated process comply with the prescriptions and objectives established by law or by the administration itself upstream of that process and to ensure that the methods and rules on which it is based are clear – and consequently reviewable*".<sup>87</sup>

Based on this interpretation of the "understandability" of the algorithmic mechanism underlying the decision, the court also asserts the prevalence of the related principle of transparency over any "*confidentiality requirements of the companies producing such IT mechanisms*"<sup>88</sup>, given the clear connection of the algorithm with the exercise of public authority.

---

<sup>86</sup> "*The so-called 'source code' of software is represented by the text of a calculation algorithm written in a programming language aimed at defining the execution flow of the program. Specifically, it is a structured sequence of all the data and commands through which the programmer designs the software and enables its execution, concretely determining its operational methods.*" T.A.R. Lazio, sez III-bis, 1.07.2020, n. 7526.

<sup>87</sup> Cons. Stato, sez. VI, 4.2.2020, n. 881

<sup>88</sup> Cons. Stato, sez. VI, 4.2.2020, n. 881

The interpretive solutions examined here, concerning practical cases of applying the right to explainability of automated AI decisions, provide an initial idea of how this new regulatory framework may be implemented.

It can already be observed that both proposed interpretations share a common functional premise: the belief that this principle aims to safeguard the fundamental rights and freedoms of the subjects affected by such decisions. This protection can primarily be ensured by providing a mechanism for human oversight of the soundness of such decisions, whether they involve the exercise of public authority or purely private activities.

However, as is evident, practical doubts remain, both regarding the content of such an explanation (as illustrated by the scope of the information requested by the court-appointed expert in the substantive case that led to Case C-203/22 before the Court of Justice of the European Union) and the balance between protecting the fundamental rights and freedoms of the subjects affected by these decisions and the rights and freedoms of economic operators who hold the commercial rights related to this technology. This analysis of the right to explainability within the framework of the GDPR highlights the crucial role of transparency in ensuring the legitimacy of automated decisions, particularly when public authority is involved. However, the practical challenges of implementing this right, especially in balancing transparency with commercial confidentiality, remain an open issue.

Building on these considerations, the following chapter will explore how the AI Act further develops and refines the right to explainability, establishing specific obligations

for AI providers and deployers to ensure that high-risk AI systems operate in a transparent, accountable, and human-centric manner.

## Chapter II

### AI Act and explainability

The theme of the "explainability" of algorithmic decisions is also central in the very recent European discipline of artificial intelligence (AI) which, addressing the matter in its entirety, has a more systematic approach and a broader perspective than the GDPR. For the analysis of the explainability within the detailed provisions of EU Regulation No. 1689 of June 13, 2024 (the AI Act), which comprises 113 articles and 13 annexes, is helpful to first outline the broader framework of "Artificial Intelligence".

#### 2.1. Artificial intelligence and the challenge of harmonizing its definition

The term Artificial Intelligence (hereafter also referred to as AI) is employed in multiple fields of human knowledge, both in the scientific and cultural realm, often with vastly different meanings.<sup>89</sup> For this reason it is very challenging to give to the

---

<sup>89</sup> The scientific community has provided multiple interpretations of this term. For instance, referring to its early stages: "*ensuring that a machine acts in ways that would be considered intelligent if a human were to behave in the same way*" in A. Turing, "*Computing Machinery and Intelligence*," Oxford University Press on behalf of the Mind Association, 1950, pag. 433-460, excerpt from: Gianluca Giannini - Antonio Pescapé (Luca Lo Sapio), "*AI e futuro di sapiens tra nuovi orizzonti ed antichi timori*," in *Scienza e Filosofia.com*, pp.28-42, <https://www.scienzae filosofia.com/2022/07/05/ai-e-futuro-di-sapiens-tra-nuovi-orizzonti-e-antichi-timori/>

More recently, according to Salmovico, artificial intelligence is "*that discipline, belonging to computer science, which studies the theoretical foundations, methodologies, and techniques that enable the design of hardware systems and software systems capable of providing electronic computers with performances that, to a common observer, would seem to pertain exclusively to human intelligence*," Marco Somalvico and Francesco Amigoni and Viola Schiaffonati, *Intelligenza Artificiale*, pp.1-17 <https://schiaffonati.faculty.polimi.it/pubblicazioni/H1.pdf>

In literature, too, we find various references to intelligent machines. Among the many, Jules Verne in the 19th century and Isaac Asimov in the 20th century are among the most well-

world AI a definition which may captures unequivocally its meaning.

Furthermore, due to its many different definitions and areas of application, the term Artificial Intelligence has, over time, acquired a sort of mystical aura, evoking, at times, a feeling of apprehension that enhances its ‘mythical status’ but perhaps distances it from its real and actual meaning. To illustrate this point, one need only recall the so-called "Frankenstein complex,"<sup>90</sup> a concept formulated by Asimov to describe "*human dread toward an intelligent machine that, having achieved a high level of autonomy, might rebel against its master.*"<sup>91</sup>

On the other hand, the very words "intelligence" and "artificial," which have been chosen to identify this particular branch of scientific research, can be misleading as they tend to imply a forced comparison: "intelligence", being an ability inherent to

---

known, or again L. Frank Baum, who in *The Wizard of Oz* described the mechanical man Tik-Tok in 1907 as "*Extra-Responsible Mechanical Man, Thinks, Speaks, Acts, and Does Everything Except Live.*"

One of the first films to feature artificial intelligence is "*Metropolis*" (1927) directed by Austrian filmmaker Fritz Lang, with the character of Maria-robot, perceived by all as human, but in reality a robot with human-like features. This was followed in 1968 by Stanley Kubrick's "*2001: A Space Odyssey*": a journey to the boundaries of space and time aboard the *Discovery One*, a spaceship guided by HAL9000, a computer endowed with artificial intelligence that interacts with humans and replicates mental activities. Lastly, "*Blade Runner*" by Ridley Scott, released in 1982, redefined the emotional and cognitive sphere of androids, no longer simple robots, but humanoids, resembling humans not only in physical appearance but also in their way of relating to themselves and reality.

<sup>90</sup> The Frankenstein complex, in psychology, is understood as the fear of what is new, unfamiliar, or not fully comprehended, and in particular the fear that humans' creations might come to life and rebel against them. Arising from the nineteenth century novel 'Frankenstein' it has been developed by the writer Isaac Asimov in his science fiction novels where it can be used to frame the complex relationship between people and automatons. A similar reflection arises from the work of Japanese engineer Masahiro Mori, *The Uncanny Valley* (1970), which describes the discomfort we experience when observing robots or human-like objects that appear almost realistic, but not entirely, causing an emotional aversion.

<sup>91</sup> Mauro G. Smiraldo, *Il dottor Frankenstein e le responsabilità nella robotica*, in *magazine.atlante.società*, Treccani.it, April 2024 <https://www.treccani.it/magazine/atlante/societa/il-dottor-frankenstein-e-le-responsabilita-nella-robotica.html>

(and exclusive to) humans, and "artificial", indicating something fabricated and material, entirely alien to anything human. It is, therefore, inevitable that the very expression, Artificial Intelligence, by its specific reference to human capability, assumes anthropomorphic connotations, alluding to a subjectivity defined as “*other*”,<sup>92</sup> foreign to the concept of “human” yet simultaneously capable of learning from humans, and potentially, not only of becoming autonomous but also (perhaps) of overpowering its own “creator.”

All this often leads to fear and unease toward AI, which, in turn, hinders the true understanding of its profound meaning.

Artificial Intelligence, therefore, represents a phenomenon that remains partially unknown, one that requires regulation<sup>93</sup> both to help us understand it and to approach it in a direct and deliberate way.

The first step to gain a full understanding of the term is to clearly define it. Before analyzing the concept of Artificial Intelligence and navigate through its multiple definitions, it is necessary to understand what is meant by "intelligence" in relation to humans. This task proves quite challenging, as the notion of "intelligence" is highly fluid and variable. Not only does common sense offer different interpretations of the meaning of intelligence, but even scientific studies on the subject are far from

---

<sup>92</sup> G. Finocchiaro, *Intelligenza artificiale. Quali regole?*, Bologna, Il Mulino, 2024, Chapter 1

<sup>93</sup> The need for precise regulation of the phenomenon of Artificial Intelligence, aimed at providing a unified and harmonious interpretative solution, as well as helping to understand what is being regulated, recognizing both its risks and potential, appears to be largely addressed by the Artificial Intelligence Regulation (so-called “AI Act”) approved by the European Parliament on March 13, 2024, and by the Council of the European Union on May 21, 2024. Available in “Regulation (EU) 2024/1689”, so-called *Artificial Intelligence Act*, in the *Official Journal of the European Union*, July 12, 2024.

unanimous, creating a controversial conceptual landscape in which the various theories on the nature of intelligence are presented in different, sometimes contradictory, ways.<sup>94</sup> The classic definition of intelligence, as reported by the Oxford Advanced Dictionary, is identified as: “*the ability to learn, understand and think in a logical way about things; the ability to do this well.*”<sup>95</sup> From this definition, the fundamental elements that characterize intelligence emerge, namely: “*the ability to do something, understand, learn.*”

However, psychology and cognitive sciences have over time partially diverged from this definition, disputing the erroneous notion that there is only one form of intelligence. Since the late 1990s<sup>96</sup>, thanks to the research of the American psychologist Howard Gardner, scientists speak of multiple intelligences. According to Gardner, intelligence should not be considered a unitary and measurable ability but a number of different abilities: he introduces the concept of emotional, linguistic, logical-mathematical, musical, bodily-kinesthetic and interpersonal intelligence. In subsequent years, Gardner expanded the types of intelligences he identified, adding

---

<sup>94</sup> Howard Gardner, *Formae mentis. Saggio sulla pluralità dell'intelligenza*, Milano, Feltrinelli, 2013, pp.1-ss

<sup>95</sup> D. L. J. Bradbery, *Oxford Advanced Learner's Dictionary Paperback*, Oxford, OUP, 2011. [https://www.oxfordlearnersdictionaries.com/definition/american\\_english/intelligence](https://www.oxfordlearnersdictionaries.com/definition/american_english/intelligence)

<sup>96</sup> The theory of multiple intelligences, proposed by psychologist Howard Gardner and addressed in Andrea Castiello D'Antonio's essay "Intelligenza Artificiale, psicologia e psicologia delle organizzazioni" (in “*Personale e Lavoro*”, November 2021, no. 638, pp-1-19), suggests that intelligence cannot be reduced to a single measure but is rather a composition of various human abilities or "forms of mental representation." Gardner highlights the uniqueness of each individual, emphasizing how cultural context, personal history, personality, and interests play decisive roles in intellectual development, providing the foundation for an educational environment that is both inclusive and stimulating, thereby responding to the diverse needs and aspirations of students. In this light, human intelligence appears inseparable from the empathetic component, arising from the ability to recognize others, self-awareness, and self-consciousness—abilities that, as Castiello D'Antonio asserts in the cited work, “*machines will find it difficult to develop in the future.*”

two more: naturalistic and existential intelligence.

Gardner's work also challenges "*the assumption that intelligence, however it may be defined, can be measured by standardized verbal instruments such as paper-and-pencil tests based on short-answer responses to question batteries*",<sup>97</sup> given that intelligence has multiple dimensions, and IQ<sup>98</sup> tests capture only a limited portion of them.

The possibility to measure "human" intelligence is certainly an important issue to address, especially within the context of this study. In some way, the idea that intelligence can be measured makes it "neutral," almost abstracted from subjective evaluation, thereby bringing it closer to the concept of standardized and objective ability.<sup>99</sup>

The criteria for identifying what constitutes "human intelligence" are, therefore, crucial for defining, or at least attempting to define, what constitutes Artificial Intelligence.

---

<sup>97</sup> Howard.Gardner, *op.cit*, pp.1-ss.

<sup>98</sup> In the study of the possible measurement of intelligence, Alfred Binet, a pioneering French psychologist, played a fundamental role in the development of experimental psychology. He devised a test and a scale to measure the intelligence of French students, aimed at identifying those who might require additional support in their educational journey. Binet's test was later improved with the introduction of the concept of mental age, which also led to the development of the IQ concept, thanks to contributions from L.M. Terman of Stanford University (hence the name Stanford-Binet scale). After the war, David Wechsler, working at Bellevue Psychiatric Hospital in New York, sought to reduce the test's excessive focus on verbal abilities. <https://doi.org/10.1590/0004-282X20170097>

In 1939, he published the first version of what would become the famous Wechsler Intelligence Scales, the most widely used IQ test by psychologists today, which is divided into the WAIS (Wechsler Adult Intelligence Scale) and WISC (Wechsler Intelligence Scale for Children). The Origins of the Stanford-Binet 5, the WAIS-IV, the WISC-V, and the WPPSI-IV Subtests Aisa Gibbons and Russell T. Warne Utah Valley University, pp.1-32 <https://osf.io/uwh2s/download/?version=1&displayName=Subtest%20origins-2018-09-18T19%3A18%3A30.846Z.pdf>

<sup>99</sup> G. Finocchiaro, *Intelligenza artificiale. Quali regole?*, Bologna, Il Mulino, 2024, Chapter 1

AI is understood to be a computer system with the ability to simulate human cognitive functions such as learning or problem-solving, or, as defined by the European Commission in its communication of April 25, 2018, AI "*refers to systems that display intelligent behavior by analyzing their environment and taking actions, with some degree of autonomy, to achieve specific goals.*" <sup>100</sup>

It is "intelligence" because it emulates human cognition; it is "artificial" because it processes information based on a computational rather than biological grounding. However, this remains a "fluid" definition, subject to change with technological evolution <sup>101</sup>. The functioning of AI relies on progressive learning algorithms, capable of teaching a machine not only how to perform a task but also how to develop its thinking based on acquired experience.<sup>102</sup>

With the rapid evolution of technology, not only does AI learning become increasingly swift, but it also becomes more autonomous. Today, precisely because of this autonomous processing capability, AI is branching out in different forms, with numerous connections and commonalities among them; the most significant of which currently are Machine Learning and Deep Learning, which can be defined as techniques of automatic learning.

In practice, anything that can be digitized (e.g. numbers, images, videos, words, likes) is stored and inserted into a Machine Learning algorithm. Through the use of such

---

<sup>100</sup> *Comunicazione della Commissione europea al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni, L'intelligenza artificiale per l'Europa*, COM (2018) 237, 25 aprile 2018, 1.

<sup>101</sup> Lucia G. Sciannella, *Intelligenza artificiale, politica e democrazia, Breve introduzione all'Intelligenza Artificiale*, DPCE Online, 51 (1), 2022 pp.337-348 <https://doi.org/10.57660/dpceonline.2022.1577>

<sup>102</sup> Lucia G. Sciannella, *op.cit* pp. 337-340

algorithms, AI can analyze data, identify patterns, classify objects, predict outcomes, and make decisions.<sup>103</sup>

More specifically, Machine Learning refers to a machine's ability to imitate human behavior and intelligence by creating models derived from direct experience. The applications of Machine Learning are manifold; think, for instance, of the daily suggestions for TV series on Netflix or the content recommendations on Instagram.<sup>104</sup>

Deep Learning, on the other hand, can be defined as a subset of Machine Learning, with a much deeper and faster level of learning, capable of learning increasingly complex concepts/visual models. While Machine Learning absorbs and processes thousands of data points, Deep Learning applications process millions of data points, generating texts and other multimedia content, such as sounds, images, and videos. Moreover, in the case of Deep Learning, algorithms are also able to self-regulate based on the experience gained from previous data analyses.<sup>105</sup>

Thus, AI not only learns quickly but also learns independently. And as Alan Turing stated in a 1951 interview: "*if a machine can think, there is the possibility that it will think more intelligently than us. And where might that lead us?*"<sup>106</sup>

This statement is particularly significant as it comes from someone who is rightly considered one of the fathers of computer science today. At this point, for a better

---

<sup>103</sup> Janiesch, C., Zschech, P., & Heinrich, K., *Machine learning and deep learning*. *Electronic Markets*, 31, 2021, pp. 685–695. <https://doi.org/10.1007/s12525-021-00475-2>

<sup>104</sup> Janiesch, C., Zschech, P., & Heinrich, K., *op.cit* pp. 685-693

<sup>105</sup> Janiesch, C., Zschech, P., & Heinrich, K., *op.cit* pp. 685-693

<sup>106</sup> E. Blakemore, *Il legame tra la nuova IA e il test di Turing: chi era davvero l'uomo che lo ha inventato?*, in *nationalgeographic.com*, 2024

understanding of the topic of AI, it seems useful to briefly outline the evolution of this technology.

## **2.2. From Turing to the AI Act: the evolution of artificial intelligence and its regulation**

Alan Turing was born in London in 1912. He was a mathematician and cryptographer. The technology related to the development of AI indeed originates from the computational model produced by Turing, known as the Automatic Computing Engine (ACE), which dates back to 1936 and is also referred to as the universal machine. This model was based on the theoretical belief that a computing tool, such as the universal machine, could simulate and reproduce human mental functions and capabilities.<sup>107</sup>

This revolutionary principle forms the basis of the modern understanding of Artificial Intelligence.

Turing's machine operated on the basis of information received, by processing it and giving the requested output. According to Turing, his machine could execute any kind of calculation.<sup>108</sup>

In 1950, Turing devised a test, later known as the Turing Test<sup>109</sup>, to determine whether a machine could be classified as “intelligent.” This test remains one of the fundamental references of modern AI.

