Perlocution vs Illocution: How Different Interpretations of the Act of Explaining Impact on the Evaluation of Explanations and XAI

Perlocution vs Illocution: How Different Interpretations of the Act of Explaining Impact on the Evaluation of Explanations and XAI

Francesco Sovrano[1,2][0000−0002−6285−1041] and Fabio Vitali[1][0000−0002−7562−5203]

[1] Department of Computer Science, University of Bologna, Italy
[2] Department of Informatics, University of Zurich, Switzerland

Abstract. This article discusses the concepts of illocutionary, perlocutionary, and locutionary acts, and their role in understanding explanations. Illocutionary acts concern the speaker's intended meaning, perlocutionary acts refer to the listener's reaction, and locutionary acts are about the speech act itself. We suggest a new way to categorise established definitions of explanation based on these speech act principles. This method enhances our grasp of how explanations work. We found that if you define explanation as a perlocutionary act, it requires subjective judgements. This makes it hard to assess an explanation objectively before the listener receives it. On the other hand, we claim that existing legal systems prefer explanations based on illocutionary acts. We propose that the exact meaning of explanation depends on the situation. Some kinds of definitions suit specific circumstances better. For example, in educational settings, a perlocutionary approach often works best, while legal settings call for an illocutionary approach. Additionally, we show how current measures of explainability can be grouped based on their theoretical support and the speech act they rely on. This categorisation helps us pinpoint which measures are best for assessing the results of Explainable AI (XAI) tools in legal or other settings. In simpler terms, we are explaining how to evaluate and improve XAI and explanations in different situations, such as in education and law. By understanding where and when to apply different explainability measures, we can make better and more specialised XAI tools. This will lead to significant improvements in AI explainability.

Keywords: Theories of explanation · Speech act theory · Evaluation of explanations · Explanations in law.

1 Introduction

Explanations are essential in various fields, such as philosophy, cognitive science, Artificial Intelligence (AI), and law, as they facilitate understanding and promote effective communication. However, the absence of a universally agreed-upon

definition of explanations poses a significant challenge [50]. One field that highlights the importance of explanations is Explainable Artificial Intelligence (XAI), which focuses on developing explainable AI systems. In XAI, explanations often focus on causality [6,64,3], providing insight into how the AI system arrived at a particular decision or recommendation. The goal is to improve transparency and trust in AI systems, making them more accessible to users.

On the other hand, cognitive science employs explanations not just to address perceived gaps in mental models [27], but to better understand and analyse the underlying cognitive processes that enable individuals to form and manipulate their mental models of the world [29]. These mental models, representations of external reality, play a significant role in cognition, reasoning, and decision-making[3] [16,17]. In contrast, legal contexts prioritise providing sufficient information while adhering to regulations. In these contexts, explanations aim to meet legal requirements while ensuring that the information is properly structured, accessible and understandable to all parties involved [7,21,50].

The lack of a universally accepted definition of explanations raises questions about the applicability of a single definition to various domains [12,38]. Different fields have varying perspectives on explanations, with each field emphasising different aspects of explanations. As such, it may be necessary to adopt distinct definitions of explanations in different contexts, but without a clear understanding of which definition may apply, it is difficult to practically assess the quality of XAI and explanations.

In this paper, we suggest that understanding the distinctions between illocutionary, perlocutionary, and locutionary acts is crucial for understanding explanations and how to evaluate them in different contexts (e.g., education, law).

These speech acts are key in evaluating explanations and influence how explanations are generated, received, and assessed by different audiences. So, by examining the relationship between speech act principles and various types of explanations, we aim to provide a more comprehensive understanding of the explanation process in different contexts, especially in XAI.

In particular, our paper discusses how current explainability measures can be categorised based on the speech act they rely on. This categorisation helps us identify which measures are best suited for evaluating the results of XAI tools and explanations in various settings, such as legal and educational contexts. We show that certain explainability metrics are more suitable than others when it comes to evaluating the legal compliance of XAI outputs. Eventually, by understanding where and when to apply different explainability measures, one can create more effective and specialised XAI tools.

In summary, this paper contributes valuable insights into the nature and evaluation of explanations, also in the XAI field. It provides a novel perspective on the nature of explanations, discusses practical tools for evaluating and comparing different kinds of explanations, and underscores the role of context

[3] The exact mechanisms and implications of mental models in various aspects of cognition and decision-making are still areas of active investigation.

in shaping explanation definitions. Ultimately, the findings and contributions of this study will have practical implications for researchers, practitioners, and decision-makers in the wider XAI community.

This paper is organised as follows. First, in Section 2, previous research on explanations is reviewed. Next, Section 3 provides background information on speech act theory and explains the concepts of illocutionary, perlocutionary, and locutionary acts. It also provides insights into explanations in different contexts, such as European law. Section 4 presents a classification of established explanation theories based on illocution, perlocution, and locution. The section highlights the unique characteristics of each category. In Section 5, the context-dependent nature of explanation definitions is discussed, emphasising the importance of adaptable approaches in different settings. Section 6 explores the impact of speech acts on the evaluation of explanations. The section discusses how illocutionary, perlocutionary, and locutionary acts can affect the assessment process. To discuss and exemplify the analytical results presented in the previous sections, Section 7 delves into the findings and observations derived from the study of two real-world XAI systems: a heart disease predictor and a credit approval system. These systems are examined using various explanation evaluation metrics and analysed through speech acts. Finally, Section 8 summarises the main findings and suggests possible avenues for future research and improvements in the understanding and development of explanations across diverse domains.

2 Related Work

The concept of speech acts, which originated from Austin's work [5] and was further developed by Searle [48], has significantly influenced the understanding of explanations and explainability across various fields. Achinstein's Ordinary Language Philosophy [1] applies speech act theory to philosophical theories of explanation, emphasising explanations as illocutionary acts that concentrate on communicative and linguistic aspects. Our work not only takes into account the illocutionary perspective, but also extends this view by comparing and contrasting it with perlocutionary acts, analysing their implications for the evaluation of explanations.

The literature contains numerous categorisations of explanations, including those based on the mechanisms to achieve explainability, as presented in [20,2,4,66,50]. These surveys explore different aspects of explainability techniques; some focus on the notion of explanation and the type of black-box system, like [20], while others, such as [66], centre on metrics to quantify the quality of explanation methods. Although these classifications provide valuable insights, our work offers a unique perspective by examining explanations through the lens of speech acts. Specifically, we focus on the distinctions between illocutionary and perlocutionary acts and their implications for the evaluation of explanations.

Our work builds upon, extends, and aligns closely with [50] in terms of its focus on the notion of explainability. Like their approach, we recognise the existence of multiple definitions of explainability, each potentially requiring its

own unique set of metrics. This acknowledgement of diversity in explainability concepts enhances our understanding of the various aspects involved in the evaluation and interpretation of explanations. However, unlike [50], we further explore the application of speech acts in the context of explanations, concentrating on the distinctions between illocutionary and perlocutionary acts and their relationship to the evaluation of explanations, suggesting that the nature and definition of explanations may be context-dependent.

