J Multidisciplinary Scientific Journal

MDPI

Article

Metrics, Explainability and the European AI Act Proposal

Francesco Sovrano [1], Salvatore Sapienza [2], Monica Palmirani [2,*] and Fabio Vitali [1]

[1] Department of Computer Science and Engineering (DISI), Università di Bologna, 40126 Bologna, Italy; francesco.sovrano2@unibo.it (F.S.); fabio.vitali@unibo.it (F.V.)
[2] CIRSFID—ALMA AI, Università di Bologna, 40126 Bologna, Italy; salvatore.sapienza@unibo.it
* Correspondence: monica.palmirani@unibo.it

Abstract: On 21 April 2021, the European Commission proposed the first legal framework on Artificial Intelligence (AI) to address the risks posed by this emerging method of computation. The Commission proposed a Regulation known as the AI Act. The proposed AI Act considers not only machine learning, but expert systems and statistical models long in place. Under the proposed AI Act, new obligations are set to ensure transparency, lawfulness, and fairness. Their goal is to establish mechanisms to ensure quality at launch and throughout the whole life cycle of AI-based systems, thus ensuring legal certainty that encourages innovation and investments on AI systems while preserving fundamental rights and values. A standardisation process is ongoing: several entities (e.g., ISO) and scholars are discussing how to design systems that are compliant with the forthcoming Act, and explainability metrics play a significant role. Specifically, the AI Act sets some new minimum requirements of explicability (transparency and explainability) for a list of AI systems labelled as "high-risk" listed in Annex III. These requirements include a plethora of technical explanations capable of covering the right amount of information, in a meaningful way. This paper aims to investigate how such technical explanations can be deemed to meet the minimum requirements set by the law and expected by society. To answer this question, with this paper we propose an analysis of the AI Act, aiming to understand (1) what specific explicability obligations are set and who shall comply with them and (2) whether any metric for measuring the degree of compliance of such explanatory documentation could be designed. Moreover, by envisaging the legal (or ethical) requirements that such a metric should possess, we discuss how to implement them in a practical way. More precisely, drawing inspiration from recent advancements in the theory of explanations, our analysis proposes that metrics to measure the kind of explainability endorsed by the proposed AI Act shall be risk-focused, model-agnostic, goal-aware, intelligible, and accessible. Therefore, we discuss the extent to which these requirements are met by the metrics currently under discussion.

Keywords: explainable artificial intelligence; explainability; metrics; standardisation; Artificial Intelligence Act

Citation: Sovrano, F.; Sapienza, S.; Palmirani, M.; Vitali, F. Metrics, Explainability and the European AI Act Proposal. *J* **2022**, *5*, 126–138. https://doi.org/10.3390/j5010010

Academic Editor: Larisa Ivascu

Received: 10 January 2022
Accepted: 10 February 2022
Published: 18 February 2022

1. Introduction

The ability and need of humans to explain has been studied for centuries, initially in philosophy and more recently also in all those sciences aiming at a better understanding of (human) intelligence. For example, Cognitive Science studies explanations as a mechanism for humans to absorb information building cognition of themselves and reality, while AI and law study how and why, respectively, to explain complex self-written software learned from data.

Measuring the degree of explainability of AI systems has become relevant in the light of research progress in the eXplainable AI (XAI) field, the Proposal for an EU Regulation on Artificial Intelligence, and ongoing standardisation initiatives that will translate these technical advancements in a de facto regulatory standard for AI systems. The AI Act will require technical documentation attached to a high-risk AI application, in order to certify both human oversight and compliance to the Regulation.

To date, standardisation entities have proposed white papers and preliminary documents showing their progress, among them we mention (an extensive list is available at [1]):

- The European Telecommunications Standards Institute (ETSI), which observed that "when it comes to AI capabilities as part of new standards, there is a need to revise these models, by identifying appropriate reference points, AI sub-functions, levels of explicability of AI, quality metrics in the areas of human-machine and machine-machine interfaces [2] (p. 23)";
- The CEN-CENELEC, which has proposed to "develop research-based metrics for explainability (to tie in with high level conceptual requirements), which can be developed into pre-standards like workshop agreements or technical [3] (p. 8)";
- ISO/IEC TR 24028:2020(E), stating that "it is important also to consider the measurement of the quality of explanations" and provides for details on the key measurements (i.e., continuity, consistency, selectivity) in paragraphs 9.3.6 and 9.3.7.

Considering that since ISO/IEC TR 24028:2020(E), the literature has started to propose new metrics and mechanisms, with this work we study and categorise the existing approaches to quantitatively assess the quality of explainability in machine learning and AI. We do so through the lenses of law and philosophy, not just computer science. This last characteristic is certainly our main contribution to the literature of XAI and law, and we believe it may foster future research to embrace an interdisciplinary approach less timidly, for the sake of a better conformity to the existing (and forthcoming) regulations in the EU.