---

<sup>107</sup> Frixione, M., & Numerico, T. , *Alan Mathison Turing*. APhEx, 7. Periodico online ISSN 2036-9972, 2013, pp.511-562  
<https://www.openstarts.units.it/server/api/core/bitstreams/0c3efe39-ae91-4779-ad54-6fda38c1205e/content>

<sup>108</sup> Frixione, M., & Numerico, *op.cit.*, pp.511-562

<sup>109</sup> Alan M. Turing, *Computing Machinery and Intelligence*, Oxford University Press on behalf of the Mind Association, 1950, p.433  
<https://phil415.pbworks.com/f/TuringComputing.pdf>

The test involves three parties: an interrogator, a human being and a machine; the conditions of the test are such that the interrogator does not know which one of respondents is the human and which is the machine. The interrogator poses a series of questions to both respondents with the purpose to determine which one is the human and which is the machine. If the interrogator attributes at least 30% of the answers given by the machine to the human, the machine has passed the test. Essentially, a machine is deemed "intelligent" if it can effectively simulate human cognitive functions and interactions, thereby "fooling" a human about its non-human nature. This represents the first standardization of what intelligence -especially as understood today- means in the context of artificial intelligence: a thinking machine that autonomously performs human-like actions, learns, and engages in dialogue as a human would.<sup>110</sup>

Turing's work on artificial intelligence, which was rooted in the belief that human mental functions could be simulated by a machine, laid the groundwork for the early forms of AI and machine learning. In the 1950s, American mathematician John McCarthy coined the term "artificial intelligence": in August 1956, McCarthy and other researchers, including Marvin Minsky, Nathaniel Rochester, and Claude Shannon, organized the Dartmouth Summer Research Project on Artificial Intelligence; the final proceedings of the conference outlined the scope of their study:

---

<sup>110</sup> Frixione, M., & Numerico, *op.cit.*, pp.511-562

*“the study will proceed on the basis of the conjecture that, in principle, every aspect of learning or any other characteristic of intelligence can be described so precisely that a machine can be built to simulate them.”*<sup>111</sup>

Thus, the Dartmouth conference marked the birth of artificial intelligence as a field of study.

Subsequently, the early explorations of language processing by machines led in 1966 to the creation by Joseph Weizenbaum, a professor at MIT, of ELIZA. Named after the main character in George Bernard Shaw's play "Pygmalion," ELIZA was a chatbot that simulated being a psychologist. In reality, ELIZA merely responded to users' statements by rephrasing them as questions, but its impact and success with the public were overwhelming. Weizenbaum, realizing the program's disruptive effects, including inducing users, who had minimal knowledge of psychology, to fully trust it, took a step back from promoting his creation, highlighting the ethical concerns it raised and the necessity for regulation.<sup>112</sup>

Today, the impact of ELIZA on the public appears remarkably relevant. Consider ChatGPT, a modern reinterpretation of ELIZA, capable of engaging in precise and detailed conversations on any topic, leveraging a network endowed with billions of parameters and unparalleled computational power.<sup>113</sup>

---

<sup>111</sup> J. McCarthy, M.L. Minsky, N. Rochester e C.E. Shannon, C.E. (2006), *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 31, 1955, AI Magazine, 27(4), pp.12-14 <https://doi.org/10.1609/aimag.v27i4.1904>

<sup>112</sup> Joseph Weizenbaum, *Il potere del computer e la ragione umana I limiti dell'intelligenza artificiale*, Edizioni Abele, 1987, pp. 35-ss

<sup>113</sup> Floyd, Christiane. *From Joseph Weizenbaum to ChatGPT: Critical Encounters with Dazzling AI Technology*. Weizenbaum Journal of the Digital Society, Research Paper, University of Hamburg, pp.1-29 <https://doi.org/10.34669/WI.WJDS/3.3.3>

Given that the term Artificial Intelligence <sup>114</sup> has multiple interpretations and it is used in many different fields, it has become necessary to reach a consensus on its definition, and mark the limits of the legitimate uses it may be put to. The AI Act can be seen as an answer to this need, aligned with the previously mentioned concept of regulations aimed at “understanding” what is being regulated.<sup>115</sup>

The AI Act is the result of an harmonization process aimed at ensuring the correct, effective, and safe use of AI systems across Europe. Its complete implementation will ensure a unified regulation and management of AI systems in all European member states, leading to a common understanding of the concept of Artificial Intelligence throughout the region.<sup>116</sup> As shown in the brief analysis above, this normative harmonization process also aims to define certain relevant concepts clearly, which are not always univocal.

---

<sup>114</sup> Among other distinctions, two macro categories of artificial intelligence can be identified: weak AI and strong AI. Weak AI is an artificial intelligence system designed to perform specific tasks. This particular type of AI is the most widespread. Generally, this AI utilizes capabilities such as machine learning, which develops by providing the machine with a series of relevant data related to the application area, allowing the machine to also make predictive analyses. Strong AI, on the other hand, also known as general artificial intelligence, is theoretically comparable to human intelligence, possessing an autonomous ability for evaluation and problem-solving.

<sup>115</sup> The authors, Alessandro Pajno, Filippo Donati, Antonio Perrucci, *Intelligenza artificiale e diritto: una rivoluzione? Diritti fondamentali, dati personali e regolazione (Vol.1)*, Torino, Il Mulino, 2022, pag. 97-ss.; spec. pag. 98, have written on the issue of the development of AI: «[...] the development of such tools, considering the social, economic, and biological implications associated with the use of artificial intelligence, does not seem to be something that can be entrusted solely to market forces. It is the role of the law to regulate this technology, in order to virtuously guide its functioning and align it with the pursuit of public interest objectives underpinning the actions of legislators and regulators.».

<sup>116</sup> European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: *A Digital Single Market Strategy for Europe*, COM(2015) 192 final, Brussels, May 6, 2015, p. 3. As of August 30, 2024: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015DC0192>

According to the definition provided in the Article 3<sup>117</sup> of the AI Act, an Artificial Intelligence system is a specific type of technology designed to operate at various levels of autonomy, utilizing its capacity for adaptation post-implementation. The AI Act addresses a “machine” designed to perform tasks and achieve specific goals, which can be categorized as either “explicit” or “implicit”. To meet these goals, AI systems analyse the data received as inputs and use this analysis to produce outputs. These outputs can include predictions, personalized content, recommendations, or decisions. Another characteristic of AI systems is autonomy, as defined by the AI Act. An AI system may function entirely autonomously or under a certain degree of human supervision. Furthermore, an AI system can modify its behavior or enhance its performance based on experience or data acquired over time. This adaptation process allows the Artificial Intelligence system to enhance its own capabilities autonomously or semi-autonomously.

It is clear that the definition provided by the AI Act encompasses all possible technical and legal aspects of AI, and as such is an essential tool to identify its object, limits, and obligations (along with the corresponding levels of responsibility); highlighting the necessity for an interpretation of the normative provisions concerning AI. Therefore, the guidelines which will be adopted by the European Commission and the legal instruments which will be enacted by the individual European states to implement the AI Act will have a crucial role to play.

---

<sup>117</sup>AI Act, Article 3: “AI system means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.”

### **2.3. The need for ethical reflection: a focus on explainability and transparency**

The revolutionary nature of "AI" technology, capable of emulating the ability to "act" and "think" like humans, in addition to having a certain degree of autonomy, has led to the emergence of a series of ethical and legal issues related to its use. This has prompted the need for reflection aimed at identifying the appropriate rules and limits to ensure the respect of the rights and freedoms of users or any individuals affected by the activities of "AI."

In this regard, it is important to remember that in the past, Artificial Intelligence was regarded as a mere tool that may not, in itself, harm or benefit its users. In practice, AI technology was viewed more as an object of intellectual speculation, philosophical and scientific in nature and, due to the fact that it may have the most diverse purposes, it was considered ethically neutral.

However, the recent increased awareness of the potential harms and threats arising from AI has sparked a profound ethical reflection on the proper use of this new technology.

Today, given the mass use and pervasive presence of digital technology in the daily lives of every individual, Artificial Intelligence can no longer be considered a neutral technology<sup>118</sup> or an accessory and an inconsequential element in human life. It is now believed that Artificial Intelligence should contribute to improve human life while

---

<sup>118</sup> *“Until recent times (the mid-2010s), artificial intelligence was considered neutral, as it provided society with tools that users were free to utilize as they wished, without involving the responsibility of the producing entities—whether companies or public laboratories—so long as these entities had properly informed potential users about the performance of their products. Furthermore, as long as these entities complied with the legal provisions governing their activities, no particular issues arose concerning the conditions under which they created, produced, and disseminated these tools.”* Daniel Andler, *Il Duplice Enigma, Intelligenza artificiale e intelligenza umana*. Torino, Piccola biblioteca Einaudi, 2024, p.345

simultaneously avoid the risk of exacerbating existing ills or creating new ones. In order to ensure that, the development of Artificial Intelligence systems should always be grounded on and accompanied by ethical debates and practices.<sup>119</sup>

Ethics is a necessary discipline for defining the limits of the application and use of Artificial Intelligence systems, serving as a support for legal and normative debate on the enactment of new regulations in this field.<sup>120</sup> AI systems cannot be regarded as neutral for human life, as it has been demonstrated that they actively and significantly influence the lives of those who come into contact with them and use their services.<sup>121</sup> In order to address these issues, the European legislator has aimed to define a comprehensive set of rights and obligations through the enactment of the AI Act for both the producers and users of artificial intelligence systems. Additionally, they have established an institutional framework with corresponding procedures to regulate the introduction of AI systems to the market.

It should be noted, for the purposes of this discussion, the absolute prominence and relevance of this new regulatory framework, of the principles of "transparency" and "explainability"; these constitute requirements that must be fulfilled by high-risks Artificial Intelligence systems<sup>122</sup> introduced in the European common market. These

---

<sup>119</sup> Bertoncini, A. L. C., Serafim, *Ethical content in artificial intelligence systems: A demand explained in three critical points*. *Frontiers in Psychology*, 14, 1074787, 2023, pp.1-10. <https://doi.org/10.3389/fpsyg.2023.1074787>

<sup>120</sup> Bertoncini, A. L. C., Serafim, *op.cit* pp.1-3

<sup>121</sup> G. Finocchiaro, *Intelligenza artificiale. Quali regole?*, Bologna, Il Mulino, 2024, Chapter 1

<sup>122</sup> The Ai Act is based on AI risk-based approach which means that classifies Ai systems into several risk categories, with different degrees of regulation applying. It is possible to analyze different types of AI systems (classified by the possible risks). Those ones are: the prohibited AI practices; the High-risk AI systems; the Transparency risks AI systems; Minimal risks AI systems; the General-purpose AI systems. In particular, the High-risk AI systems depend on

principles form the foundation of the entire AI Act, which can be considered its “cornerstone”; producers must guarantee, in accordance with the terms and conditions set by the AI Act, that AI-based systems meet these requirements. Compliance with this obligation is monitored by EU institutions and member states.

In other words, with the entry into force of the new EU regulations, producers of AI systems will be required to implement solutions that improve the "transparency" and "explainability" of their models.

The purpose of the AI Act, evidently, is to prevent and sanction, where appropriate, the improper use of Artificial Intelligence, which, given AI’s multiple functions and great power, could cause the violation of the users' fundamental rights, such as the right to privacy and confidentiality, and even harm their health and safety.

The European legislator has considered furthermore that, often, the improper use of Artificial Intelligence systems is due to an inadequate understanding of these systems by their users. Based on this observation, the AI Act grants users the right to have high-risk AI systems be 'transparent' and 'explainable,' and imposes the corresponding obligation on producers and providers to ensure these requirements are met; the respect of the above rights (and meeting of the above obligations) should guarantee to users the degree of awareness and understanding of the AI systems necessary to maintain

---

the particular purpose pursued and the modalities for which the systems are used. High-risk AI systems can be safety components of products covered by sectoral EU law or AI systems that, as a matter of principles, are considered to be high-risk when they are used in specific areas. A new test has been enshrined at the Parliament’s request (‘filter provision’) according to which AI systems will not be considered high-risk if they do not pose a significant risk of harm to the health, safety or fundamental rights of natural persons. However, an AI system will always be considered high-risk if the AI system performs profiling of natural persons. EU legislation in progress: pp.1-12 <https://www.iisf.ie/files/UserFiles/cybersecurity-legislation-ireland/EU-AI-Act.pdf>

"human" control over the "autonomous" decisions made by these systems.<sup>123</sup>

On the other hand, "transparency" and "explainability" are essential in order to establish and maintain a relationship of trust between users and AI algorithms and, ultimately, to ensure the proper development of a common market for Artificial Intelligence. It is evident that the inability to understand how an algorithm makes a particular judgment or decision makes it difficult to challenge or correct any errors or injustices, fostering an atmosphere of distrust around this tool.<sup>124</sup>

To protect users' interests and sustain their trust towards AI systems it must be ensured that users are able to understand how AI algorithms work and operate; every explanation regarding AI's working must be communicated to users in a transparent, clear, and timely manner. For this reason, any AI model must be "explainable": it must be designed and developed in such a way that it is able to explain and justify its decisions clearly and simply.

As it is well known, however, explainability is very problematic with Artificial Intelligence models which are very complex and hyper-technical, and as a consequence, users' rights are not fully guaranteed. The issue is clearly explained by referring to the so called "Black Boxes" i.e. extremely sophisticated AI systems. The term "Black Box" in fact alludes to the opacity of the internal workings of the system, a system where the inputs and outputs are known, but the processes occurring in

---

<sup>123</sup> Gyevnar B., Ferguson N., and Schafer B., *Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?* School of Informatics, University of Edinburgh; Edinburgh Law School, University of Edinburgh. 2023, pp.1-9 [arxiv.org/abs/2302.10766](https://arxiv.org/abs/2302.10766).

<sup>124</sup> Gyevnar B., Ferguson N. and Schafer B., *op.cit.* pp. 1-9

between are not understood.<sup>125</sup>

This lack of transparency leads to several critical consequences, including the difficulty of identifying biases, errors, or discrimination in the decisions made by the AI system. This problem becomes particularly critical when Artificial Intelligence operates in sensitive areas such as healthcare, justice, personnel selection, or financial credit.<sup>126</sup>

In this context, the concept of so-called "Explainable AI"<sup>127</sup> (hereinafter: XAI) is crucial, as it may provide a solution to the "black box" problem and ensure the protection of the "transparency" and "explainability" rights established by the AI Act. The concept was introduced in 2004 by Van Lent to indicate the set of processes and methods of AI which can provide explanations on the behavior of entities it controls in simulation games. Although the term is recent because the issue of "explainability"

---

<sup>125</sup> Gryz Jarek and Rojszczak Marcin, *Black box algorithms and the rights of individuals: No easy solution to the "explainability problem*, Internet Policy Review, 10(2), pp.1-6  
<https://doi.org/10.14763/2021.2.1564>

<sup>126</sup> Gryz Jarek and Rojszczak Marcin, *op.cit.* pp.1-6

<sup>127</sup> *Explainable Artificial Intelligence (XAI)* refers to the inherent ability of artificial intelligence systems to provide clear and comprehensible explanations regarding their actions and decisions. The main goal of these particular systems is to make the operations performed and the decisions made understandable to the individuals requesting them. However, often the explanations provided are accessible only to experts in the field, leaving users excluded from the understanding process. (*Tech Dispatch, Explainable Artificial Intelligence*. This publication is a brief report produced by the Technology and Privacy Unit of the European Data Protection Supervisor (EDPS); What is explainable artificial intelligence? ([https://www.edps.europa.eu/system/files/2023-11/23-11-16\\_techdispatch\\_xai\\_en.pdf](https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf))). The approach promoted by Explainable AI can be divided into two categories: *self-interpretable models*, where explainability is determined by the design of the system itself, and *post-hoc explanations*, through which the behaviour of the models is explained only after an initial observation. Emmanuel Pintelas, Ioannis E. Livieris, and Panagiotis Pintelas; *Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability, Interpretability in Machine Learning*, 2020, in Algorithm, 2020, pp. 2-17  
<https://www.mdpi.com/1999-4893/13/1/17>

dates back to the mid-1970s when researchers began investigating the explanations provided by expert systems.<sup>128</sup>

According to the “Defense Advanced Research Projects Agency” (DARPA), the purpose of XAI is to develop more comprehensible AI models while preserving high learning performance (i.e., accuracy in predictions); and also to help human users to understand, adequately trust, and effectively manage the new generation of intelligent artificial partners.<sup>129</sup> Thus, XAI aims to make Artificial Intelligence models understandable not only to experts but also to ordinary users, providing clear and accessible explanations on the internal decision-making processes of the AI systems. AI’s ability to explain its own decisions in a transparent and understandable manner should bridge the gap between technical complexity and the end user, thereby strengthening user trust and awareness, as aimed by the AI Act.<sup>130</sup>

## **2.4. The AI Act: an introduction to its complexity**

### **2.4.1. Overview of the AI Act**

On May 21, 2024, the Council of the European Union approved Regulation (EU) 2024/1689, commonly known as the *Artificial Intelligence Act (AI Act)*. This is the

---

<sup>128</sup> Ali S. Abuhmed, T. El-Sappagh, S. Muhammad, K. Alonso-Moral, J. M. Confalonieri, R. Guidotti, R. Del Seri, J. Díaz-Rodríguez, N. & Herrera, F. M *Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence*. Information Fusion, 99, 2023.p.7 <https://doi.org/10.1016/j.inffus.2023.101805>

<sup>129</sup> D. Gunning, David W. Aha, *DARPA’s Explainable Artificial Intelligence Program*, 2019, pp. 44-58 <https://doi.org/10.1609/aimag.v40i2.2850>

<sup>130</sup> Ali S. Abuhmed, T. El-Sappagh, S. Muhammad, K. Alonso-Moral, J. M. Confalonieri, R. Guidotti, R. Del Seri, J. Díaz-Rodríguez, N. & Herrera, *op.cit.* pp.7-8

world's first legislative<sup>131</sup> act regulating Artificial Intelligence systematically, providing for AI systems specifically designed to ensure the development and use of controlled, safe AI systems that respect the rights promoted and protected by the European Union (EU).

