# Towards transparent AI: how AI Act shape the future

Nídia Andrade Moreira[1]

s-njamoreira@ucp.pt

Pedro Miguel Freitas[1]

pfreitas@ucp.pt

Paulo Novais[2]

pjon@di.uminho.pt

[1] Universidade Católica Portuguesa, Faculty of Law, Católica Research Centre for the Future of Law, Porto, Portugal
[2]ALGORITMI Research Centre, LASI, University of Minho, Braga, Portugal

**Abstract.** The European Union's Artificial Intelligence Act (AIA) aims to establish legal standards for AI systems, emphasizing transparency and explainability, especially in high-risk systems. Our research it is divided into sections focusing on the legal framework established by the AIA, advancements in AI and the intersection between legal obligations and technological developments. We explore how the AIA addresses these issues and intersects with emerging research in XAI.

**Keywords:** Artificial Intelligence, AI Act, Explainable Artificial Intelligence.

## 1 Transparency in the Artificial Intelligence Act

Transparency stands as a fundamental requirement for AI systems and a core ethical principle in fostering an ecosystem of trust [1,2]. There is a need for transparency tailored to the level of expertise of stakeholders, indicating the degree of influence of an AI system in the decision-making process. This fundamental need, identified by the European Union, has been addressed by the legislator through the Artificial Intelligence Act (AI Act - AIA), which aims to establish a uniform legal framework for AI systems.

The AIA does not provide a definition of transparency[1] but the first article establishes that the regulation lays down 'harmonized transparency rules for certain AI systems" - article 1(2d), recital 26. Therefore, we conducted a qualitative analysis of the document to identify content related to the concepts of transparency and explainability. We searched for words such as 'transparent*', 'explain*', 'interpret*', and 'inform*', and

[1] The Legal Affairs Committee of the European Parliament proposed the inclusion of a definition of transparency – see [3], Article 4a.

then selected the norms that appeared to be most relevant.[2] We highlight specific transparency obligations:

i) **Obligation to inform users that they are interacting with an AI system or that the content was artificially generated**. Chapter IV outlines transparency obligations for providers and deployers of *certain* AI systems. Article 50 focuses on transparency towards individuals who interact directly with AI, stating that they must be *informed* when they are interacting with an AI system, with particular attention to systems that synthetically generate or manipulate images, audio, video, or text, or recognize emotions or biomedical categorizations. Specifically, in addition to an information duty, there is a technical obligation to ensure that synthetic content contains a mark in a machine-readable format that is detectable.

**ii) Obligations to inform deployers of high-risk AI systems *ex-ante* (article 13)**. These requirements aim to ensure that deployers (Article 3(4))[3] comprehend the functioning of the system before it enters the market. This includes ensuring that the systems comply with the obligations under this Regulation, identifying their limitations and taking necessary measures to address them, as well as understanding the output generated. These steps are crucial to guarantee that deployers use AI appropriately and under human supervision (Article 14).[4]

To ensure compliance with the Regulation, there is an obligation to provide extensive documentation before placing the system on the market or put into service (Article 11).[5] This includes a detailed description of the design of the AI system, such how it was trained and modified, as well as the 'general logic' of the system, algorithms, and data used (Annex IV).

Additionally, there is an obligation to establish 'instructions for use' (Article 3(15), Article 13(2)) to ensure that deployers understand the intended purpose and can use AI correctly. This is an *ex-ante* obligation, indicating that the 'user manual' does not necessarily require a local explanation of the parameters used and their weight in a specific case. Instead, it may be discussed whether it will be necessary an overall explainability of the model to assess which parameters are used and which weights are assigned.

Article 13(3)(b)(iv) states that 'where applicable, the technical capabilities and characteristics of the high- risk AI system to provide information that is relevant to explain its output.' The legislator should clarify the meaning of 'where applicable,' whether it means when technically possible or when required by the specific field of application or specific legislation. The same questions apply to (vii), which states 'where applicable, information to enable deployers to interpret the output of the high-risk AI system and use it appropriately.'

**iii) Obligations to explain (the output of) high-risk systems *ex post***.

---

[2] We analyse the Regulation (EU) 2024/1689 [4]. For analyses of previous versions see [5-7,20] which one with a specific categorization for transparency regulation.

