Giulio Cotogni *

GENERAL SECTION

THE EXPLAINABILITY OF AUTOMATED DECISION-MAKING: A HISTORICAL PERSPECTIVE THROUGH EU LEGISLATION

Abstract

There has been much discussion about the existence of a right to explanation of automated decision-making (ADM) in the General Data Protection Regulation (GDPR). However, little attention has been given to the evolution of the regulation of ADM, within the European Union, over the past thirty years. This paper aims to fill this gap in the literature, providing the reader with a look at this topic through the lens of a historical perspective, starting from the very first regulation of ADM in the Data Protection Directive, continuing with the GDPR and, finally, analysing how the right to explanation has ultimately been established in the Artificial Intelligence Act. We will also see how the EU has addressed the issue of transparency and explainability of ADM in other recent pieces of legislation (the 2019 reform of the EU consumer protection law, the Digital Services Act and the Platform Work Directive). Starting from this historical reconstruction of the EU regulation, a common thread will be identified: the tendency to impose increasingly stringent rules regarding the transparency and explainability of ADM. Lastly, three possible explanations for this regulatory development will be proposed.

JEL CLASSIFICATION: K10, K30, K38

SUMMARY

[*] Graduate in law from the University of Turin, Email: giulio.cotogni@edu.unito.it.

1 Introduction

The enormous amount of data available that characterises today's society[1], combined with the increase in computing capacity over the last thirty years[2], now allows Artificial Intelligence (AI) systems, especially those that exploit machine learning (ML) techniques, to render opinions, provide answers, and make decisions very quickly and accurately. Therefore, in many sectors, these systems are increasingly helping[3] (or replacing) humans, especially when it comes to making decisions. This process is called automated decision-making (ADM). This term refers to any process that allows, through the use of technological tools, to make decisions without, or at least with minimal human involvement[4]. Although ADM do not necessarily involve the use of AI technologies, most automated decisions today are made by AI systems.

Although these systems can be a booster for human prosperity, they don't come without risks[5]. In fact, it has been shown that the outputs produced by AI systems can be biased[6], erroneous[7], discriminatory[8] and can violate our privacy[9].

[1] See 'Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020' (*Statista,* 2024) <https://www.statista.com/statistics/871513/worldwide-data-created/> accessed 13 November 2024.
Camilla Tabarrini, 'Comprendere la "Big Mind": il GDPR sana il divario di intelligibilità uomo-macchina?' (2019) 2 Il diritto dell informazione e dell informatica 555, argues that the growth in the amount of available data is mainly due to the so-called "user-generated content" and the Internet of Things (IOT).

[2] See 'Computational capacity of the fastest supercomputers' (*OurWorldinData*, 2023) <https://ourworldindata.org/grapher/supercomputer-power-flops> accessed 13 November 2024.

[3] '[Artificial Intelligence] will change our lives by improving healthcare (eg making diagnosis more precise, enabling better prevention of diseases), increasing the efficiency of farming, contributing to climate change mitigation and adaptation, improving the efficiency of production systems through predictive maintenance, increasing the security of Europeans, and in many other ways that we can only begin to imagine'. See 'White Paper on Artificial Intelligence - A European approach to excellence and trust', COM(2020) 65 final 19 February 2020 1.

[4] Emiliano Troisi, 'Decisione algoritmica, Black box e AI etica: il diritto di accesso come diritto a ottenere una spiegazione' (2022) 4 Juscivile 953.

[5] 'At the same time, Artificial Intelligence entails a number of potential risks, such as opaque decision-making, gender-based or other kinds of discrimination, intrusion in our private lives or being used for criminal purposes', see White Paper on Artificial Intelligence (n 3) 1.

[6] Such bias stem mainly from the quality and choice of data with which the algorithms are trained. Kate Crawford, 'The Hidden Biases in Big Data' *Harvard Business Review* (1 April 2013) <https://hbr.org/2013/04/the-hidden-biases-in-big-data> accessed 13 November 2024, points out that 'Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big-data equation as the numbers themselves'.

[7] For example, it has been shown that it is possible to induce the detection system of a self-driving car to misperceive a traffic signal, leading it to confuse a stop sign with a speed limit. See Kevin Eykholt and others, 'Robust Physical-World Attacks on Deep Learning Models' [2018] ArXiv <https://arxiv.org/abs/1707.08945> accessed 13 November 2024.

[8] Among the most famous cases is the case of COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), an algorithm used by several US states since 2001 as a tool for judges to assess the risk of recidivism of convicted offenders and which proved, all things being equal, to discriminate against African-American criminals, predicting a higher recidivism risk for them than for white offenders, precisely because it was trained on a set of precedents that reflected this discrimination (i.e. on a set of precedents in which African-American offenders actually had a higher recidivism rate than white offenders). See Ellora T Israni, 'When an Algorithm Helps Send You to Prison' (*The New York Times*, 26 October 2017) <https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html> accessed 13 November 2024.

[9] A famous example of this mechanism is the case of an algorithm used by the Target supermarket chain, which, on the basis of the purchases made by a girl (who was, moreover, underage), correctly predicted that she was pregnant (before

Finally, AI systems suffer from an additional problem, namely that of *opacity*[10]. This term refers to the fact that, especially with regard to more sophisticated systems, it is increasingly complex for human operators to understand *how* and *why* the software has produced a certain output. To refer to these opaque systems, the term "black boxes" has been coined in the literature[11], ie systems in which "the computing operations of algorithmic systems [...] become too complex or intricate to comprehend"[12] and therefore "we can observe its inputs and outputs, but we cannot tell how one becomes the other"[13]. The problem posed by these black boxes is all the greater in the light of what has been said above: if the outputs produced by these systems are anything but objective and infallible, but, on the contrary, can be biased, erroneous and discriminatory, then the claim to make these instruments more transparent and, therefore, to obtain an explanation for their output, appears all the more legitimate.

Furthermore, the phenomenon of black boxes entails a further significant critical issue: by hindering the transparency of the various stages of the procedure, and thus compromising the possibility of verifying the validity of the reasons supporting the decision taken, black boxes pose a major obstacle to the full legitimation of the use of automated decisions and, more generally, undermines citizens' trust in AI technologies[14]. Consequently, recent years have seen the proliferation of ethical charters, guidelines and recommendations worldwide, reflecting the growing demand for greater transparency and explainability of AI systems and ADM[15]. For example, at the EU level, the Recommendation on the human rights impacts of algorithmic systems states that "[t]he use of algorithmic systems in decision-making processes that carry high risks to human rights should be

her parents even knew) and sent her vouchers for baby products. See Kashmir Hill, 'How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did' (*Forbes*, 11 August 2022) <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/> accessed 13 November 2024.

[10] Opacity seems to be at the very heart of new concerns about 'algorithms' (operating on data) among legal scholars and social scientists", Jenna Burrell, 'How the machine 'thinks': Understanding opacity in machine learning algorithms' (2016) 3(1) Big Data & Society 1.

[11] The term was coined by Frank Pasquale, '*The Black Box Society. The Secret Algorithms That Control Money and Information*' (Harvard University Press 2015). According to the Author, the black box metaphor correctly represents contemporary reality, in which people are increasingly controlled and surveilled by private corporations and governments, but are unaware of how information and data concerning their lives are disclosed and used by these entities.

[12] Sylvia Lu, 'Data Privacy, Human Rights, and Algorithmic Opacity' (2023) 110 California Law Review 2098.

[13] See Pasquale (n 11) 3.

[14] Carlo Casonato and Barbara Marchetti, 'Prime osservazioni sulla proposta di regolamento dell Unione Europea in materia di intelligenza artificiale' (2021) 3 BioLaw Journal – Rivista di BioDiritto 427.

[15] An interesting, albeit now quite dated, research of 2019, identified 84 documents globally containing ethical principles and guidelines on AI. Analyzing these documents, eleven general principles appear to emerge: transparency, justice and fairness, non-maleficence, responsibility and accountability, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity. Among these, although there is not one mentioned explicitly in all the documents, the one most referred to is the principle of transparency (found in as many as 73 documents), which is declined precisely in terms of "explainability". See Anna Jobin, Marcello Ienca and Effy Vayena, 'The global landscape of AI ethics guidelines' (2019) 1 Nature Machine Intelligence 389. The same conclusions are reached by Jessica Fjeld and others, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' (Berkman Klein Center Research Publication 2020) 1.

subject to particularly high standards as regards the explainability of processes and outputs"[16] and "[a]ffected individuals and groups should be afforded effective means to contest relevant determinations and decisions. As a necessary precondition, the existence, process, rationale, reasoning and possible outcome of algorithmic systems at individual and collective levels should be explained"[17]. The principle of explicability is also affirmed in the Ethics Guidelines for Trustworthy AI, formulated by the High Level Expert Group on Artificial Intelligence, where it is said that "[e]xplicability is crucial for building and maintaining users' trust in AI systems. This means that processes need to be transparent […] and decisions […] explainable to those directly and indirectly affected"[18].