The proposed Regulation is connected to the need of measuring the degree of explainability of AI systems, in particular high-risk. Explainability metrics shall be deemed a crucial tool to check the technical documentation required by the Proposal and, in turn, verify the extent to which the AI system is aligned the goals identified by the Proposal further discussed below. The lack of clearly-defined and standardised explainability metrics might make ineffectual the Regulation by leaving room for discretion (and, possibly, abuses) by AI systems' manufacturers, thus posing threats to individuals' rights and freedoms beyond an acceptable level. Choosing metrics to explainability that are aligned with the overall goals of the AI Act is therefore necessary to evaluate the quality of such technical documentation. Yet, we found a lack of research on this subject matter, and we believe that this contribution is timely due to the current state of the debate, which can benefit from the following study.

This paper is structured as follows. In Sections 2 and 3 we present the research background and the methodology of this paper. Then, in Sections 4–6, we explore the definitions and the properties of explainability in philosophy and in the proposed AI Act. Finally, in Sections 7 and 8 we perform an analysis of the existing quantitative metrics of explainability, discussing our findings and future research.

2. Related Work

In XAI literature there are many interesting surveys on explainability techniques [4–8], classifying algorithms on different dimensions to help researchers in finding the more appropriate ones for their own work. Practically, all these surveys focus on a classification of the mechanisms to achieve explainability rather than how to measure the quality of it, and we believe our work can help in this latter goal.

For example, ref. [6] classifies XAI methods with respect to the notion of explanation and the type of black-box system. The identified characteristics are the level of detail of explainability (from high to low: global logic, local decision logic, model properties) and the level of interpretability of the original model. Similar to [6], ref. [4] also studied XAI considering interpretability and level of detail, but does it by adding the model-specificity of the technique (model-specific or model-agnostic) to the equation. Nonetheless, their analysis also considers the need for classifying techniques according to some quality metrics, mentioning works such as [9–11]. Regardless, ref. [4] does not dive into any concrete classification of existing evaluation methods.

On the other hand, the work by Zhou et al. [8] focuses specifically on the metrics to quantify the quality of explanation methods, classifying them according to the properties they can measure and the format of explanations (model-based, attribution-based, example-based) they support. More precisely, following the taxonomy given by [9] (that separates between application-grounded, human-grounded and functionality-grounded evaluation mechanisms), ref. [8] narrows down the survey to the functionality-grounded metrics, proposing for them a new taxonomy including interpretability (in terms of clarity, broadness, and parsimony) and fidelity (as completeness, and soundness).

Among all the identified surveys, ref. [8] is certainly the closest to our work, in terms of focus of the survey. The main distinction between our work and [8] is probably our assumption that multiple definitions of explainability exist, each one possibly requiring its own type of metrics. Furthermore, differently from [8], we analyse explainability metrics on their ability to meet the requirements set by the AI Act, regardless their position in [9]'s taxonomy.

3. Methodology

We performed an exploratory literature review of existing metrics to measure the explainability of AI-related explanations, together with a qualitative legal analysis of the explainability requirements to understand the alignment of the identified metrics to the expectations of the proposed AI Act. To do so we collected all the papers cited in [8], re-classifying them. Then, we integrated with further works identified through an in-depth keyword-based research on Google Scholars, Scopus, and Web Of Science. The keywords we used were "degree of explainability", "explainability metrics", "explainability measures", and "evaluation metrics for contrastive explanations". Among the retrieved papers we selected only those following the selection criteria discussed throughout the whole paper.

An independent legal analysis was carried out on the proposed Artificial Intelligence Act. Considering the lack of case law and the paucity of studies on this novel piece of legislation, a literal assessment of its provisions is preferred for a more critical analysis based on previous enquiries.

4. Definitions of Explainability

Considering the definition of "explainability" as "the potential of information to be used for explaining", we envisage that a proper understanding of how to measure explainability must pass through a thorough definition of what constitutes an explanation and the act of explaining.

Many theories of explanations have thrived in philosophy, sometimes driven by sciences such as psychology and linguistics. Among them we cite the five most important ones in contemporary philosophy [12], coming from: Causal Realism, Constructive Empiricism, Ordinary Language Philosophy, Cognitive Science, Naturalism and Scientific Realism. Interestingly, each one of these theories devises different definitions of "explanation", sometimes in a complementary way. A summary of these definitions is shown in Table 1, shedding light on the fact that there is no complete agreement on the nature of explanations.