[3] The initial draft mentioned “users”. The term user has been replaced by the term deployers. Transparency in this area is therefore aimed at deployers, to the exclusion of consumers (non-professions). A deployer is a professional that uses an AI system under their authority.

[4] See Article 13(1) and Recital 72.

[5] By setting out how the system works, it is possible to establish risk and quality management systems (Articles 9 and 17).

Article 14(4) outlines the information that must be provided to the deployer. The legislator emphasizes the need to explain the output, especially when it generates information or recommendations for human decision-making. Considering the state of the art, understanding the output and its achievement enables the human decision-maker to: (a) be aware of potential system errors, (b) recognize automation bias, (c) make different decisions from the system, (d) choose not to use it, or (e) interrupt its operation. Therefore, this explanation is vital for evaluating decisions made by AI users, ensuring good decisions, and supporting the right to effective remedy in cases that threaten human rights.[6]

Moreover, for the first time, AIA contemplates remedies for the use of AI in Chapter IX, Section 4. The AIA allows a right to lodge a complaint with a market surveillance authority (Article 85). Therefore, if an affected person considers that the AI system does not comply with the requirements of the AIA, they could present a complaint.

Unlike the initial draft[7], the adopted text gives those subjects to an algorithmic decision the right to request an explanation - Article 86. This is particularly relevant because the regulation of the GDPR is insufficient in this domain: algorithmic decisions that affect persons may not involve personal data – for example, in the case of self-driving cars. There is a right to explain individual decision-making to any affected person when decisions are based on the output from a high-risk AI system – (recital 171). It is a right to "clear and meaningful explanations," which must be appropriate for the target audience and not overly detailed to remain understandable by a layperson and retain their utility. These explanations should address (i) the role of AI in the decision - whether auxiliary or substitutive, and (ii) the main elements of the decision. Once again, the level of detail of the explanation is not explicitly stated, but the norm stipulates that it should be sufficient to identify the "elements of the decision" without specifying their specific values.

**iv) Data transparency of high-risk models**. To ensure the quality of data used in high-risk AI (Article 10, Article 13(3)(b)(vi), Annex IV, 2(d)), and GPAI (Article 53, Annex XI, Section 1, (2 c), Annex XII, (2c) some degree of data transparency is required.

**v) Obligations for record-keeping in high-risk AI systems**. Once the system is operational, it is essential to ensure automatic recording of events to monitor its operation throughout its lifetime, in accordance with the instructions for use and the requirements of the Regulation (Articles 12). The list provided in Article 12(3) suggests that such information relates to who uses these systems and what input data they utilize, thus ensuring traceability of users.

---

[6] Moreover, this obligation is crucial for deployers to comply with the obligation established in Article 26, namely the obligation to monitor the operation of the AI system and control risks.

[7] Previously, the AIA draft did not allow a person affected by an AI system to exercise rights against the supplier or by demanding explanations. However, this has been proposed by [3]. In Article 69(c), the Committee proposed that any affected persons subject to a decision based on an output from an AI system which produces legal effect that impact health, safety, fundamental rights, socio-economic well-being should receive an explanation at the time when the decision is communicated.

**vi) Obligation for providers of general-purpose AI models to provide information to providers intending to integrate the model into their AI systems (Article 53(1b)),** including details on the data used and the technical means required for integration (Annex XII). As highlighted in [21], this obligation of information facilitates cooperation between upstream and downstream providers, which is essential in a relationship characterized by interdependence. However, this obligation should persist throughout the monitoring of systemic risk, which appears to be outlined as ""keep up-to-date" (Article 53(1b)).

## 2 Transparency in computer science

The necessity of transparency in AI systems is being addressed by AI experts. We examine the motivations behind the push for transparency, the challenges involved in achieving it, and the emerging approaches within the field of XAI that hold promise for enhancing transparency and trust in AI systems.

### 2.1 Symbolic, sub-symbolic

AI experts tend to distinguish between transparency and explainability. Transparency is seen as an intrinsic quality of an AI model, while explainability is seen as a quality of an explanation in terms of being proxy of the decision maker and comprehensible to humans [13].