From a technical standpoint, the EU opted for a *Regulation* as a normative instrument because of its immediate and *erga omnes* effectiveness in the legal systems of the Member States, making it the most suitable tool for creating a common market for AI systems in Europe. This aims to lay the groundwork for digital innovation and promote investments in this field. The intention outlined in the introductory report of the AI Act proposal is to establish the EU as "a *global leader in the development of safe, reliable, and ethical artificial intelligence*".<sup>132</sup>

The EU has defined this first regulatory approach to AI as "horizontal" because it governs these systems in general rather than in their specific applications, while also addressing particular themes such as ethical aspects, responsibility, and copyright issues.<sup>133</sup> Additionally, this new regulation fits within the broader measures introduced

---

<sup>131</sup> "The European regulation proposal on artificial intelligence aims to serve as a global reference model. It is the first legislative act that seeks to regulate the entire sector" (Giusella Finocchiaro, *Intelligenza artificiale. Quali regole? il Mulino*, 2024).

Within the Proposal for a Regulation of the European Parliament and Council establishing harmonized rules on Artificial Intelligence (AI Act) and amending certain legislative acts of the Union, 2021, it states that "*The Union's interest is to preserve the technological leadership of the EU and ensure that European citizens can benefit from new technologies developed and operating in accordance with the values, fundamental rights, and principles of the Union.*" 1. Contesto della Proposta; 1.1 Motivi e obiettivi della proposta: <https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:52021PC0206>

<sup>132</sup> Madiaga, T. *Artificial Intelligence Act*. EPRS, European Parliamentary Research Service, PE 698.792. European Union, 2024, pp. 1-13 [https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS\\_BRI\(2021\)698792\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf)

<sup>133</sup> Giusella Finocchiaro, *op.cit.*, pp. 53-55

by the EU to regulate the digital realm; therefore, its implementation will need to be coordinated with various European provisions governing the matter (e.g., the Digital Market Act, the Data Act, and the Data Governance Act).

In summary, the EU aims to define the legitimacy limits of AI systems to outline what might pose a real or potential threat to the rights of EU citizens.

Consequently, the Act seeks to establish which systems can become positive tools for addressing the economic and social challenges of the future and to contribute proactively to the development of the Union.<sup>134</sup>

#### **2.4.2. Structure and purpose of the regulation**

Structurally, Regulation 2024/1689 is a substantial legislative document, consisting of 180 Recitals and 113 Articles divided into 13 Chapters, which are further subdivided into Sections.

The dual purpose pursued is to lay down the legal foundation for developing reliable, human-centric AI systems that respect the health, safety, and fundamental rights of EU citizens while also creating a clear and certain legal framework for the development and dissemination of this new technology within the common market. These factors will increasingly determine the economic and social well-being of the Union in the future.

While this is not the venue for a detailed examination of Regulation 2024/1689, it is important to highlight the approach taken by the European legislator to regulate this

---

<sup>134</sup> European Commission, Shaping Europe's Digital Future, AI Act <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?>

phenomenon. The regulation of these new technologies is based on the “*level of risk*” that they pose, among others, in respect of the health, safety, fundamental rights of users, democracy, the environment, and the rule of law.<sup>135</sup> This approach to product safety is similar to those applied to other products, such as cars or medicines.

The AI Act defines several risk levels, ranging from the highest to the lowest risk. The logic is that the greater the risk, the more stringent the rules that must be followed, even reaching the point of outright banning the use of those Artificial Intelligence systems that pose an unacceptable risk to the interests protected by the Union. For instance, AI systems identified as a clear threat to users or designed to materially distort user behavior or cause significant harm to individuals are prohibited (Article 5 of the Regulation).

Particularly stringent rules apply to AI systems classified as “*high risk*,” such as those intended for real-time and retrospective remote biometric identification of individuals and those used for security purposes. Conversely, AI systems classified as lower risk are only subject to voluntary compliance, as the regulation is not fully binding in their respect.

In summary, Regulation 2024/1689, when regulating the introduction and use of AI systems, not only prohibits certain systems but also establishes a series of requirements for high-risk systems (Articles 6-49). It sets general transparency rules for specific AI systems (Article 50) and introduces specific regulations for general-purpose AI systems with broad capabilities (Articles 51-56). The regulatory framework is further supported by measures to encourage technological innovation within the EU (Articles

---

<sup>135</sup> Madiega, *op.cit.*, pp.1-13

57-63) and a governance structure that includes oversight of AI systems and penalties for unlawful conduct (Articles 64-101).

The complexity of the issue is such that the legislator has planned a gradual implementation of this new regulation. Indeed, while Regulation 2024/1689 came into effect on August 1, 2024, the entire regulatory framework is expected to be applied progressively, observing the steps and timelines established in the Regulation (Article 113). Finally, by August 2, 2025, Member States must designate national authorities responsible for monitoring the proper application of the AI Act's regulations. These national authorities will refer to the European AI Office as their governing body.<sup>136</sup>

## **2.5. Explainability and transparency**

As noted in the previous paragraph, the most relevant and substantial aspects of the regulation concern high-risk AI systems.

Focusing on these systems, we observe that the European legislator, with the intention of adopting a well-balanced regulation that takes into account both the need to develop the new technologies in the "common market" and the necessity to safeguard the foundational principles and values of the European Union (i.e. respect for fundamental rights, the rule of law, democracy, and the environment), introduces mechanisms and procedures that combine an *ex-ante* assessment of AI systems with *ex-post* monitoring and control activities.<sup>137</sup>

Section II of Chapter III of the Regulation is dedicated to the "requirements for high-

---

<sup>136</sup> Madiega, *op.cit.*, pp.1-13

<sup>137</sup> Madiega, *op.cit.*, pp.1-13

risk systems" and provides that the introduction into the market and the use of such AI systems may occur only on the basis of two premises (which can be considered the cornerstones of this new regulation): the system's compliance with certain mandatory requirements and the adoption of a "risk management system."<sup>138 139</sup>

It is important to emphasize that these are two interrelated factors, as the proper management of the risks associated with the use of an AI system is made possible, if not primarily, through compliance with the established mandatory requirements.

The "risk management system" is a process involving the analysis and control of the risks that the use of the AI system may adversely affect users' fundamental rights, as well as health and safety. The Regulation describes it as an "iterative, continuous process, planned and executed throughout the entire lifecycle of a high-risk AI system, requiring constant and systematic review and updating" (Art. 9, par. 2).

The declared objective of this ongoing and repeated analysis is to enable the AI system provider to adopt "*appropriate and targeted risk management measures intended to*

---

<sup>138</sup> AI Act, Section 2-3, Chapter III

<sup>139</sup> AI Act, Article 17: "Quality management systems": the areas considered by this quality system are:

1. *"Establishes quality criteria for the datasets used for training, validation, and testing of high-risk AI systems. These datasets are subject to specific governance practices and must be relevant, representative, error-checked, and aimed at the prescribed purposes.*
2. *Imposes on producers and deployers the obligation to draft technical documentation to demonstrate that the high-risk AI system complies with the established requirements and contains the necessary information to verify such compliance.*
3. *Requires AI systems to be designed in a way that allows for tracking their operation with maximum transparency. To this end, AI systems must always be accompanied by concise, accurate, easy-to-understand, and comprehensive user instructions to make users' decision-making processes as informed as possible.*
4. *Mandates that AI systems are programmed with human-machine interface tools that allow for effective human oversight, minimizing risks to health, safety, and fundamental rights.*
5. *Requires all AI systems to ensure an adequate level of safety, accuracy, robustness, and cybersecurity that lasts throughout the system's lifecycle."*

*address the identified risks"* (Art. 9, par. 2, letter d).

The provision clearly states that the focus of this analysis should be on risks that can reasonably be reduced or eliminated through the development or design of the high-risk AI system, or by providing adequate technical information. (Art. 9, par. 3). The goal is to assess whether there are risks associated with these AI systems and to adopt appropriate measures to prevent harm to users. In practice, before introducing an AI product to the European market, it must demonstrate compliance with legal requirements, such as the quality of data used, the presence of technical documentation, traceability, transparency, human oversight, and cybersecurity robustness. This means that if there are other risks that cannot be mitigated or eliminated by these measures, the AI system should not be allowed to enter the market. This seems confirmed by the fact that the Regulation does not exclude the existence of a residual risk in high-risk systems placed on the market, provided that it is an "acceptable" risk (Art. 9, par. 5). The management measures to be adopted are described through functional terms in the regulation, without going into the details of their technical characteristics, which are left to the decisions and assessments of those who provide the AI system. In particular, these measures must:

- eliminate or reduce identified risks *"as much as possible from a technical standpoint through the appropriate design and manufacturing of the high-risk AI system"*;
- implement *"appropriate mitigation and control measures where necessary to address risks that cannot be eliminated"*;
- provide *"the information required under Article 13 (a provision that, as we will*

*see later, regulates the transparency requirement) and, where necessary, training for deployers."*

Finally, a procedure has been established to test the effectiveness of the adopted measures, consisting of trials that can be conducted "*at any point during the entire development process and, in any case, before placing the system on the market or putting it into service*" (Art. 9, par. 6, 7, 8).

Recital 66 of the Regulation states that, in the risk management of high-risk AI systems, the requirements to be met include, among others, transparency and the provision of information to deployers, as well as human oversight (Articles 13 and 14 respectively).

As mentioned before, Chapter III of the Artificial Intelligence Act (AI Act) specifically focuses on high-risk artificial intelligence systems, outlining a rigorous regulatory framework to ensure their safe and ethical development and use.

The first section of this chapter is dedicated to the classification of AI systems as high-risk. Article 6 defines the specific criteria that allow for the identification of such systems, while Article 7 introduces any amendments to Annex III, a comprehensive list of AI systems that are automatically considered high-risk.

The second section, on the other hand, lists the fundamental requirements that high-risk AI systems must meet.

Key principles guiding this new regulatory framework include "human oversight" (Article 14, Chapter III) and the obligation of "transparency" for AI systems (Article 13, Chapter III and Article 50, Chapter IV).

By imposing these specific requirements on high-risk AI systems, the AI Act aims to

promote users' trust and accountability in the use of these systems. As it will be explored further, both requirements are of particular (though not exclusive) relevance in addressing the ethical issues raised by the regulation of AI systems and they are the subject of this analysis.

### **2.5.1. Article 13**

Article 13 is titled "*Transparency and the provision of information to deployers*" and contains three main provisions.

AI systems must be designed and developed to ensure transparency, enabling users to understand and properly use the results they produce. The required transparency does not only apply to the algorithm itself but also it extends to the practical and operational information that must be provided to the user. This implies an obligation for providers to supply to their customers, together with the systems, clear, comprehensive, and accessible user instructions detailing the features, capabilities, and limitations of the AI system.

Paragraph 1 establishes the obligation to design and develop high-risk AI systems: "*in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately.*"

The first paragraph focuses on the design phase of high-risk AI systems. It mandates that these systems must be designed to ensure legitimate and safe use. In other words, developers must take all necessary measures to prevent the system from being used for harmful or discriminatory purposes. This paragraph emphasizes the importance of transparency and accountability in the design of these technologies.

The second paragraph addresses the issue of instructions that must accompany high-risk AI systems.

It concerns the instructions “*for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers.*”

The goal is to ensure that users can fully understand the system's operation and use it correctly. Instructions must be written in clear and simple language, avoiding technical jargon that could confuse the user. Moreover, they must provide detailed information about the potential limitations and risks associated with using the system.

The third and final paragraph specifies the information that must be included in the user instructions.<sup>140</sup>

---

<sup>140</sup> “3. *The instructions for use shall contain at least the following information:*  
(a) *the identity and the contact details of the provider and, where applicable, of its authorized representative;*  
(b) *the characteristics, capabilities and limitations of performance of the high-risk AI system, including:*  
(i) *its intended purpose;*  
(ii) *the level of accuracy, including its metrics, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity;*  
(iii) *any known or foreseeable circumstance, related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety or fundamental rights referred to in Article 9(2);*  
(iv) *where applicable, the technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to explain its output;*  
(v) *when appropriate, its performance regarding specific persons or groups of persons on which the system is intended to be used;*  
(vi) *when appropriate, specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the high-risk AI system;*  
(vii) *where applicable, information to enable deployers to interpret the output of the high-risk AI system and use it appropriately;*  
(c) *the changes to the high-risk AI system and its performance which have been predetermined*

This includes provider identity, purpose and functioning, limitations and risks, input data, and impact on users.

In summary, Article thirteen of the AI Act represents an important step towards regulating these emerging technologies.

From the examination of this rule, some general considerations can be formulated. At the initial design stage of the system, there is a requirement to ensure its functioning is “*sufficiently transparent*,” meaning it should be transparent enough for users to understand and use the system's output consciously. The standard mentioned in the provision, in line with the EU's proportionality principle, allows providers some discretion in determining what level of transparency can be considered sufficient. On the other hand, the required transparency seems to focus primarily on understanding the system, as it is intended to enable users to “*understand how the AI system works, assess its functionality, and comprehend its strengths and limitations*” to help them “*use the system and make informed decisions*”.<sup>141</sup>

Therefore, this requirement seems to focus on the overall understanding of how the AI system works, rather than on comprehending the specific reasons behind its outputs.<sup>142</sup>

In line with this approach, paragraph 2 of the provision establishes the obligation to

---

*by the provider at the moment of the initial conformity assessment, if any;*  
*(d) the human oversight measures referred to in Article 14, including the technical measures put in place to facilitate the interpretation of the outputs of the high-risk AI systems by the deployers; (e) the computational and hardware resources needed, the expected lifetime of the high-risk AI system and any necessary maintenance and care measures, including their frequency, to ensure the proper functioning of that AI system, including as regards software updates;*  
*(f) where relevant, a description of the mechanisms included within the high-risk AI system that allows deployers to properly collect, store and interpret the logs in accordance with Article 12.”*

<sup>141</sup> Gyevnar B., Ferguson N., and Schafer B., *op.cit.* pp.1-9

<sup>142</sup> Gyevnar B., Ferguson N., and Schafer B., *op.cit.* pp.1-9

accompany high-risk AI systems with appropriate *"instructions for use"* upon their entry into the market; a list of the information that should be included in these instructions is also provided. Some of the information is prescribed as mandatory across the board, as for example *"the human oversight measures referred to in Article 14, including the technical measures implemented to facilitate the interpretation of the outputs of high-risk AI systems by deployers"*. Other information to be provided, however, is qualified by expressions such as *"where applicable"*, as for example, in the case of *"the technical capabilities of the high-risk AI system related to the provision of relevant information to explain its output"* or *"the information that enables deployers to interpret the high-risk AI system's output and use it appropriately"*.

In other cases, required information is qualified by expressions like *"where necessary"*, as for example, in the case of *"specifications for input data or any other relevant information in terms of training, validation, and testing datasets, considering the intended purpose of the high-risk AI system."*

Clearly, in these cases, the actions and steps needed to comply with the regulatory provisions allow for some discretion on the part of the system provider. This discretion will, however, require further clarification and definition when the regulation is applied in practice.

From the above it can be deduced that the AI Act aims to promote transparency through an approach that relies primarily on supplying documentation and detailed information to users, such as instructions for use and other informative materials. This provision does not impose on AI providers the obligation to use specific Explainable AI (XAI)

tools.<sup>143</sup>

In other words, rather than establishing strict obligations regarding explanatory models or techniques, the AI Act focuses on ensuring that users are informed and capable of using artificial intelligence systems appropriately and consciously.<sup>144</sup> This approach requires providers to develop and implement their models following a transparent design, possibly adopting XAI techniques, without being bound by rigid regulatory requirements in this regard.<sup>145</sup>

Finally, it's important not to overlook another key aspect of the rules on "transparency". The required documentation, which includes information about the system's features, capabilities, performance limitations, and human oversight measures, is essential not only for protecting users but also for certification and market monitoring by the relevant authorities.

### **2.5.2. Article 14**

Article 14 is dedicated to "Human oversight" and imposes a specific obligation that high-risk AI systems must be *"designed and developed, including with appropriate human-machine interface tools, so as to be effectively overseen by natural persons during their period of use"* (Art.14, par.1).

---

<sup>143</sup> Tech Dispatch, Explainable Artificial Intelligence, *op.cit.* pp.1-16

<sup>144</sup> Cecilia Panigutti, David Fernandez Llorca, Salvatore Scalzo, Ronan Hamon, Delia Fano Yela, Gabriele Mazzini, Isabelle Hupont, Henrik Junklewitz, Ignacio Sanchez, Josep Soler Garrido, Emilia Gomez, *The role of explainable AI in the context of the AI Act*. In 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA. ACM, New York, NY, USA, pp.1139-1147 <https://dl.acm.org/doi/pdf/10.1145/3593013.3594069>

<sup>145</sup> Tech Dispatch, Explainable Artificial Intelligence, *op.cit.* pp.1-16

In practice, AI systems must be structured in a way that allows for effective supervision by human operators. This means that human operators must fully understand the capabilities and limitations of the system, be able to monitor how it works, and, above all, be aware of the risks associated with over-reliance on the results generated by the system.

In this case, the right to explainability is linked to the obligation to ensure human oversight of this technology, considering, moreover, that the fundamental feature of AI is to possess a certain "level of autonomy" (Art. 3, par.1, n.1).

Article 14 is divided into 5 main paragraphs.

Paragraph (1) mandates that high-risk AI systems must be designed to facilitate human oversight, allowing for effective monitoring and control. This ensures that humans can oversee the system's operation and intervene when necessary.

Paragraph (2) outlines the purpose of human oversight, which is to prevent or mitigate potential risks to health, safety, and fundamental rights that may arise from the use of these systems. This oversight is crucial, especially in cases of unintended or unforeseen consequences.

