Legal Knowledge and Information Systems
E. Schweighofer (Ed.)

doi:10.3233/FAIA210342

A Survey on Methods and Metrics for the Assessment of Explainability Under the Proposed AI Act

Francesco SOVRANO [a] and Salvatore SAPIENZA [b] and Monica PALMIRANI [b] and Fabio VITALI [a]
[a] *University of Bologna, DISI*
[b] *University of Bologna, CIRSFID-ALMA AI*

Abstract. This study discusses the interplay between metrics used to measure the explainability of the AI systems and the proposed EU Artificial Intelligence Act. 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. This study identifies the requirements that such a metric should possess to ease compliance with the AI Act. It does so according to an interdisciplinary approach, i.e. by departing from the philosophical concept of explainability and discussing some metrics proposed by scholars and standardisation entities through the lenses of the explainability obligations set by the proposed AI Act. 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 & accessible*. This is why 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

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. 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. To date, standardisation entities have proposed white papers and preliminary documents showing their progress[1], among them we mention: the European Telecommunications Standards Institute (ETSI), the CEN-CENELEC, and ISO/IEC TR 24028:2020(E), stating that '[i]t 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; paras 9.3.6, 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

[1] An extensive list of examples is available at https://joinup.ec.europa.eu/collection/rolling-plan-ict-standardisation/artificial-intelligence

we believe it may foster future research to embrace an interdisciplinary approach less timidly, for the sake of a better conformity to existing (and new) regulations in the EU landscape.

This paper is structured as follows. In Section 2 and 3 we present the research background and the methodology of this paper. Then in Section 4, 5 and 6, we explore the definitions and properties of explainability in philosophy and in the proposed AI Act. Finally, in Section 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's literature there are many interesting surveys on explainability techniques [1,2,3,4], 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.

For example [1] classify XAI methods with respect to the notion of explanation and the type of black-box system. The identified characteristics are respectively 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. Similarly to [1], also [2] study XAI considering interpretability and level-of-detail.

On the other hand, [4] focus 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, [4] narrow 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, [4] is certainly the closest to our work, in terms of focus of the survey. The main distinction between our work and [4] is probably the assumption we do that multiple definitions of explainability exist, each one possibly requiring its own type of metrics. Furthermore, differently from [4], we analyse explainability metrics on their ability to meet the requirements set by the AI Act.

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 [4], re-classifying them. Then we integrated with further works identified through an in depth keyword-based research[2] on Google Scholars, Scopus, and Web Of Science. On the other hand, the 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 has been preferred to 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.

[2]The main keywords we used were "degree of explainability", "explainability metrics", "explainability measures", and "evaluation metrics for contrastive explanations".

Table 1. Definitions of *explanation* and *explainable information* for each theory of explanations.

Theory	Def. of Explanation	Def. of Explainable Information
Causal Realism [5]	It is a description of causality, as chains of causes and effects.	It can fully describe causality.
Constructive Empiricism [6]	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 [7]	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 [8]	Explaining is a process triggered as response to predictive failures and it is about providing information to fix that failures in a mental model (sometimes intended as a hierarchy of rules).	It can fix failures in mental models.
Naturalism and Scientific Realism [9]	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.

In 1948 Hempel and Oppenheim published their "Studies in the Logic of Explanation" [10], giving birth to what it is considered the first theory of explanation, the deductive-nomological model. After that date, many attempts followed to amend, extend or replace this first model, which is considered fatally flawed [11,5]. This gave birth to several competing and more contemporary theories of explanations [12]: i) Causal Realism, ii) Constructive Empiricism, iii) Ordinary Language Philosophy, iv) Cognitive Science, v) Naturalism and Scientific Realism. A summary of these definitions is shown in Table 1.

Interestingly, each one of these theories devises different definitions of "explanation". If we look at their specific characteristics we may find that all but *Causal Realism* are pragmatic. On the other hand, *Causal Realism* and *Constructive Empiricism* are rooted on causality, while the others not [3]. Nonetheless, *Cognitive Science* and *Scientific Realism* are more focused on the effects that an explanation has on the explainee (the recipient of the explanation).

