An Introduction to the Principle of Transparency in Automated Decision-Making Systems

L. Grisales Rendón
[*] Georg-August-Universität Göttingen/ PhD candidate, Faculty of Law, Göttingen, Germany
lauragrisalesrendon@gmail.com

Abstract - **Under current European data protection law and the approach of the European Commission regarding artificial intelligence (AI) development, the principle of transparency and explainability by design are essential to protect the data subject's rights and generate confidence in AI systems. However, a closer look reveals that the legal approach ignores limitations of the transparency principle in the practical application of automated-decision making systems. Current transparency rules also require a high standard of transparency in automated decisions due to its potential risks.**

This paper seeks to analyze the scope of the principle of transparency and its limitations in the practical field, suggesting a division of data subjects to provide automated decision-making explanations according to their level of expertise to reach the transparency principle's goal: user understanding. The paper will address the semantic discussion of whether this objective is achieved through interpretability, explainability, accountability or, transparency in the broad sense. It also maps out the analysis about guidance to address transparency through a certification mechanism, contestability by design, or supplementary post hoc explanations in AI systems.

Keywords - Transparency; Explicability; Automated decision-making; Opacity; GDPR; Artificial intelligence; Design.

I. Introduction

The study about AI has been significant since 1956 [1] [2] when John McCarthy coined the term at the Dartmouth Conferences [1]. Posterior development about AI at Dartmouth College represents the beginning of AI as a knowledge area. The contemporary AI systems work with decision mechanisms based on Big Data synthesized in complex models. Due to their level of complexity, some models are not accessible for humans, therefore considered black boxes [3].

Today, the public concern regarding AI is its opacity. The explanation of AI's behaviour through methods that experts can understand or in clean and plain language is known as Explainable Artificial Intelligence (XAI) [4]. It responds to the opacity concerns of algorithmic reasoning by seeking transparent or "explainable AI" [5]. Furthermore, the European Commission indicates trustworthiness as a prerequisite for AI uptake, highlighting transparency and accountability among the seven key requirements for building an ecosystem of trust [6]. The mentioned requirements have been developed by the High-Level Expert Group on Artificial Intelligence (AI HLEG) [7] in the Ethics Guidelines for Trustworthy AI and the Assessment list for Trustworthy Artificial Intelligence (ALTAI) [8].

The management of Big Data for governance and commercial use requires a clear regulatory framework that protects the right to privacy and a guideline that monitors how the algorithms work so that automated decisions are practical and can positively impact society.

Automated Decision-Making (ADM) algorithms are already broadly used in different societal contexts [9] and potentially impact individual's fundamental rights and freedoms. To protect privacy, the General Data Protection Regulation (GDPR) considers automated decision-making rules based on data provided directly by the individuals concerned, data observed about the individuals, and derived data such as an already created profile of the individual. It also obliges the controller to take suitable measures to guard data subject's rights, freedom, and legitimate interests, ensuring the right not to be subject to decisions based solely on automated processing, with no human involvement [10].

Although ADM is usually defined as an algorithm or as AI [11], AI is a form of algorithmic decision-making, but not all ADM systems are based on AI [12]. The present research refers to the scope of the provisions regulating ADM based on personal data that can affect the data subject and are not solely automated since contestability by design only apply to semi-automated decisions.

The EU's Declaration of cooperation on Artificial Intelligence (AI), the Commission's Communication Artificial Intelligence for Europe, its AI HLEG focused on AI, the European Parliament's resolution on robotics, and the European Economic and the Social Committee's opinion on AI, they all claim that the goal of AI is to “simultaneously maximize the benefit to society, help business, spur innovation and encourage competition.” [13] However, the concern over using ADM systems based on AI predictions persists.

In recent years there have been some alarming cases about the intrusion of ADM systems into the personal sphere:

1. the automated processing of traffic offenses in France [13];

2. the preselection of job applicants by ADM in the first stages of recruitment, to which 70% of applicants are subject to the United States and the United Kingdom, and the same policy beginning in Germany;

3. the automated process of allocating health treatment in the public system in Italy [13];

4. the automated identification of vulnerable children in Denmark [13];

5. the function of social media platforms [14] deciding what content is interesting for each user every day [15].

The field of law has also been a source of concern due to the possibility of hiring an AI attorney [16] and public policies regarding prisoner's freedom [17]. As automated decisions impact our lives, there is a need for a practical and global guideline for applying transparency in ADM systems, covering the legal, ethical, and technical/computational spectrum of algorithmic work. The binding nature of the guideline could only be achieved through its adoption in each jurisdiction and international cooperation, as has already been evidenced in the EU-U.S. Privacy Shield. Although several guidelines have been developed and different provisions indicate the obligation to comply with transparent processing from the AI design stage, the characteristics of information issuance differ depending on the type of data subject requiring the information.