In fact, if we look at their specific characteristics, we may find that all but Causal Realism are pragmatic, meaning that they envisage explanations being customised for any specific recipient. Furthermore, Causal Realism and Constructive Empiricism are rooted on causality. On the other hand, Ordinary Language Philosophy, Cognitive Science and Scientific Realism do not constrain explanations to causality, studying the act of explaining as an iterative process involving broader forms of question answering. Cognitive Science and Scientific Realism are more focused on the effects that an explanation has on the explainee (the recipient of the explanation). Interestingly, the majority of the definitions envisage the process of question answering as part, or constituent, of the act of explaining.

Table 1. In this table we summarise the definitions of explanation and explainable information for each one of the identified theories of explanations.

Theory	Definition of Explanation	Definition of Explainable Information
Causal Realism [13]	It is a description of causality, as chains of causes and effects.	It can fully describe causality.
Constructive Empiricism [14]	It is contrastive information answering WHY questions, allowing one to calculate the probability of a particular event relative to a set of (possibly subjective) background assumptions.	It provides answers to contrastive WHY questions.
Ordinary Language Philosophy [15]	Explaining is pragmatically answering to (not just WHY) questions, with the explicit intent of producing understanding.	It can be used to pertinently answer questions about relevant aspects, in an illocutionary way.
Cognitive Science [16]	Explaining is a process triggered as response to predictive failures and it is about providing information to fix these failures in a mental model (sometimes intended as a hierarchy of rules).	It can fix failures in mental models.
Naturalism and Scientific Realism [17]	Explaining is an iterative process of confirmation of truth based on inference to the best explanation. An explanation increases understanding, not simply by being the correct answer to a particular question, but by increasing the coherence of an entire belief system (e.g., a subject).	It can be used to increase understanding, i.e., by answering to particular questions.

Importantly, we assert that whenever explaining is considered to be a pragmatic act, explainability differs from explaining. In fact, pragmatism in this sense is achieved when the explanation is tailored to the specific user, so that the same explainable information can be presented and re-elaborated differently across users. It follows that for each philosophical tradition but Causal Realism, we have a definition of "explainable information" that slightly differs from that of "explanation", as shown in Table 1.

5. Explainability Desiderata

These abstract definitions show the process of answering questions as a possible common ground to explainability, without identifying the specific properties it should possess. Other works in philosophy have explored the central criteria of adequacy of explainable information. In philosophy, the most important work about the central criteria of adequacy of explainable information is likely to be Carnap's [18], specifically targeting what he calls explications. Even though Carnap studies the concept of explication rather than that of explainable information, we assert that they share a common ground making his criteria fitting in both cases. In fact, explication in Carnap's sense is the replacement of a somewhat unclear and inexact concept, the explicandum, by a new, clearer, and more exact concept called explicatum, and that is exactly what information does when it is made explainable. In this sense, our interpretation of Carnap's concept of explication is that of explainable information.

Carnap's central criteria of explication adequacy are [19] similarity, exactness and fruitfulness. Similarity means that the explicatum should be similar to the explicandum, in the sense that at least many of its intended uses, brought out in the clarification step, are preserved in the explicatum. On the other hand, exactness means that the explication should, where possible, be embedded in some sufficiently clear and exact linguistic framework, while fruitfulness means that the explicatum should be used in a high number of other good explanations (the more, the better). Notably, Carnap also discussed another desideratum: simplicity. This criterion is presented as being subordinate to the others. In fact, simplicity implies that once all other desiderata have been satisfied, the simplest available explication is to be preferred over more complicated alternatives.

Despite Carnap being arguably more aligned to Scientific Realism [20], his adequacy criteria seem to be transversal to all the identified definitions of explainability, possessing preliminary characteristics for any piece of information to be considered properly explainable. Therefore, our interpretation of Carnap's criteria in terms of measurements is the following:

- Similarity is about measuring how similar the given information is to the explanandum. This can be estimated by counting the number of relevant aspects covered by information and the number of details it can provide.
- Exactness is about measuring how clear the given information is, in terms of pertinence and syntax, regardless its truth. Differently from Carnap, our understanding of exactness is broader than that of adherence to standards of formal concept formation [21].
- Fruitfulness is about measuring how much a given piece of information is going to be used in the generation of explanations. Each explainability definitions defines fruitfulness differently.

Interestingly, the property of truthfulness (being different from exactness) is not explicitly mentioned in Carnap's desiderata. That is to say that explainability and truthfulness are complementary, yet different, as discussed also by [22]. In fact, an explanation is such regardless its truth (wrong but high-quality explanations exist, especially in science). Vice versa, highly correct information can be very poorly explainable. This naturally leads us to the following discussion about the characteristics that explanations shall have under the proposed AI Act.

6. Explainability Obligations in the Proposed AI Act

Following the EU Commission's Proposal for an Artificial Intelligence Act (AIA), it is now time to discuss how explainability is connected to the novel obligations introduced by the Act. It has to be preliminary observed that, since this piece of legislation is still undergoing an extensive debate inside and outside EU institutions, the following considerations could and will likely be subject to change.