Importantly, with the present letter, 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

In philosophy, the most important work about the central criteria of adequacy of *explainable information* is likely to be Carnap's [13]. 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 made explainable.

Carnap's central criteria of explication adequacy are [13]: *similarity*, *exactness* and *fruitfulness*[4]. *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,

[3] They study the act of explaining as an iterative process involving broader forms of question answering

[4] Carnap also discussed another desideratum, *simplicity*, but this criterion is presented as being subordinate to the others.

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).

Carnap's 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 much *similar* the given information is to the explanandum. This can be estimated by counting the number of *relevant aspects* covered by information and the *amount 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 [14].
- *Fruitfulness* is about measuring how much a given piece of information is going to be used in the generation of explanations. Consequently, each one of the explainability definitions may define *fruitfulness* differently.

Importantly, 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, but different, as discussed also by [15]. 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.

6. Explainability Obligations in the Proposed AI Act

The discussion towards "explainability and law" has departed from the contested existence of a right to explanation in the General Data Protection Regulation (GDPR) [16,17,18][5] to embrace contract, tort, banking law [19], and judicial proceedings [20]. This previous discussion focusing on legal regimes other than the AIA - yet, highly connected - constitutes a valuable background for our research. Our focus, however, shall be confined to the interaction between the nuance of explainability and obligations emerging from the Artificial Intelligence Act (AIA) already identified by these early commentators. 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 comply with certain legal regimes (tort law and contractual obligations). 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 will 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 adequacy to the many obligations set by the AIA.

[5]Explanations, including contractual ones [17] are deemed to be 'right-enabling' [19] as they is necessary and instrumental to exercises the rights enshrined in Article 22 of the GDPR, 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 [20]. Indeed, Case law on explanations is progressively becoming significant: scholars have referred to the Risk Indication System (*SyRI*) case decided by The Hague District Court in 2020 [20] on the transparency in fraud prevention systems, Case n. 8472/2019 by the Italian Consiglio di Stato concerning the allocation of teachers in public schools across the country, and the German Federal Court for Private Law BGH, Case VI ZR 156/13 = MMR 2014 on the right to access to personal data [19]

Recital 47 and art. 13(1) state that high-risk AI systems 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* (emphasis added) with the relevant obligations of the user and of the provider [...]". In our reading, this provision specifies that this explainability obligations (i.e. transparent design and development of high-risk AI systems) is *compliance-oriented*. [6].

Such compliance-oriented explainability becomes evident in the technical documentation to be provided according to Art. 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 (Chapter IV). The contents of the dossier are those detailed by Annex IV. *Inter alia*, Annex IV(2)(b) include "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. 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).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-minimisation.

Once clarified the existence of explainability obligations and their extent, 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. Therefore, metrics become useful in the course of the standardisation process: i) *ex ante*, when defining the explainability measures adopted by the standard; ii) *ex post*, when verifying in practice the adoption of a standard.

From these premises it follows that, in the light of the purposes of the AIA, any explainability metric should be at minimum: i) *Risk-focused*, ii) *Model-agnostic*, iii) *Goal-Aware*, iv) and *Intelligible & accessible*.

Risk-focused means that the metric should be functional to measure the extent to which the explanations provided by the system allows 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-enabling (e.g. art. 29) and compliance-oriented (Annex IV) explainability. While *Model-agnostic* means that the metric should be appropriate to all the AI systems regulated by the AIA[7].

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.)[8] and applicable in all the high-risk AI applications listed in Annex III. While *Intel-*

[6]The twofold 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 revers the output", art. 14(4)(d)

[7]Annex I provides a list of the AI techniques and approaches that fall within the remit of the Regulation.

[8]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.

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 the best values.