Paragraph (3) specifies the nature of oversight measures. These measures must be proportionate to the system's level of risk, autonomy, and intended use. They can be implemented either as technical safeguards built into the system ("when technically feasible") or as procedures followed by the system's deployer.

Paragraph (4) details the capabilities that human (defined as "natural persons") overseers must possess. They should be able to understand the system's limitations, monitor its operation, avoid over-reliance on its output, interpret its results, and

intervene when necessary.

Indeed, it is required that human intervention should "*properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation*" and "*correctly interpret the high-risk AI system's output*"<sup>146</sup>. Furthermore, the AI system must also ensure that the "natural person" responsible for "human oversight" "*remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system*", which refers to the so-called "automation bias." This even extends to the possibility of deciding "*not to use the high-risk AI system*" or "*to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure*".

The paragraph (5) of the EU AI Act introduces specific safeguards for high-risk AI systems used for biometric identification.

It mandates that any identification made by such a system must be independently verified by at least two human operators with the necessary expertise before any action or decision is taken based on it. However, this requirement may be waived in certain specific contexts, such as law enforcement, migration, border control, or asylum,

---

<sup>146</sup> The requirement of human oversight is outlined in Article 14 of the AI Act and further elaborated in Recital 48. Article 14 establishes that AI systems must be designed in a way that allows for effective human oversight during their operation. Specifically, individuals assigned to human oversight should be able to fully understand the capacities and limitations of high-risk AI systems, correctly interpret their outputs, and remain aware of the risks of over-reliance on automated results, known as "automation bias."

Recital 48 further clarifies that, where appropriate, human oversight measures should ensure that AI systems are responsive to human operators and that those responsible for oversight have the necessary competence, training, and authority to intervene when needed, including the ability to stop or interrupt the system's operation if required. Cecilia Panigutti, David Fernandez Llorca, Salvatore Scalzo, Ronan Hamon, Delia Fano Yela, Gabriele Mazzini, Isabelle Hupont, Henrik Junklewitz, Ignacio Sanchez, Josep Soler Garrido, Emilia Gomez *op.cit.* p.1144

where it is deemed disproportionate.

Also, with reference to this rule some general considerations can be formulated. This oversight measure, also conceived to prevent or minimize the risks arising from the use of high-risk AI systems, which must be designed and managed in such a way that, "*where appropriate and proportionate*", the person tasked with oversight not only has the ability (a) to monitor properly the functioning of the AI system, taking into account the risk of malfunction caused by the so-called "automation bias" (i.e., the phenomenon of excessive human reliance on the machine), but also (b) to interpret correctly the system's output. For this reason, and to ensure the effectiveness of the measure, the Regulation requires that oversight be entrusted "*to natural persons who possess the necessary competence, training, and authority, as well as the necessary support*" (Art. 26, par. 2).

It is significant that, since this is "human oversight", the Regulation expressly provides that this measure should intervene particularly when the application of other requirements is ineffective and the risks arising from the use of the AI system persist (Art. 14, par. 2). Among the measures that the person charged with the oversight may adopt is not to use the AI system or even to switch it off (Art. 14, par. 4, letters (d) and (c)). This should manifestly and concretely affirm human control over artificial intelligence.

Explainability in this context materializes as the act of understanding by the human supervisor, whose task is to ensure that the artificial intelligence system operates appropriately and without risk. This process of understanding is essential to ensure the system's compliance, enabling the supervisor to consciously evaluate the AI's

decisions and behaviors.<sup>147</sup>

In this way, explainability is not limited to the system's technical ability to make its mechanisms understandable, but it includes the power conferred to a human to verify and intervene so that the system operates ethically, safely, and in accordance with the law. For instance, the AI system must "*guide and inform the natural person entrusted with human oversight so that they can make informed decisions regarding the possibility, timing, and manner of intervention, in order to avoid negative consequences or risks, or to stop the system if it is not functioning as intended*" (Recital 73).

### **2.5.3. Article 86**

In the context of explainability, further consideration must be given to the provision contained in Article 86 of the Regulation, titled "right to an explanation of individual decision-making processes."

Article 86<sup>148</sup> is part of Section IV (Remedies) of Chapter IX (Post-Market Monitoring,

---

<sup>147</sup> Cecilia Panigutti, David Fernandez Llorca, Salvatore Scalzo, Ronan Hamon, Delia Fano Yela, Gabriele Mazzini, Isabelle Hupont, Henrik Junklewitz, Ignacio Sanchez, Josep Soler Garrido, Emilia Gomez, *op.cit.* p.1142

<sup>148</sup> Article 86 "Right to Explanation of Individual Decision-Making": "1. Any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system listed in Annex III, with the exception of systems listed under point 2 thereof, and which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.

2. Paragraph 1 shall not apply to the use of AI systems for which exceptions from, or restrictions to, the obligation under that paragraph follow from Union or national law in compliance with Union law.

3. This Article shall apply only to the extent that the right referred to in paragraph 1 is not otherwise provided for under Union law."

Information Sharing and Market Surveillance). This section of the EU AI Act focuses on the measures to be taken after a high-risk AI system has been placed on the market. It outlines the procedures for addressing any potential issues or risks that may arise.

The provision grants the right to any individual affected by a decision made based on a high-risk AI system, and who suffers harm to their health, safety, or other fundamental rights as a result of that decision, "*to obtain clear and meaningful explanations from the deployer regarding the role of the AI system in the decision-making process and the main elements of the decision taken*".

The importance of this provision is that “explainability”, understood as the transparency of the AI's algorithmic process, focuses not so much on the overall functioning of the system, such as its capabilities, characteristics and limitations, but rather on the specific decision made in the individual case of the user who requests an explanation.<sup>149</sup>

This measure thus appears to operate on a somewhat different level compared to previous provisions and, at least based on an initial examination, seems to have a more profound impact. Indeed, according to Recital 171, in such cases, "*the explanation should be clear and meaningful and provide a basis on which affected persons can exercise their rights.*"<sup>150</sup>

---

<sup>149</sup>Cecilia Panigutti, David Fernandez Llorca, Salvatore Scalzo, Ronan Hamon, Delia Fano Yela, Gabriele Mazzini, Isabelle Hupont, Henrik Junklewitz, Ignacio Sanchez, Josep Soler Garrido, Emilia Gomez, *op.cit.* p.1142

<sup>150</sup> AI act, Recital 171: “*Affected persons should have the right to obtain an explanation where a deployer’s decision is based mainly upon the output from certain high-risk AI systems that fall within the scope of this Regulation and where that decision produces legal effects or similarly significantly affects those persons in a way that they consider to have an adverse impact on their health, safety or fundamental rights. That explanation should be clear and meaningful and should provide a basis on which the affected persons are able to exercise their*

#### **2.5.4. Preliminary considerations on the right of explainability under AI Act**

An analysis of the AI Act's provisions on transparency and explainability reveals, first and foremost, a different approach compared to the GDPR. The AI Act is essentially based on “transparency by design” and a system of control over AI products, both at the time of their market entry and throughout their lifecycle, with the aim of ensuring AI transparency.

On the other hand, compared to the GDPR, the AI Act introduces a specific provision regarding the “right to an explanation.” Indeed, Article 86 explicitly regulates the “*right to an explanation of individual decision-making processes*”.

In my opinion, this alone represents an innovation compared to the GDPR, as it states more clearly (and verbatim) the existence of such a right for AI users within the European legal framework.

The issue that will need to be addressed in the near future concerns the scope of this right and the manner in which such an “explanation” will be provided taking into account the peculiarities of AI.

From a textual perspective, two observations can be made regarding this provision.

The wording of Article 86 differs from that of Article 15 of the GDPR. While the latter refers to “*meaningful information about the logic involved, as well as the significance and envisaged consequences of such processing for the data subject*”, Article 86 concerns “*clear and meaningful explanations about the role of the AI system in the decision-making process and the key elements of the adopted decision*”.

---

*rights. The right to obtain an explanation should not apply to the use of AI systems for which exceptions or restrictions follow from Union or national law and should apply only to the extent this right is not already provided for under Union law.”*

The scope of the AI Act’s provision appears more specific: it refers to explanations rather than mere information.

Moreover, there is no doubt that such “explanations” must pertain to the specific decision that, according to the affected party, has infringed upon their rights. Therefore, this does not concern general information on how the system operates (as provided for in Articles 13 and 14 of the GDPR).

Additionally, the explanation must clarify the “role” played by AI in the decision-making process. For example, it must indicate whether the decision was fully automated or whether, and to what extent, a human intervened in its making. This aspect is absent from the GDPR provisions and, in my view, could be a key factor in understanding how and why a given decision was reached.

However, the provision is also clear in limiting the scope of the explanation to the key “elements” of the decision.

Considering that Article 86 is the result of an amendment introduced by the European Parliament to the original proposal from the Commission, it is worth noting that the initial wording of the amendment, as presented for discussion in the assembly<sup>151</sup> was

---

<sup>151</sup> Position of the European Parliament, Text Presented A9-0188/2023: 13.06.2023 Procedure: 2021/0106(COD) Amendment 630 – Proposal for a Regulation with a View to Adopting the Regulation of the European Parliament and of the Council Establishing Harmonized Rules on Artificial Intelligence and Amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797, and (EU) 2020/1828 (Artificial Intelligence Regulation). Article 68quater (new): *Right to Explanation of Individual Decision-Making Processes*

1. *Individuals subject to a decision made by an operator based on the output of a high-risk AI system that produces legal effects or similarly significant impacts on them, in a way they perceive as negatively affecting their health, safety, fundamental rights, socioeconomic well-being, or any other rights derived from the obligations*

slightly different from the final version approved by assembly.<sup>152</sup> The original version referred to: “*clear and meaningful explanations, pursuant to Article 13(1), on the role of the AI system in the decision-making process, the key parameters of the decision taken, and the relevant input data.*”

The approved version refers “*the right to obtain clear and meaningful explanations from the deployer regarding the role of the AI system in the decision-making process and the key elements of the adopted decision*”.

---

*established in this regulation, have the right to request clear and meaningful explanations, pursuant to Article 13(1), regarding the role of the AI system in the decision-making process, the main parameters of the decision made, and the related input data.*

2. *Paragraph 1 does not apply to the use of AI systems for which Union or national law provides exceptions or limitations to the obligation under paragraph 1, insofar as such exceptions or limitations respect the essence of fundamental rights and freedoms and constitute a necessary and proportionate measure in a democratic society.*
3. *This article applies without prejudice to Articles 13, 14, 15, and 22 of Regulation 2016/679.”* [https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236\\_IT.html](https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_IT.html)

<sup>152</sup> European Parliament, *Position approved at first reading on March 13, 2024, in view of the adoption of the regulation of the European Parliament and of the Council establishing harmonized rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144, and Directives 2014/90/EU, (EU) 2016/797, and (EU) 2020/1828 (Artificial Intelligence Regulation)*. EP-PE\_TC1-COD(2021)0106,

Art. 86, Right to an Explanation of Individual Decision-Making Processes. “*Any affected person subject to a decision made by the deployer based on the output of a high-risk AI system listed in Annex III, except for the systems listed under point 2 of the same Annex, that produces legal effects or similarly significant impacts on that person in a manner they perceive as negatively affecting their health, safety, or fundamental rights, has the right to obtain clear and meaningful explanations from the deployer regarding the role of the AI system in the decision-making process and the key elements of the adopted decision. Paragraph 1 does not apply to the use of AI systems for which exceptions or limitations to the obligation established in this paragraph are provided under Union law or national law, in accordance with Union law. This article applies only to the extent that the right referred to in paragraph 1 is not otherwise provided for under Union law*”. [https://www.europarl.europa.eu/doceo/document/TC1-COD-2021-0106\\_IT.pdf](https://www.europarl.europa.eu/doceo/document/TC1-COD-2021-0106_IT.pdf).

Given that the final legislative text adopts a more generic (and it seems to me broader) wording it seems reasonable to conclude that the explanation should encompass both the “key parameters” and the “input data”.

Unfortunately, even in this case, the provision is not entirely clear, for example, it remains doubtful whether or not the "explanation" should also include the algorithm code or, on what assumptions, it is possible for the user to know it.

This creates an interpretative issue that partially mirrors the one arising from the GDPR provisions (Articles 13, 14, 15, and 22). In this regard, it will be necessary to consider that the AI Act states in its recitals that the right to an explanation under Article 86 is functionally linked to the ability of users or affected parties to protect their rights. Indeed, Recital 171 explicitly states “*that explanation should be clear and meaningful and should provide a basis on which the affected persons are able to exercise their rights*”.<sup>153</sup>

Thus, an interpretative challenge may arise, with ethical implications, regarding the scope and content of this “explanation.” This is particularly important because the rights in question concern individuals' health, safety, and fundamental rights, as explicitly stated in the first paragraph of Article 86.<sup>154</sup>

---

<sup>153</sup> During the discussion on the proposal for AI regulation in the European Parliament on 13.06.2023, Danish MEP Karen Melchior (Renew) recalled the will of the assembly: “*We ensure that users of AI have the right to an explanation, because for citizens to trust AI, they have to understand the decisions that it makes. And finally, we ensure that citizens have a right to recourse because decisions made by AI should never be final.*”[https://www.europarl.europa.eu/doceo/document/CRE-9-2023-06-13-ITM-008\\_IT.html](https://www.europarl.europa.eu/doceo/document/CRE-9-2023-06-13-ITM-008_IT.html)

<sup>154</sup> “The Right to an Explanation under the GDPR and the AI Act” Bjørn Aslak Juliussen, Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway, pp. 1-14 : “*One relevant question is whether the deployer is only required to explain the role of the output of the AI system in the later decision process or whether the logic of the AI system*

Similarly to the issue of data processing, it remains unclear how Article 86 of the AI Act will be applied in cases where decisions are made by AI systems that qualify as “black box” models, which, as is well known, are currently not explainable given the current state of technology.

Unless it is established that such AI systems cannot be placed on the European market<sup>155</sup>, it will be necessary to monitor both the technological evolution of these intelligent systems, including “meta-systems” for automated ethical-legal assessment of AI-driven decisions<sup>156</sup>, and the future directions taken by the judiciary in addressing

---

*needs to be explained. The wording of Article 86 (1) only requires explaining the role of the AI system in the decision-making process. However, if Article 86 (1) is interpreted in line with Recital (171) of the AI Act, it becomes evident that the explanation should be "clear and meaningful" and provide "a basis on which the affected persons are able to exercise their rights". The right to an explanation under Article 86 (1) does, thus, not only cover the simple role the output from the AI system had in the decision process but also – as far as it is feasible – the data, algorithm type, and other relevant aspects of the inferred output from the AI system. Moreover, the right to an explanation under the AI Act "shall apply only to the extent" that the right is not covered under Union law, according to the AI Act Article 86 (3). If a data subject under the GDPR has the right to meaningful information about the logic involved in automated individual decision-making under Article 15 (1) (h) of the GDPR, the right to an explanation under Article 86 of the AI Act does not apply, according to Article 86”*  
<https://munin.uit.no/bitstream/handle/10037/36406/article.pdf?sequence=2&isAllowed=y>

<sup>155</sup> Paolo Gaggero, Calogero Alberto Valenza, *Le moderne tecniche di credit scoring tra GDPR, disciplina di settore e AI Act*, in *Rivista di diritto bancario*, July/September 2024, p. 847: "An effective protection of the data subject would therefore require adopting a substantial, rather than merely formal, approach to compliance with privacy regulations, ensuring the actual explainability and interpretability of decisions made by the models. This would mean, in practice, explicitly banning algorithms that generate inference criteria that are not identifiable even by the programmers or structured according to non-deterministic logic."

Forbes stated, “Likely, EU AI regulations won’t permit an opaque system”  
<https://www.forbes.com/sites/glenngow/2021/10/10/the-eu-is-regulating-your-ai-five-ways-to-prepare-now/> ;

Chief AI scientist at Meta Yann Lecun interprets the European position as “deep learning must be banned”; <https://twitter.com/ylecun/status/1545210275237953537>

<sup>156</sup> Gianfranco Basti, *La sfida etica dell'intelligenza artificiale e il ruolo della filosofia*, Aquinas, Year LXV 2022/II, p. 316, discusses AI systems that use higher-order logic

this delicate issue, which will likely become a matter of legal scrutiny in the coming years.

## **2.6. European Office for AI**

As mentioned earlier, the implementation of this new discipline is accompanied by the establishment, at both the European and individual Member State levels, of ad hoc authorities dedicated to overseeing its correct application. In particular, in Europe, the establishment of the European AI Office<sup>157</sup> is foreseen, which is charged with ensuring that AI systems comply with European regulations and respect the fundamental rights of citizens. To ensure a smooth and transparent decision-making process among the 27 member states of the European Union, the European AI Office cooperates directly with national authorities specializing in the field and with the European Centre for Algorithmic Transparency, thus creating an interconnected and interdisciplinary community of experts.

The AI Office consists of five units and two advisors: the "Excellence in AI and Robotics" unit; the "Regulation and Compliance" unit; the "AI Security" unit; the

---

algorithms for meta-control over the consistency of complex mathematical and logical proofs made by human operators, or by the so-called "automatic theorem provers," or even functional programming aimed at creating AI systems for the meta-control of the reliability and safety of programs from other automated systems, whose malfunction could have catastrophic effects and thus cannot be tested in their application field like any other software. Among these AI systems used for meta-control activities are also the so-called "ethical reasoners," designed for meta-evaluation, or for the automatic ethical/legal justification/explanation using higher-order deontic logics, of decisions made by an AI system with its "ethical competence," before the system's decisions turn into actions in the external environment.