Given that the concept of transparency adopted in this chapter is that of the reciprocal relationship between the sender and receiver of the information, the data subject should also be divided into two categories (expert and non-expert users), depending on their level of expertise on the ADM system-related topic of the specific program.

II. THE CONCEPT OF TRANSPARENCY

Taking as a basis the relational notion provided by Meijer [18], transparency demands a relation between an agent and a recipient. This notion applies to both the transparency and the accountability principle since the communication in the relationship agent- recipient involves, in addition to the issuance of information, its proper reception, and understanding [18]. As transparency and accountability reinforce each other, this paper discusses transparency in ADM systems, concluding that accountability strategies are essential to ensure transparency in those.

The concept of transparency cannot be reduced to the simple communication of information. It also encompasses accountability to help identify strategies to improve communication. That is the detailed explanation of decisions to satisfy the understanding of the data subjects or stakeholders, including data on the functioning of the systems [19].

To comply with the relational function of transparency, which requires data subjects truly understands the functioning of the process that goes from collecting data to making a decision that affects them, some guidelines on transparency have emerged. These guidelines also serve as a starting point for data processors or controllers to correctly apply personal data protection rules and comply with transparency principles.

Even though some authors suggest that algorithmic transparency is only achievable through open-access schemes [20], this possibility would bring disadvantages that are developed below, and others assert that no algorithmic model can be transparent [20]. Users are asking for the rendering of information regarding their data and the automated decision-making process. This demand is reflected in the various searches for guidelines on transparency for responsible data management.

A. Algorithm Transparency Guidelines

Some guidelines, such as the Santa Clara's principles on Transparency and Accountability in Content Moderation [21] and the FACT principles [20], seek to reinforce the confidence of the data subject. They are based on an understanding of the results of specific algorithmic decisions [19], and focus on providing indisputable answers.

Contrary to this can we find the EU-US Privacy Shield that, as explained in the EC report on Automated decision-making based on personal data, that has been transferred from the EU to certified companies [22]. It is a framework for transfers of personal data between the EU and the US that provides protection more in line with EU data protection legislation. The Privacy Shield is based on a voluntary self-certification system whereby US companies agree to comply with a set of privacy principles, which become enforceable under US law. The Privacy Shield does not contain any principles that offer protections similar to those in Article 22 of the GDPR. However, in most data transfer cases, the contract for which the data transfer occurs establishes that the controller is subject to the GDPR [22]. Therefore, European legislation would also apply.

Likewise, the Government of Canada provides a guideline, its Algorithmic Impact Assessment (AIA), which, according to the Directive Automated Decision-Making, is a measure to evaluate AI that helps designers mitigate the risks associated with AI by transparently designing algorithms [23]. A practical guide that integrates a Transparency by Design model to design transparent AI systems was published in the paper Towards Transparency by Design for Artificial Intelligence [18] by the Science and Engineering Ethics journal. Its great advantage is to explain, in an orderly and concise manner, by phases, the process of application of the principles of transparency, employing a set of nine principles to cover "contextual, technical, informational, and stakeholder-sensitive considerations." However, its problem is common to the different guidelines: there is no precise technical instruction for applying the principles of transparency.

On the other hand, the broader scope and generality of the Assessment List for Trustworthy AI [8] is the feature that makes it the most practical guide available. This is a guideline developed by the AI HLEG and established by the European Commission to promote trustworthiness on AI by addressing the concepts of Lawful AI, Ethical AI, and Robust AI. Its purpose is to ask for the accomplishing with seven key requirements to achieve trustworthy AI that can be implemented by technical or non-technical means. The guideline contains a non-exhaustive Trustworthy AI assessment list for operationalizing the key requirements. It addresses qualitative and quantitative processes and, allows its guidelines to have a broad scope of application, filling gaps in the interpretation of the transparency principles, as required.

B. Challenges of the Claim for Transparency concerning automated decision-making practices

The principal concerns regarding machine learning algorithms are unfairness, discrimination, and opacity [24]. Accordingly, the objectives to be achieved are fairness, accountability, and transparency.

The literature has pointed out problems associated with the ADM. The main concern, without which the others cannot be solved, is opacity, which can be countered with "transparency by design." Growing awareness of personal data's value and the adverse effects of automated decision's potential biases has led to a call to reduce or eliminate the effects of opacity.

Burrel has categorized three types in ADM regarding opacity: intentional opacity, illiterate opacity, and intrinsic opacity. Intentional opacity is mitigated by the GDPR's regulation of the right to explanation; illiterate opacity responds to the relational notion of transparency already described: the lack of understanding of the information by the data subject though it has been released. The intrinsic opacity refers to the nature of the methods with which the ADM system works [24]. ADMs are decisions based on the predictions resulting from BigData collection and analysis through ML models or hand-crafted rules [12]. ML's most common methods, deep learning and artificial neural networks are hard to interpret, thus being considered opaque.