The possible risks posed by automated decisions, have created a debate about the need for a new right: on the assumption that, except in cases provided by law, a decision made by a human being does not give rise in the person concerned to a right to an explanation of the decision, the question was raised whether, if the decision is instead made by an algorithm, it is necessary to configure in the person concerned a *right to explanation*[19].

This paper examines the EU regulation of automated decisions and is structured as follows. Sections 2-3 look at EU regulation of automated decisions from a historical perspective, starting with the very first regulation of the subject with the Data Protection Directive (DPD), continuing with the General Data Protection Regulation (GDPR), and ending with the recent Artificial Intelligence Act (AI Act). Section 4 shows how transparency and explainability of automated decisions have also made their way into specific areas of EU legislation. Section 5 highlights what, in the writer's opinion, is the common thread that has characterised the evolution of EU regulation. Section 6 proposes three possible explanations of this thread. Finally, in Section 7, some conclusions are drawn.

[16] Recommendation of the Committee of Ministers to member States on the human rights impacts of algorithmic systems CM/Rec(2020)1, 8 April 2020, para 4.1.

[17] ibid para 4.3.

[18] 'Ethics Guidelines for Trustworthy AI', 8 April 2019, 13. See, also, para 1.4., where it is stated that "Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions […] technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. […] Whenever an AI system has a significant impact on people s lives, it should be possible to demand a suitable explanation of the AI system s decision-making process. Such explanation should be timely and adapted to the expertise of the stakeholder concerned (eg layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available".
See also the OECD AI Policy Observatory definition of the principle of transparency and explainability: OECD, 'Recommendation of the Council on Artificial Intelligence' (2024), available at <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449> accessed 15 October 2024 and, on the same topic, the UNESCO 'Recommendation on the Ethics of Artificial Intelligence' (*UNESCO*, 23 November 2021) <https://unesdoc.unesco.org/ark:/48223/pf0000381137> accessed 16 November 2024, (III) para 40.

[19] Jacopo Dirutigliano, 'Trasparenza a spiegabilità degli algoritmi' in Ugo Pagallo and Massimo Durante (eds), *La politica dei dati* (Mimesis edizioni 2022) 282.

2 The evolution of ADM regulation from the Data Protection Directive to the General Data Protection Regulation

The decision to analyse the DPD and the GDPR together derives from two reasons. First, the GDPR stands as the successor to the Directive in the field of personal data protection, since, with its entry into force in May 2018, it repealed the latter. Second, Article 15 of the DPD, which first regulated the topic of automated decisions, is taken up almost identically by Article 22 of the GDPR. The structure of the two Articles is, in fact, very similar: both enshrine the right of the individual not to be subjected to automated decisions[20], both provide for exceptions to this prohibition in specific cases[21] and both, in such cases, provide a number of safeguards for the person subjected to ADM.

Although Article 22 GDPR has not much changed from Article 15 of the DPD, a few changes are still noteworthy and, moreover, the practical importance of the provision has increased with augmented use of ADM in our society.

Firstly, although both mention, in the same words, “the right not to be subject to a decision", this “right” has been interpreted in two different ways[22]: whereas in the DPD it was considered to all intents and purposes a right, so that the person unfairly subjected to an automated decision has the burden of exercising it[23], in the GDPR, on the other hand, it is not a right, but a literal prohibition, so it is not necessary for the person concerned to take action[24].

[20] Article 15(1) states that ‘Member States shall grant the right to every person not to be subject to a decision which produces legal effects concerning him or significantly affects him and which is based solely on automated processing of data intended to evaluate certain personal aspects relating to him, such as his performance at work, creditworthiness, reliability, conduct, etc.’ Article 22(1) states that ‘The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her’.

[21] Article 22(2) states that paragraph 1 shall not apply if the decision (a) is necessary for a contract between the data subject and a data controller; (b) is authorised by Union or Member State law; (c) is based on the data subject's explicit consent.

Article 15(2) allows derogations from paragraph 1 if the decision (a) is necessary for a contract, requested by the data subject, between the data subject and a data controller; (b) is authorised by a law.

[22] This divergence in interpretation is partly the result of the different regulatory nature of the two acts. In the case of the GDPR, in fact, the choice of the regulatory source instead of the directive entails the creation of uniform constraints that are directly applicable throughout the entire territory of the EU and removes from the Member States those margins of discretion that instead characterised the interpretation of the DPD and that most likely weakened it. See Barbara Marchetti and Leonardo Parona, ‘La regolazione dell’Intelligenza Artificiale: Stati Uniti e Unione Europea alla ricerca di un possibile equilibrio’ (2022) 51(1) DPCE online 237.

[23] Lee A Bygrave Dr., ‘Minding the machine: Article 15 of the EC Data Protection Directive and automated profiling’ (2001) 17(1) Computer Law & Security Report 17, 18 ‘Article 15(1) does not take the form of a direct prohibition on a particular type of decision making (profile application). Rather it directs each EU Member State to confer on persons a right to prevent them being subjected to such decision making [...]. This would leave the actual exercise of the right to the discretion of each person’.

[24] As clarified by the Court of Justice of the European Union (CJEU) in the ‘SCHUFA case’ (Case C-634/21 *OQ v Land Hessen* [2023] ECLI:EU:C:2023:957), para 52: ‘Article 22(1) of the GDPR confers on the data subject the 'right' not to be the subject of a decision solely based on automated processing, including profiling. That provision lays down a prohibition in principle, the infringement of which does not need to be invoked individually by such a person’.

See also Maja Brkan, ‘Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond’ (2019) 27(2) International Journal of Law and Information Technology 91, 99 where it is stated

Secondly, in the GDPR, explicit consent is included as a case in which ADM is allowed[25] and, finally, as opposed to the provisions in Article 15 of the directive, it is no longer necessary that the data subject requests the contract in order for the automated decision to be lawful[26].

However, the key distinction between the two provisions lies in the safeguards afforded to the person subjected to an automated decision, which, as has already been said, is admissible only when one of the exceptions outlined in Article 22(2) of the GDPR or Article 15(2) of the DPD applies.

Under the DPD, the only safeguard available to the person subjected to an automated decision, is the opportunity to “put his point of view”, enshrined in Article 15. This provision is linked to Article 12 (right of access) which provides for the right to obtain from the controller: “knowledge of the logic involved in any automatic processing of data concerning him at least in the case of the automated decisions referred to in Article 15(1)”.

In the GDPR, on the other hand, the rules dealing with the accountability of automated decisions are more numerous and are contained in Articles 13, 14, 15, 22(3) and Recital 71. All together, these provisions put in place “a broader, stronger, and deeper algorithmic accountability regime than what existed under the EU’s Data Protection Directive”[27]. Article 22(3) of the GDPR, in fact, in addition to the right to express his or her point of view” (similar to the possibility to put his point of view” enshrined in Article 15 DPD), also guarantees the right to obtain human intervention on the part of the controller” and, above all, to contest the decision”.

Articles 13(2)(f) and 14(2)(g) establish the data subject's right to be informed, while Article 15(1)(h) guarantees the data subject's right of access. All three provisions, with identical wording, require the data controller to provide the data subject with a range of information, including the existence of an ADM under Article 22 and, at least in those cases, “meaningful information about the logic involved” and “the significance and the envisaged consequences” of the decisions. In the GDPR, unlike the DPD, there is the addition of the term “meaningful”, which means that the controller should convey information about the rationale and the criteria relied upon in reaching the decision,

that ‘Interpreting Article 22(1) as giving data subject the right that she has to actively exercise could in consequence lead to detrimental effects for her and run contrary to the purpose of this provision […]. A systematic interpretation of Article 22 implies that only automated decisions fulfilling the requirements of paragraph 2 and allowing for safeguards from paragraph 3 of this provision are authorised by the GDPR. Therefore, […] it is more appropriate to construct the data subjects' 'right' as a prohibition of fully automated decision-making that the data controllers have to comply with’. This position is also confirmed by the EC ‘Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679’ (WP29 Guidelines), 22 August 2018, 20, where it is stated that Article 22(1) ‘establishes a general prohibition’ of automated decision-making meaning that ‘individuals are automatically protected from the potential effects this type of processing may have’.

[25] See Article 22(2) GDPR (n 21).

[26] Sandra Wachter, Brent Mittelstadt and Luciano Floridi, ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’ (2017) 7(2) International Data Privacy Law 76, 82.