So, attending to this distinction AI models can be divided into two categories: (i) Inherently transparent models, whose inner workings can be understood by humans. Although they can be easily understood by experts, it may be necessary to use explainability techniques [5] to allow laypeople to understand them [13]. In these systems, we can use *ante-hoc* techniques that generate the explanation from the internal process of the model, being faithful to the decision process. (ii) Opaque models (black box), which, despite their opacity, are models that allow greater precision in the results. In these cases, it is necessary to use a *post-hoc* explicability technique that develops a method to explain the model's decisions. In these systems, the explanation is pursued attending to the output [5].

This classification attends to an opacity as a consequence of the high-complex or sub-symbolic design.[8] Considering symbolic AI which aims to represent human knowledge declaratively, i.e., in the form of rules [11, p. 13], although it might be difficult to turn the logic programming into a representation that is (easily) understandable, the model can be explained reconstructing all the steps that lead to the outcome [29].[9]

[8] This is the most relevant category of opacity, but there are also other types of opacity, namely because of technical illiteracy and as deliberate corporate or state secrecy [30]. Thus, the source of opacity can be human cognition, technical or legal, depending on the target.

[9] This approach uses comprehensible (logical) languages that allow us to check how the machine arrived at a particular result. However, this logic may not be easily understood by non-specialists [12], which requires the use of techniques that allow different stakeholders to better understand it.

On the other hand, sub-symbolic does not assume an explicit representation of knowledge [11, p. 13] and its behaviour might be impossible to represent.

### 2.2. The role of explainability

Explainable AI (XAI) appears as a solution to help find explanations. There are different types of methods that can provide different types of explanations – e.g., global explanation (overall system) or local explanations (specific output); textual or visual.

The choice of a model depends mainly on precision. Thus, there is a dichotomy between precision and transparency, which is overcome by XAI [13,17]. Despite the developments in the field, some authors prefer a *transparent-by-design approach*.

Attending to limitations of XAI, namely the problem of reliability and robustness of explanations, some authors mentions that explanation methods are inaccurate and do not fully correspond to the original model. For example, [14] points out that when there is a correlation between data, the model's explanation may indicate a determining factor in the decision that was not considered by the model (confounding variables). Additionally, if the explanation identifies the factors that have been considered, it may not explain how they were determinant. So, focus on analysing the factors influencing the prediction and neglect to explain the decision-making process.

Thus, the core of the argument is that explanatory power is mostly based on correlations and not on causality. On this point, [13, p. 83] notes that interpretability can guarantee the underlying causality in the model's reasoning by identifying which relevant variables affect the outcome. It seems to us that the same argument could be made for explainability. Indeed, one of the aims of XAI is to find causality between the variables in a model. On its own, it doesn't guarantee causality, but it can validate results obtained by causal inference techniques or help to understand possible causal effects in the data [13, p. 86]. In addition, the analysis of correlations could be relevant to detect hidden correlations that produce discrimination. Attending to this, [14] suggests that opaque models should be preferred whenever there is an interpretable model with the same level of performance.[10].

Although this dichotomy, it seems to us that the relevance of interpretability/explainability in high-risk models opens the door to a new type of AI: precise and interpretable/explainable [15].

## 3 Intersection

The concept and dimension of transparency and explainability differ between legal experts and AI experts [5]. This can lead to a mismatch, with legal experts having unfounded expectations that ignore technical aspects, and XAI experts failing to provide solutions that address legal and ethical concerns.

In discussions about AI regulation, the term transparency is used as an *umbrella* term that encompasses various notions beyond the scope of AI literature. It not only relates

[10] According to the author, there are usually no differences in performance.

to the transparency or opacity of the model or with explaining black-box models as discussed in XAI literature [13] or constraining a model to be transparent by design as Interpretable AI does [10]. The obligations of transparency are also extended to the data used, how the system is designed and implemented, the awareness of stakeholders interacting with AI, cybersecurity and quality assessments. Therefore, transparency includes other requirements, such as traceability, explainability, and communication [25].

### 3.1. Degrees of explainability

Article 50 only requires that anyone who interacts with certain AI systems or content generated by them must be informed. This is an *informative* statement, so it is simplified. However, Article 13(1) talks about an "appropriate type and level of transparency," and Article 13(2) mentions the need for "concise, complete, correct, and clear information that is relevant, accessible and comprehensible to deployers" in relation to the instructions for use identified in Article 13(3). Finally, Article 86 mentions "clear and meaningful explanations."