Considering the nature and the characteristics of the requirements posed by the AIA, it is worth questioning how such explainability metrics could be designed to fulfil the necessities of all the entities whose behavior will be regulated by the AIA. Notably, the

Act defines a significant number of undertakings, including AI "provider" (art. 3(1)) and "small scale provider" (art. 3(3)), "user" (art. 3(3)), "importer" (art. 3(6)), "distributor" (art. 3(7)), "operator" (art. 3(8)), and authorities, including "notifying authority" (art. 3(19)), "conformity assessment body" (art. 3(21)), "notified body" (art. 3(22)), "market surveillance authority" (art. 3(26)), "national supervisory authority" (art. 3(42)), "national competent authority" (art. 3(43)).

The discussion towards "explainability and law" has departed from the contested existence of a right to explanation in the General Data Protection Regulation (GDPR) [23,24] to embrace contract, tort, banking law [25], and judicial proceedings [26]. While these legal sectors present significant differences in the applicable law and jurisdiction, they all try to cast light on the significance of algorithmic transparency within existing legal sectors. Contextually, the most recent discussion has included the proposed AIA and the obligations set by this forthcoming piece of legislation linked to the explainability. Differently from other domains, the AIA is specific to AI systems and requires an ad hoc discussion rather than the framing of these systems in the discussion of other legal domains. This is because AI technologies are not placed within an existing legal framework (e.g., banking), but the whole legal framework (i.e., the AIA) is built around AI technologies. However, the previous discussion focusing on other legal regimes constitutes a valuable background for our research and thus it contributes to our discussion. The interpretations proposed by recent commentators [25,26] identify several nuances of algorithmic transparency. Our focus, however, shall be confined to the interaction between the nuance of explainability and obligations emerging from the AIA already identified by these early commentators.

As regards the GDPR, scholars have extensively discussed whether or not the right to receive an explanation for "solely automated decision-making" processes exists in the GDPR. Regardless of the answer, the data controller has an obligation to provide "meaningful information about the logic involved" in the automated decision (art. 13(2)(f)), art. (14(2)(g)), art. (15(1)(h)). This information is deemed to be "right-enabling" [1] as it is necessary and instrumental to exercise the rights enshrined in Article 22, namely, to express views on the decision and to contest it. The same goes with the kind of transparency that is necessary to ensure the right to a fair trial in the context of judicial decision-making [9]. Then, the discussion identified a "technical" necessity of explainability, that is necessary to improve the accuracy of the model. In legal terms, it is echoed by the "protective" transparency that is needed to minimise risks and to comply with certain legal regimes, namely tort law and contractual obligations, especially in consumer and banking law. As with data protection law, these varieties are instrumental to improve a product and protect its users or the persons affected by the system from damages.

If explainability is often instrumental to achieve some legislative goals, it is likely that it could be meant to foster certain regulatory purposes also under the AIA. From the joint reading of a series of provisions, it can be argued that explainability in the AIA is both *user-empowering* and *compliance-oriented*: on the one hand, it serves to enable users of the AI system to use it correctly; on the other hand, it helps to verify the adequacy of the system to the many obligations set by the AIA, ultimately contributing to achieve compliance.

Recital 47 and art. 13(1) state that high-risk AI systems listed in Annex III shall be designed and developed in such a way that their operation is comprehensible by the users. They should be able (a) to interpret the system's output and (b) to use it in an appropriate manner. This is a form of user-empowering explainability. Then, the second part of art. 13 specifies that "an appropriate type and degree of transparency shall be ensured, with a view to achieving compliance with the relevant obligations of the user and of the provider [. . .]". In our reading, this provision specifies that this explainability obligation (i.e., transparent design and development of high-risk AI systems) is compliance-oriented.

The two-fold goal of art. 13(1) is then echoed by other provisions. As regards the user-empowering interpretation, art. 14(4)(c) relates explainability to "human oversight" design obligations. These measures should enable the individual supervising the AI system to correctly interpret its output. Moreover, this interpretation shall put him or her in the

position to decide whether it might be the case to "disregard, override or reverse the output" (art. 14(4)(d). User-empowering explainability is also needed for the general evaluation of the decisions taken by AI system users, in particular by qualified entities, such as law enforcement authorities and the others listed in Recital 38. In these cases, not only is the technical documentation relevant for the compliance assessment, but also the contribution of the AI system to the final decision shall be clear. Such clarity is necessary to ensure a right to effective remedy in cases that threaten fundamental rights and for which the law already provides redress mechanisms.