[27] Margot E Kaminski, ‘The Right to Explanation, Explained’ (2019) 34(1) Berkeley Technology Law Journal 190, 193.

therefore the quality of being "meaningful" must be evaluated from the perspective of the data subject[28]. In order to make this information meaningful and understandable, the Guidelines on Automated individual decision-making and Profiling, drawn up by the Article 29 Working Party (WP29 Guidelines), state that "real, tangible examples of the type of possible effects should be given"[29]. The reference to the "significance" and the "envisaged consequences" of the decision refer back to the idea that, for the purposes of contestation (Article 22), it is essential to fully understand the concrete results and the risks emanating from the contextual use of the data[30]. In fact, the WP29 Guidelines clarify that "[t]he data subject will only be able to challenge a decision or express their view if they fully understand how it has been made and on what basis"[31].

Lastly, Recital 71 takes over the content of Article 22(3) and, although it has no legal effect[32], constitutes the only provision in which the Regulation expressly mentions the term *explanation*: in fact, the Recital states that "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".

While it is common ground that Article 15 of the DPD did not enshrine any right to an explanation of automated decisions, the existence of such a right in the GDPR, on the other hand, has been the subject of a lengthy debate in the doctrine, between those who, on the one hand consider that the GDPR enshrines a genuine right to an explanation of the specific decision[33] and those who, on the other hand, argue for the existence of a much more limited "right to be informed"[34].

Regardless of this debate, the regulation certainly makes some important steps forward with respect to the discipline contained in the DPD on ADM accountability regime[35]: as

[28] Emre Bayamlıoglu, 'The right to contest automated decisions under the General Data Protection Regulation: Beyond the so-called right to explanation"' (2022) 16(4) Regulation & Governance 1058, 1067.

[29] WP29 Guidelines 26.

[30] See Bayamlıoglu (n 28) 1067.

[31] WP29 Guidelines 27.

[32] CJEU in Case C-355/95 *P Textilwerke Deggendorf GmbH (TWD) v Commission of the European Communities and Federal Republic of Germany* [1997] ECLI:EU:C:1997:241, para 21 states that 'the operative part of an act is indissociably linked to the statement of reasons for it, so that, when it has to be interpreted, account must be taken of the reasons which led to its adoption'.

[33] See Troisi (n 4); Emiliano Troisi, 'AI e GDPR: L'automated decision making, la protezione dei dati e il diritto alla "intelligibilità" dell'algoritmo' (2019) 1 European Journal of Privacy Law & Technologies 41; Gianclaudio Malgieri and Giovanni Comandè, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (2017) 7(4) International Data Privacy Law 243; Bryce Goodman and Seth Flaxman, 'European Union regulations on algorithmic decision-making and a right to explanation' (2017) 38(3) AI Magazine 50.

[34] Sandra Wachter and others (n 26); Lilian Edwards and Michael Veale, 'Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for' (2017) 16(1) Duke Law & Technology Review 18.

[35] About GDPR, Kaminski (n 27) 208, states that "[...] this regime, if enforced, has the potential to be a sea change in how algorithmic decision-making is regulated in the EU". About the DPD, Lee A Bygrave Dr. (n 23) 21 says that the right in Article 15(1) "resembles a house of cards [which], in the context of currently common data-processing practices, [...] is quite easy to topple'. Interestingly, he also notes

already mentioned, the DPD simply configured the right to obtain “knowledge of the logic involved in any automatic processing of data” and the right to "put his point of view”, whereas GDPR instead provides the data subject with three stronger tools: the right to obtain human intervention, the right to contest the decision and the right to obtain meaningful information.

Despite this, the GDPR has proven insufficient to fully ensure the explainability of automated decisions. Two main criticisms have been made. First, the regulation, does not elaborate much beyond suggesting the existence (never established by the CJEU case law) of a right to an explanation of automated decisions[36]. Moreover, the practical relevance of such a right has been almost meaningless, given the absence of litigation on the merits. Second, the scope of Article 22 is rather limited, since, for a decision made by automated means to fall under it, it must be based *solely* on automated processing (including profiling): hence, all those decision-making processes in which there is human intervention, albeit minimal, remain excluded from the scope of Article 22[37] and, therefore, from access to the guarantees offered by the GDPR. Moreover, as pointed out in doctrine[38], also the requirement that the decision produces “legal effects” on the individual or affects him or her “in a similar significant way” poses some interpretive problems that contribute to undermining the scope of the provision[39].

3 The right to explanation in the Artificial Intelligence Act

The AI Act explicitly recognises, for the first time in EU legislation, the existence of a right to an explanation of automated decisions.

Paragraph 1 of Article 86 states that any affected person subject to a decision which is taken by the deployer “on the basis of the output from a high-risk AI system” and which “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” has the

that: 'Nevertheless, this situation might well change in the future if, as is likely, automated profiling becomes more extensive”.

[36] Themistoklis Tzimas, ‘Algorithmic Transparency and Explainability under EU Law’ (2023) 29(4) European Public Law 385, 400.

[37] Sandra Wachter and others (n 26). Other Authors, on the other hand, have preferred a broader interpretation of this requirement, deeming included in the definition all those decisions that are ‘automated in substance’, that is, those in which human intervention, while present, is essentially irrelevant in determining the final decision, see Emiliano Troisi, ‘AI e GDPR: L’automated decision making, la protezione dei dati e il diritto alla “intelligibilità” dell’algoritmo’ (2019) 1 European Journal of Privacy Law & Technologies 41, 47; Iole Pia Di Ciommo, ‘La prospettiva del controllo nell’era dell’Intelligenza Artificiale: alcune osservazioni sul modello Human In The Loop’ (2023) 9 Federalismi.it 68, 75.

[38] See Sandra Wachter and others (n 26) 92-93; Tal Zarsky, ‘Incompatible: The GDPR in the Age of Big Data’ (2017) 47 4(2) Seton Hall Law Review 995.

[39] However, we can refer to what the WP29 Guidelines state with respect to Article 22 of the GDPR, i.e. that a decision producing ’legal effect’ is a decision affecting data subject's legal rights, legal status or her rights under a contract, while a decision producing ‘similarly significantly affects’ does not mean that this effect needs to have any legal implications for the data subject; rather, ‘similar’ refers to the significance and not the nature of the effect. Also, the Guidelines provide some examples of such significant effect: automated decisions affecting data subject's financial circumstances, access to health services or education. See WP29 Guidelines 21.

right to obtain from the deployer[40] "clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken".

This provision applies to decisions made on the basis of the output from a "high-risk AI system". In the AI Act the different applications of AI technologies have been classified into three categories (prohibited AI practices, high-risk systems, low or minimal risk systems), based on the risk they may pose to the framework of fundamental values and rights of the EU. The choice of the EU regulator was to limit the right to an explanation to high-risk systems only. The rationale behind this choice is to avoid requiring the deployer to provide an explanation for outputs produced by AI systems that pose less risk to fundamental values and rights, because, as it has been pointed out in the literature, there is a certain trade-off between the explainability of an AI system and its degree of accuracy[41]. At the current stage, there are eight categories of AI systems considered to be high-risk: systems used for biometrics, systems involved in the management of critical infrastructures (e.g. the water, gas or electricity supply system), systems used for education and vocational training, systems used for employment and the management of workers, systems that determine access to essential private and public services, systems used for law enforcement, systems used in immigration management and border control, and systems used in the administration of justice and democratic processes (e.g. the elections)[42].

The scope of Article 86 is broader than that of Article 15 of the DPD or Article 22 of the GDPR, since the requirement that the decision be "based solely" on automated processing has disappeared[43], so Article 86 also applies in all those cases where the AI system is used merely as a support for the decision made by a human being. This is certainly an important step forward in the regulation of ADM, since, by equating fully automated decisions and those in which the AI system simply acts as a support to the human decision maker, the EU legislator is showing awareness of the tendency of the human decision maker to conform to algorithmic reasoning and not to deviate from it, considering it to tend to be

[40] According to Article 4(4) 'deployer' means any 'natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity'.

[41] '[U]nfortunately, in many contexts, the better-performing systems are the less explainable ones. In particular, neural networks are often the most effective approach to deal with pattern recognition and natural language processing. Thus, predictive performance and transparency are often conflicting objectives and there will have to be a trade-off between the two.', Mateusz Grochowski and others, 'Algorithmic Transparency and Explainability for EU Consumer Protection: Unwrapping the Regulatory Premises' (2021) 8(1) Critical Analysis of Law 43, 48.
Also, the 'Ethics Guidelines for Trustworthy AI' (n 18), 18, admits that 'trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability)'. For more on this topic see also Alex A Freitas, 'A Critical Review of Multi-objective Optimization in Data Mining: A Position Paper' (2004) 6(2) ACM SIGKDD Explorations Newsletter 77; Philipp Hacker and others, 'Explainable AI under Contract and Tort Law: Legal Incentives and Technical Challenges' (2020) 28 Artificial Intelligence and Law 415, 430-431.

[42] See Annex III AI Act.

[43] Article 86, in fact, speaks of a decision which is taken by the deployer "on the basis of the output from a high-risk AI system".

infallible ("moutunnier effect"[44]). In addition, unlike previous legislation, the AI Act regulates the issue of the explainability of the ADM, regardless of whether or not personal data processing takes place.