In contrast, logical methods are more comfortable to interpret in natural language. Its disadvantage lies in its limited predictive performance. They are used mainly in high-risk data. One way to combat its opacity is to use machine learning models that are more transparent to humans [24].

Unfortunately, there is an inverse correlation between the AI capability working with ML and its transparency level. Methods that can achieve high predictive performance are usually difficult to interpret[1], and more transparent methods[2] have a lower predictive performance [25, 26]. As acknowledged by the AI HLEG, “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)." [7]

A widely mentioned solution to obtain algorithmic transparency is the openness of its source code. However, this solution has various disadvantages. Firstly, complete code openness may lead to an abuse of algorithms by malicious third parties. For this reason, transparent companies keep at least parts of their code closed to prevent abuse [20]. Secondly, it would affect trade secret protection and, in some cases, privacy law itself [24]. Finally, even with the disclosure of documentation, procedures, and code, if the relational concept of transparency is felt, thus the data subjects understanding is not achieved, this will not constitute transparency [20].

In order not to sacrifice the predictive performance of AI systems, ways to make ML methods more transparent have been investigated, especially for systems that process sensitive data. Still, these methods have not been fully developed, making them not very applicable and unpopular [26].

Another solution to increase transparency in ADM is for data controllers to consider group-related issues in applying the GDPR provisions on Data Protection Impact Assessments (DPIAs) to the ML algorithms explanations process [27, 28] due to the high risk of biases and discrimination posed by ADM. This can also be done by applying data protection by design principles at the design stage [28].

The nature of algorithmic autonomy makes it challenging to evaluate. Besides, algorithms and programming will advance at an ever-increasing rate. For this reason, algorithmic technical models are needed that allow a close relationship between the public and the algorithm. In this way, the transparency principle, the relevant guidelines, and data handling guidelines could be brought into line with practice.

C. The Nature of the Black Box

Machine learning is a subset of AI that can produce, from data sets, systems capable of building models and of improving themselves through experience, without the need for additional human programming [29] or monitoring [30, 31]. Predictive algorithms derive their rules from processed data. Some predictive algorithms models, such as neural networks, prevent a detailed examination of the data processing rules. Data processing systems working with machine learning can be observed by their input, output, and transfer characteristics [26], but they evolve in an opaque way so that programmers cannot inspect how the algorithm manages the data (the black box). There is, thus, a problem of transferring knowledge and opacity.

Since the emerging of the AI study, the concept of opacity has changed considerably. In 1956, W. Ross Ashby claimed that everything we observe is a black box and that human intelligence can monitor the black box's inputs and outputs to make them white [32]. Some authors are skeptical about ADM's opacity concerns since human decision-making also has intrinsic bias and risks. After all, the human mind is also a black box [33].

Neither HDM nor ADM can be inspected in-depth [5]. However, as stated in the White Paper on AI, the same AI bias could have a much more significant effect [6]. Thus, the imposition of a higher transparency standard to automated algorithmic decisions than HDM appears to be fair.

The search for solutions to guarantee user privacy protection is based on accepting the biases of the ADM algorithms. As affirmed by Bozdag, and Naik, and Bhide [34–36], the theory of the lack of bias in algorithms contradicts what was demonstrated in research before and after their investigations. The AI HLEG has confirmed the existence of algorithmic bias and its unintentional effects in

[1] Support vector machines or artificial neural networks.

[2] Rule systems, linear regression, logistic regression or decision trees.

its Ethics Guidelines for Trustworthy Artificial Intelligence [7].

D. The Idea of Total Transparency

Absolute transparency is often defined as the ideal. However, the optimal level of transparency must be related to the communication between the data controller and the data subject [18]. Consequently, algorithmic transparency is limited by the type of user seeking an explanation [27]. Data subject's ignorance on computational and data importance makes ADM transparency challenging to assess [24, 37].

The disclosure of more information than required by the data subject to have a higher level of transparency in the processing of personal data is not a problem, as long as the provision of the GDPR to communicate the data processing to the data subject in an exact way, to ensure their understanding, is complied with. It is considered essential to include educational measures as a tool for consumer advocacy and as a tool for citizen empowerment, in the case of ADM systems used in public administration, so that they can defend their rights.

III. Application of transparency provisions for an effective communication with the data subject

In addition to the constitutional protection of privacy in western societies [18], at least 130 jurisdictions have a comprehensive data privacy law. Surprisingly, despite the complexity of privacy regulations, 80% of organizations are inclined to comply with them [38], as compliance positively affects organizations [39]. Regulations are different from one country to another. The regulations diversity challenges organization to abide by the laws of each country where they operate [40].