It seems that the European legislator was also concerned, as we have learned from studies in the field of XAI [13], with the need for the explanation to be valuable and comprehensive for the target audience of a particular domain. Thus, the necessary information will depend on (i) the foreseeable purpose and (ii) the stakeholders.

Simple and meaningful explanations should be prioritized. As suggested by [6, p. 363] "simple and concise explanations could be given first, with more detailed, expert-oriented explanations provided on a secondary level upon demand." This may be an appropriate way to provide step-by-step explanations so that (i) there is no initial information overload for the user/deployer and (ii) there is no information gap if the simplified information is deemed insufficient, especially in the case of AI experts.[11]

### 3.2. Metrics for explainability

Explainability cannot be defined in the abstract. It depends on the complexity of the model, the context in which it is used, the data, the purpose and the intended audience.

System documentation is important for *static explainability* of the model [23]. It should detail data flows and processing steps of AI, addressing seven main aspects: (i) purpose, (ii) data sources, (iii) algorithms used, (iv) reasoning processes, (v) main risks, (vi) stakeholders, (vii) performance. The European Union should consider adopting "model cards" as suggested by [24] for model reporting in each area, clarifying to developers what essential information is required. Once in the market, *runtime explainability* is important to "identify the role of active processes and data in achieving the system's purpose" [23].

In terms of metrics to understand the output of the model, in the case of a black box system, explaining how the output was generated may be difficult or impossible. But understanding the output is only one measure of explainability, so other measures of explainability could be applied, such as transparency of the data used. Furthermore,

[11] In addition, particular attention should be paid to the way information is communicated.

while some solutions can be achieved by XAI, other measures could be adopted as an alternative or in addition to XAI, such as a requirement to explain AI-based decision-making in high-risk categories. Transparency of the model is different from transparency of the decision [12, p.5].

Moreover, explanations can have different *objectives* [13, p. 86], including identifying causality, informativeness, trust, fairness, privacy, etc., all of which require different explanations. An explainability metric does not necessarily have to be a causality metric.

In cases where the human decision-maker and the person affected by the decision *need to be convinced by the justification* of the machine decision or recommendation, transparent models should be preferred. Even so, opaque models can be used if XAI tools are employed, in which case the system should be considered as a whole and not as separate systems [5].

The level of transparency will not depend on a general regulation, as the AI Act claims, but on the area in which the system operates, its principles, *legis artis*, and, above all, the level of trust in such a system. Explainability is useful to ensure human oversight and guarantee a good decision. Moreover, metrics should be defined not only for AI systems but also for explanations.[12] As suggested by [5], we could consider integrating XAI methods into the AI system to extend their obligations. Thus, regulatory requirements imply verification of the XAI methods, guaranteeing their quality. Otherwise, ignoring the quality of explainability is equivalent to ignoring this transparency requirement and may even be harmful, as a false explanation may cause more damage than no explanation at all.

### 3.3. Types of transparency obligations

Analysing the AIA there are different types of transparency obligations for AI systems. In fact, the regulation lists some information that must be provided, which should be proportional to the risk category.

**(1) User-transparency:** This type of transparency allows users of AI to use it correctly or be aware of the content generated by AI. It is crucial to address the gaps in technical literacy among users. In some cases, it enables users to decide whether to engage with models or services that use this technology. The idea is to empower users.

**(2) Compliance-transparency (or deployer-explanation)[13]:** These explanations make it possible to verify that the system complies with the requirements laid down in the AI Act, such as non-discrimination. Information is essential for providers, organizations - such as supervisory authorities - or individuals affected by a decision to con-

[12] Therefore, explainability metrics must be established, which should not be confused with the quality of the system's results [19, p.2715]. A model can be highly explainable but perform inadequately. However, the quality of the explanation can help understand the quality of the model.

[13] [7] divided explainability in the AIA in *user-empowering* and *compliance oriented*. However, with recent changes now referring "deployers", we should consider deployer-empowering to allow them to comply with AIA; so, in fact, is a compliance-oriented measure.

duct legal reviews and inspect whether the AI system is in line with specific requirements for high-risk systems. Along with that, transparency could help deployers of GPAI systems fine-tune the performance of the model. These explanations are targeted at an audience that is technically more capable, although they can also be used by individuals affected by a decision to contest the system itself.