Nevertheless, in other respects, the scope of the provision is more specific. In fact, since under Article 3 of the AI Act[45] an AI system is only defined as such if it possesses a certain degree of autonomy, this means that if a system does not possess a minimum degree of autonomy, it will not fall within the scope of Article 86.

For the person subject to the automated decision to be able to assert this right, the decision must produce "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". The rationale behind this requirement is to avoid requiring the deployer to provide an explanation for outputs that do not substantially affect the individual, because of the aforementioned trade-off between the explainability of an AI system and its degree of accuracy[46]. Nevertheless, the provision bases the existence of "legal effects" or "similarly significantly affects" on the subjective perception of the individual, which will probably make it quite easy to prove.

Turning, finally, to the content of the right to an explanation, Article 86 states that the person is entitled to obtain from the deployer "clear and meaningful explanations of (i) "the role of the AI system in the decision-making procedure and" (ii) "the main elements of the decision taken". To clarify the practical content of the right to explanation, we can appeal to other provisions contained in Section 2 of the AI Act, which sets out the requirements for high-risk AI systems.

Article 11 states that before a high-risk AI system is placed on the market or put into service, detailed technical documentation must be prepared (and kept updated during the entire lifetime of the AI system). This documentation serves to prove that the system complies with the requirements set out in Section 2, with a view to ensuring a form of *ex ante* transparency. The technical documentation should include certain key elements[47], including information on: (i) the general logic of the AI system and of the algorithms, (ii) the main classification choices and the relevance of the different parameters, (iii) the description of the expected output and output quality of the system, (iv) the training methodologies and techniques and the training data sets used, including a general description of these data sets and (v) information of the human oversight measures needed in accordance with Article 14.

[44] The expression, invoked with regard to the justice sector, is due to Antoine Garapon and Jean Lassègue, *Justice digitale. Révolution graphique et rupture anthropologique* (PUF 2018) 239.

[45] Article 3(1) defines an AI system as "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".

[46] See Grochowski and others (n 41).

[47] See Annex IV AI Act.

Article 12 deals with ensuring the traceability of actions performed by the AI system during its operation. In fact, it states that high risk AI systems “shall technically allow for the automatic recording of events (logs) over the lifetime of the system”. The rationale behind this requirement is to ensure greater transparency *during* the operation of the AI system. The importance of the principle of traceability is highlighted both by Recital 27 of the AI Act, which links transparency with traceability and explainability[48], and by the Ethics Guidelines for Trustworthy AI, which state that traceability “facilitates auditability as well as explainability”[49].

Article 13 deals with ensuring that AI systems are designed and developed in such a way as to ensure that their operation is sufficiently transparent to allow deployers to interpret the system's output and use it appropriately. To this end, Article 13 affirms that these systems shall be accompanied by “instructions for use”[50]. These instructions shall contain, at least, (i) information about the intended purpose of the AI system, (ii) the level of accuracy, robustness and cybersecurity of the AI system, (iii) any known or foreseeable circumstance which may lead to risks to the health and safety or fundamental rights, (iv) its technical capabilities to provide information to explain its output (so the deployer knows whether the system is a “black box” or not), (v) its performance regarding specific individuals or groups of individuals on which the system is intended to be used and (vi) specifications for the input data[51]. This provision is particularly relevant for the purpose of ensuring that the right to an explanation is effective, since, under Article 86, the deployer is the party responsible for providing an explanation to the person subject to the automated decision made through a high-risk AI system.

The last important provision regarding high-risk AI systems is Article 14, which enforces the principle of human oversight. In particular, natural persons to whom human oversight is assigned should be enabled to (i) monitor its operation (e.g. to detect anomalies); (ii) be aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system […] in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons” (“moutunnier effect”[52]); (iii) correctly interpret system s output; and (iv) decide not to use the high-risk AI system” or to otherwise disregard, override or reverse the output of the high-risk AI system” and to interrupt the system through a stop” button[53]. This

[48] Recital 27 affirms that ‘[t]ransparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability’.
[49] Ethics Guidelines for Trustworthy AI 18.
[50] ‘In an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers’, see Article 13(2) AI Act.
[51] See Article 13(3) AI Act.
[52] See Garapon and Lassègue (n 44).
[53] Furthermore, paragraph 5 of Article 14, strengthens the principle of human in the loop with regard to the outputs produced by remote biometric identification systems (see point 1(a) of Annex III), which are considered particularly dangerous to the fundamental rights of individuals. It is stipulated that in order for the deployer to take a decision/action on the basis of the identification resulting from the system, that identification must be confirmed by at least two natural persons with the necessary competence, training and authority.

provision aims to ensure the principle of human in the loop and requires that high-risk systems be developed with a design that allows for human oversight (principle of transparency-by-design). The rationale behind this principle is that human oversight may prevent or minimise the risks to health, safety or fundamental rights arising from the use of the AI system.

Finally, although it does not directly deal with the issue of transparency and interpretability, Article 10 is also worth mentioning, which requires that the training, validation and testing data of the AI system meet certain quality criteria[54]. The principle of data quality (already enunciated in Article 5 of the GDPR with references to personal data), assumes particular relevance in the field of ML, given that decision-making algorithms learn and make decisions on the basis of the data they are provided with and, moreover, as mentioned earlier (see Section 1), much of the bias that afflicts the outputs of AI systems derives precisely from poor quality data.

In conclusion, the provisions we have analysed aim to impose greater transparency and explainability of the decisions produced by AI systems at three different stages:

-in the *ex ante* phase (i.e., before the high-risk system is placed on the market) with the obligation to drawn up the technical documentation (Article 11);

-during the operation of the system, both through the obligation to keep log files of the AI system (Article 12) and through the principle of human oversight and human in the loop (Article 14);

-in the *ex post* phase (i.e., after the system has produced the output) by ensuring that the deployer correctly interprets and uses the system's output (Article 13) and is therefore able to provide an explanation to the person subjected to the automated decision (Article 86).

The right to explanation, along with the other rules on transparency and explainability of ADM outlined in the AI Act, hold significant practical relevance in contemporary society. In fact, Recital 171 emphasises that the explanation mandated by Article 86 “should be clear and meaningful and should provide a basis on which the affected persons are able to exercise their rights”. Therefore, within the AI Act, transparency and explainability are more than just broad principles: they are seen as essential tools to enable the exercise of fundamental rights and to safeguard key principles of the legal system, which are also guaranteed by the Charter of Fundamental Rights of the European Union (CFREU)[55]. Recitals from 54 to 61 identify, for each of the eight categories of high-risk AI systems, the fundamental rights and legal principles safeguarded by the rules on transparency and explainability. Recital 54, which deals with AI systems used for remote biometric identification, emphasises the principle of non-discrimination[56]; Recital 55, which deals with AI systems used in critical infrastructures, highlights the protection of human life and

[54] See paragraphs 2-5 Article 10 AI Act.

[55] Charter of Fundamental Rights of the European Union 2012/C 326/02 of 26 October 2012 [2012] OJ C326/391 (CFREU).

[56] Article 21 CFREU.

health and the protection of social and economic activities; Recital 56 recalls the right to education and training[57]; Recital 57 refers to workers' rights[58]; Recital 58, which deals with AI systems used to determine an individual's access to essential public and private services[59], focuses on the right to social protection, human dignity and the right to an effective remedy; Recital 59 stresses that transparency and explainability of ADM are necessary to enable the individual to exercise important procedural fundamental rights, such as the right to an effective remedy and to a fair trial[60], as well as the right of defence[61] and the presumption of innocence[62]; Recital 60, which deals with the AI systems used in the management of migration, asylum and border control, refers to the rights to free movement, non-discrimination, protection of private life and personal data, international protection and good administration[63] and, finally, Recital 61 emphasises the importance of transparency in the AI systems used in the administration of justice as a necessary condition for safeguarding democracy, the rule of law, individual freedoms as well as the right to an effective remedy and to a fair trial.

In light of the above, it can be affirmed that the right to explanation, as well as the other rules of the AI Act on transparency and explainability, are considered by the EU legislator to be instruments of fundamental importance for protecting a number of concrete rights of the individual, as well as key principles of the EU legal system.

4 Explainability of ADM in other pieces of EU legislation

In Sections 2 and 3 it has been described the historical path that led the EU legislator to finally recognise the right to an explanation in the AI Act, nevertheless, the issue of transparency and explainability of automated decisions is increasingly present within EU legislation, and has also been addressed in other recent pieces of EU legislation. This Section briefly recalls some of the regulations on the subject, contained in the 2019 reform of the EU consumer protection law, in the Digital Services Act[64] (DSA) and in the Platform Work Directive[65].