A. A European Union Based Approach

As the European Union works on fostering fundamental freedoms and the rule of law at an international level [41], and as the GDPR is the pioneer legislation in comprehensive extra-territorial protection of data privacy, a study based on the EU provisions serves as a basis for the construction of international guidelines.

Although transparency regarding data processing was already regulated in the Treaty on the European Union [42] and the Charter of Fundamental Rights of the European Union [43], the GDPR included it as a principle, ensuring data subjects can control their data. Transparency is mentioned in the GDPR in the rights of the data subject of chapter III regarding the data subjects' rights. Transparency is also addressed in art. 5, regarding the principle of lawfulness, fairness, and transparency. It contains the controller or data processor obligation to protect access to transparent data processing information for the data subject.

Furthermore, recital 58, explains the transparency principle, and in recital 71, the importance of transparent processing of personal data in ADM algorithms. However, they are not legally binding [44]. Concluding, the controller or processor's legal obligation is to provide information in a "concise, transparent, intelligible, and easily accessible form, using clear and plain language." [10]

Data controllers must explain separately and using unambiguous language the most important consequences of the data treatment: providing an overview of the types of processing that could have the most significant impact on the fundamental rights and freedoms of data subjects concerning the protection of their data [45].

GDPR provisions relating to ADM systems, such as data protection by design, data protection impact assessments, corporate rules, and the appointment of a data protection officer, may provide the controller with guidance to identify risks directly and thus ensure minimum quality standards while safeguarding individual and, indirectly, group rights and freedoms [28]. Through their disclosure and access rights, data protection authorities can reinforce controller obligations by examining ADM processes and conducting data protection impact assessments during their audits [28], as seen in prominent cases such as Foodinho [46] and Deliveroo [47] in Italy, and Mercadona [48] in Spain.

The transparency principle seeks to impulse the willingness of data collectors or processors to communicate with the data subject assertively. Some companies have announced the forthcoming communication policies [19].

While the debate on GDPR has indicated challenges of achieving transparent ADM systems, different guidelines on transparency for responsible data management have been created in the industry, academics, and governments in the last years. In any case, the GDPR succeeds in reinforcing ADM system provider's awareness about data subject's rights and freedoms [28]. Consequently, transparency is a legal obligation that impacts AI algorithm's design to help individuals understand how the algorithm works to meet the GDPR's purpose.

B. The Meaningless Transparency Paradigm

The spectrum of data protection law in Europe is apprehensive. Transparency being considered a principle means that any event that is not expressly regulated must also comply with the transparency requirement. However, the application of the transparency principle is limited by detaching it from its practical meaning. Despite the best intents of the data controller, the complexity of the data processing carries a sort of inherent, practical obscurity from the perspective of the data subject if:

1. There is a lack of understanding by the data subject: for this reason, the GDPR requires a clear explanation of the data processing [10]. Furthermore, the controller should not use a mathematical and elaborate explanation regarding data processing through machine learning. Instead, he should simplify the way the algorithm or machine learning works for the data subject [49].

2. The public is not informed: the materialization of the principle of transparency requires the socio-technical component, in which transparency is valuable because it reaches a critical and informed public [20].

Although the commissioner for Justice Vera Jourová announced more awareness among European citizens about data protection, of the 60% of Europeans who read their privacy statements, only 13% read them thoroughly [50]. In the context of ADM, due to the way it works, we are faced with information that is more complicated to transmit since it must be translated from algorithmic language to natural language. As Kroll stated [51], transferring knowledge from the algorithm to the human would be done under algorithmic logic. Furthermore, the transfer of clear information about the algorithm's operation is especially problematic in domains where the tagged data is limited [52].

Is greater transparency or greater information disclosure better to improve transparency and AI trustworthiness? Treating information disclosure as the ultimate goal does not appear sufficient for user's understanding of the decision-making process, at least regarding automated decisions.

In addition to the obligation for data processing policies to be reported in an understandable way to the data subject, this information's importance and the data's value must be clear. It is, therefore, necessary to classify the types of users according to their level of expertise. If the user's particular needs are not addressed, transparency is not feasible, and we could be in a "meaningless transparency paradigm." [27]

Beyond the reflection on the role of the data subject's interest in applying the principle of transparency, a dvision of the data subject between average knowledge users or non-expert and field expert users would allow practical strategies application of transparent provisions in ADM systems.

IV. Strategies for practical application of transparency provisions in ADM systems for field expert users: accountable ADM

Three appealing strategies for ensuring transparency in ADM systems come through accountability: explainability or post hoc transparency, contestability by design, and certification mechanism. Although accountability is a different concept and principle, both - transparency and accountability - are complementary.