Metric	Information Format	Supporting Theory	Subject - based	Covered Criteria	Name
[21]	Rule-based	Causal Realism	**No**	Similarity, Fruitfulness	Fidelity, Completeness
[22]	Feature Attribution	Causal Realism	**No**	Similarity, Fruitfulness	Monotonicity, Non-sensitivity, Effective Complexity
[23]	Rule-based	Causal Realism	**No**	**Similarity, Exactness, Fruitfulness**	Fidelity, Unambiguity, Interpretability, Interactivity
[24]	**All**	Causal Realism, Cognitive Science, Scientific Realism	Yes	Exactness, Fruitfulness	Causability
[25]	**All**	Cognitive Science, Scientific Realism	Yes	Exactness, Fruitfulness	Satisfaction, Trust, Mental Models, Curiosity, Performance
[26]	Example-based	Constructive Empiricism	**No**	Exactness	Proximity, Sparsity, Adequacy (Coverage)
[22]	Example-based	Constructive Empiricism	**No**	Similarity, Fruitfulness	Non-Representativeness, Diversity
[27]	Natural Language Text	Ordinary Language	**No**	**Similarity, Exactness, Fruitfulness**	Aspects Coverage, Degree of Explainability

ligible & accessible means that if information on the metrics is not accessible (e.g. due to intellectual property reasons) or the results of a metric are not reproducible (e.g. due to a subjective evaluation), explainees will confront with a situation of uncertainty, as an *ignotum per ignotius*. This would contradict the risk minimisation principle.

7. Discussing Existing Quantitative Measures of Explainability

In this section we identify some pros and cons of existing metrics (and measures) to quantitatively estimate the degree of explainability of information, with the aim of understanding their range of applicability across different needs and interpretations of explainability. We do it by performing a qualitative classification of these measures based on Carnap's desiderata, the theories of explanation presented in Section 4 and the main principles identified in Section 6.

More precisely, in Table 2 we classified the metrics on the following dimensions: the *format of information* supported by the metric (i.e. rule-based, example-based, natural language text, etc.); the *supporting theory of the metric* (i.e. cognitive science, constructive empiricism, etc.); *subjectivity* (whether the metric requires evaluations given by humans subjects); the *covered criteria of adequacy*. Then, in Table 3 we aligned the *supporting theories* (hence also the metrics) to the properties identified with the analysis of the AI Act carried out in Section 6.

Doing so, we considered only a part of the dimensions adopted by [4]. More precisely, we kept *clarity*, *broadness* and *completeness*, aligning the first two to Carnap's *exactness* and the latter to *similarity*. In fact, we deemed *soundness* to be as *truthfulness*, a complementary characteristic to explainability and not a characteristic of explainability, as discussed in Section 5. While *broadness* and *parsimony* were 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, so we selected methods possibly

Table 3. Explainability definitions alignment to the properties identified in Section 6.

	Risk-Focused	Model-Agnostic	Goal-Aware	Intelligible & 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 and hard to reproduce, this makes them less accessible
Naturalism and Scientific Realism	Yes, it can be	Yes, the evaluation is subject-based	Yes	Unlikely. It relies on (usually) expensive subject-based metrics

applicable also on ex-ante or more generic types of explanations.

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. What is common to all the metrics based on Cognitive Science is that they require humans subjects for performing the measurement, therefore they tend to be more expensive than the others, at least in terms of human effort. Furthermore, the metrics proposing heuristics to measure all Carnap's desiderata are just two, one for Causal Realism [23] and the other for Ordinary Language Philosophy [27]. Interestingly, [23] evaluates the three desiderata separately, while [27] propose a single metric combining all of them.

Finally, the results shown in Table 3 indicate that the metrics supported by both Causal Realism and Constructive Empiricism might struggle at being model-agnostic and goal-aware, this probably limits their applicability to very specific contexts.

8. Final Remarks

With this work we proposed an interdisciplinary analysis of explainability metrics in Artificial Intelligence. More specifically, through the lens of the obligations enshrined by the proposed Act, we identified that explainability metrics should be *risk-focused*, *model-agnostic*, *goal-aware*, *intelligible & accessible*. We found that these characteristics pose some constraints on the scope of explainability metrics, suggesting that different metrics may be complementary, serving different roles, depending on the context. In fact, as shown in Table 3, while the majority of *supporting theories* have the potential to result in *risk-focused* metrics, some of them might have important issues with *goal-awareness*, *intelligibility* and *accessibility*.