**(3) Decision-explanation**[14]: In this case, we are referring to decision support systems that recommend a decision that has an impact on an individual. Here, explanations need to justify the decision itself, not just the design of the system, so that (i) people charged with reviewing them, and (ii) those affected by the decision can understand the output, ensuring informed decision-making and guaranteeing the right to contest AI recommendations or decisions. These explanations ensure a good decision, determining the level of interaction of humans with AI systems. They empower decision-makers and users[15], i.e., people affected by the decision.

### 3.3.1. Technical solutions?

As we have seen, there is a more robust obligation placed on high-risk AI systems. However, no distinction is made between the *specific functions* of such systems. Looking at these obligations, we can question whether such requirements dictate the use of specific technical solutions or if they are even technically feasible.

It may be difficult or even impossible to impose a complete technical obligation for black box systems. If such an obligation is imposed, we are faced with two scenarios: either we are dealing with an obligation that is destined to fail, or only symbolic or inherently transparent (albeit less complex) models are allowed. There is also a third hypothesis: transparent complex models and opaque models can be used if they can be explained using XAI techniques. To understand if there is a technical obligation, we should consider the objective of the regulation.

According to [10, p.1147], the regulation intends to "ensure that users of high-risk AI systems understand *how to use* the outputs of the AI system appropriately and can effectively oversee the system." For this reason, [10] considers that AIA does not dictate a specific technical solution and XAI techniques or transparent-by-design approaches may not be needed.

In fact, transparency measures such as instructions of use and training of the human overseeing the system are presented to address these obligations. We tend to partially agree. If a transparency obligation intended to guarantee information for users or deployers – type 1 and 2 – is sufficient, then the existence of documentations and instructions of use might be adequate. We consider *transparency as information*. Transpar-

[14] The second and the third are important to address two types of opacity: (i) intentional opacity, where companies deliberately hide information from public scrutiny; and (ii) opacity due to complex models that are difficult tto understand. Our concern is with how the AIA addresses the the second type. As noted in [16], addressing this type of opacity requires careful selection and design of the algorithms.

[15] Not only to give them knowledge, but also to allow the exercise of their rights.

ency is essential to understand the purpose and the interaction between human and machine, ensuring compliance with regulation and performance instructions. Design solutions could be needed, for example, to prevent automation bias.

However, the situation is different when transparency focuses on the *output* – type 3 - and instead of explaining how to use the output, aims to *explain it.* Here we have *transparency as explainability*. To guarantee informed decision-making, it could be important for the decision-maker to understand how the systems work internally and not only how to use them. This may even be imposed by the specific area in which the system is applied, but that does not result from the AI Act.

### 3.3.2. Right to an explanation?

Before the latest changes to the AI Act (AIA), there was a long debate surrounding a possible subjective right to an explanation under the General Data Protection Regulation (GDPR) – *v.g,*[16], which remains unresolved. In fact, only Recital 71 of the GDPR explicitly mentions the right to an explanation [9]. However, despite Article 22(3) of the GDPR not explicitly mentioning the right [6, p. 348; 26, p. 80], it could be considered a prerequisite to exercise the right to contest. This argument is also used in the context of Articles 13(2)(f), Article 14(2)(g), and Article 15(1)(h) of the GDPR. In case of automated decision as described in Article 22(1) GDPR the controller must provide "meaningful information" about the logic involved.[16] Information will only be meaningful if it attends to the individual rights the GDPR confers, namely, the right to contest [6,27].

Article 86 of the AIA explicitly mentions the right to an explanation. In these cases, there is an obligation to explain "the role of the AI system in the decision-making process". This requires identifying the relevance of the AI system for the decision, including the purpose and role of the machine and the human in the decision-making process. Furthermore, the regulation requires identifying the "main elements of the decision taken". However, there is some ambiguity regarding whether this entails the right to know (i) the feature values used to make the prediction, (ii) the weight of the specific features used in a specific case (local explanation), or (iii) the weights of the features in the model more generally (global explanation). As some authors consider attending to GDPR, we believe that this should be balanced with trade secrets and incentives to innovation, that why generally local explanations are enough [6][17].