<sup>157</sup> Within Section 1, Chapter 7 of the AI Act, the essential characteristics and requirements that the AI Office and its board must adhere to are outlined. Specifically, Article 64 (Section 1, Chapter 7, AI Act) states: 1. The Commission shall develop Union expertise and capabilities in the field of AI through the AI Office. 2. Member States shall facilitate the tasks entrusted to the AI Office, as reflected in this Regulation. For a complete overview of Section 2 of Chapter 7

"Innovation and Policy Coordination in AI" unit; and the "AI for the Benefit of Society" unit, along with a chief scientific advisor and an international affairs advisor. Leveraging its full range of expertise, the primary goal of the AI Office is the correct implementation of the rules and principles provided for by AI legislation. To this end, it undertakes various activities, including verifying the proper application of European regulations concerning AI and defining codes of conduct, guidelines, and executive acts to monitor compliance with Regulation 2024/1689 within the European Union. It also conducts investigations into potential violations of Regulation 2024/1689 and develops new tools and parameters to assess the capabilities of general AI models and to classify models with systemic risks.

It should be noted that verifying the correct application of Regulation 2024/1689 means setting limits on the legitimacy of using systems based on artificial intelligence. However, this does not necessarily mean limiting the development of artificial intelligence itself; in fact, the Union's goal is to promote the safe and thus effective and functional use of such systems in order to harness their social and economic benefits. For this reason, the AI Office has the additional task of creating experimentation spaces for Artificial Intelligence to ensure that it may be adopted safely and that it may prove useful; it is a matter of controlling, for instance, that the system is developed by the interested party following an appropriate path of literacy and awareness.

Moreover, the AI Office supervises the AI Pact, which allows businesses to engage with the European Commission and other stakeholders. As we will see below the AI Pact was introduced before the entry into force of the AI Act in order to help businesses

to plan ahead and prepare for the implementation of AI legislation.<sup>158</sup>

## **2.7. AI Pact**

Finally, it is worth noting that, within the scope of the application of this new discipline, given its particularly innovative nature the European Commission has promoted the AI Pact (“Anticipated Voluntary Compliance for AI”) to proactively encourage voluntary adherence to the principles and rules set out in the AI Act, which, as we have seen, will be implemented gradually over time. Many entities and organizations have expressed interest in the initiative following the Commission's publication of a call for proposals in November 2023 for its promotion.<sup>159</sup>

The AI Pact serves two key purposes: first, to create a network and a conducive space for the exchange of ideas; second, to encourage stakeholders to swiftly adopt the necessary measures to comply with the principles and obligations set forth by the new European legislation. This network creates an environment where stakeholders can collaborate, share experiences, and exchange knowledge about AI. Such exchanges are particularly valuable in defining best practices in the field of AI, and stakeholders are encouraged to engage with one another regarding the processes and practices implemented to ensure compliance with the AI Act, such as transparency and human oversight. Furthermore, it is expected that the results achieved by the various participants in the AI Pact will be collected and published by the AI Office. This will

---

<sup>158</sup> European AI Office, Shaping Europe’s digital future <https://digital-strategy.ec.europa.eu/en/policies/ai-office>

<sup>159</sup> AI Pact, Shaping Europe’s digital future <https://digital-strategy.ec.europa.eu/it/policies/ai-pact>

increase visibility and enhance trust in the technologies developed by the participants. Ultimately, the goal of this initiative, promoted by the European Union, is to establish a shared understanding of the objectives, principles, and ideological and normative foundations of the AI Act, while also defining concrete steps for its implementation across different professional fields.

Additionally, this initiative has the potential to positively influence public opinion by increasing trust in Artificial Intelligence, promoting its credibility, and raising broader awareness of the protections and guarantees provided by the AI Act.

To further explore these concepts, it is essential to address the ethical and societal implications of the right to explainability in the context of both the GDPR and the AI Act. These regulations aim not only to protect individuals' rights but also to ensure that the application of Artificial Intelligence aligns with fundamental ethical principles. With this in mind, Chapter III will delve into the ethical reflections on the principle of explainability in both frameworks, examining how these regulations respond to the societal need for understanding in the face of increasingly complex AI systems. Specifically, this chapter will explore the intersection of ethics, scientific development, law, and the need for clarity, all of which contribute to mitigating fear and promoting trust in AI technologies.

## Chapter III

### Ethical reflections on the principle of explainability in the GDPR

#### and in the AI Act

#### 3.1. Ethics, society, scientific development, and law: the need to understand to avoid fear

In the preceding paragraphs, the essential content of regulations on Artificial Intelligence and personal data processing (GDPR Regulation 2016/679) was examined, with particular focus on the right to "explainability." For both regulations, explainability concerns the result of an automated process derived from mechanical computation.

The end user of the automated service has the right to understand and receive an explanation about the result or decision produced by the machine in order to safeguard their fundamental rights, whether their personal data are involved.

As outlined in previous chapters, both regulations include some provisions aimed at safeguarding this right. Therefore, it can be argued that there is a degree of "formal" protection for the right to explainability.

This leads to an initial ethical reflection on this topic, prompted specifically by the presence of such regulatory provisions.

Why did the legislator deem it essential to include the user's right to explainability of computational outcomes?

To attempt to answer this question, it is essential to start with the context, specifically the reality as shaped and influenced by the widespread adoption of AI technology.

Our daily lives are increasingly shaped by mechanical tools and technologies that

enhance and support human activities. Consider the use of transportation, landline telephony, and household appliances. For none of these mechanical tools has there ever been a recognized need to establish specific regulations that protect the right to explain the services provided by machines.<sup>160</sup>

In the case of AI, however, the situation is different.

I believe this is because the spread of new technologies has had, and continues to have, an impact that is absolutely incomparable to the transformations caused by past technological innovations.<sup>161</sup>

The technological transformation related to AI is overwhelming, revolutionary, and inevitable. It is especially tied to services, activities, tasks, and functions that are entirely delegated to machines, machines that operate on data provided by humans but that also independently process results.<sup>162</sup>

To summarize with a simple example: until now, parking a car relied solely on auditory sensors indicating the presence of an obstacle. With artificial intelligence, the car parks itself. What changes is not just the entity that acts but also the one that decides<sup>163</sup>.

---

<sup>160</sup> High-level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, Bruxelles, 2019, pp.1-39

<sup>161</sup> Luisa Torchia, *Lo Stato digitale. Una introduzione*, Bologna, Il Mulino, 2023, p.21: "It is not merely the global scale of the transformation—(...)—but the speed of the ongoing transformation and the continuous production of new inventions that quickly render the rules and control mechanisms being designed and implemented ineffective or obsolete. This is not about a single invention—such as electricity, the internal combustion engine, or television—but rather about a process of transformation in individual habits and modes of social interaction, whose outcomes never fully stabilize but instead continue to evolve."

<sup>162</sup> European Parliamentary Research Service, *The ethics of artificial intelligence: Issues and Initiatives*, Scientific Foresite Unit, March 2020, pp. 5-18 [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS\\_STU\(2020\)634452\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf)

<sup>163</sup> In this regard, we can recall the definition given by the Charter of Robotics, approved in February 2017 by the European Parliament: "the autonomy of a robot can be defined as the ability to make decisions and implement them in the external world, independently of external

The machine autonomously parks, and the machine "decides" how and where.

It is a simple example, but one that clearly highlights the focal point of the issue: today, and increasingly in the future, AI will autonomously perform various functions, independently making the related decisions.<sup>164</sup>

From this, we can derive an initial consideration about the need, first and foremost emotional and distinctly "human", to understand why and how a given decision is made. In other words, there is a fundamental need to understand and receive an explanation of how the machine operates and the decisions it produces, highlighting the importance of the right to explanation<sup>165</sup>.

However, this need alone does not justify the provision of such a right (and obligation) to explainability, as described above.

Given the revolutionary nature of AI technology, its social impact, and its rapid global proliferation, this topic requires an analysis that goes beyond the technical and regulatory aspects.

An examination of the influence of ethics on social and scientific evolution, particularly in the field of informatics, is necessary.

---

*control or influence; (...) such autonomy is of a purely technological nature, and its level depends on the degree of complexity with which a robot's interaction with the environment has been designed."* Resolution of the European Parliament of February 16, 2017, containing recommendations to the Commission concerning civil law rules on robotics (2015/2103(INL) (2018/C 252/25), see in particular pf. AA, in OJEU 18.7.2018, C 252/239.

<sup>164</sup> An interesting case, to stay with this simple example, was the news that appeared in national newspapers, accompanied by a related video, about an incident that occurred to a driver in China. The driver, after getting out of the vehicle, lets the car park itself in "autonomous driving" mode, but the situation takes an unexpected and dangerous turn when the vehicle "decides" to abandon the parking process and speeds towards the roadway; [video.corriere.it/video-virali/fa-parcheggiare-l-auto-al-pilota-automatico-ma-finisce-malissimo/27ce0bb9-b146-44b5-95ed-24bec12faxlk](https://www.corriere.it/video-virali/fa-parcheggiare-l-auto-al-pilota-automatico-ma-finisce-malissimo/27ce0bb9-b146-44b5-95ed-24bec12faxlk).

<sup>165</sup> Vaassen, B. *AI, Opacity, and Personal Autonomy*. *Philosophy & Technology*, 35(4), 2022, pp. 1–20. <https://doi.org/10.1007/s13347-022-00577-5>

It is essential to investigate the connection between this phenomenon and ethical reflection, as analyzing AI solely as a technical, objective, and regulatory tool proves insufficient. Therefore, an ethical-humanistic exploration of the regulations under review is needed, both in relation to the preliminary phase of drafting the law and in its subsequent application.<sup>166</sup>

In fact, for these regulations, as with the drafting and preparation of any legal framework, the legal reasoning underlying them is always, or at least should always be anchored, in ethical-philosophical reflection. This reflection serves as a fundamental element for defining the key principles to uphold, the limits not to be exceeded, and the values and interests to protect within the legal text.

As the French philosopher Henri Bergson argued in *The Two Sources of Morality and Religion*, ethics is indeed the most important source of lawmaking. Ethics serves as the tool for overturning the foundations of a society that has closed in on itself, clinging too old and unjust principles. Law, in turn, transforms ethical intuition into norms, rules, and customs, but it must be ready to revolutionize itself again when a new moral vision emerges.<sup>167 168</sup>

From this definition, it is clear that law is never a static, immobile discipline but, on the contrary, it is in constant motion.

The law, in other words, changes in response to societal transformations, with the two

---

<sup>166</sup> Raffini, D., & Gorrieri, L. *Intelligenza artificiale e questioni etiche*. *Bioetica*, 31(2) 2023, pp.332–350

<sup>167</sup> Henri Bergson, *Le due fonti della morale e della religione* (1932), Milano, Edizioni di Comunità, 1973, pp. 65-66

<sup>168</sup> Carlo Cardia, *Il fondamento etico del diritto*, *Rivista telematica* <https://www.statoechiese.it>, fascicolo n. 7 del 2021, pp. 19-21 [https://d1vbhhqv6ow083.cloudfront.net/contributi/Cardia.M\\_Il\\_fondamento.pdf](https://d1vbhhqv6ow083.cloudfront.net/contributi/Cardia.M_Il_fondamento.pdf)

becoming interdependent elements where the evolution of one lead to the transformation of the other.

It is precisely within ethics and the principles underpinning it that the primary engine of such change can be identified. Ethics, as Bergson aptly noted in his work, represents the personal and collective orientation towards action. This behavior is then defined within the broad categories of good and evil and synthesized into rules aimed at steering society in a specific direction based on ethical principles.<sup>169</sup>

Thus, law becomes the means and tool to regulate human behavior in accordance with ethical principles. These principles, however, are not universal but are themselves influenced by the historical, cultural, and social context in which they arise. Despite their mutable nature, they exert a significant influence on human conduct in all practical areas, including the development of science.<sup>170</sup>

Indeed, the trajectory of scientific studies has always been significantly shaped by a preliminary ethical reflection aimed at contributing to and guiding the choices and priorities of research.

Since the mid-20th century, following the end of World War II, and particularly in the aftermath of the so-called "Manhattan Project" (a research and development program sponsored by the U.S. government in collaboration with Britain and Canada, which led to the design and construction of the first atomic bomb), it has become evident not only to the scientific community but also to the general public, how science can exceed human understanding and control, generating not only benefits but also significant

---

<sup>169</sup> Henri Bergson, *op.cit.* pp.65-67

<sup>170</sup> Carlo Cardia, *op.cit.* pp.19-21

negative and risky consequences for humanity and future generations.<sup>171</sup>

As a result, the sensitivity to ensuring the ethicality of scientific research has increased and continues to grow, aiming to uphold specific values such as the integrity of research, the responsibility of scientists, the development of science that respects diverse cultural or religious values, and the establishment of a shared morality among different populations with varied cultures and traditions.<sup>172</sup>

Naturally, even in the specific realm of information sciences and technologies, ethical reflection has accompanied (and continues to accompany) the research and application of these new technologies as well as the related legal frameworks. Noteworthy in this context is the contribution of Deborah Johnson, a professor in the Department of Engineering and Society at the University of Virginia. As early as 1985, she authored the manual *Computer Ethics*, identifying ethics as a fundamental element for understanding how computers function. She argued that they "*pose new versions of standard moral problems and moral dilemmas, exacerbating the old problems, and forcing us to apply ordinary moral norms in uncharted realms.*"<sup>173</sup>

---

<sup>171</sup> Carl Mitcham, *Ethics and Science: An Introduction*, Cambridge University Press, 4, Kindle Edition, 2012, pp. 1-22

<sup>172</sup> Carl Mitcham, *op.cit.*, pp.4-5: *It is common to think of science as objective and value neutral. If this is true, then ethics - as the systematic study of norms and values in human conduct - would seem to have only an external relationship to science. But the value neutrality of science is a myth that critical reflection readily challenges. Even as we assert the value neutrality of science, we often claim that science is a morally admirable enterprise that frees from superstition, discloses reality, speaks truth to power, and opens new pathways to material progress. Investments in science are justified by the goods science is alleged to bring, including not just knowledge but increased health and wealth, along with serving as a basis for better personal and public decision-making. Indeed, scientific knowledge is linked to moral imperatives for action. Once we know from science that smoking is harmful, is it not the case that there is an obligation to do something personal behavior and public policy with regard to smoking?*

<sup>173</sup> Deborah G. Johnson, *Computer Ethics*, Prentice-Hall. (Second Edition 1994), pp.1-181

The series of conferences known as ETHICOMP, organized by Simon Rogers -who, in 1998, became the first European professor specializing in Computer Ethics at Montford University in the UK- played a significant role in fostering the development of computing ethics. These conferences focused on promoting interdisciplinary dialogue among fields such as philosophy, computer science, and law, with the aim of advancing the study of computing ethics.<sup>174</sup>

Following these early examples of spaces fully dedicated to research in the field of computing ethics, the evolution of science, the development and widespread global use of the Internet, and other technological advancements have contributed to the emergence of Computer Ethics as an independent discipline. This is evidence of the ongoing and pressing need for ethics in the realm of computing.<sup>175</sup>

The aforementioned concepts, scientific progress, societal change, ethical principles, and law, are highly relevant to the new information technologies, their applications, their regulation, and their social impact and development.

In line with these needs, the European Commission, as early as June 2018, established a High-Level Expert Group with the express goal of promoting reliable AI based on principles of legality, ethics, and robustness, in preparation for subsequent European regulations on AI. This group developed a specific contribution entitled *Ethics Guidelines for Trustworthy AI*.

The document, aptly named *Ethics Guidelines for Trustworthy AI*, defines “ethics” as:

---

<sup>174</sup> Terrell Ward Bynum, *The Foundation of Computer Ethics*, Melbourne, Australia, keynote address at the AICEC99 Conference, July 1999, pp.6-13 <https://doi.org/10.1145/572230.572231>

<sup>175</sup> Terrell Ward Bynum, *op.cit.* pp. 6-13

*an academic discipline and a branch of philosophy. Broadly speaking, ethics deals with questions such as "What is a good action?", "What is the value of a human life?", "What is justice?" or "What is happiness?" At the academic level, there are four main fields of ethical research: i) meta-ethics, which primarily concerns the connotation and denotation of normative statements and how their truth values (if any) are determined; ii) normative ethics, which focuses on practical means of determining the morality of actions by examining the norms underlying just or unjust behavior and attributing value to specific actions; iii) descriptive ethics, which involves the empirical investigation of individuals' moral behaviors and beliefs; and iv) applied ethics, which addresses what should (or could) be done in particular circumstances (often historically new) or specific areas of action (often unprecedented in history). Applied ethics deals with real-life situations where decisions must be made quickly and often with limited rationality. AI ethics is generally considered a form of applied ethics, focusing on the normative issues raised by the design, development, implementation, and use of AI.<sup>176</sup>*

Regarding ethics, the terms "moral" and "ethical" often recur. The term "moral" refers to concrete and factual patterns of behavior, customs, and conventions specific to certain cultures, groups, or individuals at a particular time. The term "ethical" refers to the evaluative judgment of such actions and behaviors from a systematic and academic perspective.

Furthermore, *AI ethics* is defined as:

*the development, distribution, and use of AI in a way that ensures compliance with*

---

<sup>176</sup> High-level Expert Group on Artificial Intelligence, *op.cit.* p.37

*ethical norms, including fundamental rights as special moral rights, ethical principles, and related fundamental values. It is the second of the three essential and necessary components of Trustworthy AI (8).*<sup>177</sup>

The document also specifies that:

*A specific sectoral code of ethics, no matter how consistent, developed, or refined through successive versions, can never replace ethical reasoning itself, which must always remain sensitive to contextual details that cannot be captured by general guidelines. Beyond developing a normative framework, ensuring Trustworthy AI requires building and maintaining an ethical culture and mindset through public debate, education, and practical learning (9).*<sup>178</sup>

From this document, we can derive several insights into the ethical reasons that led the European legislator to include the right to explain automated decisions within the regulatory framework for AI (and, indeed, even earlier, in the GDPR).