The compliance-oriented explainability interpretation becomes evident in the technical documentation to be provided according to Article 11. Compliance is based on a presumption of safety if the system is designed according to technical standards (art. 40) to which adherence is documented, whereas third-party assessment appears only post-market or on specific sectors. This compliance framework is inspired by the New Legislative Framework (NLF) [27]. First, the legislation details the essential requirements for product safety, then they become best practice or standards through standardisation entities. Chapter IV of the AIA regulates this process in the case of high-risk AI systems. Inter alia, Annex IV(2) (b) includes "the design specifications of the system, namely the general logic of the AI system and of the algorithms" among the information to be provided to show compliance with the AIA before placing the AI system in the market. Hence, the system should be explainable in a manner that allows an evaluation of conformity by the provider in the first instance and, when necessary, by post-market monitoring authorities, not only with regards to the algorithms used, but also—and most crucially—to the general reasoning supporting the AI system, as the use of a conjunctive "and" seems to suggest. Since the general approach taken by the proposed AIA is a risk-reduction mechanism (Recital 5), this form of explainability is ultimately meant to contribute to minimising the level of potential harmfulness of the system.

User-empowering and compliance-oriented explainability overlap in art. 29(4): the user shall be able to "monitor the operation of the high-risk AI system on the basis of the instructions of use". When a risk is likely to arise, the user shall suspend the use of the system and inform the provider or the distributor. This provision entails the capability of understanding the working of the system (real-time) and making previsions on its output. Suspending in the case of likely risk is the overlapping between the two nuances of explainability: the user is empowered to stop the AI system to avoid contradicting the rationale behind the AIA, i.e., risk minimization. Differently from the GDPR, no explicit provision enables the person affected by the system to exercise rights against the provider or the user of the system or to access explanations about the system's working. Moreover, explainability obligations are solely limited to high-risk AI systems: medium-risk (art. 52) follows "transparency obligations" that consist of disclosing the artificial nature of the system in the case of chat-bots, the exposition to certain recognition systems, the "fake" nature of image, audio or video content.

Once the existence of explainability obligations and their extent is clarified, let us discuss the requirements that metrics should have to ease compliance with the AIA. Let us remind that, under the Proposal, adopting a standard means certifying the degree of explainability of a given AI system: if the system is compliant to the standard and such compliance is documented according to Annex IV, then it can enjoy the presumption of safety and can be placed in the market. Therefore, metrics become useful in the course of the standardisation process (i) ex ante, when defining the explainability measures adopted by the standard. Standardisation entities will be allowed to measure and compare the explainability across different systems (ii) ex post, when verifying in practice the adoption of a standard. Market surveillance authorities can verify whether the explainability of AI system under scrutiny conforms to the one mandated by the standard by comparing the two.

In the light the purposes of the AIA, the legislative technique adopted, and the needs of standardisation entities, explainability metrics should have at minimum the

following characteristics: risk-focused, model-agnostic, goal-aware, and intelligible and accessible. Risk-focused means that the metric should be functional to measure the extent to which the explanations provided by the system allow for an assessment of the risks to the fundamental rights and freedoms of the persons affected by the system's output. This is necessary to ensure both user-empowering (e.g., art. 29) and compliance-oriented (Annex IV) explainability. Model-agnostic means that the metric should be appropriate to all the AI systems regulated by the AIA, hence machine learning approaches, logic- and knowledge-based approaches, and statistical approaches (Annex I). Goal-aware means that the metric should be flexible towards the different needs of the potential explainees (i.e., AI system providers and users, standardisation entities, etc.) and applicable in all the high-risk AI applications listed in Annex III. Since it might be hard to determine ex ante the nature, the purpose, and the expertise of the explainee, the metrics should consider the highest possible number of potential explainees, included qualified users listed in Recital 38. Intelligible and accessible means that if information on the metrics is not accessible (e.g., due to intellectual property reasons), explainees will be confronted with an ignotum per ignotius situation: they will be informed about the degree of explainability of a system, yet without understanding how this result is achieved. This would contradict the principle of risk minimisation.

7. Discussing Existing Quantitative Measures of Explainability

In this section we analyse and categorise existing metrics and measures to quantitatively estimate the degree of explainability of information. The goal of this paper is to give a quick overview to researchers and practitioners of the pros and cons of each quantitative metric, to understand its applicability across different needs and interpretations of explainability. For the classification we use the following set of dimensions specifically designed to evaluate the characteristics defined in Sections 4 and 6:

- Evaluation Type: evaluations can be quantitative or qualitative. Intuitively, quantitative evaluations are harder to achieve;
- Model-Specificity: whether the metric is model-agnostic, working on every possible AI, or not;
- Supported Media: whether the metric can evaluate only textual information, or images, or both;
- Supported Format: whether the metrics support case-based explanations, or rule-based, or both;
- Information Format: the information format supported by the metric, i.e., rule-based, example-based, natural language text, etc.;
- Supporting Theory: the theory to which the metric is inspired to, i.e., Cognitive Science, Constructive Empiricism, etc.;
- Subject-based: whether the evaluation requires human subjects or not.
- Measured Criteria: the combination of Carnap's adequacy desiderata that is measuring, i.e., similarity, exactness, fruitfulness.