Our paper emphasises the importance of context and the distinction between perlocutionary and illocutionary acts, arguing that different situations necessitate different types of explanations. This approach aligns with the pluralist perspective presented in [12], which supports the idea of multiple concepts of explanation. While our paper categorises existing explanation definitions using speech act concepts and focuses on the implications of these distinctions for various domains, Colombo [12] delves into the philosophy and psychology of explanation, drawing on experimental philosophy and advocating for pluralism based on results from psychology and philosophy.

In summary, our work sets itself apart from related research by offering a comprehensive analysis of speech acts in the context of explanations and explainability, particularly focusing on the comparison of illocutionary and perlocutionary acts. We investigate their relationships with various philosophical theories of explanation and present a novel perspective that can inform the design and evaluation of explainable AI systems.

3 Background

The study of explanations spans various disciplines, including philosophy, cognitive science, artificial intelligence, and law. However, there is no consensus on a shared definition of explanation, with different fields emphasising different aspects. To better understand explanations, it is essential to explore the perspectives on the nature of explanations across different domains. This section provides background information on speech act theory, European law, and other contexts where explanations are commonly used.

3.1 Speech Act Theory: Illocution, Perlocution, and Locution

Speech Act Theory is a branch of pragmatics that aims to explain how people use language to perform actions in the world. According to this theory, when people use language, they not only convey information, but also perform actions such as making requests, giving orders, or making promises. These actions are known as speech acts, and they can have different effects on the listener depending on the context and the speaker's intention.

In the 1950s and 1960s, philosophers J.L. Austin [5] and John Searle [48] developed Speech Act Theory as a way to understand the complex nature of language use. Their work highlights the importance of considering the intention

and context behind language use, rather than simply focusing on the literal meaning of words.

Speech Act Theory is essential for understanding communication because it provides a framework for analysing the various ways in which language is used to perform actions in the world. By considering the speaker's intention, the listener's interpretation, and the context in which a speech act is performed, Speech Act Theory enables a more nuanced understanding of communication. This deeper understanding can lead to better comprehension of how language is used in various domains, such as law, education, and XAI. Illocutionary, perlocutionary, and locutionary acts are key components of speech acts:

- **Illocutionary acts** refer to the speaker's intended meaning or purpose behind a speech act. For example, when someone says 'I promise to be there at 7 pm', they are performing the illocutionary act of making a promise.
- **Perlocutionary acts** pertain to the effect that a speech act has on the listener. Perlocutionary acts are the actual effects of a speech act, which can vary depending on the listener's interpretation, expectations, and context. For instance, if the listener believes the speaker is not trustworthy, the perlocutionary effect of the promise may be scepticism or doubt.
- **Locutionary acts** involve the actual words and sentences used in a speech act. For example, the locutionary act of the sentence 'I promise to be there at 7 pm' is the act of producing that particular string of words.

In everyday communication, one can easily find examples of illocutionary, perlocutionary, and locutionary acts. For instance, consider the scenario where a teacher says, 'Please open your textbooks to page 42'. Here, the illocutionary act is the teacher's intention to request that the students turn to a specific page in their textbooks. This act can differ from the literal meaning of the words spoken. Other examples of illocutionary acts include utterances such as:

- 'Can you please close the window?' (request),
- 'I would like to buy this dress' (proposal),
- and 'I forbid you to leave the house' (prohibition).

On the other hand, the perlocutionary act occurs when the students respond to the teacher's request by opening their textbooks to page 42. Examples of perlocutionary acts include expressions such as:

- 'Thank you for your help' (the listener feels appreciated),
- 'You're fired' (the listener feels upset),
- and 'I'm sorry for your loss' (the listener feels comforted).

Finally, the locutionary act is the teacher's actual utterance of the sentence, 'Please open your textbooks to page 42'. Other examples of locutionary acts include phrases like:

- 'The sky is blue',
- 'I am hungry',

– and 'The book is on the table'.

Understanding the differences between these three types of acts can help us gain a deeper understanding of language and its roles in communication. By analysing and interpreting how language is used to convey meaning and perform actions, we can refine our comprehension of communication across various domains and situations.

3.2 Explaining Automated Decision-Making: Regulatory Landscape and Challenges

Automated decision-making systems have become increasingly common, impacting various aspects of our lives. As a result, regulatory frameworks have emerged to ensure transparency, fairness, and accountability in these systems. This section discusses the right to explanation in the General Data Protection Regulation (GDPR)[4], the proposed AI Act[5], and the Platform-to-Business (P2B) Regulation[6], focusing on the challenges and opportunities they present for explaining automated decision-making systems.

The Right to Explanation in the GDPR. The GDPR introduced the *right to explanation*, allowing individuals to obtain explanations when their legal status is affected by a solely automated decision-making process. The GDPR outlines two types of explanations: *ex-ante* and *ex-post*. Minimal explanations required under the GDPR include causal, descriptive, and justificatory explanations [56]. While it is unclear whether user-centred personalized explanations are required, the GDPR mandates that data controllers provide 'meaningful information about the logic involved' in an automated decision. The Think Tank of the European Parliament suggests that a reasonable explanation should possess various qualities, including intelligibility and understandability [13].

The Proposed AI Act and its Role in Explainability. The proposed AI Act seeks to address the risks posed by AI systems by setting new obligations to ensure transparency, lawfulness, and fairness for high-risk AI systems listed in Annex IV. It aims to establish mechanisms to ensure quality throughout the AI system's life cycle while preserving fundamental rights and values. The AI Act promotes user-empowering and compliance-oriented explainability, enabling users to understand the AI system's operation and helping verify compliance with the many obligations set by the AI Act [51].

According to [51], user-empowering explainability is related to human oversight design obligations, while compliance-oriented explainability is evident in the technical documentation required by Article 11. The AI Act aims to minimize potential harmfulness through these explainability measures.

[4] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679`
[5] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206`
[6] `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32019R1150`

Transparency and Fairness in the P2B Regulation. The aim of the P2B Regulation is to address imbalances between online platforms and businesses by ensuring transparency in their terms and conditions. According to Article 5 of the 2019/1150 P2B Regulation, online marketplaces are required to provide explanations about the main parameters that determine the ranking of products and the reasons for their relative importance [21].

To meet these requirements, online marketplaces must provide easily accessible, up-to-date explanations in plain and intelligible language. If the possibility exists for any direct or indirect remuneration to influence ranking, providers must describe those possibilities and their effects on ranking. The explanations should help business and corporate website users understand how the ranking mechanism takes into account the characteristics of goods and services, their relevance to consumers, and website design characteristics. However, providers are not required to disclose algorithms or information that could enable consumer harm or deception.