[57] Article 14 CFREU.
[58] Articles 15 and 31 CFREU.
[59] Articles 34 and 36 CFREU.
[60] Article 47 CFREU.
[61] Recital 59 states that 'The impact of the use of AI tools on the defence rights of suspects should not be ignored, in particular the difficulty in obtaining meaningful information on the functioning of those systems and the resulting difficulty in challenging their results in court, in particular by natural persons under investigation'.
[62] Article 48 CFREU.
[63] Article 41 CFREU.
[64] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L277.
[65] The Proposal for a Directive of the European Parliament and of the Coucil on improving working conditions in platform work COM(2021) 762 final was approved by the European Parliament in April 2024 and is still awaiting Council's 1st reading position.

4.1 Explainability of ADM in the 2019 reform of the EU consumer protection law

As recognised by the EU Resolution 2019/2915[66], the development of ADM in the business-consumer relation, on the one hand, "is expected to make a significant contribution to the knowledge economy and offers benefits [...] for consumers through innovative products and services and for businesses through optimised performance", but, on the other hand, it "also presents challenges for consumer trust and welfare, especially in terms of empowering consumers to identify such processes, to understand how they function, to make informed decisions on their use, and to opt out"[67].

In this field, ML algorithms and ADM can be used to profile consumers, enabling businesses to personalise the prices of goods and services offered to them, a practice known as price discrimination. More generally, these tools can be used to alter consumers' freedom of choice and manipulate their decisions in a way that, before the advent of these technologies, was unthinkable. In fact, even though such attempts at manipulation are not new in the context of the business-consumer relation, the use of AI technologies offers significant possibilities for enhancing these practices, as these tools make it possible to predict consumer behaviour more accurately, in real time, and based on huge amounts of data, which can be derived both from online interactions (through, for example, clicks, likes or purchase history) and from the offline world (Internet of Things)[68]. The new possibilities introduced by AI have therefore transformed previous, "static and undifferentiated"[69] manipulation strategies into "dynamic, interactive, intrusive, and incisively personalisable choices architectures-decision-making contexts that can be specifically designed to adapt and to exploit each individual user's particular vulnerabilities"[70]. Consequently, the use of these tools has increased the information asymmetry between consumer and business, which was already historically present in this field[71].

In light of what has been said so far, the need to enforce greater transparency and explainability of the work of ADM systems has also arisen in the consumer discipline[72], so, in 2019, the EU intervened by amending the regulation.

[66] European Parliament resolution of 12 February 2020 on automated decision-making processes: ensuring consumer protection and free movement of goods and services [2020] OJ C294.

[67] ibid letters B) and C).

[68] Nathalie De Marcellis-Warin and others, 'Artificial intelligence and consumer manipulations: from consumer's counter algorithms to firm's self-regulation tools' (2022) 2(4) AI and Ethics 259, 260.

[69] ibid 261.

[70] Daniel Susser, Beate Roessler and Helen Nissenbaum, 'Online manipulation: Hidden influences in a Digital World' (2019) 4(1) Georgetown Law Technology Review 1, 3-4.

[71] Martin Ebers, 'Liability For Artificial Intelligence And EU Consumer Law' (2021) 12(2) Journal of Intellectual Property, Information Technology 204, 208.

[72] The European Parliament resolution (n 66) paragraph 1, states that consumers 'should be properly informed about how [ADM] function, about how to reach a human with decision-making powers, and about how the system s decisions can be checked and corrected'. See, also paragraph 13, which stresses that 'in light of the significant impact that automated decision-making systems can have on consumers, especially those in vulnerable situations, it is important for those systems not only to use high-quality and unbiased data sets but also to use explainable and unbiased algorithms'.

First, with Directive 2019/2161, Article 7 of the Unfair Commercial Practices Directive 2005/29/EC was amended: the new paragraph 4(a) requires traders to disclose the "main parameters determining the ranking of products presented to the consumer […] and the relative importance of those parameters". Also in 2019, the EU regulator intervened, through Regulation 2019/1150, to impose a similar obligation on online search engine providers: Article 5(2) of the Regulation requires online search engine providers to set out the "main parameters, which individually or collectively are most significant in determining ranking" and "the relative importance of those main parameters". Recital 22 of the 2019/2161 Directive and Recital 24 of the 2019/1150 Regulation both clarify that "main parameters" means any "general criteria, processes, specific signals incorporated into algorithms or other adjustment or demotion mechanisms used in connection with the ranking". Finally, the EU legislator, through Directive 2019/2161, has specifically addressed the practice of price-discrimination implemented by means of algorithms. With the aim of imposing greater transparency on this tool, the new Article 6(1)(ea) of the Consumer Rights Directive[73], stipulates that the trader must inform the consumer "that the price was personalised on the basis of automated decision-making".

Therefore, in the field of business-consumer relations, transparency and explainability of automated decisions are essential to protect consumer rights, such as the right to make free and informed decisions and to rebalance, also guaranteed by Article 38 CFREU.

4.2 Explainability of ADM in the Digital Services Act

Also, in the DSA there is a number of provisions that deal with enforcing the transparency and explainability of automated decisions, particularly those used by online platforms in content moderation processes and recommender systems (RSs).

Both these processes are carried out by algorithms[74], since the huge growth of so-called user-generated content[75] and online user interactions (through clicks, likes, shares, etc.), on the one hand makes it impossible for providers to delegate content moderation activities to human operators alone, and, on the other hand, provides a huge amount of

[73] Directive 2011/83/EU of the European Parliament and of the Council of 25 October 2011 on consumer rights, amending Council Directive 93/13/EEC and Directive 1999/44/EC of the European Parliament and of the Council and repealing Council Directive 85/577/EEC and Directive 97/7/EC of the European Parliament and of the Council [2011] OJ L304.

[74] See Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic content moderation: Technical and political challenges in the automation of platform governance' (2020) 7(1) Big Data & Society 1.
For instance, algorithms currently control more than 95% of content removal and bans on Facebook (up from 23% in 2017); YouTube now reports that 98% of the videos removed for violent extremism are flagged by machine-learning algorithms", and Twitter revealed that 93% of "terrorist content" is reported by proprietary internal tools (i.e., algorithms for detecting terrorist content) and removed. See Sergio Sulmicelli, 'Algorithmic content moderation and the LGBTQ+ community s freedom of expression on social media: insights from the EU Digital Services Act' (2023) 2 BioLaw Journal – Rivista di BioDiritto 471, 478; see also 'Twitter Transparency Report' (5 April 2018) <https://blog.twitter.com/official/en_us/topics/company/2018/twitter-transparency-report-12.html> accessed 14 November 2024.

[75] In 2022, about 500 hours of videos were uploaded every minute on Youtube. See 'Hours of video uploaded to YouTube every minute' (*Statista*, 2024) <https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/> accessed 14 November 2024.

information about the users themselves, which can be used to profile them and show them personalised content that may interest them (through the RSs). The two phenomena (moderation and recommender systems) are thus connected, since when there is a need to work with large amounts of data, algorithms become an indispensable tool. Nevertheless, the use of algorithms in these fields poses a number of ethical and legal problems that the EU has sought to address by requiring greater transparency and explainability.

Beginning with the activity of moderation[76], in this field there is a need to balance the efficiency of the moderation activity performed by algorithms with the principle of freedom of expression online, especially because, even for the most sophisticated algorithms, it is difficult to understand the context behind a certain sentence[77] (with the risk of causing numerous false positives), so if the moderation activity is performed by an algorithm, it is necessary both to make explicit the role it plays and to make the reasons behind its intervention understandable.

To this end, Articles 14 and 15 of the DSA stipulate, respectively, the obligation for providers of intermediary services to outline (in the terms and conditions) information on any policies, procedures, measures and tools used for the purpose of content moderation, “including algorithmic decision-making and human review[78]” and the obligation to, at least once a year, make publicly available a report on the moderation activity carried out on their platform, which must contain, among other things, a disclosure on “any use made of automated means for the purpose of content moderation”[79]. Even though these provisions do not deal with the explicability of algorithmic outputs (i.e., the reasons behind the decision to moderate or not moderate a piece of content), they nonetheless impose a general obligation of transparency on the use of automated systems, similar to what is enshrined in consumer protection law (as reformed in 2019) with respect to the use of ADM in price-discrimination.

[76] Article 3(t) of the Regulation clarifies that ‘content moderation means “the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient s account”.

[77] Natasha Duarte and Emma Llansò, ‘Mixed messages? The limits of automated social media content analysis’ (*Center for Democracy and Technology*, 28 November 2017) <https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/> accessed 14 November 2024, 5: Among studies using NLP to judge the meaning of text (including hate speech detection and sentiment analysis), the highest accuracy rates reported hover around 80%, with most of the high-performing tools achieving 70 to 75% accuracy. These accuracy rates may represent impressive advancement in NLP research, but they should also serve as a strong caution to anyone considering the use of such tools in a decision-making process. An accuracy rate of 80% means that one out of every five people is treated wrong in such decision-making; depending on the process, this would have obvious consequences for civil liberties and human rights”.