A. Transparency Post Hoc: Explainability

There are technical limits to AI transparency, which relate to its traceability. The traceability of the reasoning behind an automated decision requires the algorithm to be designed to make its decision-making process explainable without exposing the decision-making process [53]. Sometimes, it is impossible to decipher the black box of automated decision systems so that the algorithm's decision process is difficult to disclose from the design stage [18]. As described by Dr. David Leslie, a transparent AI implies the justifiability of both the processes involved in its design and application and its results. In a strict sense of the standard, ADM systems have justification if the design and implementation processes that went into the decision and the decision itself are ethically permissible, nondiscriminatory, and worthy of public trust [54].

An ex-ante explanation of an automated decision may describe only the system's functionality. In contrast, an ex-post explanation may address both the system's functionality and the reasoning for the specific decision [44].

Apart from this, one can think of relying on the powers that the GDPR grants to data protection authorities to fulfill their inspection duty. In case of a data protection audit, the authorities will have access to the DPIAs of any company or data processor, in cases where a DPIAs is mandatory. As there must always be a period for the data subject to access the data on how the algorithm processed his information, an audit will explain its operation to the user or data subject. Additionally, it could broaden the scope of their control function by executing awareness activities through individual audits. Apart from qualitative reasoning, black box or specification-based testing can be analyzed [55]. The data sets and processes that result in the AI system decision should be documented in the most accurate way. This documentation process makes it possible to identify why an ADM was wrong, anticipate unintended consequences, and prevent future errors [18].

B. Certification Mechanism

A Post hoc explanation is not the only scenario in which DPIAs can ensure transparency and reduce the risks of infringement of data subject's rights. Since DPIAs are mandatory when the processing could result in a high risk to the rights and freedoms of natural persons [10], any automated processing that works with ML is likely to require a DPIA, even if the decision is not fully automated [56]. In this sense, the impact of data processing for automated decision-making can be assessed from the GDPR, applying the principle of transparency through the DPIA.

However, DPIAs are mainly based on the security of personal data and less on the protection of human rights[57]. For this reason, Mantelero proposes The Human Rights, Ethical and Social Impact Assessment (HRESIA), a data security model based on the protection of human rights and the achievement of values, while proposing the formation of an ad hoc committee of experts to engage the data subjects or communities that would be possibly affected by automated decisions [57, 58].

A general quality insurance or a certification assessment mechanism, regardless of which, must, as mentioned in the Ethics Guidelines for Trustworthy AI, “be properly aligned with the industrial and social norms of the different contexts”, and, as recommended by Mantelero, be operated by professionals.

C. Contestability by Design

Despite the general rule that AI systems must meet the criteria of data protection by design, existing research has indicated a few ways for privacy and computational non-expert users to understand algorithmic system's complex and unpredictable nature. A strategy that has emerged to comply with the transparency obligations considering data protection goals is applying contestability by design [59].

Contestability is the possibility of acting on an automated decision by correcting or making suggestions to improve the ADM process. It represents one way to achieve the optimal level of transparency, in which the desired and achieved transparency levels coincide, at least regarding field expert users [18, 60].

By using predictive contestability, the user would be engaged with the operation of the algorithm, knowing the explanation of each prediction, being aware of adversarial machine learning and the risks to the security of his privacy, considering possible attacks, identifying potential violations of his rights or interests, and acting directly or indirectly on the predictive algorithm. An ADM model designed with a contestability system would benefit only field expert users since the contestable design will give them the possibility to influence the algorithm working with which he/she disagrees and the system capacity to respond to multiple user's influence since Data-driven systems and machine learning can have serious deficiencies that lead to unwanted and unfair outcomes [18].

Ensuring a contestability by design system from the early stages of software that uses machine learning to make automated decisions could be a way of allowing human intervention to co-create algorithm performance and enhance transparency. Algorithms would then be understandable for users, which allows them to provide corrections and refine them [60, 61].

When determining the means for processing and at the time of processing itself, the right to human intervention should be pursued through an appropriate organization [10]. Adequate provisions for human intervention could reduce user liability since personal data subjects can contain the harmful effects of automated decisions if detected on time. Using the predictive contestability, the user would have a greater engagement with the operation of the algorithm than in the current structure because he would know the explanation of each prediction and may be aware of adversarial machine learning and the risks to the security of his privacy, considering possible attacks, identifying potential violations of their rights or interests, and acting on the predictive algorithm.

V. CONCLUSION

The determination of an appropriate level of transparency depends on a set of legal, ethical, and technical characteristics of the ADM system and its users.

From a legal standpoint, each jurisdiction in which data privacy is regulated has its unique peculiarities. Through international cooperation and diplomacy, it is advisable to adopt global parameters to defend the right to privacy and drive AI development. The GDPR has been a pioneer in applying a principle of extraterritoriality to protect personal data. Thus, it is essential to study its regulations, which can serve as a basis for regulations worldwide. In addition, the Assessment List for Trustworthy AI appears to be the ideal guideline for the principle of transparency given its broad scope.