3.3 Explaining as Answering Questions in XAI and Computer Science

Computer science, through XAI, has long studied the topic of explanations and how to generate them (e.g., for explaining complex software computations, for law compliance), frequently drawing from philosophy and social sciences [39]. Two distinct types of explainability are predominant in the literature of XAI: rule-based and case-based. Rule-based explainability is when explainable information is a set of formal logical rules describing information related to cause and effects. For example, the inner logic of a model, its causal chain, how it behaves, why that output gave the input, and what would happen if the input were different. While case-based explainability is when the explainable information is a set of input-output examples (or counter-examples) meant to give an intuition of the model's behaviour. For example, counterfactuals, contrastive explanations, or prototypes[7].

The idea of answering questions as explaining is familiar to XAI and compatible with everyone's intuition of what constitutes an explanation. In fact, despite the different types of explainability one can choose, it is always possible to frame the information provided by explainability with one or (sometimes) more questions. In particular, it is common to many works in the field [44,34,39,19,14,62,43,30,37] the use of generic (e.g., `why`, `who`, `how`, `when`) or more punctual questions to clearly define and describe the characteristics of explainability [32]. For example, Lundberg et al. [36] assert that the local explanations produced by their TreeSHAP (a XAI algorithm for estimating the importance of features as input to an AI model) might enable 'agents to predict `why` the customer they are calling is likely to leave' or 'help human experts understand `why` the model made a specific recommendation for high-risk decisions'. Similarly,

[7] Prototypes are instances of the ground-truth considered similar to a specific input-output for which the similarity explains the model's behaviour.

Dhurandhar et al. [14] state that they designed the Contrastive Explanation Method (CEM; a XAI algorithm for the generation of counterfactuals and other contrastive explanations) to answer the question '`why` is input x classified in class y?'.

Several authors [34,39,19], analysing XAI literature, were able to hypothesise that a good explanation, about an automated decision-maker, answers at least the following questions:

- What did the system do?
- Why did the system do P?
- Why did the system not do X?
- What would the system do if Y happens? ,
- How can one get the system to do Z, given the current context?
- What information does the system contain?

In particular, from a preliminary analysis, it appears that most classical XAI algorithms focus more on the production of explainable software and explanations that generally follow a one-size-fits-all approach, answering one (or sometimes a few) predefined questions well. However, one-size-fits-all explanations tend to lack user-centrality, usually failing to answer all the questions an explainee might have. This is also suggested by Liao et al. [33], who show that no single XAI seems to be able to cover all identified user needs and that various XAI algorithms may be needed to explain a system better. Indeed, users' needs in terms of explainability are multiple and challenging to capture [32], e.g., they may concern terminology, system performance, system outputs, and inputs.

3.4 Explanations in Education and Pedagogy

Explanations play a crucial role in education and pedagogy, contributing to effective teaching and learning, pedagogical practice, and understanding human behaviour and mental processes. Designing intelligent agents capable of providing explanations to people can draw upon models of how humans explain decisions and behaviour to each other, as the primary function of explanation is to facilitate learning [35,65,39]. Heider [22] suggests that people seek explanations to enhance their understanding of someone or something, developing stable models for prediction and control. Research supports this hypothesis, indicating that people tend to inquire about events or observations they consider abnormal or unexpected from their perspective [25,24,39].

Incorporating the principles of learner-centred education (LCE), promoted by UNICEF[8], can help create effective explanations that facilitate learners' understanding of complex concepts, ideas, and theories [58,18]. These explanations must be well-crafted, engaging, and tailored to learners' understanding and learning objectives, as LCE requires a high level of active control from learners over the content and process of learning [47]. Pedagogical practice involves designing

[8] `https://www.unicef.org/esa/sites/unicef.org.esa/files/2019-08/ThinkPiece_9_LearnerCentredEducation.pdf`

and implementing learning experiences that enable the acquisition of knowledge, skills, and attitudes while considering learners' needs, capacities, and interests [8]. Aligning learning objectives and explanations with LCE principles enables educators to create more personalised and meaningful learning experiences, empowering learners to actively engage in their education.

Crafting effective explanations in education and pedagogy can be challenging due to the need to adapt explanations to the person's understanding and the concepts being explained. However, advances in technology, such as multimedia and interactive tools, offer opportunities for delivering engaging and interactive explanations. Interdisciplinary collaborations can lead to innovative approaches for explaining complex concepts, ideas, and theories. Additionally, effective explanations can foster critical thinking, creativity, and problem-solving skills, promoting deeper understanding and lifelong learning.

4 An Explained Classification of Theories of Explanations in Terms of Illocution, Perlocution, and Locution

In this section, we provide an overview of various theories of explanation and explainability, highlighting key philosophical perspectives that shape our understanding of these concepts. Understanding the nature of explanations and explainable information is crucial for measuring explainability effectively.

In 1984, Hempel and Oppenheim published their 'Studies in the Logic of Explanation' [23], giving rise to what is considered the first theory of explanation: the deductive-nomological model. After this work, many modified, extended, or replaced this model, which was considered fatally flawed [9,46]. Indeed, Hempel's epistemic theory of explanations is not empiricist: it is concerned (mistakenly) only with logical form, so an explanation can be such regardless of the actual processes and entities conceptually required to understand it. Several more modern and competing theories of explanation have been the result of this criticism [38]. For example, Salmon's realist theory [46], called Causal Realism, emphasises that actual processes and entities are conceptually necessary to understand precisely why an explanation works. Instead, the Constructive Empiricism of Van Fraassen [59] relies more on a Bayesian interpretation of probability, framing explanation as a creative process of building models that are likely true.

In contrast to these theoretical and primarily scientific approaches, other philosophers have favoured a theory of explanation that is more grounded in how people perform explanations [38]. For example, Achinstein's theory [1], based on Ordinary Language Philosophy, emphasises the communicative or linguistic aspect of an explanation and its usefulness in answering questions and fostering understanding between individuals. The theory of Holland et al. [27] instead, based on Cognitive Science, frames the process of explaining as a purely cognitive activity and explanations as a certain kind of mental representation. Conversely, Sellars [49] suggests a different way of thinking about the epistemic meaning of the explanatory act, making it more of a utilitarian process of constructing a coherent belief system.

In particular, Hempel's, Salmon's, and Van Fraassen's theories frame the act of explaining more as a *locutionary act* [5], whereby an explanation is such because it utters something about causality. Differently, Achinstein's theory explicitly frames explaining as an *illocutionary act* [5] so that an explanation is such because of the intention to explain. The theories of Holland and Sellars, on the other hand, frame explaining more as a *perlocutionary act* [5], thus with an explanation being such because of the effects it produces in the interlocutor. For more details about locution, illocution and perlocution read Section 3.1.

Table 1: **Philosophical definitions of explanation and explainable information.**[9] In this table, we summarise the definitions of *explanation* for each one of the identified theories of explanations. We also indicate which *speech act* they mostly refer to.