[78] Article 14(1) DSA.

[79] “[I]ncluding a qualitative description, a specification of the precise purposes, indicators of the accuracy and the possible rate of error of the automated means used in fulfilling those purposes, and any safeguards applied”, Article 15(1)(e) DSA.

The explainability of moderation activity is addressed in Article 17 DSA, which requires providers of hosting services to provide a clear and specific “statement of reasons” to the affected uploader for each content moderation decision. This statement shall include “information on the use made of automated means in taking the decision, including information on whether the decision was taken in respect of content detected or identified using automated means” and “the facts and circumstances relied on in taking the decision”. The purpose of Article 17 is twofold: on the one hand, it aims to make moderation activity knowable to the user (with a view to countering practices such as shadow banning[80]); on the other hand, it seeks to make moderation activity explainable[81]. Although even Article 17 does not explicitly mention a right to an explanation of the algorithmic decision (i.e., the decision to moderate a piece of content), such a right could perhaps be derived on the basis of the general obligation to justify the moderation activity performed, since this is, in the vast majority of cases, carried out through algorithms.

As for recommender systems, the DSA defines them as “a fully or partially automated system used by an online platform to suggest in its online interface specific information to recipients of the service or prioritise that information, including as a result of a search initiated by the recipient of the service or otherwise determining the relative order or prominence of information displayed”[82]. This definition highlights the method (“fully or partially automated”), aim (“to suggest”), content (“specific information”), target (“recipients of the service”), input (“as a result of a search initiated by the recipient”) and output (“determining the relative order or prominence of information displayed”) of a recommendation process[83].

Recommender systems thus influence a central aspect of the user experience on the online platform, namely what content is shown to them[84]. Moreover, because algorithm-based recommender systems often rely on implicit personal data, such as browsing and click-through history, their functioning is not explained to users, and their influence is not

[80] “Shadow banning” is a term used to refer to a moderation action that allows a particular user to be hidden from an online community, or to make content posted by him invisible to other users. It differs from “banning” proper in that the profile of the affected user is not banned and/or deleted from the platform, and his or her content is not deleted, but is, instead, rendered unavailable to other users. As a result, the user in question remains completely unaware of the sanction and continues to behave normally. For more on this topic, see Paddy Leerssen, ‘An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation’ (2023) 48 Computer Law & Security Review 1.

[81] ibid 6.

[82] Article 3(s) DSA.

[83] Matteo Fabbri, ‘Self-determination through explanation: an ethical perspective on the implementation of the transparency requirements for recommender systems set by the Digital Services Act of the European Union’, *AIES '23: AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2023) 653 <http://dx.doi.org/10.1145/3600211.3604717> accessed 14 November 2024.

[84] These systems allowed people to “filter what they want to read, see, and hear”, not coming “across topics and views that you have not sought out”, Cass R Sunstein, Republic: *Divided democracy in the age of social media* (Princeton University Press 2018). Furthermore, “automated recommendations determine not only what we see on platforms, but also our potential interest for new or different categories of content. This influencing potential can be interpreted as an instance of the “new emerging grey power” of tech companies, which is exercised about which questions can be asked, when and where, how and by whom and hence what answers can be received in principle”, Luciano Floridi, ‘The new grey power’ (2015) 28 Philosophy & Technology 329, 332.

accountable. Furthermore, the use of RSs raises a number of problems: it can lead to the creation of so-called “echo chambers” and the consequent polarization of online debate[85]; it can foster nudging practises[86] and, finally, it incentivises the circulation of viral content[87] that can prove harmful[88].

With the DSA, therefore, an attempt was made to implement greater transparency and explainability of these systems in order to reduce the impact of those harms by increasing users' awareness. Article 27 DSA states that providers of online platforms “shall set out in their terms and conditions, in plain and intelligible language, the main parameters used in their recommender systems”. The aim of this provision is to “explain why certain information is suggested to the recipient of the service”[89]; therefore, the parameters need to include, at least, “the criteria which are most significant in determining the information suggested to the recipient of the service” (i.e., content) and the reasons for its “relative importance”[90] (i.e., ranking). Also, non-binding Recital 70 DSA states that online platforms “should clearly present the main parameters for such recommender systems in an easily comprehensible manner to ensure that the recipients understand how information is prioritised for them”. According to some, a right to explanation for RSs’ outputs could be identified in this formulation: in fact, the “easily comprehensible manner” of presenting the parameters of RSs so that “the recipients understand how information is prioritised for them” can come to effect only if RSs are explainable[91].

Therefore, with respect to content moderation and the activity of recommender systems, the transparency and explainability of ADM are functional to the protection of important individual rights, such as the freedom of expression and information, both guaranteed by Article 11 CFREU.

4.3 Explainability of ADM in the Platform Work Directive

The issue of transparency and explainability of automated decisions has also arisen with reference to the world of labour and, in particular, to the regulation of work on digital

[85] The polarisation of online debate occurs as opinions are no longer exposed to confrontation (since the individual is only exposed to messages that confirm his or her opinions), but, on the contrary, are continually reinforced within “bubbles” in which the same ideas are always circulating.

[86] According to the original definition proposed in behavioural economics, nudges are the features of a choice architecture “that have an influence on which decisions people make”, Richard H Thaler, Cass R Sunstein, ‘*Nudge: Improving Decisions about Health, Wealth, and Happiness*’ (Penguin Books 2009).

[87] Recital 70 DSA states that RSs “play an important role in the amplification of certain messages, the viral dissemination of information and the stimulation of online behaviour”.

[88] A case of absolute harm of inclusion covered by the international press concerns the blackout challenge” on TikTok, which encourages users to film themselves as they choke themselves to the point of fainting and then regain consciousness on camera: various cases emerged in which minors died while trying the challenge, Kari Paul, ‘Families sue TikTok after girls died while trying ‘blackout challenge’ (*The Guardian*, 6 July 2022) <https://www.theguardian.com/technology/2022/jul/05/tiktok-girls-dead-blackout-challenge> accessed 14 November 2024.

[89] Article 27(2) DSA.

[90] ibid.

[91] Fabbri (n 83) 657.

platforms. The EU Commission addressed this issue with its recent proposal for a Directive on the improvement of working conditions in digital platform work (PWD), which was approved by the EU Parliament in April 2024.

Recital 4 of the PWD clarifies the link between the coming of “algorithm-based technologies, including automated monitoring or decision-making systems” and “the emergence and growth of digital labour platforms”. Recital 8, on the one hand, recognises the increasingly central role played by algorithm-based ADM and monitoring systems, which “increasingly replace functions that managers usually perform in businesses, such as allocating tasks, the pricing of individual assignments, determining working schedules, giving instructions, evaluating the work performed, providing incentives or imposing sanctions”, on the other hand, also recognises that persons performing platform work “often do not have access to information on how the algorithms work, which personal data are being used and how their behaviour affects decisions taken by automated systems [and] often do not know the reasons for decisions taken or supported by automated systems and lack the possibility to obtain an explanation for those decisions”.

The PWD devotes the entire Chapter III to the topic of algorithmic management and, in particular, deals with regulating the use of automated monitoring systems[92] and ADM systems by digital labour platforms.

The latter are defined in Article 2(9) as “systems which are used to take or support, through electronic means, decisions that significantly affect persons performing platform work including the working conditions of platform workers”. The PWD, like the AI Act, expressly includes in the definition of "ADM systems" also those cases where these systems are used simply as a support for the final decision. Moreover, as in previous legislation, there is the expression “significantly affect”, but the directive gives some examples of what this term means, namely those decisions “affecting their recruitment, access to and organisation of work assignments, their earnings including the pricing of individual assignments, their safety and health, their working time, their access to training, promotion or its equivalent, their contractual status, including the restriction, suspension or termination of their account”[93].

In the regulation of automated decisions, the PWD has three objectives: (i) to impose transparency on the use of automated monitoring systems and ADM systems (Article 9), (ii) to ensure the principle of human in the loop (Article 10), and (iii) to guarantee the data subject's right to an explanation of the automated decision (Article 11).

With a view to fostering transparency on the use of ADM, Article 9 requires digital labor platforms to inform platform workers and platform workers' representatives (and also, upon request, competent national authorities) about the use of automated monitoring

[92] Which are, according to Article 2(8), “systems which are used for, or support monitoring, supervising or evaluating the work performance of persons performing platform work or the activities carried out within the work environment, including by collecting personal data, through electronic means”.