As mentioned also before, in our analysis we decided to consider Carnap's criteria of adequacy as the main properties of explainability. Despite the presence of many works in XAI discussing or proposing metrics for measuring the truthfulness of explanations (i.e., [11]) we decided to focus only on those ones measuring the degree of explainability of information. Others measure explainability indirectly, through metrics to estimate the user-centrality of explanations. We will consider these ones as supporting the interpretation of explainability from Cognitive Science.

Therefore, our whole analysis is based on the alignment of the different sub-properties identified by literature (i.e., coherence, fidelity, etc.) to Carnap's desiderata. As consequence, we consider only a part of the dimensions adopted by [8]. More precisely, we keep clarity, broadness and completeness, aligning the first two to Carnap's exactness and the latter to similarity. In fact, we deem soundness to be as truthfulness, a complementary

characteristic to explainability. On the other hand, broadness and parsimony are considered as characteristics to achieve pragmatic explanations rather than properties of explainability.

Furthermore, differently from ISO/IEC TR 24028:2020(E) we did not focus on metrics specific to ex-post Feature Attribution explanations, selecting methods possibly applicable also on ex-ante or more generic types of explanations. In fact, Feature Attribution is only for explainability about causality, hence being more centred on Causal Realism, while our investigation tries to compare different metrics across the supporting philosophical theories.

As shown in Table 2, we were able to find at least one example of metric for each supporting philosophical theory, with a majority of metrics focused on Causal Realism and Cognitive Science. The lack of metrics clearly aligned to Scientific Realism is probably due to the fact that the property of being the best explanation cannot be an objective measure of the likelihood that it is true. What is common to all the metrics based on Cognitive Science is that they require human subjects for performing the measurement, making them more expensive with respect to the others, at least in terms of human effort. Furthermore, the metrics proposing heuristics to measure all the three main Carnap's desiderata are just two, one for Causal Realism [28] and the other for Ordinary Language Philosophy [29]. Interestingly, ref. [28] evaluates the three desiderata separately, while [29] propose a single metric combining all of them.

Table 2. Comparison of different explainability metrics. The column "Metric" points to reference papers, while column "Name" points to the names used by the authors of the metric to describe it. Elements in bold are column-wise, indicating they are the best ones.

Metric	Information Format	Supporting Theory	Subject-Based	Measured Criteria	Name
[30]	Rule-based	Causal Realism	**No**	Similarity, Fruitfulness	Fidelity, Completeness
[31]	Feature Attribution	Causal Realism	**No**	Similarity, Fruitfulness	Monotonicity, Non-sensitivity, Effective Complexity
[28]	Rule-based	Causal Realism	**No**	**Similarity, Exactness, Fruitfulness**	Fidelity, Unambiguity, Interpretability, Interactivity
[32]	**All**	Causal Realism, Cognitive Science, Scientific Realism	Yes	Exactness, Fruitfulness	Causability
[33]	**All**	Cognitive Science, Scientific Realism	Yes	Exactness, Fruitfulness	Satisfaction, Trust, Mental Models, Curiosity, Performance
[34]	Example-based	Constructive Empiricism	No	Exactness	Proximity, Sparsity, Adequacy (Coverage)
[31]	Example-based	Constructive Empiricism	No	Similarity, Fruitfulness	Non-Representativeness, Diversity
[29]	Natural Language Text	Ordinary Language	No	**Similarity, Exactness, Fruitfulness**	Aspects Coverage, Degree of Explainability

Let us now discuss the extent to which philosophy-oriented metrics measure explainability and can match the requirements set by the proposed AIA. First, under the AIA, metrics should allow the measurement of the capability of the system to provide information related to the risks posed to fundamental rights and freedoms of the persons affected by the system. This is a form of goal-oriented explainability, thus calling for a pragmatic interpretation of explanations as that of all the theories identified in Section 4 but Causal Realism. Then, metrics shall be appropriate to the list of AI approaches listed in

Annex I. This entails that only those based on a model-agnostic approach to explainability can ease compliance to the AIA, unless a combination of different model-specific metrics is envisaged.

Furthermore, metrics shall be also adaptive to the several market sectors which can observe a substantial deployment of high-risk AI systems. Therefore, given the horizontal application of the AIA and the contextual applicability of sectoral legal frameworks, we have that any explainability metric should be flexible enough to adapt to different technological constraints and explanation objectives. Considering that explanation objectives can be framed in terms of questions to answers, definitions of explanations being focused solely on specific enquiries as Causal Realism and Constructive Empiricism may struggle in meeting the adaptivity requirement.