Nonetheless, our analysis of these metrics was qualitative and not quantitative. In fact, all of the considered metrics were tested by their authors on very specific applications and technologies, raising the issue of whether they can be seemingly effective under different implementation scenarios. Hence, we envisage that a more quantitative analysis should be carried on, perhaps by defining a proper benchmark on which metrics can be thoroughly evaluated from a legal perspective.

Therefore, we believe that more academic contributions and new benchmarks for quantitative legal analysis are needed, to better understand the pros and cons of existing technologies, for any standardisation process to be finalised and effectively deployed in the EU panorama. For example, 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 at least to consolidate the interpretation of the Act in the light of its future changes.

References

[1] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys (CSUR). 2018;51(5):1-42.
[2] Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access. 2018;6:52138-60.
[3] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115.
[4] Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics. 2021;10(5):593.
[5] Salmon WC. Scientific explanation and the causal structure of the world. Princeton University Press; 1984.
[6] Van Fraassen BC, et al. The scientific image. Oxford University Press; 1980.
[7] Achinstein P. The Nature of Explanation. Oxford University Press; 1983.
[8] Holland JH, Holyoak KJ, Nisbett RE, Thagard PR. Induction: Processes of Inference, Learning, and Discovery. Bradford books. MIT Press; 1989.
[9] Sellars WS. Philosophy and the Scientific Image of Man. In: Colodny R, editor. Science, Perception, and Reality. Humanities Press/Ridgeview; 1962. p. 35-78.
[10] Hempel CG, Oppenheim P. Studies in the Logic of Explanation. Philosophy of science. 1948;15(2):135-75.
[11] Bromberger S. Why-questions. na; 1966.
[12] Mayes GR. Theories of Explanation; 2001. Available from: https://iep.utm.edu/explanat/.
[13] Leitgeb H, Carus A. Rudolf Carnap; 2021. Available from: https://plato.stanford.edu/archives/sum2021/entries/carnap/.
[14] Brun G. Explication as a method of conceptual re-engineering. Erkenntnis. 2016;81(6):1211-41.
[15] Hilton DJ. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. Thinking & Reasoning. 1996;2(4):273-308.
[16] Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law. 2017;7(2):76-99.
[17] Wachter S, Mittelstadt B, Russell C. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. Harv JL & Tech. 2017;31:841.
[18] Selbst A, Powles J. "Meaningful Information" and the Right to Explanation. In: Conference on Fairness, Accountability and Transparency. PMLR; 2018. p. 48-8.
[19] Hacker P, Passoth JH. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. From the GDPR to the AIA, and Beyond (August 25, 2021). 2021.
[20] Ebers M. Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework (s). An Overview of the Current Legal Framework (s)(August 9, 2021) Liane Colonna/Stanley Greenstein (eds), Nordic Yearbook of Law and Informatics. 2020.
[21] Villone G, Rizzo L, Longo L. A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence; 2020. .
[22] Nguyen Ap, Martínez MR. On quantitative aspects of model interpretability. arXiv preprint arXiv:200707584. 2020.
[23] Lakkaraju H, Kamar E, Caruana R, Leskovec J. Interpretable & explorable approximations of black box models. arXiv preprint arXiv:170701154. 2017.
[24] Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the system causability scale (SCS). KI-Künstliche Intelligenz. 2020:1-6.
[25] Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for explainable AI: Challenges and prospects; 2018. Available from: https://arxiv.org/abs/1812.04608.
[26] Keane MT, Kenny EM, Delaney E, Smyth B. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. arXiv preprint arXiv:210301035. 2021.
[27] Sovrano F, Vitali F. An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability. arXiv preprint arXiv:210905327. 2021. Available from: https://arxiv.org/abs/2109.05327.