[93] See Article 2(9) PWD.

systems or decision-making systems. In particular, regarding ADM systems, the directive requires digital labor platforms to inform workers about (i) “the categories of decisions that are taken or supported by such systems”, (ii) the “main parameters that such systems take into account” together with their “relative importance” and “the way in which the platform worker’s personal data or behaviour influence the decisions”, and (iii) the “grounds” for a subset of especially significant decisions including refusal of remuneration, termination of the worker’s account, or any decision of “equivalent or detrimental effect”[94]. Nevertheless, the information required by Article 9 is quite general and assumes only an explanation of the general operation of the system (“global” explanation), rather than the specific decision made (“local” explanation), especially because this information must be provided on the first day of work[95] and because it is generic to all workers.

Article 10 deals with ensuring human oversight of the operation of these systems. Firstly, paragraph 5 prohibits the use of ADM for making certain particularly significant decisions, such as any decision to “restrict, suspend or terminate the contractual relationship or the account of a person performing platform work or any other decision of equivalent detriment”. These decisions, under Article 10, can only be made by human beings. Furthermore, Article 10 requires platforms to staff themselves with the necessary competence, training and authority to, at least every two years, oversee and evaluate “the impact of individual decisions” taken or supported by automated monitoring and decision-making systems, used by the digital labour platform, on workers, including “their working conditions and equal treatment at work”[96]. The PWD is concerned with making sure that this control activity is effective and not merely formal: in fact, it stipulates that controllers must have the authority and expertise to be able also to override automated decisions and must be protected from disciplinary or other "adverse treatment" for exercising their functions[97].

Nevertheless, the most relevant provision with respect to automated decisions is Article 11, which states that platform workers “have the right to obtain an explanation from the digital labour platform for any decision taken or supported by an automated decision-making system without undue delay”. This explanation (in oral or written form) shall be presented in a “transparent and intelligible manner”. Article 11 regulates in detail the procedure by which the worker can obtain this explanation. Digital labour platforms have to provide platform workers with access to a contact person (who has to possess the necessary competence, training and authority to exercise that function) “to discuss and to clarify the facts, circumstances and reasons having led to the decision”. For particularly

[94] See Article 9(1)(c) PWD.

[95] See Article 9(3) PWD.

[96] See Article 10(1) PWD.

[97] See Article 10(2) PWD.

significant decisions (such as the decision to terminate the worker's account[98]), the worker must be provided also with, at the latest on the day which the decision takes effect, a "written statement of the reasons"[99]. If then, the worker is not satisfied with the reasons given to him by the contact person or the written statement, he shall have the right to request the digital labour platform to review that decision, to which the platform will have to respond with a "a sufficiently precise and adequately substantiated reply" within two weeks of receipt of the request. Finally, Article 11(3) states that if the decision "infringes the rights" of the worker, the platform shall rectify that decision within two weeks of the adoption of the decision and shall take the necessary steps, including, if appropriate, a modification of the ADM system or a discontinuance of its use, in order to avoid such decisions in the future.

In conclusion, the EU legislator is concerned about the impact that automated monitoring systems and ADM systems, used by digital labour platforms, may have on the world of work. Consequently, imposing greater transparency and explainability on the functioning of these systems is functional to the protection of workers' rights, which are also recognised by Articles 15 ("Freedom to choose an occupation and right to engage in work") and 31 ("Fair and just working conditions") CFREU.

5 The common thread in EU legislation regarding ADM regulation

Sections 2 and 3 analysed the evolution of EU ADM legislation from the DPD to the AI Act; then, Section 4 demonstrated the growing necessity of regulating ADM across three distinct fields. Two considerations can be drawn from this.

First, the heterogeneity of the areas in which the EU legislator has intervened to regulate the use of ADM (consumer protection law, online content moderation, the use of recommender systems and the regulation of work on digital platforms) demonstrates, in the writer's opinion, the increasing pervasiveness of ADM in today's society: wherever it is necessary, or convenient, to operate with large amounts of data, there is a need to use algorithmic decision-making processes and, as a result, the regulator intervenes to impose greater transparency and explainability on ADM.

Second, looking at the evolution of ADM regulation that took place between the DPD, GDPR and the AI Act, the chronological factor would seem to play a role in the standard of ADM transparency and explainability required by regulation: more recent regulatory acts have higher standards. In fact, the DPD only recognised the data subject's right to

[98] Any decision "to restrict, suspend or terminate the account of the person performing platform work, any decision to refuse the payment for work performed by the person performing platform work, any decision on the contractual status of the person performing platform work, any decision with similar effects or any other decision affecting the essential aspects of the employment or other contractual relationships", see Article 11(1) PWD.

[99] Even if this written statement of reasons "may give the sense of human involvement, it seems plausible that such a statement could be pro forma—or even created by a text generation system", Michael Veale and Michael Six Silberman, Reuben Binns, 'Fortifying the algorithmic management provisions in the proposed Platform Work Directive' (2023) 14(2) European Labour Law Journal 308, 316.

obtain "knowledge of the logic involved in any automatic processing of data" and the right to "put his point of view"; subsequently, the GDPR also recognised the data subject's right to obtain human intervention, the right to contest the decision and the right to obtain meaningful information and, finally, the AI Act expressly provided for the right to an explanation of the decision made through a high-risk AI system. This analysis suggests the existence of a common thread in EU regulation of ADM, namely the tendency to impose increasingly stringent rules on transparency and explainability.

6 The reasons behind this common thread

Sections 2 and 3 highlighted how the call for greater transparency and explainability of automated decisions has been translated by the EU legislator into EU law, starting from the very first regulation of automated decisions in 1995 with the Data Protection Directive, to the most recent EU legislation on the subject, the AI Act. It has also been shown (Section 4) how demands for transparency and explainability of ADM have made their way into the regulation of specific sectors at the EU level. Adopting this historical perspective, a common thread in EU legislation was highlighted, namely the tendency to impose increasingly stringent rules on the transparency and explainability of ADM. It now remains to understand the rationale behind this intervention, i.e., what reasons have prompted the EU regulator to address this issue with increasing frequency.

Here, three possible causes are suggested: (i) the increased pervasiveness of ADM in our society due to the technological progress occurred in the field of AI and ML, (ii) the EU regulator attempt to strengthen citizens' trust in AI technologies and (iii) the overconfidence in human decision-making process (HDM).

The first reason is that automated decisions, since 1995, have occupied an increasingly central place in our society. In turn, the increased pervasiveness of ADM is due to the technological progress in the field of AI and ML. In fact, the term "Artificial Intelligence" encompasses several techniques and approaches (symbolic AI, ML, neural networks, decision tree, deep learning, etc.), which differ not only in their accuracy and predictive ability, but also in their degree of explainability[100]. In recent years, the most widely used AI systems are those based on ML. What distinguishes these systems from the rest is their ability to learn automatically from the data provided to them: the software, in order to produce the output, does not follow pre-specified rules of behaviour in an operator-

[100] For example, the field of so-called Symbolic AI" requires software to provide pre-defined, step-by-step specifications of the rules, facts, and structures that define the characteristics of the evolving calculations of probabilities made by the computer programme. Symbolic AI is tied to representations provided by humans undertaking the programming, consequently, it allows, in principle, for explanations on the outcome of specific calculations as well as documenting programme specific requirements. See Herwig C H Hofmann, 'An Introduction to Automated Decision-Making (ADM) and Cyber-Delegation in the Scope of EU Public Law' [2021] University of Luxembourg Law Research Paper No. 2021-008 1.

defined way (i.e. it does not rely on explicit "if-then rules[101]), but “it autonomously and dynamically develops the decision rule by applying learning and adaptive algorithms[102]”. This feature, on the one hand, makes the system more efficient and capable of performing more complex tasks, but on the other hand, makes it less understandable[103]. The success of ML[104] is due both to the increase in the amount of data available and to the increase in computational capacity that has occurred in the last thirty years[105]. The combination of these two factors has made ML-based AI systems more accurate and efficient and, therefore, more popular. Nevertheless, because these systems are less explainable and act with a greater degree of autonomy than other AI technologies, this has increased the demand for transparency and explainability.

The second reason that may explain the choice of the EU regulator to devote more attention to the discipline of ADM could be civil society's distrust of AI and, consequently, of automated decisions made through it: this distrust may have increased the demand for transparency and explainability. In fact, whereas a decision made by a human, except in cases expressly provided for by law, does not give rise to a right to an explanation on the part of the person concerned, on the other hand, when the decision is made by an “artificial” decision-maker, a right to an explanation has been established. It could be said that this choice is legitimised by the fact that the outputs produced by AI systems are still imperfect[106] (as discussed in Section 1). Nevertheless, to this assertion, which is certainly true, it could be objected that also human decisions can be erroneous, can be affected by bias[107], can lead to episodes of discrimination and can be opaque[108]. Moreover,

[101] The system is given only rules about how to learn from data”, as pointed out by Yavar Bathaee, ‘The Artificial Intelligence Black Box And The Failure Of Intent And Causation’ (2018) 31(2) Harvard Journal of Law & Technology 890, 898.

[102] Troisi (n 4) 954.