In this regard, the contribution of the High-Level Expert Group particularly emphasizes the need for a human-centered configuration of AI, aimed at ensuring that human values play a central role in the development, deployment, use, and monitoring of AI systems. It also highlights the importance of enabling the evaluation and verifiability of AI algorithms, data, and design processes.

The concept of verifiability in the functioning of AI was, therefore, included as one of the key and necessary characteristics for "trustworthy" AI.

The European Expert Group believed that the "explainability" of AI *was fundamental*

---

<sup>177</sup> High-level Expert Group on Artificial Intelligence, *op. cit.*, p. 46

<sup>178</sup> High-level Expert Group on Artificial Intelligence, *op. cit.*, p. 10

*to creating and maintaining user trust in AI systems. This principle implies that processes must be transparent, the capabilities and purpose of AI systems must be communicated openly, and decisions, wherever possible, must be explainable to those directly or indirectly affected. Without such information, a decision cannot be properly challenged. However, it is not always possible to explain why a model generated a particular result or decision (and which combination of input factors contributed to it). This is the so-called "black box" case, where algorithms require special attention. In such circumstances, other measures may be necessary to ensure explainability (for example, traceability, verifiability, and transparent communication about the system's capabilities), provided that the system as a whole respect fundamental right. The level of explainability required largely depends on the context and the severity of the consequences if the result is incorrect or otherwise imprecise.*<sup>179</sup>

The considerations outlined in the previous point are entirely agreeable, as well as summarizing the main ethical principle that justifies the right to explainability, which is reiterated and enshrined in the two regulations examined in this study.

As mentioned earlier at the beginning of this paragraph, the explainability of computational processes is primarily necessary to create and maintain user trust in AI systems.

Referring to the example of the self-driving car: in the event of an accident, it is evident that both the car owner's and the third party's interest, as well as their right, is to understand the reasons behind the behavior of the automated car, as this is a key element in reconstructing the entire sequence of events that led to the damaging

---

<sup>179</sup> High-level Expert Group on Artificial Intelligence, *op. cit.*, p. 14

incident.

From the same document by the European Expert Group, still referring to the principle of explainability in AI, however, a caveat also emerges: *"it is not always possible to explain, however, why a model generated a particular result or decision (and which combination of input factors contributed to it)."*

And also:

*These principles, while undoubtedly indicating the direction towards possible solutions, remain abstract ethical prescriptions. It cannot therefore be assumed that operators in the AI sector will find the right solution based solely on the principles outlined above; however, they should address ethical dilemmas and trade-offs by rationally reflecting based on evidence rather than relying on intuition or random decisions. It may, however, happen in certain situations that it is not possible to identify ethically acceptable compromises. Some fundamental rights and the principles related to them are absolute and non-negotiable (for example, human dignity).<sup>180</sup>*

The same Expert Group, therefore, left open huge ethical dilemmas, which could only be resolved, according to the June 2018 document, upon the actual occurrence of cases, and, therefore, as far as this study is concerned, through the application of the regulations to the various situations provided for and protected.

It also declared, with a sort of "warning" to future generations, that "it may, however, happen in certain situations that it is not possible to identify ethically acceptable compromises", highlighting the existence of absolute and inviolable human rights.

---

<sup>180</sup> High-level Expert Group on Artificial Intelligence, *op. cit.*, p. 15

## 3.2. Explainability: reflections on practical ethics

### 3.2.1. Ethical necessity of explainable artificial intelligence: importance of the human-AI relationship of trust

The need for transparency regarding algorithmic decisions, as declared by the European Expert Group in its June 2018 report- arises, even being statutory or ethical requirement, from a deeply “human” sentiment of fear and distrust toward anything that appears uncontrollable. This is especially true considering that such automated decisions increasingly affect (even) fundamental human rights.

Today, AI and machines have assumed such a central role in people's daily lives that algorithms (their essential component) can be characterized as a "*cultural machine: operating both within and beyond the self-referential barrier of effective calculability, producing culture on a macro-social level while simultaneously creating objects, processes, and cultural experiences*".<sup>181 182</sup>

Given the omnipresence and impact of artificial intelligence systems in individuals' daily lives, their understanding (that is effective and practical knowability) is a fundamental element in establishing systems that reinforce user trust, as well as secure systems that respect fundamental human rights and principles.

The comprehensibility of decisions made by an explainable artificial intelligence system is, therefore, the fundamental condition for creating and maintaining trust in AI

---

<sup>181</sup> Ed Finn, *Che cosa vogliono gli algoritmi. L'immaginazione nell'era dei computer*, Torino, Einaudi, 2018, pp. 1-264

<sup>182</sup> Cristiana Benetazzo, *Intelligenza artificiale e diritto: la sfida etica ed antropologica*; in *Journal of Ethics and Legal Technologies* – Volume 6(2) – December 2024, pp.69-70 : *The cultural questions raised by the relationship between artificial intelligence and the human person invite reflection on a technology that pertains to the practical knowledge of humanity, but above all, evokes the imaginative capacity necessary to transform an object into an artifice, embedding it in our actions through its distinct potentialities.*

applications, fostering greater acceptance and use of AI across various domains.

That said, there is no doubt that the aforementioned legal provisions respond to an ethical need and that the "ethics" of automated decision-making processes is the primary foundation for the desired trust-based relationship between humans and machines.<sup>183</sup>

The ethical foundation of the principle of "explainability" for automated decisions lies in the conviction that, to create the necessary conditions for ensuring the protection of individuals' rights and freedoms, it is essential to guarantee their ability to understand the data processed and generated by the machine that makes the automated decision, as well as the decision-making processes leading to the outcome. Consequently, this provides tools to identify and address potential abuses or malfunctions of the machine and to safeguard fundamental individual rights.<sup>184</sup>

### **3.2.2. Explainability, a tool in service of the human being**

Explainability is directed toward the human user of the information system, who must always be guaranteed the ability to verify the machine's operations.

The human being is the sole recipient of this right, which aligns with the ethical demand for transparency and a return to the anthropocentric essence of AI systems.

These purposes are evident in both regulations, where the intention to prioritize the "human", element and particularly the decision's recipient, emerges. They ensure the

---

<sup>183</sup> The Alan Turing Institute, Dr David Leslie, *Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector*, 2019 pp. 5-8

<sup>184</sup> Hendrik Kempt, *(Un)explainable Technology*, Aachen, Germany, Palgrave Macmillan, 2024, pp. 54-62

recipient's role in the control, formation, and concrete verification of the data, tools, and results generated by the "machine," establishing the user's right to know how the machine operates and the logical steps that lead to a given outcome.

The power of "control" is the cornerstone of the system: only in the presence of real control power, especially by the user subject to the automated decision, can we speak of a "democratic" system or, at least, measures aimed at protecting individuals' rights and freedoms. Explainability, therefore, constitutes an ethical value and a true safeguard for human beings who use AI services or are otherwise affected by its decisions.

Such a principle, with its intended purpose, is supported by the belief that while machines are entrusted with cognitive functions previously considered exclusive to humans, the ability -and thus the right- to comprehend remains uniquely human.<sup>185</sup>

As Professor Luciano Floridi stated: "*True intelligence is not algorithmic; it is the ability to understand, that is, to intus-legere, to 'read within,' to deeply grasp and find unexpected connections between different areas of knowledge. ... Machines can never do these things because, if they were as free as we are, they would be more dangerous than useful. Machines function but do not understand. And understanding cannot be reduced to an algorithm.*"<sup>186</sup>

Indeed, algorithmic decision-making does not draw inspiration from human action but is based on probability calculations rather than logical-deductive processes, where the final result is a consequence of evaluating premises not just in utilitarian but also in

---

<sup>185</sup> Fabio Grigenti, *La prospettiva etica europea sull'intelligenza artificiale*, 2019, pp.1-25

<sup>186</sup> Luciano Floridi, *L'etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*, Milano, Raffaello Cortina Ed., 2022, pp. 1-384

ethical terms.

For this reason, the maintenance of human control over artificial intelligence actions is also justified in order to "protect" human autonomy. Illuminating in this regard are the words of Professor Floridi, who notes that *"when we adopt AI and its smart actions, we voluntarily give up part of our decision-making power to technological artifacts (...) the risk is that the growth of artificial autonomy could undermine the flourishing of human autonomy (...)* It is clear, therefore, that human autonomy must be promoted, and the autonomy of machines must be limited and made inherently reversible, should human autonomy need to be protected or restored (consider, for example, a pilot being able to deactivate the autopilot and regain full control of the aircraft). This introduces a notion that can be defined as meta-autonomy, or the model of the delegation decision. Humans should retain the power to decide which decisions to make, exercising the freedom of choice where necessary and delegating it in cases where reasons of primary importance, such as effectiveness, might prevail over the loss of control. But any delegation should, in principle, remain reversible, adopting as a final guarantee the power to decide to decide again."<sup>187</sup>

### **3.2.3. Explainability and the specific needs of users: explainability as a relative concept**

Given that ethical AI must be explainable and that explainability is intended for the (human) user of the system, the question then arises: how can the principle of algorithmic transparency, as established by the regulations examined, be concretely

---

<sup>187</sup> Luciano Floridi, *op.cit*, p.99

implemented? Consider, for instance, the reference to "meaningful information about the logic involved" contained in Article 15, paragraph 1, letter h of the GDPR, or the references to AI transparency in the European regulation of 2024, which are intended to enable users to adequately understand such systems with a general knowledge of their functioning.

In this regard, it should first be noted that explainability is a relative concept since it hinges on the ability to answer the question, "How does this artificial intelligence system work?" This question is closely tied to the objective and subjective context in which it is posed, and consequently, the answer may and must vary.<sup>188</sup>

In the medical field, for example, it is not always sufficient to explain how an artificial intelligence system operates in exclusively technical terms, limiting the explanation to how the algorithm reaches a decision. Instead, additional information may be needed, such as how the system analyses the available data to understand results based on correlations (e.g., how a certain medical treatment performed in similar cases). In the healthcare sector, it is therefore necessary for the information provided by the AI system to be sufficient for making responsible decisions for the patient's benefit.

In other cases, the question might not concern a specific detail but rather a general description of procedural mechanisms, the dataset used, the data's origin, and the algorithmic processes employed to reach an outcome.

Explainability, therefore, must assume different forms and content depending on the case: to achieve a general understanding of an artificial intelligence system, it is crucial

---

<sup>188</sup> Hendrik Kempt, Jan-Christoph Heilinger, Saskia K. Nagel, *Relative explainability and double standards in medical decision-making*, Ethics and Information Technology, Volume 24, Article number 20, (2022), pp.1-5

to determine what information and which effective standard methods should be communicated to users to explain the system.<sup>189</sup>

This issue is central to analyzing the phenomenon, especially considering the lack of a clear and uniform legal definition of "explanation" and the difficulty of identifying criteria and tools to assess how comprehensible an explanation is to its recipient.

From a subjective perspective (i.e., considering the recipient of the automated decision), assuming that the demand for transparency correlates with the principles of freedom and the protection of fundamental human rights, it is undeniable that the above-mentioned criteria must aim to protect all users democratically.

Thus, explainability must have different content not only depending on the sector in which it is applied but also on the audience to whom the automated service is directed.

To satisfactorily understand an artificial intelligence system, every user requires an appropriate explanation.

In this regard, it cannot be overlooked that, in the case of AI, the technological aspect is so complex that even the recipient of the regulation is required to have specific (and prior) competence or "knowledge" to utilize the tool of machine "knowability." The difficulty of understanding algorithmic languages by the average citizen risks becoming an obstacle that effectively limits citizens' freedom and equality, hindering the full development of the human person and meaningful participation in various social and economic spheres.

This creates unacceptable disparities between those who possess the cognitive tools

---

<sup>189</sup> Mohammad Amir Khusru Akhtar, Mohit Kumar, Anand Nayyar, *Towards Ethical and Socially Responsible Explainable AI Challenges and Opportunities*, Springer Nature Switzerland, 2024, p.15

needed to understand algorithmic languages and those who do not. For the latter, it may even lead to a potential limitation of active participation in democratic life and possible harm to their personal dignity or privacy, even without their awareness or ability to effectively use the tools of protection.

This raises the ethical question of whether such a first, and hardly avoidable, discrimination is acceptable.

Given the significance of artificial intelligence systems, it is crucial that everyone, regardless of their background or profession, can understand why a system has made a particular decision. This transparency ensures that the human-selected mechanisms driving the technology's behavior are clear and comprehensible.<sup>190</sup> In essence, the comprehensibility of operation should be inclusive and democratic, accessible to all, regardless of cultural background, technical expertise, or individual abilities. Given the inherently "persuasive" nature of artificial intelligence, it is essential for humans to be aware of this influence and avoid placing excessive reliance on automated decisions.<sup>191</sup>

---

<sup>190</sup> Europe Council, Committee on Equality and Non-Discrimination. *Preventing discrimination caused by the use of artificial intelligence* (Report). Rapporteur: Mr Christophe Lacroix, Belgium, Socialists, Democrats and Greens Group, (2020), <https://pace.coe.int/en/files/28715/html>

<sup>191</sup> Regarding the "persuasiveness" of AI, the doctrine points out that the only regulatory reference is contained in paragraph 4, letter b) of Article 14 of the AI Act, which establishes "*that the individuals responsible for oversight must remain aware of the potential tendency to automatically rely on or place excessive trust in the output produced by a high-risk AI system ('automation bias), particularly for high-risk AI systems used to provide information or recommendations for decisions to be made by individuals.*" "*Automation bias*" refers to the tendency of humans interacting with machines to rely on their outputs, even to the point of neglecting or ignoring other information coming from different sources: a form of cognitive laziness that everyone has experienced in various situations, anchoring, for example, to initial information and/or intuitions in a decision-making process. Taking this tendency into account and, therefore, recognizing that the norms address not robots, but real people with their

### 3.2.4. Ethical need of transparency of data

This requires a system design that not only functions technically but is also socially responsible and ethical: transparency and explainability thus become the fundamental principles of human-centered design.<sup>192</sup>

In this regard, we add that, to develop a transparent, ethical, and fair artificial intelligence system, it is essential to implement it with a broad and diverse dataset that is impartial and representative, to avoid biases that may emerge from the lack of equitable representation of various social groups.<sup>193</sup> It is well known that the output of an AI system directly depends on the quality and characteristics of the training data and that potential "errors" or "unfairness" in the system, which could lead to discriminatory outcomes or violations of fundamental freedoms and rights, may stem from using insufficiently diversified datasets.<sup>194</sup>

For example, using data from an unrepresentative group of people, excluding or underrepresenting certain individuals, can generate models that do not accurately reflect the population's diversity, risking discrimination or unfair outcomes.

For this reason, another fundamental requirement for system transparency will be knowledge and understanding of the data used to implement artificial intelligence systems.

---

*cognitive and behavioral biases, is already a turning point, in line with the European Union's push to consider behavioral sciences in the development of "future-proof rules."* Germana Lo Sapio, *L'Artificial Intelligence Act e la prova di resistenza per la legalità algoritmica*, in *Federalismi.it*, *Rivista di diritto pubblico italiano, comparato, europeo, Osservatorio trasparenza* 10.07.2024, n. 16/2024, p. 280 <https://www.federalismi.it/nv14/articolo-documento.cfm?Artid=5086>

<sup>192</sup> Mohammad Amir Khusru Akhtar, Mohit Kumar, Anand Nayyar, *Towards*, *op.cit.* pp.13-16

<sup>193</sup> Mohammad Amir Khusru Akhtar, Mohit Kumar, Anand Nayyar, *op.cit.* pp.21-25

<sup>194</sup> Mohammad Amir Khusru Akhtar, Mohit Kumar, Anand Nayyar, *op.cit.* pp.21-25

In other words, it will be necessary to ensure transparency regarding the origin, selection, and nature of the data to provide users with a more complete view of the process leading to a given decision.

This approach will not only foster trust but will also help identify and correct potential biases in the data, improving the system's accountability and reliability.<sup>195</sup>

In conclusion, to have ethical and responsible artificial intelligence systems, it will be essential to implement ethical governance that, considering the aforementioned needs and challenges, aims to ensure that artificial intelligence is developed transparently, fairly, and responsibly.

This should minimize risks of bias, discrimination, and collateral harm while incorporating specific monitoring and periodic review mechanisms necessary to adapt to technological and regulatory evolution.

Only in this way users can trust the system.

### **3.2.5. Conflicting needs: the ethics of explainability and technological and economic development**

As stated above, the goal of explainable AI is to bridge the trust gap between humans and AI, thus enabling greater acceptance and proper implementation of these technologies in various sectors.

Given the rapid and continuous evolution of the AI sector, it is essential to first question whether European regulations are sufficiently equipped to effectively govern it.

---

<sup>195</sup> Mohammad Amir Khusru Akhtar, Mohit Kumar, Anand Nayyar, *op.cit.* pp. 21-25

There is a risk that, day by day, the gap between the provisions of trustworthy AI regulations and technical advancements, tools, and practices in related fields will widen.

It is worth considering whether such rapid and overwhelming development could ultimately result in the forced abandonment, or at least the inevitable lowering, of certain ethical standards and principles.

Ensuring that users are informed and aware of how the system operates undoubtedly promotes more transparent and responsible interactions with AI. However, striking a balance between the complexity of the model and its ability to be easily understood and explained remains a challenging issue.

This issue can significantly influence the performance and functionality of AI systems.<sup>196</sup>

At this point, it should not be overlooked that the need to make systems intelligible often clashes with the objective, often tied to economic interests, of achieving high performance.