On the technical side, many AI systems are opaque, but some others have a high level of transparency. The problem comes from the duality between choosing to obtain a better output or better explainability of the algorithm performance. This also involves an ethical sphere. Therefore, ML (deep learning or artificial neural networks) is used when the algorithm is preferred. Still, in the case of high-risk data handling, the explainability provided by logical methods is preferred.

A division of data subjects into categories will facilitate communication and information about the algorithm's data processing and decision-making. Considering that there must be a relationship of effective communication between the data controller and the data subject for the practical application of transparency, it is paramount to consider how the data subject receives the information in the best possible way for its understanding. If the relational concept of transparency is felt, the data subject`s understanding is not achieved, which will not constitute transparency.

A solution for balancing the right to privacy, information, and transparency on the one hand, and technological development and resource efficiency through the use of metadata on the other, is to apply data protection by design on the algorithm, particularly contestability by design, at least to field expert ADM users. Their expertise will enable them to identify bias effectively and participate in joint work, which may occur directly or indirectly, improve the algorithmic work, and achieve future fairer decisions, thus achieving data protection goals. However, this option is limited because it can only be applied in its technical sense to expert users. Therefore, an in-depth and comparative analysis is needed to use contestable design and certification mechanisms, which would respond to the application of the current rules governing automated decision-making and data protection and apply to all data subjects without differentiation.

REFERENCES

[1] J. McCarthy, M. Marvin, R. Nathaniel, and S. Claude E, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *AI Magazine*, vol. 27, no. 4, 2006.

[2] J. Moor, "The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years," *AI Magazine*, vol. 27, no. 4, pp. 87–91, 2006.

[3] S. Rinzivillo, *A Simple Guide to Explainable AI.* [Online]. Available: https://www.ai4europe.eu/research/simple-guide-explainable-ai (accessed: Feb. 1 2022).

[4] R. Calegari, G. Ciatto, V. Mascardi, and A. Omicini, "Logic-based technologies for multi-agent systems: a systematic literature review literature review," *Auton Agent Multi-Agent Syst*, vol. 35, no. 1, 2021, doi: 10.1007/s10458-020-09478-3.

[5] J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan, "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?", *Philos. Technol.*, vol. 32, no. 4, pp. 661–683, 2019, doi: 10.1007/s13347-018-0330-6.

[6] European Commission, "White Paper on Artificial Intelligence A European approach to excellence and trust," Brussels, 2020.

[7] High Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI,", 2019.

[8] High Level Expert Group on Artificial Intelligence, "The Assessment List for Trustworthy Artificial Intelligence (ALTAI)," European Commission, Brussels, 2020.

[9] Information Commissioner's Office, *Automated decision-making and profiling.* [Online]. Available: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/automated-decision-making-and-profiling/ (accessed: Feb. 1 2022).

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

[11] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. de Vreese, "In AI we trust? Perceptions about automated decision-making by

artificial intelligence," *AI & Soc*, vol. 35, no. 3, pp. 611–623, 2020, doi: 10.1007/s00146-019-00931-w.

[12] B. E. Thapa, "Predictive Analytics and AI in Governance: Data-driven government in a free society: Artificial Intelligence, Big Data and Algorithmic Decision-Making in government from a liberal perspective," The European Liberal Forum, Brussels. Accessed: Feb. 1 2022. [Online]. Available: https://liberalforum.eu /wp-content/uploads/2021/07/PUBLICATION_AI-in-e-governance.pdf

[13] Bertelsmann Stiftung and AlgorithmWatch, *Automating Society Taking Stock of Automated Decision-Making in the EU.* [Online]. Available: https://algorithmwatch.org/de/wp-content/uploads/2019/02/Automating_Society_Report_2019.pdf (accessed: Feb. 1 2022).

[14] K. Murnane, "Which Social Media Platform Is The Most Popular In The US?," *Forbes*, 03 Mar., 2018. https://www.forbes.com/sites/kevinmurnane/2018/03/03/which-social-media-platform-is-the-most-popular-in-the-us/?sh=685458f71e4e (accessed: Feb. 1 2022).

[15] J. Orlowski, *The Social Dilemma*. The technology that connects us, 2020. Accessed: Feb. 1 2022. [Online]. Available: https://www.thesocialdilemma.com/

[16] K. Turner, "Meet 'Ross,' the newly hired legal robot," *The Washington Post*, 16 May., 2016. https://www.washingtonpost.com /news/innovations/wp/2016/05/16/meet-ross-the-newly-hired-legal-robot/?noredirect=on (accessed: Feb. 1 2022).

[17] L. Jaume-Palasí and M. Spielkamp, "Ethics and algorithmic processes for decision making and decision support," AlgorithmWatch Working Paper, vol. 2, pp. 1–19.