Theory	**Definition of Explanation**	**Speech Act**
Causal Realism	Descriptions of causality, expressed as chains of causes and effects.	Locution
Constructive Empiricism	Contrastive information that answers `why` questions, allowing one to calculate the probability of a particular event relative to a set of (possibly subjective) background assumptions.	Locution
Ordinary Language Philosophy	Answers to questions (not just why ones) given with the explicit intent of producing understanding in someone, i.e., the result of an illocutionary act.	Illocution
Cognitive Science	Mental representations resulting from a cognitive activity. They are information which fixes failures in someone's mental model.	Perlocution
Naturalism and Scientific Realism	Information which increases the coherence of someone's belief system, resulting from an iterative process of confirmation of truths aimed at improving understanding.	Perlocution

Thus, each of these theories devises different definitions of explanation and explainability, 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. Nevertheless, according to [38], fundamental disagreements on the nature of explanations are just of two types, metaphysical and meta-philosophical, and mainly unrelated to their *logical* and *cognitive* structure. This gives room to understandings of 'explanations' that may be complementary, some focusing more on cognition and others on logic.

[9] This table extends a similar one in [50].

We observe that when explaining is not considered a locutionary act and thus it must satisfy someone's needs, explainability differs from explaining. Indeed, pragmatically satisfying someone (e.g., user-centrality) is achieved when explanations are designed for a specific person or audience. This implies that the same explainable pieces of information can be presented and re-elaborated differently across different individuals as different explanations. The type and order of explainable information matter and directly impact the quality of the resulting explanations. In simpler terms, not every combination of explainable information qualifies as an explanation according to illocutionary and perlocutionary theories.

The distinction between illocutionary and perlocutionary explainability lies in the explainer's intent and its impact on the explainee. Illocutionary explanations, as in Ordinary Language Philosophy, focus on the explainer's intention to create understanding in someone. For instance, in Ordinary Language Philosophy, an 'explanation' is used to answer questions about pertinent aspects, intending (i.e., illocution) to generate understanding in someone. On the other hand, perlocutionary explanations, as in Cognitive Science, emphasise the actual effects explanations have on explainees. In short, illocutionary explanations aim for a perlocutionary effect, but the intention doesn't always guarantee the desired outcome. Although Ordinary Language Philosophy highlights user-centrality, it doesn't focus on the user as much as Cognitive Science or other theories that treat explanations as perlocutionary acts.

When designing explainable systems from a prescriptive perspective, it is important to align illocutionary and perlocutionary explainability more closely. Designers should create systems that not only intend to provide meaningful explanations but also effectively achieve the desired impact on users. However, it is crucial to recognise that the explainer's intention may not always lead to the desired effect on the explainee. By considering both illocutionary and perlocutionary aspects, AI researchers and practitioners can develop more adaptable and user-centred systems, addressing the diverse needs and expectations of users.

5 Exploring the Contextual Nature of Explanations

To fully grasp the nature of explanations, it is important to understand not only their philosophical underpinnings but also the influence of the different kinds of speech acts on them. The distinctions between these speech acts have a significant impact on the evaluation of explanations, as they determine what is considered a valid or effective explanation in various scenarios. In this section, we will explore the nuances of these speech acts in different contexts, such as legal, educational, and XAI settings, and highlight the importance of tailoring explanations to better align with the specific needs and requirements of the domain or audience.

As discussed in Section 3.2, explanations in the legal context have a more justificatory connotation [56,50]. They must be clear but do not necessarily need to be personalised for a child or different types of people (e.g., those who do

not speak the official language). The essential aspect is that the necessary information to understand (for instance) the logic of an AI model is present and expressed in a common language that is easily understandable according to the law [61].

Conversely, in the educational context (cf. Section 3.4), an explanation is considered valid when it has an effect on the person. Thus, an 'explanation' may be suitable for an adult but not for a 3-year-old child. Therefore, it is not possible to explain Einstein's general relativity to a 3-year-old child in the same way it would be done with an undergraduate physics student or a high school student.

In general, in XAI (cf. Section 3.3), an explanation is considered valid when it reveals something about a black-box. Although it may not be user-centred (and therefore not optimal), it remains an explanation because it clarifies an aspect (any) of the black box. Consequently, this type of explanation is accepted regardless of its effects on the person or its ability to answer as many (archetypal) questions as possible.

Thus, in the three different contexts mentioned above, three different speech acts and therefore different definitions of explanation are applied. For laws, an illocutionary sense is used; for education, a perlocutionary sense is applied, while for XAI, a locutionary sense is adopted.

However, this does not imply that explanations in the context of law or XAI cannot have a perlocutionary effect. For instance, one notable sub-field of XAI, called Human-Centred XAI [33], focuses on developing tools that offer customised explanations to meet users' requirements. Therefore, the distinction lies not in the absence of perlocutionary explanations in XAI, but rather in the minimum act required for information to qualify as an explanation. For XAI, a locutionary act is sufficient, while the law necessitates an illocutionary act, and education and similar domains call for perlocutionary acts.

We observe a complexity hierarchy among the different explanatory acts. The occurrence of a locutionary act is necessary for the performance of an illocutionary act, and both of these acts can be crucial for the realisation of a perlocutionary act. In particular, the locutionary act can be better associated with the production of explainable information, while the other two acts can be better associated with the production of explanations. In the case of illocutionary acts, these explanations should aim to be useful for a majority of people by addressing a wide range of questions (cf. Section 4). On the other hand, perlocutionary acts should strive to be effective for a specific individual, tailored to their needs.

Based on the classification reported in Section 4, we find that different definitions of explanations are applicable to different contexts. Indeed, the definitions from Cognitive Science and Naturalism seem to align better with education and pedagogy, as they are perlocutionary. In contrast, the definition taken from Ordinary Language Philosophy is illocutionary and aligns better with the law (as suggested also by [50]). Meanwhile, the definitions from the remaining theories align more with most XAI practices, which often involve designing systems that

provide understandable insights into their decision-making processes, answering well only one or few specific questions.

In essence, the definition of explanation depends on the context and the type of speech act that is deemed sufficient for a satisfactory explanation. Consequently, the evaluation criteria which can be used to assess the quality of explanations and XAI tools cannot be independent from the context, particularly in situations where adherence to the law is critical.

In the subsequent section, we will present a methodology to establish a connection between evaluation metrics for explanations and philosophical theories, and thereby, speech acts. Subsequently, in Section 7, we will engage in a comprehensive discussion of our findings and their implications for the evaluation of XAI.

6 Linking of Explanation Evaluation Metrics to Philosophical Theories

In this section, we explore the alignment of various explainability metrics with philosophical theories to provide a comprehensive understanding of their applicability across different needs and interpretations of explainability. We have reviewed the literature and selected a diverse range of metrics applicable to various types of explanations, including ex-ante and ex-post explanations. Our selection is based on their relevance, novelty, and applicability in the context of XAI.