Flexibility towards all the potential explainees entails that subject-based metrics would necessarily require a significant number of explainees to be tested and standardised. While this requirement is not impossible in theory, it might be hard to reach in practice. Finally, as with the requirement of flexibility towards the potential explainees, the intelligibility, accessibility and understandability of metrics require the metrics to be economically accessible. However, all the subject-based metrics may be very expensive, thus making the metric less accessible for potential competitors. The same goes with the intelligibility and those metrics developed under standards that are not open to public scrutiny. An additional problem of subject-based metrics is that they rely on the behaviour of subjects and they lack reproducibility for a subject acting in an indeterministic way, so that running the same measurement twice might lead to completely different results.

In Table 3, we show how the different explainability definitions are aligned to the main principles identified in Section 6. The results show the pros and cons of each approach, suggesting that different metrics may be complementary, serving different roles depending on the context.

Table 3. This table shows how the different philosophical definitions of explainability are aligned to the properties we identified in Section 6.

Theory	Risk-Focused	Model-Agnostic	Goal-Aware	Intelligible and Accessible
Causal Realism	Yes, if understanding risks implies understanding causality	Not available yet	No, it's not pragmatic and it considers only goals related to causality	Yes, it can be
Constructive Empiricism	Yes, if explaining risks is about answering WHY questions	Not available yet	No, it focuses only on WHY questions	Yes, it can be
Ordinary Language Philosophy	Yes, it can be	Maybe, only if all the explanations can be represented in a natural language	Yes	Yes, it can be
Cognitive Science	Yes, it can be	Yes, the evaluation is subject-based	Yes	Unlikely, all the subject-based metrics may be very expensive, thus making the metric less accessible for potential competitors
Naturalism and Scientific Realism	Yes, it can be	Yes, the evaluation is subject-based	Yes	Unlikely, it relies on usually expensive subject-based metrics

8. Final Remarks

Our work encompasses different disciplines proposing an interdisciplinary analysis of explainability metrics in Artificial Intelligence. This is why we classified the existing metrics in XAI, to assess the conformity of explanations under the AI Act, with the final goal of pointing to possible incompatibilities and directions for future literature to take.

To this end, our own approach started with the identification of the major definitions of explanation, in philosophy, from which we also framed the meaning of explainability. Then, inspired by Carnap's criteria of adequacy and the proposed AI Act, we selected a set of properties that explainable information should possess, thus classifying the metrics we found in literature accordingly. The properties emerging from the proposed AI Act were identified starting from a literal interpretation of the Proposal and how explainability is conceptualised by the Act, i.e., as a "user-empowering" and as a "compliance-oriented" design requirement. More specifically, through the lens of the obligations enshrined by the proposed Act, we identified that, to ease compliance, explainability metrics should be risk-focused, model-agnostic, goal-aware, intelligible, and accessible.

None of the proposed metrics seem to match all these requirements. Some theories are more concerned with the effects of explanations on explainees and his/her needs (e.g., Cognitive Science), whereas others are more concerned with the language used to explain (e.g., Ordinary Language Philosophy). Therefore, we have that the former tends to require metrics that are less accessible, both economically and epistemically, whilst the latter is more dependent on the format of the explanation, potentially being less model-agonistic, unless we assume that every explanation can be represented in natural language.

Considering the current level of discussion and that our findings might be subject to change due to the institutional debate about the Proposal, further research is needed to consolidate the interpretation of the Act in the light of its future changes and to define metrics that can be used in the course of the standardisation process while being respectful of the Proposal and, in particular, its Annexes.

Another open question to be left to further research is the translation of explainability metrics requirements into practice by selecting one or more concurring metrics that are compliant with the legislative goals of the AI Act and, eventually, the thresholds that meet the Act's requirements and expectations. Considering the current level of discussion and the relevance of standardisation processes in the current debate, it is likely that explainability levels will become crucial in the evaluation process of the admissibility of high-risk AI systems in the EU. At the same time, such measurements will allow judges and other authorities to fruitfully evaluate the degree of interpretability of AI systems in light of the goals pursued by the AI Act, in concrete cases. Such scrutiny will affect their decisions, when these systems are used by such authorities as with the case of Recital 38, and the degree of interaction with AI systems, e.g., in the case of refuse of their use. For instance, a judge could reject the support of an AI system if the metrics that are used to describe the explainability of its decision-making are not aligned with the goals of the Act. Finally, the metrics will constitute a benchmark in the evaluation of compliance to the AI Act by the systems' manufacturers in courts.

