Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Formalising Explainability and Interpretability: A Conceptual Framework and Taxonomical Dimensions

Stefano Pieri
University of Huddersfield
George Baryannis
g.bargiannis@hud.ac.uk
University of Huddersfield
Emmanuel Papadakis
University of Huddersfield

Research Article

Keywords: Explainable artificial intelligence, Interpretability, Explainability, Conceptual framework, Taxonomy, Human-centred AI

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9267494/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Formalising Explainability and Interpretability: A Conceptual Framework and Taxonomical Dimensions

Stefano Pieri[1], George Baryannis[1*], Emmanuel Papadakis[1]

[1*]School of Computing and Engineering, University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, United Kingdom.

*Corresponding author(s). E-mail(s): g.bargiannis@hud.ac.uk;
Contributing authors: stefano.pieri@hud.ac.uk; e.papadakis@hud.ac.uk;

Abstract

Research on explainable artificial intelligence (XAI) has expanded rapidly, yet the concepts of interpretability and explainability remain theoretically unstable and inconsistently defined. This paper addresses that weakness directly through a systematic literature review focused specifically on published definitions of the two concepts. We analyse these definitions by decomposing them into semantic units and reorganising the recurring conceptual material into a structured framework centred on four entities: *AI system*, *Human user*, *Interaction*, and *Context*. On this basis, we derive a set of high-level taxonomical dimensions along which existing accounts differ, including their formalism, measurability, specificity, and dependence on users, domains, and contexts. Our analysis shows that interpretability and explainability are not cleanly separable in the literature, but are better understood as overlapping regions of a broader multidimensional conceptual space. The paper therefore contributes a methodological approach for semantic analysis in fragmented research areas, a novel conceptual framework for interpretability and explainability, and a taxonomical basis for comparing existing definitions. More broadly, it takes a concrete step towards the formalisation of core XAI concepts and towards a more mature theoretical foundation for the field.

Keywords: Explainable artificial intelligence, Interpretability, Explainability, Conceptual framework, Taxonomy, Human-centred AI

1 Introduction

The increasing complexity and scale of contemporary artificial intelligence (AI) systems have made it progressively harder for human users to understand the reasons behind their outputs and decisions (Castelvecchi, 2016; Holzinger, Zatloukal, & Müller, 2024). This loss of comprehensibility, often described as the black-box problem, has become a central concern in modern AI because highly capable systems may produce decisions whose internal logic is not readily accessible to users, developers, or affected stakeholders (Carvalho, Pereira, & Cardoso, 2019; Castelvecchi, 2016; Von Eschenbach, 2021). Although this concern is especially prominent in data-driven machine learning, questions of explanation and understanding are not limited to sub-symbolic approaches, and have also emerged in symbolic and knowledge-based AI, including planning and reasoning (De Bruijn, Warnier, & Janssen, 2022; Fox, Long, & Magazzeni, 2017; Swartout & Moore, 1993).

The importance of this problem is both technical and societal. The opacity of AI systems can undermine trust, hinder debugging and improvement, and restrict adoption in high-stakes domains such as healthcare, finance, security, and transport (Adadi & Berrada, 2018; Ali et al., 2023; Arrieta et al., 2020; Zednik, 2021). These concerns are further reinforced by growing regulatory pressure. In the European context, both the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act (EU AI Act) have intensified demands for greater transparency and explanation in algorithmic decision-making, particularly in high-risk applications (Bibal, Lognoul, De Streel, & Frénay, 2021; Kaminski & Malgieri, 2025). In response, research in explainable AI (XAI) has expanded rapidly, especially through the development of computational methods for extracting information from AI models, including widely used approaches such as SHAP and LIME (Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016; Schneeberger et al., 2023).

Despite this methodological growth, the conceptual foundations of the field remain unsettled. In particular, there is still no clear consensus on what *interpretability* and *explainability* mean, what elements they should include, or how they should be distinguished. Existing literature reveals a fragmented and at times contradictory landscape in which the two concepts are treated variously as synonyms, overlapping notions, hierarchical concepts, or domain-dependent constructs. Although a small number of studies have addressed terminological and conceptual clarification, this line of work remains limited in scope (Cabitza et al., 2023; Clinciu & Hastie, 2019; Graziani et al., 2023; Palacio et al., 2021; Roundtree, 2023). Most existing surveys focus on methods, applications, or evaluation, rather than on the semantic and conceptual structure of definitions themselves.