Although the AIA does not expressly impose *a specific* technical solution, it can be useful in guaranteeing the right to an explanation. If an AI system "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" (Article 86), an opaque system cannot be used unless post hoc explainability is granted. That's the cases where the instructions of use shall contain "technical capabilities and characteristics of the

[16] According to [22] this is not a complex explanation of algorithms, but sufficient information to allow that the data subject to understand the reasoning behind the decision.

[17] As identified by [28], transparency costs "can be mitigated by choosing where and how transparency interventions are necessary".

high-risk AI system to provide information that is relevant to explain its output" (Article 13(3)(b)(iv)) and "information to enable deployers to interpret the output" (Article 13(3)(b)(vii)). This necessitates XAI techniques to be more targeted, and it may not be worthwhile to use an opaque system where the output is generated not by a human or even by human-legible rules.

Given the concern about the 'accuracy' of models reflected in Article 15 should transparency models be favoured even if this reduces accuracy, or is it sufficient to use ex post techniques? The model accepted will depend on two factors: (i) the system's domain of action and (ii) the expectations of its users, including explanation metrics. But black box models cannot be introduced if they are not accompanied by XAI techniques, making it necessary to have a transparent or explainable model.

## 4 Conclusions

AIA emphasizes transparency as a crucial aspect for certain AI systems, particularly high-risk ones. Transparency is envisioned as a vital mechanism to ensure accountability, trustworthiness[18], and fairness[19] in AI applications. This should be achieved through an environment of transparency throughout the value chain that extends beyond the obligations outlined in GDPR[20].

Transparency as explicability could require the adoption of both technical and non-technical measures [1]. The AIA does not mandate specific technical rules for high-risk systems, so providers retain the freedom to develop models if they ensure users can utilize the system appropriately and understand the output.

This is a significant step towards algorithmic transparency, but it also poses risks. Explaining black box decisions can be challenging and may lead to misplaced trust [18], information overload, incorrect or incomplete explanations, and increased automation bias [19]. Some authors argue that instead of requiring AI systems to explain themselves, we should opt to require the use of interpretable models for high-risk decisions [14,18]. In fact, the demand for transparency may incentivize providers to adopt explainable AI (XAI) techniques or prioritize transparency-by-design approaches.

In the future is essential demystify the metrics for each type of explainability, beginning with the distinction between ex ante and ex post governance tools. This idea of transparency can be illusory and empty if the requirements remain abstract [20].

Transparency always serves a purpose (it is a mean to an end). It is essential to understand the purposes outlined in the AIA, each corresponding to different metrics. Once the objectives are defined, it is necessary to analyse which techniques are best suited to each objective and which metrics meet those objectives.

---

[18] Trust in a system is bolstered when stakeholders comprehend its design and behaviour.

[19] Explainability plays a crucial role in understanding and correcting any flaws in the system, ensuring fairness. Moreover, a fair system allows affected individuals to challenge decisions made by AI systems, which is only possible if individuals have the right to an explanation.

[20] GDPR primarily applies to automated decisions with significant or legal effects [8], leaving out AI systems that could have significant impact such as disseminating fake news.

However, mere knowledge of the AIA is insufficient to meet legal requirements. While the AIA adopts a risk-based approach, it needs to be complemented and interpreted based on the *specific domain* of application of the system. The AIA establishes key frameworks to be further specified by European Standardization Organizations (ESOs), necessitating clarification by the European regulator through directives or guidelines tailored to each country's specificities. Also, it needs to balance the benefits and costs that its use entails, exploring alternative measures to achieve the goals.

**Acknowlegments**

The work of Nídia Andrade Moreira has been supported by FCT - Fundação para a Ciência e Tecnologia within the Grant 2021.07986.BD.