Author Contributions: Conceptualization, F.S., M.P., S.S., F.V.; methodology, F.S., M.P., S.S., F.V.; validation, F.S., M.P., S.S., F.V.; formal analysis, F.S., M.P., S.S., F.V.; investigation, F.S., M.P., S.S., F.V.; writing—original draft preparation, F.S., M.P., S.S., F.V.; writing—review and editing, F.S., M.P., S.S., F.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. European Commission. Artificial Intelligence Rolling Plan 2021. Available online: https://joinup.ec.europa.eu/collection/rolling-plan-ict-standardisation/artificial-intelligence (accessed on 9 January 2022).
2. European Telecommunications Standards Institute. Artificial Intelligence and Future Directions for ETSI. Available online: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp34_Artificial_Intellignce_and_future_directions_for_ETSI.pdf (accessed on 9 January 2022).
3. CEN-CELENEC. CEN-CENELEC Response to the EC White Paper on AI. Available online: https://ftp.cencenelec.eu/EN/News/PolicyOpinions/2020/CEN-CLC_AI_FG_ (accessed on 9 January 2022).
4. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
5. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
6. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42. [CrossRef]
7. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honululu, HI, USA, 25 April 2020.
8. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **2021**, *10*, 593. [CrossRef]
9. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
10. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
11. Mohseni, S.; Block, J.E.; Ragan, E.D. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv* **2018**, arXiv:1801.05075.
12. Randolph Mayes, H. Theories of Explanation. 2001. Available online: https://iep.utm.edu/explanat/ (accessed on 9 January 2022).
13. Salmon, W.C. *Scientific Explanation and the Causal Structure of the World*, 1st ed.; Princeton University Press: Princeton, NJ, USA, 1984.
14. Van Fraassen, B.C. *The Scientific Image*, 1st ed.; Oxford University Press: Oxford, UK, 1980.
15. Achinstein, P. *The Nature of Explanation*, 1st ed.; Oxford University Press: Oxford, UK, 1983.
16. Holland, J.H.; Holyoak, K.J.; Nisbett, R.E.; Thagard, P.R. *Induction: Processes of Inference, Learning, and Discovery*, 1st ed.; MIT Press: Cambridge, MA, USA, 1989.
17. Sellars, W.S. Philosophy and the Scientific Image of Man. In *Science, Perception, and Reality*; Colodny, R., Ed.; Humanities Press/Ridgeview: New York, NY, USA, 1962; pp. 35–78.
18. Dutilh Novaes, C.; Reck, E. Carnapian explication, formalisms as cognitive tools, and the paradox of adequate formalization. *Synthese* **2017**, *194*, 195–215. [CrossRef]
19. Leitgeb, H.; Carus, A. Rudolf Carnap. 2021. Available online: https://plato.stanford.edu/archives/sum2021/entries/carnap/ (accessed on 9 January 2022).
20. Alai, M. Scientific Realism, Metaphysical Antirealism and the No Miracle Arguments. *Found. Sci* **2020**, 1–24. [CrossRef]
21. Brun, G. Explication as a method of conceptual re-engineering. *Erkenntnis* **2016**, *81*, 1211–1241. [CrossRef]
22. Hilton, D.J. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Think. Reason.* **1996**, *2*, 273–308. [CrossRef]
23. Selbst, A.; Powles, J. "Meaningful Information" and the Right to Explanation. In Proceedings of the Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23 February 2018.
24. Wachter, S.; Mittelstadt, B.; Floridi, L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *Int. Data Privacy Law* **2017**, *7*, 76–99. [CrossRef]
25. Hacker, P.; Passoth, J.K. Varieties of AI Explanations under the Law. From the GDPR to the AIA, and Beyond. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3911324 (accessed on 9 January 2022).
26. Ebers, M. Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s). In *Nordic Yearbook of Law and Informatics*; Colonna, L., Greenstein, S., Eds.; Poseidon Förlag: Stockholm, Sweden, 2020.
27. Veale, M.; Zuiderveen Borgesius, F. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Comput. Law Rev. Int.* **2021**, *22*, 97–112.
28. Lakkaraju, H.; Kamar, E.; Caruana, R.; Leskovec, J. Interpretable & explorable approximations of black box models. *arXiv* **2017**, arXiv:1707.01154.
29. Sovrano, F.; Vitali, F. An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability. *arXiv* **2021**, arXiv:2109.05327.
30. Villone, G.; Rizzo, L.; Longo, L. A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence. In Proceedings of the 28th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, 7 December 2020.
31. Nguyen, A.; Rodrıéguez Martıénez, M. On quantitative aspects of model interpretability. *arXiv* **2020**, arXiv:2007.07584.
32. Holzinger, A.; Carrington, A.; Müller, H. Measuring the quality of explanations: The system causability scale (SCS). *Künstl. Intell.* **2020**, *34*, 193–198. [CrossRef] [PubMed]

33. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for explainable AI: Challenges and prospects. *arXiv* **2018**, arXiv:1812.04608.
34. Keane, M.T.; Kenny, E.M.; Delaney, E.; Smyth, B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), Montreal, QC, Canada, 21 August 2021.