This is the case with the most advanced computational models, which can make accurate and sophisticated predictions but whose functioning is often difficult to interpret due to their complex structure. This effectively creates various "black boxes" in their operations, making it difficult, if not impossible, to understand how decisions are made, thereby reducing transparency in the process.<sup>197</sup>

In such cases, one must ask whether it is ethically acceptable to compromise the right

---

<sup>196</sup> McDermid JA, Jia Y, Porter Z, Habli I., *Artificial intelligence explainability: the technical and ethical dimensions*. Phil. Trans. R. Soc., 2021. <https://doi.org/10.1098/rsta.2020.0363>

<sup>197</sup> Gryz Jarek and Rojszczak Marcin, *op.cit* pp.1-6

to knowability in favor of the equally legitimate and constitutionally guaranteed right (as per Article 4 of the Italian Constitution) to development and progress.

Another question is whether users will still have the option to decline AI-provided services and choose a fully "human" service instead.

The points of reflection are many, and the possible solutions to the listed doubts appear complex, involving an inevitable confrontation between "intelligent machines" and "human intelligence," between human intervention and mechanical operation, between guaranteeing human free will and the autonomous development of artificial intelligence, between ethical evaluation and the opportunities for scientific growth.

On one hand, it is clear that technological progress has amplified humans' potential for knowledge and action. On the other hand, this technological evolution exposes humanity to new risks that must be addressed in ethical terms, i.e., in terms of the ability to respect or consciously evolve the principles of reference.

The concept of AI "explainability" should (at least partially) address these needs.

However, the use of the conditional is essential, as the questions raised remain unresolved and dependent on how the rules are applied in practice and how relevant practices and case law evolve over time (including jurisprudence).

## Conclusion

The growing influence of artificial intelligence systems and big data represents one of the most complex and pressing challenges contemporary society faces. This technological revolution, while offering extraordinary opportunities for progress and innovation, also brings with it far-reaching ethical, legal, and social implications that are often difficult to foresee or manage.

In this context, the principle of explainability emerges as a cornerstone for ensuring not only the protection of fundamental rights but also for fostering a more balanced and informed relationship between humans and advanced technologies. As observed, it is not a static or rigidly defined concept but rather a constantly evolving rule that requires, and will continue to require in the future, ongoing adaptation. It is indispensable for preserving human dignity and respecting individual freedom in an era dominated by digitalization.

Throughout this analysis, it has become clear that the GDPR and the AI Act are among the European Union's most ambitious regulatory responses to the increasing proliferation and, at the same time, the opacity of automated technologies. These regulations, by imposing obligations of transparency and accountability, aim to rebalance the relationship between individuals and systems, restoring control to citizens, who are often relegated to passive roles when faced with complex and difficult-to-understand decision-making mechanisms. In this framework, technology takes on a dual role: on the one hand, it serves as a tool for progress and development; on the other, it represents a potential source of exclusion and discrimination, capable of amplifying pre-existing inequalities and creating new ones.

One of the most fascinating and, at the same time, problematic aspects of this issue lies in the intrinsic tension between technological efficiency and democratic principles. On one hand, artificial intelligence systems promise rapid and precise results and, in general, ever-increasing efficiency. On the other hand, they risk sacrificing fundamental values such as transparency, fairness, and the right to informed participation. Explainability, in this sense, cannot be seen solely as a regulatory obligation or technical requirement but must be recognized and promoted as a fundamental ethical need. Only through understanding automated decisions can individuals contest them, modify them, and, in cases of errors or injustices, oppose and defend themselves if necessary. This right is closely linked to informational self-determination, which represents the ability of each individual to maintain control over their personal data, an essential element for ensuring freedom and autonomy in the digital context.

We know that artificial intelligence can be highly beneficial in many sectors, but if poorly governed, it can also be the source of significant distortions. For example, a particularly sensitive issue concerns the relationship between explainability and algorithmic discrimination. Algorithms, if not properly monitored and designed, can perpetuate and amplify social, economic, and cultural inequalities. One example of this phenomenon is found in recruitment processes or the granting of credit, where the use of biased or incomplete data can lead to discriminatory or unfair decisions. In such contexts, explainability plays a crucial role, as it allows for the identification and correction of potential biases, contributing to the creation of fairer and more inclusive systems.

The increasing complexity of the legislative and regulatory framework, which combines the GDPR with the AI Act, highlights several structural gaps and practical difficulties that risk undermining effective protection. An example is the distinction between the right to *ex-ante* information, which aims to provide preventive explanations about how systems operate, and the right to *ex-post* explanation, which allows individuals to understand a decision after it has been made. The tangible risk is that such information may either be too difficult for citizens, who lack specialized expertise, to understand or lack practical utility. This underscores the importance of further investigation and refinement to ensure that explainability is not only formal but also substantive, translating it into accessible, intuitive language capable of creating real value for end users.

To ensure the effective protection of fundamental rights, a more robust and detailed approach is required. This could include strengthening the provisions on explainability by introducing clear guidelines to ensure that automated decisions are transparent and comprehensible; promoting investments in research and development of explainable AI (XAI) technologies aimed at enhancing the transparency and understandability of algorithms; implementing enhanced human oversight systems to ensure that automated systems are subject to meaningful human intervention, particularly in critical sectors such as healthcare and justice; and harmonizing internal policies across European countries and among corporate entities working in this specific area. Specifically, the creation of a harmonized regulatory framework that integrates the GDPR with other relevant regulations, such as the AI Act, is necessary to ensure comprehensive protection of fundamental rights.

Through these measures, it will be possible to strike a balance between technological innovation and the protection of rights, safeguarding the dignity and freedoms of individuals in the digital age.

The ethical component plays a crucial role in this scenario. Technological progress cannot be conceived as a neutral process or one separated from human responsibilities. On the contrary, it must be guided by constant ethical reflection, capable of steering choices toward solutions that safeguard the fundamental values of society. The principle of explainability stands out as an essential tool for balancing the aspiration for innovation with the need to preserve democratic rights and freedoms. It is not merely about ensuring transparency but about building a system that promotes justice and equity, where technologies serve as tools for the collective good.

It is important to emphasize that the principle of explainability is not limited to regulating the relationship between citizens and technologies but is part of a broader reflection on collective responsibility. Explainability is not only an individual right but also a social and institutional duty. Ensuring the transparency of automated decision-making processes means strengthening the relationship of trust between citizens, companies, and institutions, fostering informed and democratic participation in decision-making processes.

Looking to the future, it is evident that digital technological progress represents an inevitable and, in some ways, necessary reality, but this cannot disregard its ethical dimension. The path outlined by the GDPR and the AI Act is certainly promising but requires further improvements to address the new challenges posed by this true technological revolution. Proactive regulation, capable of anticipating the social and

ethical implications of new technologies, represents the only way to ensure that innovation becomes a tool of empowerment rather than a factor of discrimination and a violation of rights and freedoms.

In conclusion, it is believed that, in this domain, the principle of explainability -while still imperfect in its practical application- represents an indispensable element for ensuring a more just, transparent, and respectful society of human rights in the digital era. The defense and constant renewal of rights are a responsibility that concerns everyone, both as individuals and as a collective. Only by accompanying technological progress with profound and ongoing ethical reflection will it be possible to build a future that reflects the highest aspirations of humanity.

## Chapter I

### References:

- Norberto Bobbio, *L'età dei diritti*, Torino, Einaudi, 1990
- Hannah Arendt, *The Human Condition*, London, The University of Chicago Press, 1958, pp. 68-73
- Paolo Guarda, Giorgia Bincoletto, *Diritto comparato della privacy e della protezione dei dati personali*, Ledizioni, Marzo 2023
- Janice Richardson, *Law and the Philosophy of Privacy*, New York, Routledge, 2017
- Giusella Finocchiaro, *La protezione dei dati personali in Italia, Regolamento UE N.016/679 e d.lgs. 10 Agosto 2018, n.101*, Bologna, Zanichelli, 2019
- Buttarelli, *Banche dati e tutela della riservatezza*, Milano, Giuffrè Editore, 1997
- Jody R. Westby, *International Guide to Privacy*, American Bar Association, 2004
- Giuseppe Cassano, Vincenzo Colarocco, Giovanni Battista Gallus, Francesco Paolo Micozzi, *Il Processo di Adeguamento al GDPR*, Milano, Giuffrè Editore, 2022
- D.F Wallace, *The Pale King, An Unfinished Novel*, Little, Brown and Company, 2011
- Frank Pasquale, *The Black Box Society: The secret Algorithms that control Money and Information*, Cambridge-London, Harvard University Press, 2015
- T. Tridimas, *The Court of Justice and Judicial Activism*, in *European Law Review*, 1996

### Web references:

- Gianmarco Cifaldi, *Evolution of Concepts of Privacy and Personal Data Protection under the Influence of Information Technology Development*, in *Sociology and Social Work Review*, Volume 7 (Issue 1), 2023 : <https://ricerca.unich.it/retrieve/7798a6bc-f30c-46c0-8397-19cbb84ee841/Evolution-of-Concepts-of-Privacy-and-Personal-Data-Protection-under-the-Influence-of-Information-Technology-Development.pdf>
- Bryce Goodman, Seth Flaxman, *European Union regulation on algorithmic decision-making and a “right to explanation”*, Oxford, United Kingdom, Oxford Internet Institute and Department of Statistics, University of Oxford, 31 August 2016: <https://doi.org/10.48550/arXiv.1606.08813>
- Samuel Warren and Louis Brandeis, *The right of Privacy*, Harvard Law Review, Vol.4, No.5 (Dec. 15, 1890): <https://docenti.unimc.it/benedetta.barbisan/teaching/2017/17581/files/the-right-to-privacy-warren-brandeis>
- The history of data protection regulation [https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation\\_en](https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en)
- Luigi Rendina, *Privacy vs protezione dati personali: attenti alla differenza, ne va della nostra identità*, Agenda Digitale, 30 October 2019

- [:https://www.agendadigitale.eu/sicurezza/privacy/privacy-e-protezione-dati-personali-cosa-sono-quali-differenze-cosa-e-cambiato-col-gdpr/?](https://www.agendadigitale.eu/sicurezza/privacy/privacy-e-protezione-dati-personali-cosa-sono-quali-differenze-cosa-e-cambiato-col-gdpr/)
- United Nation Information Centre, Italy, *Universal Declaration of Human Rights* <https://www.ohchr.org/en/human-rights/universal-declaration/translations/italian?>
  - United Nations, *Universal Declaration of Human Rights*: <https://www.un.org/en/about-us/universal-declaration-of-human-rights>
  - *The Universal Declaration of Human Rights and the European Union*, 2018: [https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/757559/EPRS\\_A\\_TA\(2023\)757559\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/757559/EPRS_A_TA(2023)757559_EN.pdf)
  - Council of Europe Portal, *The European Convention on Human Rights, The Convention in 1950* <https://www.coe.int/en/web/human-rights-convention/the-convention-in-1950?>
  - *Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data*, 1981: <https://rm.coe.int/1680078b37>
  - *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data*: <https://eur-lex.europa.eu/eli/dir/1995/46/oj>
  - *Direttiva (UE) 2016/680 del Parlamento Europeo e del Consiglio del 27 April 2016*: <https://eur-lex.europa.eu/legal-content/IT/TXT/PDF/?uri=CELEX:32016L0680>
  - *GDPR Text* : <https://gdpr-text.com/it/>
  - European Data Protection Supervisor, *The History of the General Data Protection Regulation* [https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation\\_en?](https://www.edps.europa.eu/data-protection/data-protection/legislation/history-general-data-protection-regulation_en?)
  - European Council, Council of the European Union, *How Maastricht changed Europe*, October 2024 <https://www.consilium.europa.eu/en/maastricht-treaty/>
  - Gianclaudio Malgieri , Giovanni Comandé, *Why a Right to Legibility of Automated Decision- Making Exists in the General Data Protection Regulation*, in *International Data Privacy Law*, Volume 7, Issue 4, November 2017 : <https://doi.org/10.1093/idpl/ipx019>
  - Giovanna De Minico, *Justice and artificial intelligence: a changing balance*, in *Rivista AIC Associazione italiana dei costituzionalisti*, n. 2/2024, del 29.04.2024, <https://www.rivistaaic.it/it/rivista/ultimi-contributi-pubblicati/giovanna-de-minico/giustizia-e-intelligenza-artificiale-un-equilibrio-mutevole>
  - See Court of Justice of the EU, Grand Chamber, December 2, 2009, No. 89; Court of Justice of the EU, Grand Chamber, December 22, 2008, No. 333: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62008CJ0089>
  - R. Messinetti, *The protection of the human person versus Artificial Intelligence. Decision-making power of the technological apparatus and right to explanation of the automated decision*, in *Contract and Business*, n. 2, 2019 <https://doi.org/10.57230/EJPLT221ET>
  - Mortaji, S. T. H., & Sadeghi, M. E. *Assessing the reliability of artificial*

- intelligence systems: Challenges, metrics, and future directions*. International Journal of Innovation in Management Economics and Social Sciences, 4(2), 2024 [www.ijimes.ir](http://www.ijimes.ir)
- Paul De Hert, Vagelis Papakonstantinou, *The proposed data protection Regulation replacing Directive 95/46/EC: A sound system for the protection of individuals*, Sciencedirect.com, Computer Law and Security Review, April 2012 [10.1016/j.clsr.2012.01.011](https://doi.org/10.1016/j.clsr.2012.01.011)
  - European Commission, Legal framework of EU data protection [https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection\\_en?](https://commission.europa.eu/law/law-topic/data-protection/legal-framework-eu-data-protection_en?)
  - Yuhong Yan, *The Risk-Based Approach to Personal Data Protection and the Response of the International Trade Law*, Beijing Law Review, Vol.14. No.3, September 2023 [10.4236/blr.2023.143067](https://doi.org/10.4236/blr.2023.143067)
  - Andrea Ferrario and Michele Loi. *How Explainability Contributes to Trust in AI*. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, <https://doi.org/10.1145/3531146.3533202>
  - Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M. B., & Kang, B., *Survey on Explainable AI: From Approaches, Limitations, and Applications Aspects*, 2023, pp. 162-163 <https://doi.org/10.1007/s44230-023-00038-y>
  - Chen, Z. *Ethics and discrimination in artificial intelligence-enabled recruitment practices*. Humanit Soc Sci Commun 10, 2023 <https://doi.org/10.1057/s41599-023-02079-x>
  - Giuseppe Mobilio, *L'Intelligenza Artificiale e le Regole Giuridiche alle prova: il caso paradigmatico del GDPR*, in *Federalismi.it*, 16, 2020: [https://www.federalismi.it/nv14/articolodocumento.cfm?Artid=43539&content=&content\\_author=](https://www.federalismi.it/nv14/articolodocumento.cfm?Artid=43539&content=&content_author=)
  - Simon Chandler, *How Explainable AI is Helping Avoid Bias*, Forbes, 2020 <https://www.forbes.com/sites/simonchandler/2020/02/18/how-explainable-ai-is-helping-algorithms-avoid-bias/>
  - Lucia G. Scianella, *Intelligenza artificiale, politica e democrazia*, DPCE online
  - Court of Justice of the EU, Grand Chamber, December 2, 2009, No. 89; Court of Justice of the EU, Grand Chamber, December 22, 2008, No. 333: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:62008CJ0089>
  - Nicol Turner Lee, *Making AI more explainable to protect the public from individual and community harms*, Brookings, November 29, 2023 <https://www.brookings.edu/articles/making-ai-more-explainable-to-protect-the-public-from-individual-and-community-harms/>
  - Erica Palmerini, *Decisioni algoritmiche e diritto dei dati*, giudicedonna.it, Numeri 1-2/2023
  - Francesco Sovrano, Fabio Vitali, Monica Palmirani, *Modeling GDPR-Compliant Explanations for Trustworthy AI*, September 2021 <https://arxiv.org/pdf/2109.04165>
  - Castets-Renard, C. *Accountability of algorithms in the GDPR and beyond: A European legal framework on automated decision-making*. Fordham Intellectual

- Property, Media and Entertainment Law Journal, 2019, Article 3  
<https://ir.lawnet.fordham.edu>
- Carlo Colapietro, *Algorithms between transparency and protection of personal data" in Federalismi.it* - Observatory on transparency, 2023,  
<https://www.federalismi.it/nv14/articolo-documento.cfm?Artid=48430>
  - T. Tridimas, *The Court of Justice and Judicial Activism*, in *European Law Review*, 1996 :  
[https://www.researchgate.net/publication/228151038\\_Judicial\\_Activism\\_and\\_the\\_Court\\_of\\_Justice\\_How\\_Should\\_Academics\\_Respond](https://www.researchgate.net/publication/228151038_Judicial_Activism_and_the_Court_of_Justice_How_Should_Academics_Respond)
  - Wachter, S., Mittelstadt, B., & Floridi, L. *Why A Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*. *International Data Privacy Law*, 2017, pp. 76–99  
<https://doi.org/10.1093/idpl/ix005>
  - Castets-Renard, C. *Accountability of algorithms in the GDPR and beyond: A European legal framework on automated decision-making*. *Fordham Intellectual Property, Media and Entertainment Law Journal*, 2019, Article 3
  - Bryan Casey, Ashkon Farhangi, Roland Vogl, *Rethinking Explainable Machine: The GDPR's "Right to Explanation" Debate and the rise of algorithmic audit enterprise*, 2019, *Berkeley Technology Law Journal*, Vol.34:143,  
<https://doi.org/10.15779/Z38M32N986>