This gap motivates the present study. We argue that progress on interpretability and explainability requires closer attention to the definitional layer of the field. If the concepts remain unstable at the level of meaning, then comparison, evaluation, operationalisation, and standardisation all become more difficult. More specifically, the literature still lacks a systematic analysis of how these concepts are defined, what recurring semantic components they contain, how they relate to one another, and whether they can be organised within a more coherent conceptual structure.

In response, this paper conducts a systematic literature review focused specifically on published definitions of interpretability and explainability. We analyse these definitions by decomposing them into semantic units and reorganising the recurring conceptual material into a structured framework. The resulting framework is organised around four entities, namely *AI system*, *Human user*, *Interaction*, and *Context*. On this basis, we also derive a set of high-level taxonomical dimensions along which existing accounts differ, including their formalism, measurability, specificity, and dependence on users, domains, and contexts.

The contribution of the paper is therefore threefold. First, it introduces a methodological approach for the semantic analysis of concepts in a fragmented research area. Second, it proposes a novel conceptual framework that reconstructs the definitional landscape of interpretability and explainability in a more explicit and integrated form. Third, it uses that framework to compare the two concepts and to show that, although they are not identical, they are also not cleanly separable in the literature as it currently stands. Instead, they appear to occupy overlapping regions of a broader shared conceptual space.

The remainder of the paper is structured as follows. Section 2 reviews the background and related work, including the historical development of the two concepts, the main definitional positions in the literature, and the conceptual gap addressed by this study. Section 3 presents the methodology, including the review protocol, data extraction process, and synthesis procedure. Section 4 reports the main findings of the review and identifies the recurring conceptual patterns that emerge from the corpus. Section 5 introduces the proposed conceptual framework and the taxonomical dimensions derived from it, while also comparing interpretability and explainability within the framework and outlining the implications of their overlap and divergence. Section 6 discusses the broader significance of the framework, including its implications for conceptual clarification and human-centred XAI design. Finally, Section 7 concludes the paper and outlines directions for further research.

2 Background and related work

2.1 Historical development of interpretability and explainability

Interpretability and explainability did not emerge in identical ways, even though they are now often discussed together. In the interpretability literature, a major line of work framed the issue in terms of the gap between model optimisation and real-world evaluation. Doshi-Velez and Kim (2017) argued that interpretability becomes necessary when a problem is incompletely formalised, so predictive performance alone is not sufficient to judge whether a system is acceptable. In this view, interpretability is linked to broader desiderata such as trust, causality, fairness, robustness, and usability. Subsequent work reinforced the idea that interpretability is not a single, context-free property. Tomsett, Braines, Harborne, Preece, and Chakraborty (2018) proposed a role-based perspective in which different actors require different forms of understanding, while Lipton (2018) highlighted the conceptual vagueness of the term and its use as a proxy for distinct goals and technical properties. Survey work further reflected

this diversity by classifying different ways in which black-box models may be made more understandable (Guidotti et al., 2018).

Explainability developed along a partly different path. In areas such as planning, the focus was on enabling systems to answer queries about their behaviour, for example why a plan was chosen, why alternatives were rejected, or why certain actions could not be performed (Fox et al., 2017). Other influential work drew on psychology, philosophy, and cognitive science to argue that explanation is fundamentally human-centred and social rather than merely a technical property of a model (Miller, 2019). This view was reinforced by work on user-centred explainability (Ribera & Lapedriza, 2019) and by research relating explanation to human judgement, psychometrics, and evaluation practices (Malik et al., 2022; Vilone & Longo, 2021). At the same time, critiques of post hoc explanation methods, especially in high-stakes settings such as medical imaging sharpened the distinction between transparent model design and explanatory add-ons (Rudin, 2019). Broader reviews such as Arrieta et al. (2020) consolidated these debates within the wider XAI literature.

Overall, interpretability was more often associated with model understanding, transparency, and the limits of formal evaluation, whereas explainability increasingly incorporated communication, justification, and user-centred considerations. This historical divergence helps explain why the two terms are still used inconsistently and why a more systematic conceptual analysis is needed.

2.2 Existing definitions and positions in the literature

No widely accepted account of interpretability and explainability currently exists. Instead, the literature reveals a fragmented landscape, with disagreement about the semantic scope of the two concepts, the entities involved, their mutual relationship, and even the possibility of formulating general definitions.