[18] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Towards Transparency by Design for Artificial Intelligence," *Science and engineering ethics*, vol. 26, no. 6, pp. 3333–3361, 2020, doi: 10.1007/s11948-020-00276-4.

[19] N. P. Suzor, S. Myers West, A. Quodling, and J. York, "What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation," *International Journal of Communication*, vol. 13, no. 0, pp. 1526–1543, 2019. [Online]. Available: https://ijoc.org/index.php/ijoc/article/view/9736/2610

[20] J. Kemper and D. Kolkman, "Transparent to whom? No algorithmic accountability without a critical audience," *Information, Communication & Society*, vol. 22, no. 14, pp. 2081–2096, 2019, doi: 10.1080/1369118X.2018.1477967.

[21] "The Santa Clara Principles on Transparency and Accountability in Content Moderation," Santa Clara, 2020. [Online]. Available: https://santaclaraprinciples.org/

[22] G. Bodea, K. Karanikolova, D. K. Mulligan, and J. Makagon, "Automated decision-making on the basis of personal data that has been transferred from the EU to companies certified under the EU-U.S. Privacy Shield: Fact-finding and assessment of safeguards provided by U.S. law," European Commission, Brusselss, 2018. Accessed: Feb. 1 2022. [Online]. Available: https://ec.europa.eu/info/sites/default/files/independent_study_on_automated_decision-making.pdf

[23] Treasury Board of Canada, *Algorithmic Impact Assessment Tool.* [Online]. Available: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html#shr-pg0 (accessed: Feb. 1 2022).

[24] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, Transparent, and Accountable Algorithmic Decision-making Processes," *Philos. Technol.*, vol. 31, no. 4, pp. 611–627, 2018, doi: 10.1007/s13347-017-0279-x.

[25] Sebastian Nusser, "Robust Learning in Safety-Related Domains : machine learning methods for solving safety-related application problems," PhD Dissertation, Otto-von-Guericke-Universität Magdeburg, Magdeburg, 2009. [Online]. Available: https://www.researchgate.net/publication/40220479_Robust_Learning_in_Safety-Related_Domains_machine_learning_methods_for_solving_safety-related_application_problems

[26] L. Muehlhauser, *Transparency in Safety-Critical Systems.* [Online]. Available: https://intelligence.org/2013/08/25/transparency-in-safety-critical-systems/ (accessed: Feb. 1 2022).

[27] L. Edwards and M. Veale, "Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for," *Duke Law & Technology Review*, vol. 16, no. 1, 2017, doi: 10.2139/ssrn.2972855.

[28] S. Dreyer, W. Schulz, and Bertelsmann Stiftung, *The General Data Protection Regulation and Automated Decision-making: Will it deliver?*

[29] A. Guadamuz, "Artificial intelligence and copyright," *Wipo Magazine*, vol. 5, 2017. [Online]. Available: https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html

[30] J. Thulin, "Machine Learning-based Classifiers for the Direkt Profil Grammatical Profiling System," Master´s Thesis, Lund University, Lund, 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.142.4723

[31] L. Grisales Rendón, "Attribution of Copyright to Artificial Intelligence Generated Works," Master´s Thesis, Georg-August-Universität Göttingen, 2019. [Online]. Available: https://ediss.uni-goettingen.de/handle/21.11130/00-1735-0000-0003-C1A6-7?locale-attribute=en

[32] R. Glanville, "Inside every white box there are two black boxes trying to get out," *Syst. Res.*, vol. 27, no. 1, pp. 1–11, 1982, doi: 10.1002/bs.3830270102.

[33] L. Nizami, "Glanville's 'Black Box': what can an observer know?," *Rivista Italiana di Filosofia del Linguaggio*, vol. 14, no. 2, 2020, doi: 10.4396/AISB201905.

[34] G. Naik and S. S. Bhide, "Will the future of knowledge work automation transform personalized medicine?," *Applied & translational genomics*, vol. 3, no. 3, pp. 50–53, 2014, doi: 10.1016/j.atg.2014.05.003.

[35] E. Bozdag, "Bias in algorithmic filtering and personalization," (in En;en), *Ethics Inf Technol*, vol. 15, no. 3, pp. 209–227, 2013, doi: 10.1007/s10676-013-9321-6.

[36] Brent Mittelstadt, "Auditing for Transparency in Content Personalization Systems," *International Journal of Communication*, vol. 10, p. 12, 2016. [Online]. Available: https://www.researchgate.net/publication/309136069_Auditing_for_Transparency_in_Content_Personalization_Systems

[37] F. Pasquale, *The black box society: The secret algorithms that control money and information*. Cambridge, Massachusetts, London, England: Harvard University Press, 2016.