We established a **method for correlating explanation metrics with the respective philosophical theories** that underpin them. This method hinges on the assumptions and viewpoints that guide the design and interpretation of each metric. Specifically, it involves the following steps:

1. **Determining the metric's goal**: First, we need to ascertain the objective of the metric. It could be to assess the quality of explanations, the fidelity of the model, the user satisfaction, or some other facet of explainability.
2. **Exploring the metric's methodology**: Next, we examine the strategy the metric adopts towards explainability. For instance, does it attempt to quantify causal relationships, or does it focus on user cognitive processes? Does it strive to build likely accurate models or emphasise the communicative aspects of the explanation?
3. **Associating the metric with the relevant philosophical theory**: Based on the metric's goal and methodology, we can then link it to the suitable philosophical theory. This includes:
   - Linking the metric to *Causal Realism* if it aims to quantify causal relationships.
   - Associating it with *Cognitive Science, Naturalism, and Scientific Realism* if it centres on cognitive processes and user understanding.
   - Relating it to *Constructive Empiricism* if it prioritises the development of likely accurate models.

– Connecting it to *Ordinary Language Philosophy* if it focuses on the linguistic or communicative facets of explanation.

Table 2 presents an overview of the metrics and their associated philosophical theories, which encompass Causal Realism, Cognitive Science, Naturalism, Scientific Realism, Constructive Empiricism, and Ordinary Language Philosophy. The sources of these metrics include recent papers that propose novel explainability metrics or evaluate existing ones. We discuss each metric in more detail below.

[45] introduce objective metrics to quantify explainability. It argues that many explanations are generated post-hoc, resulting in limited meaning due to their lack of transparency and fidelity. To address this issue, the paper proposes four metrics: Performance Difference between the explanation's logic and the agent's actual performance, Number of Rules outputted by the explanation, Number of Features used to generate the explanation, and Stability of the explanation. These metrics are grounded in Causal Realism as they focus on quantifying the causal relationships between variables to provide more meaningful explanations, emphasising the necessity of actual processes and entities for understanding.

[60] introduce a novel comparative approach to evaluate and compare rule sets produced by post-hoc rule extractors using six quantitative metrics. The goal is to identify superior methods capable of successfully modelling explainability. This work is connected to Causal Realism, as it assumes that an explanation can be evaluated in terms of how it discloses causal relationships between input and output, representing them as a set of inference rules that explain the underlying causal mechanisms of the system, taking into account the actual processes and entities required for understanding.

The metrics proposed by [31] for quantifying fidelity, ambiguity, and interpretability can be seen as aligned with Causal Realism because they are designed for evaluating the quality of compact decision sets that explain the black box model.

[28] propose the System Causability Scale, which combines aspects of causality, mental representations, and iterative understanding improvement. System Causability Scale is aligned to Causal Realism, Cognitive Science, and Naturalism and Scientific Realism theories. That is because this metric evaluates explanations in terms of their capacity to represent causal relationships, align with mental models, and facilitate the user's understanding of the underlying system, emphasising the importance of actual processes, entities, and cognitive aspects in creating meaningful explanations.

[26] propose Satisfaction, Trust, Mental Models, Curiosity, and Performance metrics, grounded in Cognitive Science, Naturalism and Scientific Realism theories. It focuses on the cognitive aspects of explanations and their impact on users' understanding and behaviour. These metrics assess how well explanations address user expectations, foster trust, align with mental models, stimulate curios-

[10] This table extends a similar one in [50].

Table 2: **Comparing Explainability Metrics**[10]. The column labelled 'Source' provides references to the papers, while the 'Metrics' column lists the names of the metrics mentioned in the papers. The 'Subject-based' column indicates whether the metrics require subjective feedback from human subjects. Bold elements denote the best values within each column. Additional information includes what explanations are considered by the metric (e.g., rules) and the *Supporting Theory* of the metric, which refers to the philosophical theory that underlies the metric (e.g., Cognitive Science, Constructive Empiricism).

Source	**Explanations are:**	**Closest Supporting Theory**	**Subject based**	**Metrics**
[45]	Rules	Causal Realism	**No**	Performance Difference, Number of Rules, Number of Features, Stability
[60]	Rules	Causal Realism	**No**	Fidelity, Completeness
[31]	Rules	Causal Realism	**No**	Fidelity, Unambiguity, Interpretability, Interactivity
[28]	**Any text or image**	Causal Realism, Cognitive Science, Naturalism & Co.	Yes	System Causability Scale
[26]	**Any text or image**	Cognitive Science, Naturalism & Co.	Yes	Satisfaction, Trust, Mental Models, Curiosity, Performance
[15,40] [63,57] [42,10]	**Any text or image**	Cognitive Science, Naturalism & Co.	Yes	Usability: Effectiveness, Efficiency, Satisfaction
[41]	Contrastive Examples	Constructive Empiricism	**No**	Non-Representativeness, Diversity
[55]	**Any Natural Language Text**	Ordinary Language Philosophy	**No**	Degree of Explainability

ity, and improve overall performance, taking into account the cognitive processes involved in understanding and explaining.

[15,40,63,57,42,10] propose usability metrics, such as Effectiveness, Efficiency, and Satisfaction connected to Cognitive Science, Naturalism and Scientific Realism theories. These metrics evaluate explanations from the perspective of users' cognitive processes and their overall experience with the system, assessing the quality of the explanations in terms of their utility, ease of use, and user satisfaction, emphasising the role of cognitive aspects in understanding and engaging with explanations.

[41] propose two metrics for example-based contrastive or counterfactual explanations: Non-Representativeness and Diversity. Non-Representativeness is a measure of fidelity of the explanation, and high non-representativeness can indicate factual inaccuracy. Diversity, on the other hand, is used to demonstrate the degree of integration of the explanation by measuring the spread of examples in the input space. These metrics align with Constructive Empiricism as they aim to quantify how likely is a contrastive explanation to explain something accurately. By monitoring the fidelity and diversity of the examples provided by the explanation method, they assess the quality of the explanation generated by the model in terms of its representativeness and variety.

[55,53] present a novel model-agnostic metric called Degree of Explainability, which objectively measures the explainability of correct information within complex systems. The metric is explicitly based on Achinstein's Theory of Explanations from Ordinary Language Philosophy, and it leverages deep language models for knowledge graph extraction and information retrieval.

Our analysis shows that most tools for evaluating XAI software align with causal realism and constructive empiricism, while tools for evaluating explanatory user interfaces are more aligned with the interpretation from Cognitive Science. In general, as shown in Table 2, theories such as Cognitive Science that define explanations as perlocutionary acts require an evaluation of the explanation that is inseparable from the user's opinion or subjective outcome, as the explanation is valid when it produces an effect on the user.

On the other hand, illocutionary definitions based on Ordinary Language Philosophy do not bind the explanation to the user's effects and therefore allow for a more objective evaluation in line with the law's requirements. Similarly, locutionary definitions do not require subjective evaluations but also do not require the explanatory tool (for example, an XAI) to answer as many questions as possible in a cohesive and coherent manner. It is sufficient that they correctly explain what is causing something, even if such explanation does not provide an in-depth understanding of the explanandum.