## References

[1] Independent High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI ( 2019).
[2] European Commission. White Paper on Artificial Intelligence – A European approach to excellence and trust (2020).
[3] European Parliament. Compromise AMs – JURI AI Act – FINAL 30/08/2022 (2022).
[4] Regulation 2024/1689 (EU) of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).
[5] Gyevnar, B., Ferguson, N., & Schafer, B. Bridging the transparency gap: What can explainable AI learn from the AI Act? In K. Gal, A. Nowé, G. J. Nalepa, R. Fairstein, & R. Rădulescu (Eds.), Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications; vol. 372. pp. 964 – 971 (2023).
[6] Hacker, P. and Passoth, J.H.. "Varieties of AI Explanations Under the Law. From GDPR to the AIA, and Beyond". In Holzinger, Goebel, Fong, Moon, Müller and Samek (eds.), Lecture Notes on Artificial Intelligence 13200: xxAI - beyond explainable AI, Springer, pp. 343-373 (2022).
[7] Sovrano, F., Sapienza, S., Palminari , M. , Vitali, F "Metrics, Explainability and the European AI Act Proposal", J, 581, 126-138, (2022).
[8] Edwards, L. and Veale, M, "Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?. IEEE Security & Privacy, 16(3), pp. 46-54, (2018).
[9] Ebers, M., Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s). In: Colonna , Liane ; Greenstein, Stanley (eds.), Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence (2021)
[10] Panigutti, C., Hamon, R., Hupont, I., Llordca, D.F., Yela, D.F., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Garrido, J.S., Gomez, E. The role of explainable AI in the context of the AI Act. In 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT'23). pp. 1139–1150, (2023).
[11] Dignum, V.. Responsible Artificial Intelligence. How to Develop and Use AI in a Responsible Way. Springer (2019).

[12] Bruijnm, H., Warnier, M., Janssen, M.: The perils and pitfalls of explainable AI: strategies for explaining algorithmic decision-making. Gov. Inf. . 39(2), 101666 (2022).
[13] Arrieta, A., et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115 (2020).
[14] Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, *1*(5), 206–215.
[15] Petkovic, D. "It is Not "Accuracy vs. Explainability"—We Need Both for Trustworthy AI Systems," in IEEE Transactions on Technology and Society, vol. 4, no. 1, pp. 46-53 (2023).
[16] Goodman, B. and Flaxman, S "European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation", *AI Magazine* 38(3), 50-57, (2017).
[17] Belle, V., Papantonis, I.: Principles and practice of explainable machine learning. Front. Big Data 4, 688969 (2021).
[18] KU Leuven Homepage, Cuypers, A., The right to an explanation in the AI Act: a right to interpretable models? 2024, https://www.law.kuleuven.be/citip/blog/the-right-to-explanation-in-the-ai-act-a-right-to-interpretable-models/, last accessed 15/07/2024.
[19] Mulder, W.D, Valcke, P. The need for a numeric measure of explainability. 2021 IEEE International Conference on Big Data (Big Data), pp. 2712-2720 (2021).
[20] Walke, F.; Bennek, L.; and Winkler, T.J., "Artificial Intelligence Explainability Requirements of the AI Act and Metrics for Measuring Compliance". Wirtschaftsinformatik 2023 Proceedings. 77 (2023).
[21] Moreira, N.A., Freitas, P.M., Novais, P. The AI Act Meets General Purpose AI: The Good, The Bad and The Uncertain. In: Moniz, N., Vale, Z., Cascalho, J., Silva, C., Sebastião, R. (eds) Progress in Artificial Intelligence. EPIA 2023. Lecture Notes in Computer Science(), vol 14116. Springer, Cham (2023).
[22] Article 29 Data Protection Working Party. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (WP251rev.01) (2018).
[23] European Telecomunications Standards Institute. Securing Artificial Intelligence (SAI); Explicability and transparency of AI processing (2023).
[24] Mitchell, M., Wu, S. Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Zpitzer, E., Raji, I.D. ,, Gebru, T. "Model Cards for Model Reporting", FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency, p. 220-229 (2019).
[25] European Commission. Building Trust in Human-Centric Articial Intelligence. Technical Report COM(2019) 168 final (2019).
[26] Wachter, S. ; Mittelstadt, B.; Floridi, L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in General Data Protection Regulation. International Data Privacy Law,Vol. 7(2) (2017).
[27] Selbst, A., Powles, J. Algorithmic Accountability and Public Reason. International Data Privacy Law 7(4), pp. 233-242 (2017).
[28] EPRS. A governance framework for algorithmic accountability and transparency (2019).
[29] Facchini, A.; Termine, A.. Towards a Taxonomy for the Opacity of AI Systems. In: Müller, V.C. (eds) Philosophy and Theory of Artificial Intelligence 2021. PTAI 2021. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 63. Springer, Cham, pp.73-89 (2021).
[30] Burrel, J.. How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society (2006).