[38] R. Waitman, "Privacy Comes of Age During the Pandemic," blog entry, Jan. 2021. Accessed: Feb. 1 2022. [Online]. Available: https://blogs.cisco.com/security/privacy-comes-of-age-during-the-pandemic

[39] Cisco, "Forged by the Pandemic: The Age of Privacy," 2021. Accessed: Feb. 1 2022. [Online]. Available: https://www.cisco.com /c/dam/en_us/about/doing_business/trust-center/docs/cisco-privacy-benchmark-study-2021.pdf?CCID=cc000160&DTID=esootr000515&OID=rptsc024690

[40] M. Peitz and J. Waldfogel, Eds., *The Oxford Handbook of the Digital Economy*. New York: Oxford University Press, 2012.

[41] European Union, *Aims and values of the EU.* [Online]. Available: https://european-union.europa.eu/principles-countries-history/principles-and-values/aims-and-values_en (accessed: Feb. 1 2022).

[42] P. Office, "Consolidated Version Of The Treaty On European Union: OJ C 202/1," in *Official Journal of the European Union*, 2016. Accessed: Feb. 1 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C:2016:202:FULL

[43] *Charter of Fundamental Rights of the European Union*: OPOCE, 2012.

[44] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017, doi: 10.1093/idpl/ipx005.

[45] Article 29 Working Party, "Guidelines on transparency under Regulation 2016/679" Apr. 2018.

[46] Garante per la Protezione dei Dati Personali, *Abstract of Italian SA's order as issued against Foodinho S.r.l*: Garante per la Protezione dei Dati Personali, 2021. Accessed: Feb. 8 2022. [Online]. Available: https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/9677611

[47] DataGuidance, *Italy: Garante fines Deliveroo €2.5M for unlawful processing of riders' personal data.* [Online]. Available: https://www.dataguidance.com/news/italy-garante-fines-deliveroo-%E2%82%AC25m-unlawful-processing (accessed: Feb. 8 2022).

[48] Agencia Española de Protección de Datos, *Resolución de procedimiento # PS/00120/2021 contra Mercadona S.A.* Accessed: Feb. 8 2022. [Online]. Available: https://www.aepd.es/es/documento/ps-00120-2021.pdf

[49] Article 29 Data Protection Working Party, "Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679," Article 29 Data Protection Working Party, Feb. 2018.

[50] Council of Europe, "Data Protection Day: 28 January 2021: 40th anniversary of Convention 108 &15th Data Protection Day," Council of Europe, Strasbourg, 2021. Accessed: Feb. 17 2021. [Online]. Available: https://www.coe.int/en/web/data-protection/data-protection-day

[51] J. A. Kroll *et al.*, "Accountable Algorithms," *University of Pennsylvania Law Review*, vol. 165, no. 3, 2016. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2765268

[52] G. Bansal, B. Nushi, E. Kamar, D. S. Weld, W. S. Lasecki, and E. Horvitz, "Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff," *AAAI*, vol. 33, no. 01, pp. 2429–2437, 2019, doi: 10.1609/aaai.v33i01.33012429.

[53] F. Doshi-Velez *et al.*, "Accountability of AI Under the Law: The Role of Explanation," Ethics and Governance of Artificial Intelligence Iniciative, Harvard University, Nov. 2017. [Online]. Available: https://dash.harvard.edu/handle/1/34372584

[54] D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector," The Alan Turing Institute, London, 2019. Accessed: Feb. 1 2022. [Online]. Available: https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf

[55] J. Gao, H. Tsao, and Y. Wu, *Testing and Quality Assurance for Component-based Software*: Artech House, 2003. [Online]. Available: https://books.google.de/books?id=oUEwDwAAQBAJ

[56] Information Commissioner's Office, Data protection impact assessments, https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/ (accessed: Feb. 1 2022).

[57] A. Mantelero, "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment," *Computer Law & Security Review*, vol. 34, no. 4, pp. 754–772, 2018, doi: 10.1016/j.clsr.2018.05.017.

[58] A. Siapka, "The Ethical and Legal Challenges of Artificial Intelligence: The EU response to biased and discriminatory AI," Master´s Thesis, The Panteion University of Social and Political Sciences, Athens, 2019.

[59] Kars Alfrink, T. Turel, A.I. Keller, N. Doorn, and G.W. Kortuem, "Contestable City Algorithms," pp. 1–5, 2020. [Online]. Available: https://research.tudelft.nl/en/publications/contestable-city-algorithms

[60] T. Hirsch, K. Merced, S. Narayanan, Z. E. Imel, and D. C. Atkins, "Designing Contestability: Interaction Design, Machine Learning, and Mental Health," *DIS. Designing Interactive Systems (Conference)*, vol. 2017, pp. 95–99, 2017, doi: 10.1145/3064663.3064703.

[61] J. Krause, A. Perer, and K. Ng, "Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose California USA, 05072016, pp. 5686–5697.