7 Findings and Discussion

This section synthesises the insights gathered in Sections 6 and 5, analysing their practical implications for the actual evaluation of XAI systems. Our focus is on two specific XAI systems: a heart disease predictor and a credit approval

system. These systems have been evaluated using various explainability metrics associated with distinct speech acts. The analytical tools introduced in this paper were instrumental in this analysis.

We dissect and discuss the results of three studies [52,54,55], centred on usability and XAI-specific metrics. The goal is to gain insights into the effectiveness and applicability of these metrics across various contexts.

The first XAI system, the heart disease predictor, is designed for healthcare applications. Its foundation is the XGBoost [11] and TreeSHAP [36] models, as detailed in Section 3.3. The second system, the credit approval system, employs a basic Artificial Neural Network in tandem with the Counterfactual Explanations Method (CEM) [14], also discussed in Section 3.3. Both systems fall under the category of *conventional XAI Explainers*. They generate explanations by supplementing the XAI output with comprehensive contextual information.

In two user studies [52,54], these conventional systems were compared with *Enhanced XAI Explainers* and *Interactive XAI Explainers*, verifying their *usability*, a measure influenced by Cognitive Science (cf. Section 6). The latter two types of explainers are able to provide expanded information about the XAI systems, with the Interactive Explainers specifically offering user-driven interactive explanations. Conversely, the third study [55] applied the *DoX* metric, based on Ordinary Language Philosophy (cf. Section 6), on all these XAI systems.

A fascinating pattern emerged when comparing usability metrics and DoX scores, aligning with our insights from Section 6: the tests showed that Interactive XAI Explainers excelled in usability over others, yet both the Interactive and the Enhanced XAI Explainers achieved identical DoX scores. This difference demonstrates a distinct divergence in focus: usability metrics measure user interaction and understanding (perlocution), while DoX measures information depth and completeness (illocution). This distinction aligns with the legal implications outlined by Wachter et al. [61], according to which the legal acceptability of an explanation depends more on the depth and breadth of information given, rather than on customisation or interactivity.

An additional observation arose in these tests, when it was found out that TreeSHAP, even when combined with XGBoost (which theoretically ensures accurate, high-quality explanations [36]), does not necessarily yield superior usability or DoX scores. This emphasises the argument that good explanations from a locutionary standpoint (i.e., those generated by TreeSHAP) may not meet illocutionary or perlocutionary goals, as demonstrated by the lower-scoring conventional XAI Explainers using TreeSHAP.

In this regard, TreeSHAP explanations, despite their mathematical rigour, do not solely determine effectiveness in explanations. This underscores that locutionary metrics may not be enough in all those contexts where effective explanations need to meet illocutionary or perlocutionary goals, not just locutionary ones. For example, usability might be crucial in user-centred environments such as education, while DoX scores might provide valuable insights into the illocutionary force of explanations in legal contexts.

In summary, the case studies highlight the usefulness of contemplating philosophical theories when selecting and applying explanation evaluation metrics. As suggested in Sections 6 and 5, this approach can indeed augment our comprehension and practice of XAI.

8 Conclusions and Future Work

Throughout this paper, we have examined the distinctions between perlocutionary, illocutionary, and locutionary acts in the context of explanations and their implications for the evaluation of explanations. We have argued that the evaluation of explanations should be context-dependent and tailored to the specific needs of different situations and users, and our user testings [52,54,55] have indeed essentially confirmed our arguments.

In educational settings, a perlocutionary understanding of explanations may be more appropriate, as the primary goal is to facilitate understanding and learning for the individual. This necessitates considering the listener's background, prior knowledge, and cognitive abilities when crafting explanations.

Conversely, in other contexts such as the legal one, an illocutionary understanding of explanations is typically more desirable. In these settings, it is crucial to establish facts and determine the speaker's intended meaning while maintaining a certain level of objectivity. The focus is less on the listener's subjective understanding and more on the overall consistency and clarity of the explanation itself.

By acknowledging the distinctions between perlocutionary, illocutionary, and locutionary acts in the context of explanations, we can develop more effective strategies for evaluating and generating explanations in various domains. This understanding can be applied to enhance the quality of explanations in educational materials, legal documents, and scientific research.

Moreover, this contextual approach to explanations also has implications for the development of artificial intelligence and natural language processing systems. As these systems increasingly undertake tasks that involve generating explanations, comprehending the nuances of perlocutionary and illocutionary acts will be essential in creating more human-like and effective explanations tailored to specific situations and users.

For future research, we intend to explore the practical applications of our findings in various domains and investigate methods for systematically adapting explanations based on the context and the needs of the listener. This will enable the development of more effective and contextually appropriate explanations, ultimately benefiting a wide range of fields, from education and law to artificial intelligence and natural language processing.

References

1. Achinstein, P.: The Nature of Explanation. Oxford University Press (1983), https://books.google.it/books?id=0XI8DwAAQBAJ
2. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (9 2018). https://doi.org/10.1109/ACCESS.2018.2870052
3. Antoniadi, A.M., Du, Y., Guendouz, Y., Wei, L., Mazo, C., Becker, B.A., Mooney, C.: Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. Applied Sciences **11**(11) (2021). https://doi.org/10.3390/app11115088, https://www.mdpi.com/2076-3417/11/11/5088
4. Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012, https://doi.org/10.1016/j.inffus.2019.12.012
5. Austin, J., Urmson, J., Sbisà, M.: How to Do Things with Words. William James lectures, Clarendon Press (1975), https://books.google.it/books?id=XnRkQSTUpmgC
6. Beckers, S.: Causal explanations and XAI. In: Schölkopf, B., Uhler, C., Zhang, K. (eds.) 1st Conference on Causal Learning and Reasoning, CLeaR 2022, Sequoia Conference Center, Eureka, CA, USA, 11-13 April, 2022. Proceedings of Machine Learning Research, vol. 177, pp. 90–109. PMLR (2022), https://proceedings.mlr.press/v177/beckers22a.html
7. Bibal, A., Lognoul, M., de Streel, A., Frénay, B.: Legal requirements on explainability in machine learning. Artif. Intell. Law **29**(2), 149–169 (2021). https://doi.org/10.1007/s10506-020-09270-4, https://doi.org/10.1007/s10506-020-09270-4
8. Brandes, D., Ginnis, P.: A Guide to Student-centred Learning. Stanley Thornes (1996), https://books.google.ch/books?id=MTJSGGTAN3MC
9. Bromberger, S.: Why-questions. In: Colodny, R.G. (ed.) Mind and Cosmos – Essays in Contemporary Science and Philosophy, pp. 86–111. University of Pittsburgh Press (1966)
10. Buçinca, Z., Lin, P., Gajos, K.Z., Glassman, E.L.: Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In: Paternò, F., Oliver, N., Conati, C., Spano, L.D., Tintarev, N. (eds.) IUI '20: 25th International Conference on Intelligent User Interfaces, Cagliari, Italy, March 17-20, 2020. pp. 454–464. ACM (2020). https://doi.org/10.1145/3377325.3377498, https://doi.org/10.1145/3377325.3377498
11. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 785–794. ACM (2016). https://doi.org/10.1145/2939672.2939785, https://doi.org/10.1145/2939672.2939785
12. Colombo, M.: Experimental philosophy of explanation rising: The case for a plurality of concepts of *Explanation*. Cogn. Sci. **41**(2), 503–517 (2017). https://doi.org/10.1111/cogs.12340, https://doi.org/10.1111/cogs.12340