[103] As early as the 1950s, Alan Turing noted that a machine capable of learning could operate in ways that were not anticipated by its creators and trainers, even without their understanding of the machine's internal workings. See Alan M Turing, ‘Computer Machinery and Intelligence’ (1950) LIX(236) Mind 433.

[104] In fact, these systems are now used in a very wide plurality of areas: to make loans, select candidates for a job, set the premium for an insurance policy, target advertising to individual consumer preferences, but also to detect tax evaders, drug trafficking, as well as to conduct the fight against terrorism, see Jenna Burrell (n 10). Furthermore, it is unthinkable for certain activities (such as content moderation) to be performed by human beings, both because the amount of data to be processed is incomputable for a human, and because technological progress (i.e., the increase in the computational capacity of computers and the consequent lowering of costs) has made it inefficient to entrust these processes to humans.

[105] See Tabarrini (n 1).

[106] One might wonder whether, if we were somehow certain that algorithmic decisions were always perfect”, we could dispense with an explanation of the decision. Carlo Casonato, ‘AI and Constitutionalism: The Challenges Ahead’ in Bertrand Braunschweig and Malik Ghallab (eds), *Reflections on Artificial Intelligence for Humanity* (Springer 2021) 138.

[107] It has been amply demonstrated in the psychological literature that human reasoning can be affected by bias and can be conditioned in many ways. See Cameron Buckner, ‘Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour’ (2023) 74(3) The British Journal for the Philosophy of Science 681.

[108] Buckner (n 108) points out how human decision-making can also be, like algorithmic decision-making, affected by bias and can take the form of a real “black-box”. See also Vincent Chiao, ‘Transparency at Sentencing: Are Human Judges More Transparent Than Algorithms?’ in Jesper Ryberg and Julian V Roberts (eds), *Sentencing and Artificial Intelligence* (Oxford Academic 2022) 34, who explores the topic the transparency and explicability of human decision-making in reference to judicial decisions and comes to the same conclusions.

algorithmic decisions are, to a certain extent, exposed to less risk than human decisions: when making decisions, AI systems are not influenced by feelings, they cannot lie, and, for example, they are not intrinsically racist; on the contrary, they make decisions based on statistical probability and the analysis of large amounts of data. Conversely, none of this can be said across the board with regard to human decisions. Therefore, one of the causes of this distrust, far from being based on rational grounds, could be precisely the artificial nature of the decision-maker.

Another part of civil society's distrust of AI could stem from the fact that humans often do not understand them, both because of the computational gap between humans and AI and because of the aforementioned black-box problem. Lastly, another source of this distrust may stem from our (perhaps excessive) fear of these technologies, probably, in part, because we are conditioned by the science fiction literature of the last century[109], which has often depicted robots (think Terminator) or “supercomputers” (such as the famous HAL 9000) as dangers to humans.

Whatever the cause of this distrust, the EU regulator has sought to fight it by making ADM more transparent and by explaining the reasons behind the outputs of AI systems, with the aim of strengthening EU citizens' trust in these technologies[110]. This approach recalls to that taken by the EU with the GDPR: the purpose of the latter, in fact, far from being to restrict the circulation of data, was precisely to strengthen the confidence of EU citizens in such a way as to increase the circulation of data within the EU market[111]. After all, the “trustworthiness” of AI, is one of the central themes of EU regulation: it is identified, on several occasions[112], as the key to spreading more trust in this technology and encouraging its use by citizens and businesses.

The third possible reason behind the increasing regulation on transparency and explainability of ADM could be the overconfidence placed in the transparency and explainability of HDM[113] that, perhaps, raises the standards required of ADM explainability. This overconfidence in HDM leads people to expect standards of transparency and explainability from ADM that, in reality, are not guaranteed even in human decision-making[114]. As a result, there is a “double standard” between the level of explainability demanded of ADM and that demanded of HDM[115]. This double standard, then, would result

[109] Giorgio Buttazzo, ‘Artificial Consciousness: Utopia or Real Possibility?’ (2002) 34(7) Computer 24.

[110] On the link between transparency and trust, see Heike Felzmann and others, ‘Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns’ (2019) 6(1) Big Data & Society 1.

[111] In fact, not surprisingly, the very name of the GDPR speaks of the ‘free movement’ of such data.

[112] See White Paper on Artificial Intelligence 3, which expresses the EU will to create an ‘ecosystem of trust’ because ‘Building an ecosystem of trust is a policy objective in itself, and should give citizens the confidence to take up AI applications and give companies and public organisations the legal certainty to innovate using AI’. Moreover, in the text of the AI Act, the word ‘trust’ (in the form of ‘trust’, ‘trustful’, ‘trustworthy’ and ‘trustworthiness’) appears 24 times.

[113] John Zerilli and others, ‘Transparency in Algorithmic and Human Decision- Making: Is There a Double Standard?’ (2019) 32 Philosophy & Technology 661.

[114] ibid.

[115] ibid.

not so much from an excessive distrust of machines but, rather, from an excessive trust placed in the transparency and explainability of decisions made by humans[116], which sets the bar for ADM explainability very high.

7 Conclusions

The objectives of this contribution were twofold. The first was to analyse EU regulation of automated decisions from a historical perspective to highlight a certain trend: the increasing focus on transparency and explainability of automated decisions. Moreover, it was also highlighted how the issue of transparency and explainability in ADM has also emerged in EU regulation of specific sectors (consumer protection law, online content moderation, recommender systems and the regulation of work on digital platforms).

The second objective of this contribution was to suggest some explanation for such a normative development. Three causes were proposed in Section 6: (i) the increased pervasiveness of ADM in our society due to the technological progress in the field of AI and ML, (ii) the EU regulator attempt to strengthen citizens' confidence in AI technologies and (iii) the overconfidence placed in human decision-making that, perhaps, raises the standards required of ADM explainability.

Although the intent of the EU legislator to enhance the transparency and explainability of ADM is reasonable, it is not without consequences. This topic cannot be discussed in depth here, nevertheless, the growing demand for explainability of AI systems poses a number of balancing problems with other interests. First, there is the issue of protecting trade secrets and intellectual property rights (IPRs) involved[117]. Second, keeping an AI system opaque can also be important for ensuring its effectiveness (for example to prevent spambots from using the disclosed algorithm to attack the system or prevent people from cheating the system by tilting the outputs of an AI system in a desired direction)[118]. Moreover, it has been pointed out in the literature that there is a certain trade-off between the degree of explainability of an AI system and its accuracy[119]: the more explicable one makes the AI system, the more one reduces its accuracy and vice versa. This raises a critical dilemma: either one pursues the goal of making AI systems (and their decisions) as transparent and explainable as possible (while reducing their accuracy) or

[116] 'While we do not deny that transparency and explainability are important desiderata in algorithmic governance, we worry that automated decision-making is being held to an unrealistically high standard here, possibly owing to an unrealistically high estimate of the degree of transparency attainable from human decision-makers', ibid 662. Nevertheless, other authors have pointed to different explanations for this 'double standard', see, for example, Mario Günther and Atoosa Kasirzadeh, 'Algorithmic and human decision making: for a double standard of transparency' (2022) 37(1) AI & SOCIETY 375.

[117] See Paul B de Laat, 'Algorithmic decision-making employing profiling: will trade secrecy protection render the right to explanation toothless?' (2022) 24(17) Ethics and Information Technology 16.

[118] Martin Ebers, 'Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s)' in Liane Colonna and Stanley Greenstein (eds) *Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence* 103.

[119] See Grochowski and others (n 41).

one pursues the goal of making them as accurate as possible, while giving up explaining their outputs.

Finally, another potentially negative effect of transparency concerns privacy and data protection: making available the training data of the ML algorithm (which is a way of making them more transparent) may violate privacy law and the GDPR, if the dataset enables identification of personal data[120].

In light of this tension between the explainability of ADM and other conflicting interests, this study could contribute to a dual purpose: (i) raising greater awareness about the direction taken by the EU legislator over the past thirty years regarding ADM regulation, and (ii) proposing some possible explanations behind this regulatory development.

[120] In 2020, a Swedish administrative court of second instance granted a journalist access to the source code of the algorithm (despite the fact that it was protected by trade secret) used by the town of Trelleborg to automate decisions in welfare services. A particularly interesting aspect of the case was that when the disclosure of the source code took place, at the same time the personal data of some 250 citizens (first name, last name, and social security code) who had had dealings with the municipality to access welfare services were made public, because these data were included in the source code. This highlights one of the possible problems with source code disclosure: in addition to the infringement of the economic freedom of companies, this solution may also infringe on the privacy rights of third parties. See Katarina Lind, 'Central authorities slow to react as Sweden s cities embrace automation of welfare management' (*Algorithm Watch*, 17 March 2020) <https://algorithmwatch.org/en/trelleborg-sweden-algorithm/> accessed 14 November 2024.