## Chapter II

### References:

- L. Frank Baum, *Il Mago di Oz*, Torino, Einaudi, 2012
- Giusella Finocchiaro, *Intelligenza artificiale. Quali regole?*, Bologna, Il Mulino, 2024
- D. L. J. Bradbery, *Oxford Advanced Learner's Dictionary Paperback*, Oxford, OUP, 2011
- Howard Gardner, *Formae mentis. Saggio sulla pluralità dell'intelligenza*, Milano, Feltrinelli, 2013
- Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, W. H. Freeman and Company, 1976
- Alessandro Pajno, Filippo Donati, Antonio Perrucci, *Intelligenza artificiale e diritto: una rivoluzione? Diritti fondamentali, dati personali e regolazione (Vol.1)*, Bologna, Il Mulino, 2022
- Daniel Andler, *Il Duplice Enigma, Intelligenza artificiale e intelligenza umana*, Torino, Piccola biblioteca Einaudi, 2024

### Web references:

- A. M. Turing, *Computing Machinery and Intelligence*, Oxford University Press on behalf of the Mind Association, 1950 :  
<https://phil415.pbworks.com/f/TuringComputing.pdf>

- Frixione, M., & Numerico, T. , *Alan Mathison Turing*. APhEx, 7. Periodico online ISSN 2036-9972, 2013 <https://www.openstarts.units.it/server/api/core/bitstreams/0c3efe39-ae91-4779-ad54-6fda38c1205e/content>
- Mauro G. Smiraldo, *Il dottor Frankenstein e le responsabilità nella robotica*, in magazine.atlante.società, Treccani.it, April 2024 <https://www.treccani.it/magazine/atlante/societa/il-dottor-frankenstein-e-le-responsabilita-nella-robotica.html>
- Antonio Pescapé (Luca Lo Sapiro), "AI e futuro di sapiens tra nuovi orizzonti ed antichi timori," in Scienza e Filosofia.com, pp.28-42, <https://www.scienzae filosofia.com/2022/07/05/ai-e-futuro-di-sapiens-tra-nuovi-orizzonti-e-antichi-timori/>
- Gianluca Giannini - Antonio Pescapé (Luca Lo Sapiro), "AI e futuro di sapiens tra nuovi orizzonti ed antichi timori," in Scienza e Filosofia.com, 2022: <https://www.scienzae filosofia.com/wp-content/uploads/2022/07/03-GIANNINI.pdf>
- The Origins of the Stanford-Binet 5, the WAIS-IV, the WISC-V, and the WPPSI-IV Subtests Aisa Gibbons and Russell T. Warne Utah Valley University, pp.1-32 <https://osf.io/uwh2s/download/?version=1&displayName=Subtest%20origins-2018-09-18T19%3A18%3A30.846Z.pdf>
- Hélio A. G. Teive, *Alfred Binet: Charcot's pupil, a neuropsychologist and a pioneer in intelligence testing*, Arquivo de Neuro-Psicquiatria, September 2017 <https://doi.org/10.1590/0004-282X20170097>
- Floyd, Christiane. *From Joseph Weizenbaum to ChatGPT: Critical Encounters with Dazzling AI Technology*. Weizenbaum Journal of the Digital Society, Research Paper, University of Hamburg <https://doi.org/10.34669/WI.WJDS/3.3.3>
- European Commission, AI Act <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?>
- Marco Somalvico and Francesco Amigoni and Viola Schiaffonati, *Intelligenza Artificiale*: <https://schiaffonati.faculty.polimi.it/pubblicazioni/H1.pdf>
- Masahiro Mori, *The Uncanny Valley*, 1970 : <https://www.almendron.com/tribuna/wp-content/uploads/2018/01/morunc.pdf>
- M. G. Smiraldo, *Il dottor Frankenstein e le responsabilità nella robotica*, Treccani.it, April 2024: <https://www.treccani.it/magazine/atlante/societa/il-dottor-frankenstein-e-le-responsabilita-nella-robotica.html>
- Andrea Castiello d'Antonio, *Intelligenza Artificiale, psicologia e psicologia delle organizzazioni. Su alcuni aspetti dell'Intelligenza Artificiale negli ambienti di lavoro*, 2021: <https://www.castiellodantonio.it/sites/default/files/andrea-castiello-dantonio-intelligenza-artificiale-psicologia-e-psicologia-delle-organizzazioni.pdf>
- *Comunicazione della Commissione europea al Parlamento europeo, al Consiglio, al Comitato economico e sociale europeo e al Comitato delle regioni, L'intelligenza artificiale per l'Europa*, COM (2018) 237, 25 aprile 2018, 1: <https://eur-lex.europa.eu/legal-content/IT/ALL/?uri=CELEX:52018DC0237>
- European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the

- Committee of the Regions: *A Digital Single Market Strategy for Europe*, COM(2015) 192 final, Brussels, May 6, 2015, p. 3. As of August 30, 2024: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52015DC0192>
- Lucia G. Sciannella, *Intelligenza artificiale, politica e democrazia, Breve introduzione all'Intelligenza Artificiale*, DPCE Online, 51 (1), 2022: <https://doi.org/10.57660/dpceonline.2022.1577>
  - Janiesch, C., Zschech, P., & Heinrich, K., *Machine learning and deep learning*. *Electronic Markets*, 31, 685–695, 2021. <https://doi.org/10.1007/s12525-021-00475-2>
  - E. Blakemore, *Il legame tra la nuova IA e il test di Turing: chi era davvero l'uomo che lo ha inventato?*, 01/2024: <https://www.nationalgeographic.it/la-nuova-ia-potrebbe-superare-il-test-di-turing-chi-era-l-uomo-che-lo-ha-inventato>
  - Gyevnar B., Ferguson N., and Schafer B., *Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?* School of Informatics, University of Edinburgh; Edinburgh Law School, University of Edinburgh. 2023 [arxiv.org/abs/2302.10766](https://arxiv.org/abs/2302.10766).
  - J. McCarthy, M.L. Minsky, N. Rochester e C.E. Shannon, *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, August 31, 1955, *AI Magazine*, 27(4), 12 <https://doi.org/10.1609/aimag.v27i4.1904>
  - Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Seri, J., Díaz-Rodríguez, N., & Herrera, F. M *Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence*. *Information Fusion*, 99, 2023. <https://doi.org/10.1016/j.inffus.2023.101805>
  - Madiega, T. *Artificial Intelligence Act*. EPRS | European Parliamentary Research Service, PE 698.792. European Union, 2024, [https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS\\_BRI\(2021\)698792\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf)
  - D. Gunning, David W. Aha, *DARPA's Explainable Artificial Intelligence Program*, 2019, pp. 44-58 <https://doi.org/10.1609/aimag.v40i2.2850> European Commission, *Shaping Europe's Digital Future, AI Act* <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai?>
  - Bertoincini, A. L. C., Serafim, *Ethical content in artificial intelligence systems: A demand explained in three critical points*. *Frontiers in Psychology*, 14, 1074787, 2023. <https://doi.org/10.3389/fpsyg.2023.1074787>
  - EU Legislation in Progress, *Artificial Intelligence Act*, 2024: <https://www.iisf.ie/files/UserFiles/cybersecurity-legislation-ireland/EU-AI-Act.pdf>
  - Technology and Privacy Unit of the European Data Protection Supervisor (EDPS) *TechDispatch. Explainable Artificial Intelligence* , 2023: [https://www.edps.europa.eu/system/files/2023-11/23-11-16\\_techdispatch\\_xai\\_en.pdf](https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf)
  - Emmanuel Pintelas, Ioannis E. Livieris, Panagiotis Pintelas, *A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White Box Intrinsic Interpretability*, 2020, in *Algorithm*, 2020, 13(1), 17:

- <https://www.mdpi.com/1999-4893/13/1/17>
- *Proposal for a Regulation of the European Parliament and Council establishing harmonized rules on Artificial Intelligence (AI Act) and amending certain legislative acts of the Union, 2021:* <https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:52021PC0206>
  - EU Artificial Intelligence Act: <https://artificialintelligenceact.eu/>
  - High-Level Expert Group on Artificial Intelligence, EU Commission, *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment*, 2019: <https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1>
  - *The role of explainable AI in the context of the AI Act.* In 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA. ACM, New York, NY, USA <https://dl.acm.org/doi/pdf/10.1145/3593013.3594069>
  - Gryz Jarek; Rojszczak Marcin, *Black box algorithms and the rights of individuals: No easy solution to the “explainability problem*, Internet Policy Review, 10(2), <https://doi.org/10.14763/2021.2.1564>
  - European AI Office, Shaping Europe’s digital future <https://digital-strategy.ec.europa.eu/en/policies/ai-office>
  - AI Pact, Shaping Europe’s digital future <https://digital-strategy.ec.europa.eu/it/policies/ai-pact>
  - Francesca Mattassoglio, *La Corte di giustizia europea, algoritmi e credit scoring. L’apertura del vaso di Pandora delle società che si “limitano” a elaborare gli scoring*, in *DB non solo diritto bancario.it, Dialoghi di diritto dell’Economia, Note*, 10 gennaio 2025
  - Cecilia Panigutti, David Fernandez Llorca, Salvatore Scalzo, Ronan Hamon, Delia Fano Yela, Gabriele Mazzini, Isabelle Hupont, Henrik Junklewitz, Ignacio Sanchez, Josep Soler Garrido, Emilia Gomez, *The role of explainable AI in the context of the AI Act.* In 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA. ACM, New York, NY, USA <https://dl.acm.org/doi/pdf/10.1145/3593013.3594069>
  - European Parliament, *Position established at first reading on March 13, 2024, in view of the adoption of the regulation of the European Parliament and of the Council es*
  - *tablishing harmonized rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144, and Directives 2014/90/EU, (EU) 2016/797, and (EU) 2020/1828 (Artificial Intelligence Regulation).* EP-PE\_TC1-COD(2021)0106, Art. 86 [https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236\\_IT.html](https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_IT.html)
  - Discussion on the proposal for AI regulation in the European Parliament on 13.06.2023, Danish MEP Karen Melchior (Renew) [https://www.europarl.europa.eu/doceo/document/CRE-9-2023-06-13-ITM-008\\_IT.html](https://www.europarl.europa.eu/doceo/document/CRE-9-2023-06-13-ITM-008_IT.html)
  - Paolo Gaggero, Calogero Alberto Valenza, *Le moderne tecniche di credit scoring tra GDPR, disciplina di settore e AI Act*, in *Rivista di diritto bancario*,

July/September 2024

- Forbes stated, “Likely, EU AI regulations won’t permit an opaque system”  
<https://www.forbes.com/sites/glenngow/2021/10/10/the-eu-is-regulating-your-ai-five-ways-to-prepare-now/>
- Chief AI scientist at Meta Yann Lecun interprets the European position as “deep learning must be banned”;  
<https://twitter.com/ylecun/status/1545210275237953537>
- Luigi Carbone, *L’algoritmo e il suo giudice*, presentation at the conference “*Digital Administration – Daily Efficiency and Smart Choices*,” University of Naples Federico II, May 9–10, 2022, available on [www.giustizia-amministrativa.it](http://www.giustizia-amministrativa.it)
- Gherardo Carullo, “*Decisione amministrativa e intelligenza artificiale*”, *Diritto dell’informazione e dell’informatica*, fasc. 3, 2021:  
<https://air.unimi.it/retrieve/dfa8b9a8-cc5d-748b-e0533a05fe0a3a96/OA%20Carullo%20-%20Decisione%20amministrativa%20e%20intelligenza%20artificial.pdf>

#### Filmography:

- Fritz Lang, *Metropolis*, 1927
- Stanley Kubrick, *2001: A Space Odyssey*, 1968
- Ridley Scott, *Blade Runner*, 1982

### **Chapter III**

#### References:

- Luisa Torchia, *Lo Stato digitale-Una introduzione*, Bologna, Il Mulino, 2023
- Henri Bergson, *Le due fonti della morale e della religione (1932)*, Milano, Edizioni di Comunità, 1973
- Carl Mitcham, *Ethics and Science: An Introduction*, Cambridge University Press, 4, Kindle Edition, 2012
- Deborah G. Johnson (1985), *Computer Ethics*, Prentice-Hall. (Second Edition 1994)
- Terrell Ward Bynum (1999), *The Foundation of Computer Ethics*, keynote address at the AICEC99 Conference, Melbourne, Australia, July 1999
- Ed Finn, *Che cosa vogliono gli algoritmi. L’immaginazione nell’era dei computer*, Torino, Einaudi, 2018
- Hendrik Kempt, *(Un)explainable Technology*, Palgrave Macmillan, 2024
- Luciano Floridi, *L’etica dell’intelligenza artificiale. Sviluppi, opportunità, sfide*, Milano, Raffaello Cortina Ed., 2022
- Mohammad Amir Khusru Akhtar, Mohit Kumar, Anand Nayyar, *Towards Ethical and Socially Responsible Explainable AI Challenges and Opportunities*, Springer Nature Switzerland, 2024

## Web references:

- European Parliamentary Research Service, *The ethics of artificial intelligence: Issues and Initiatives*, Scientific Foresite Unit, March 2020 [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS\\_STU\(2020\)634452\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf)
- Raffini, D., & Gorrieri, L. *Intelligenza artificiale e questioni etiche*. *Bioetica*, 31(2) 2023, 332–350.
- Vaassen, B. *AI, Opacity, and Personal Autonomy*. *Philosophy & Technology*, 35(4), 2022, 1–20. <https://doi.org/10.1007/s13347-022-00577-5>
- Carlo Cardia, *Il fondamento etico del diritto*, Rivista telematica <https://www.statoechiese.it>, fascicolo n. 7 del 2021: [https://d1vbhhqv6ow083.cloudfront.net/contributi/Cardia.M\\_Il\\_fondamento.pdf](https://d1vbhhqv6ow083.cloudfront.net/contributi/Cardia.M_Il_fondamento.pdf)
- Resolution of the European Parliament of February 16, 2017, containing recommendations to the Commission concerning civil law rules on robotics (2015/2103(INL) (2018/C 252/25), see in particular pf. AA, in OJEU 18.7.2018, C 252/239
- Terrell Ward Bynum, *The Foundation of Computer Ethics*, Melbourne, Australia, keynote address at the AICEC99 Conference, July 1999, pp.6-13 <https://doi.org/10.1145/572230.572231>
- Video, Corriere, October 2024: [video.corriere.it/video-virali/fa-parcheggiare-l-auto-al-pilota-automatico-ma-finisce-malissimo/27ce0bb9-b146-44b5-95ed-24bec12faxlk](https://video.corriere.it/video-virali/fa-parcheggiare-l-auto-al-pilota-automatico-ma-finisce-malissimo/27ce0bb9-b146-44b5-95ed-24bec12faxlk).
- High-level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, Bruxelles, 2019: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Cristiana Benetazzo, *Intelligenza artificiale e diritto: la sfida etica ed antropologica*; in *Journal of Ethics and Legal Technologies* – Volume 6(2) – December 2024 : <https://jelt.padovauniversitypress.it/system/files/papers/JELT-2024-2-5.pdf>
- The Alan Turing Institute, Dr David Leslie, *Understanding artificial intelligence ethics and safety. A guide for the responsible design and implementation of AI systems in the public sector*, 2019: [https://www.turing.ac.uk/sites/default/files/2019-08/understanding\\_artificial\\_intelligence\\_ethics\\_and\\_safety.pdf](https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf)
- Committee on Equality and Non-Discrimination.. *Preventing discrimination caused by the use of artificial intelligence* (Report). Rapporteur: Mr Christophe Lacroix, Belgium, Socialists, Democrats and Greens Group, (2020)
- Fabio Grigenti, *La prospettiva etica europea sull'intelligenza artificiale*, 2019: [https://www.academia.edu/72039360/Fabio\\_Grigenti\\_La\\_prospettiva\\_etica\\_sull\\_intelligenza\\_artificiale](https://www.academia.edu/72039360/Fabio_Grigenti_La_prospettiva_etica_sull_intelligenza_artificiale)
- Hendrik Kempt, Jan-Christoph Heilinger, Saskia K. Nagel, *Relative explainability and double standards in medical decision-making*, *Ethics and Information Technology*, Volume 24, Article number 20, (2022) : <https://link.springer.com/article/10.1007/s10676-022-09646-x>

- Germana Lo Sapiro, *L'Artificial Intelligence Act e la prova di resistenza per la legalità algoritmica*, in *Federalismi.it*, Rivista di diritto pubblico italiano, comparato, europeo, Osservatorio trasparenza 10.07.2024, n. 16/2024, p. 280, available at <https://www.federalismi.it/nv14/articolo-documento.cfm?Artid=5086>
- Akhtar, M.A.K., Kumar, M., Nayyar, A. (2024), *Ensuring Fairness and Non-discrimination in Explainable AI*. In: *Towards Ethical and Socially Responsible Explainable AI. Studies in Systems, Decision and Control*, vol 551. Springer
- Gryz Jarek; Rojszczak Marcin, *Black box algorithms and the rights of individuals: No easy solution to the “explainability problem*, *Internet Policy Review*, 10(2), <https://doi.org/10.14763/2021.2.1564>
- Sandra Wachter, Brent Mittelstadt, Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, in *International Data Privacy Law*, Volume 7, Issue 2, May 2017, <https://dx.doi.org/10.2139/ssrn.2903469>
- McDermid JA, Jia Y, Porter Z, Habli I., *Artificial intelligence explainability: the technical and ethical dimensions*. *Phil. Trans. R. Soc.*, 2021. <https://doi.org/10.1098/rsta.2020.0363>
- Opinion of Advocate General P. Pikamäe - CGUE, delivered on 16 March 2023 - Case C-634/21 OQ v Land Hesse, Joined party: SCHUFA Holding AG: <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A62021CC0634&qid=1737582781025>
- Opinion of Advocate General Richard De La Tour delivered on 12 September 2024, Case C-203/22 CK. Interested parties: Dun & Bradstreet Austria GmbH, Magistrat der Stadt Wien: <https://eur-lex.europa.eu/legal-content/IT/TXT/?uri=CELEX%3A62022CC0203&qid=1737582973197>