13. DG, E.: Understanding algorithmic decision-making: Opportunities and challenges (2019), https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624261
14. Dhurandhar, A., Chen, P., Luss, R., Tu, C., Ting, P., Shanmugam, K., Das, P.: Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 590–601 (2018), https://proceedings.neurips.cc/paper/2018/hash/c5ff2543b53f4cc0ad3819a36752467b-Abstract.html
15. Dieber, J., Kirrane, S.: A novel model usability evaluation framework (muse) for explainable artificial intelligence. Inf. Fusion **81**, 143–153 (2022). https://doi.org/10.1016/j.inffus.2021.11.017, https://doi.org/10.1016/j.inffus.2021.11.017
16. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Hum. Factors **37**(1), 32–64 (1995). https://doi.org/10.1518/001872095779049543, https://doi.org/10.1518/001872095779049543
17. Gary, M.S., Wood, R.E.: Mental models, decision rules, and performance heterogeneity. Strategic Management Journal **32**(6), 569–594 (2011). https://doi.org/https://doi.org/10.1002/smj.899, https://onlinelibrary.wiley.com/doi/abs/10.1002/smj.899
18. Geelan, D.: Teacher Explanations, pp. 987–999. Springer Netherlands, Dordrecht (2012), https://doi.org/10.1007/978-1-4020-9041-7_65
19. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M.A., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: Bonchi, F., Provost, F.J., Eliassi-Rad, T., Wang, W., Cattuto, C., Ghani, R. (eds.) 5th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2018, Turin, Italy, October 1-3, 2018. pp. 80–89. IEEE (2018). https://doi.org/10.1109/DSAA.2018.00018, https://doi.org/10.1109/DSAA.2018.00018
20. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019). https://doi.org/10.1145/3236009, https://doi.org/10.1145/3236009
21. Hacker, P., Passoth, J.: Varieties of AI explanations under the law. from the GDPR to the aia, and beyond. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K., Samek, W. (eds.) xxAI - Beyond Explainable AI - International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. Lecture Notes in Computer Science, vol. 13200, pp. 343–373. Springer (2020). https://doi.org/10.1007/978-3-031-04083-2_17, https://doi.org/10.1007/978-3-031-04083-2_17
22. Heider, F.: The psychology of interpersonal relations. Psychology Press (1982)
23. Hempel, C.G., Oppenheim, P.: Studies in the logic of explanation. Philosophy of Science **15**(2), 135–175 (1948). https://doi.org/10.1086/286983
24. Hilton, D.J.: Mental models and causal explanation: Judgements of probable cause and explanatory relevance. Thinking & Reasoning **2**(4), 273–308 (1996). https://doi.org/10.1080/135467896394447, https://doi.org/10.1080/135467896394447
25. Hilton, D.J., Slugoski, B.R.: Knowledge-based causal attribution: The abnormal conditions focus model. Psychological review **93**(1), 75 (1986)

26. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. CoRR **abs/1812.04608** (2018), http://arxiv.org/abs/1812.04608
27. Holland, J., Holyoak, K., Nisbett, R., Thagard, P.: Induction: Processes of Inference, Learning, and Discovery. Bradford books, MIT Press (1986), https://books.google.it/books?id=Z6EFBaLApE8C
28. Holzinger, A., Carrington, A.M., Müller, H.: Measuring the quality of explanations: The system causability scale (SCS). Künstliche Intell. **34**(2), 193–198 (2020). https://doi.org/10.1007/s13218-020-00636-z, https://doi.org/10.1007/s13218-020-00636-z
29. Horne, Z., Muradoglu, M., Cimpian, A.: Explanation as a cognitive process. Trends in Cognitive Sciences **23**(3), 187–199 (2019). https://doi.org/https://doi.org/10.1016/j.tics.2018.12.004, https://www.sciencedirect.com/science/article/pii/S1364661318302857
30. Jansen, P., Balasubramanian, N., Surdeanu, M., Clark, P.: What's in an explanation? characterizing knowledge and inference requirements for elementary science exams. In: Calzolari, N., Matsumoto, Y., Prasad, R. (eds.) COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan. pp. 2956–2965. ACL (2016), https://aclanthology.org/C16-1278/
31. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. CoRR **abs/1707.01154** (2017), http://arxiv.org/abs/1707.01154
32. Liao, Q.V., Gruen, D.M., Miller, S.: Questioning the AI: informing design practices for explainable AI user experiences. In: Bernhaupt, R., Mueller, F.F., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjøn, P., Zhao, S., Samson, B.P., Kocielnik, R. (eds.) CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020. pp. 1–15. ACM (2020). https://doi.org/10.1145/3313831.3376590, https://doi.org/10.1145/3313831.3376590
33. Liao, Q.V., Varshney, K.R.: Human-centered explainable AI (XAI): from algorithms to user experiences. CoRR **abs/2110.10790** (2021), https://arxiv.org/abs/2110.10790
34. Lim, B.Y., Dey, A.K., Avrahami, D.: *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. In: Jr., D.R.O., Arthur, R.B., Hinckley, K., Morris, M.R., Hudson, S.E., Greenberg, S. (eds.) Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4-9, 2009. pp. 2119–2128. ACM (2009). https://doi.org/10.1145/1518701.1519023, https://doi.org/10.1145/1518701.1519023
35. Lombrozo, T.: The structure and function of explanations. Trends in Cognitive Sciences **10**(10), 464–470 (2006). https://doi.org/https://doi.org/10.1016/j.tics.2006.08.004, https://www.sciencedirect.com/science/article/pii/S1364661306002117
36. Lundberg, S.M., Erion, G.G., Chen, H., DeGrave, A.J., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**(1), 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9, https://doi.org/10.1038/s42256-019-0138-9

37. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: A grounded interaction protocol for explainable artificial intelligence. In: Elkind, E., Veloso, M., Agmon, N., Taylor, M.E. (eds.) Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, Montreal, QC, Canada, May 13-17, 2019. pp. 1033–1041. International Foundation for Autonomous Agents and Multiagent Systems (2019), http://dl.acm.org/citation.cfm?id=3331801
38. Mayes, G.R.: Theories of explanation (2001), https://iep.utm.edu/explanat/
39. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007, https://doi.org/10.1016/j.artint.2018.07.007
40. Mohseni, S., Block, J.E., Ragan, E.D.: Quantitative evaluation of machine learning explanations: A human-grounded benchmark. In: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (eds.) IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021. pp. 22–31. ACM (2021). https://doi.org/10.1145/3397481.3450689, https://doi.org/10.1145/3397481.3450689
41. Nguyen, A., Martínez, M.R.: On quantitative aspects of model interpretability. CoRR **abs/2007.07584** (2020), https://arxiv.org/abs/2007.07584
42. Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Vaughan, J.W., Wallach, H.M.: Manipulating and measuring model interpretability. In: Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., Drucker, S.M. (eds.) CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021. pp. 237:1–237:52. ACM (2021). https://doi.org/10.1145/3411764.3445315, https://doi.org/10.1145/3411764.3445315
43. Rebanal, J.C., Combitsis, J., Tang, Y., Chen, X.A.: Xalgo: a design probe of explaining algorithms' internal states via question-answering. In: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (eds.) IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021. pp. 329–339. ACM (2021). https://doi.org/10.1145/3397481.3450676, https://doi.org/10.1145/3397481.3450676
44. Ribera, M., Lapedriza, À.: Can we do better explanations? A proposal of user-centered explainable AI. In: Trattner, C., Parra, D., Riche, N. (eds.) Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), Los Angeles, USA, March 20, 2019. CEUR Workshop Proceedings, vol. 2327, p. 38. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf
45. Rosenfeld, A.: Better metrics for evaluating explainable artificial intelligence. In: Dignum, F., Lomuscio, A., Endriss, U., Nowé, A. (eds.) AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021. pp. 45–50. ACM (2021). https://doi.org/10.5555/3463952.3463962, https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p45.pdf
46. Salmon, W.: Scientific Explanation and the Causal Structure of the World. Book collections on Project MUSE, Princeton University Press (1984), https://books.google.it/books?id=2ug9DwAAQBAJ
47. Schweisfurth, M.: Learner-centred Education in International Perspective: Whose pedagogy for whose development? Education, Poverty and International De-

velopment, Taylor & Francis (2013), https://books.google.ch/books?id=dT4jLusPp9AC
48. Searle, J.R.: Austin on locutionary and illocutionary acts. The Philosophical Review **77**(4), 405–424 (1968), http://www.jstor.org/stable/2183008
49. Sellars, W.: Science, Perception and Reality. New York: Humanities Press (1963)
50. Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F.: A survey on methods and metrics for the assessment of explainability under the proposed AI act. In: Erich, S. (ed.) Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021. Frontiers in Artificial Intelligence and Applications, vol. 346, pp. 235–242. IOS Press (2021). https://doi.org/10.3233/FAIA210342, https://doi.org/10.3233/FAIA210342
51. Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F.: Metrics, explainability and the european ai act proposal. J **5**(1), 126–138 (2022). https://doi.org/10.3390/j5010010, https://www.mdpi.com/2571-8800/5/1/10
52. Sovrano, F., Vitali, F.: From philosophy to interfaces: an explanatory method and a tool inspired by achinstein's theory of explanation. In: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (eds.) IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021. pp. 81–91. ACM (2021). https://doi.org/10.1145/3397481.3450655, https://doi.org/10.1145/3397481.3450655
53. Sovrano, F., Vitali, F.: An objective metric for explainable AI: how and why to estimate the degree of explainability. CoRR **abs/2109.05327** (2021), https://arxiv.org/abs/2109.05327
54. Sovrano, F., Vitali, F.: Explanatory artificial intelligence (yai): human-centered explanations of explainable ai and complex data. Data Mining and Knowledge Discovery (2022). https://doi.org/10.1007/s10618-022-00872-x, https://doi.org/10.1007/s10618-022-00872-x
55. Sovrano, F., Vitali, F.: How to quantify the degree of explainability: Experiments and practical implications. In: 31th IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2022, Padova, July 18-23, 2022. pp. 1–9. IEEE (2022)
56. Sovrano, F., Vitali, F., Palmirani, M.: Modelling gdpr-compliant explanations for trustworthy AI. In: Ko, A., Francesconi, E., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Electronic Government and the Information Systems Perspective - 9th International Conference, EGOVIS 2020, Bratislava, Slovakia, September 14-17, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12394, pp. 219–233. Springer (2020). https://doi.org/10.1007/978-3-030-58957-8_16, https://doi.org/10.1007/978-3-030-58957-8_16
57. Szymanski, M., Millecamp, M., Verbert, K.: Visual, textual or hybrid: the effect of user expertise on different explanations. In: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (eds.) IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021. pp. 109–119. ACM (2021). https://doi.org/10.1145/3397481.3450662, https://doi.org/10.1145/3397481.3450662
58. Thagard, P.: Analogy, explanation, and education. Journal of Research in Science Teaching **29**(6), 537–544 (1992). https://doi.org/https://doi.org/10.1002/tea.3660290603, https://onlinelibrary.wiley.com/doi/abs/10.1002/tea.3660290603
59. Van Fraassen, B.C.: The Scientific Image. Clarendon Library of Logic and Philosophy, Clarendon Press (1980), https://books.google.it/books?id=VLz2F1zMr9QC

60. Vilone, G., Rizzo, L., Longo, L.: A comparative analysis of rule-based, model-agnostic methods for explainable artificial intelligence. In: Longo, L., Rizzo, L., Hunter, E., Pakrashi, A. (eds.) Proceedings of The 28th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Republic of Ireland, December 7-8, 2020. CEUR Workshop Proceedings, vol. 2771, pp. 85–96. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2771/AICS2020_paper_33.pdf
61. 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 **7**(2), 76–99 (2017)
62. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law and Technology **31**(2) (2018). https://doi.org/10.2139/ssrn.3063289, http://dx.doi.org/10.2139/ssrn.3063289
63. Wang, X., Yin, M.: Are explanations helpful? A comparative study of the effects of explanations in ai-assisted decision-making. In: Hammond, T., Verbert, K., Parra, D., Knijnenburg, B.P., O'Donovan, J., Teale, P. (eds.) IUI '21: 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, April 13-17, 2021. pp. 318–328. ACM (2021). https://doi.org/10.1145/3397481.3450650, https://doi.org/10.1145/3397481.3450650
64. Warren, G., Keane, M.T., Byrne, R.M.J.: Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. In: Heyninck, J., Meyer, T., Ragni, M., Thimm, M., Kern-Isberner, G. (eds.) Proceedings of the Workshop on Cognitive Aspects of Knowledge Representation co-located with the 31st international join conference on artificial intelligence (IJCAI-ECAI 2022), Vienna, Austria, July 23, 2022. CEUR Workshop Proceedings, vol. 3251. CEUR-WS.org (2022), https://ceur-ws.org/Vol-3251/paper1.pdf
65. Williams, J.J., Lombrozo, T., Rehder, B.: The hazards of explanation: Overgeneralization in the face of exceptions. Journal of Experimental Psychology: General **142**(4), 1006 (2013)
66. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5) (2021). https://doi.org/10.3390/electronics10050593, https://www.mdpi.com/2079-9292/10/5/593