Definitions of interpretability often foreground human understanding, but differ in what exactly is to be understood and how. Doshi-Velez and Kim (2017) define interpretability as the ability to explain or present information in understandable terms to a human, while Guidotti et al. (2018) similarly emphasise providing reasons or details in human-understandable form. Other definitions shift attention towards model behaviour or representation. Gilpin et al. (2018) describe interpretability as the science of understanding what a model did or should have done, and Montavon, Samek, and Müller (2018) characterise it as a mapping from an abstract representation to a human-understandable domain. Miller (2019) links interpretability to the degree to which an observer can understand the cause of a decision, whereas Rudin (2019) argues that interpretability is domain-dependent and resists any universal definition.

Definitions of explainability vary just as much. Arrieta et al. (2020) define it as the capacity to provide reasons or details that clarify a system's functioning, while also stressing adaptation to the target audience. Pearl (2019) likewise associates explainability with providing reasons behind predictions and recommendations, whereas Gilpin et al. (2018) links it to the promotion of user trust. Other accounts make its relational and contextual character more explicit. Langer et al. (2021) tie explainability to the satisfaction of stakeholders' desiderata in a given context of use, while Köhl et al. (2019) argue that explainable systems should clarify relevant aspects

of the explanandum, adapt to their audience, and remain context-aware. Guidotti et al. (2018) define explanation as an interface that faithfully and understandably mediates between a human and a system's decisions. By contrast, Rudin (2019) uses explainable machine learning to refer to explaining a black box through a second model.

The relationship between interpretability and explainability is equally unsettled. Some authors use the two terms interchangeably or treat one as a form of the other. Lipton (2018) links them closely and views post hoc explanations as enabling interpretation, while Miller (2019) accepts the idea of explanation as post hoc interpretability and in places treats the terms as near-synonymous. Others adopt hierarchical or set-theoretic views. Gilpin et al. (2018), Antoniou, Papadakis, and Baryannis (2022) and Dib (2024) suggest that explainability may subsume interpretability, whereas Zablocki, Ben-Younes, Pérez, and Cord (2022) take the opposite view. Clinciu and Hastie (2019) instead represent the two as intersecting but non-identical sets. The absence of agreement on this basic relation is itself evidence of conceptual instability.

A further disagreement concerns whether general definitions are possible at all. Lipton (2018) argues that interpretability lacks a formal technical definition and has often been used in a quasi-scientific way. Rudin (2019); Rudin et al. (2022) similarly question whether a universal definition of interpretability can be formulated, given its dependence on domain and use case. Carvalho et al. (2019) likewise stress the dependence of the concept on users and contexts of application. At the same time, some proposals move towards more formal or measurable definitions. Dhurandhar, Iyengar, Luss, and Shanmugam (2017) introduce a formal account grounded in robustness, and Jaakkola, Janhunen, Kuusisto, Rankooh, and Vilander (2025) explore logic-based definitions of interpretability using formal languages such as propositional logic (Baryannis et al., 2018).

2.2.1 Synthesis and rationale for the present study

Taken together, these definitions reveal substantial conceptual fragmentation. The disagreement concerns not only wording, but also the scope of the concepts, the entities and relations involved, and the extent to which the concepts can be expressed in general, formal, or measurable terms. At the same time, a recurring pattern can still be discerned across the literature, involving an AI system, a human recipient, and some form of interaction oriented towards particular goals. This suggests the need for a multidimensional framework that makes these recurring elements explicit and structurally comparable.

2.3 Related work on conceptual clarification

To the best of our knowledge, a systematic analysis of definitions of interpretability and explainability remains lacking. A small number of studies have addressed terminological, semantic, or conceptual clarification, but this line of work remains limited in scope.

The closest contribution to the present study is Roundtree (2023), who compare definitions of explainability, interpretability, fairness, and privacy in order to identify

recurrent terminology and ideas. Clinciu and Hastie (2019) survey definitions of transparency, intelligibility, interpretability, and explainability, representing their relations with Venn diagrams. Habiba, Habib, Bogner, Fritzsch, and Wagner (2025) interview AI professionals about their understanding of explainability, while Graziani et al. (2023) provide a broader cross-disciplinary analysis of related concepts across law, philosophy, sociology, cognitive science, labour, and technological sciences. A related stream compiles and compares definitions in tabular form, for example in Cabitza et al. (2023), Palacio et al. (2021), and Graziani et al. (2023).

These studies are valuable steps towards conceptual clarification, but they do not provide a systematic account of how definitions of interpretability and explainability are structured and related.

2.4 Interactional perspectives on XAI

The idea that explanation is fundamentally interactional predates XAI. This matters for the present study because it shows that modelling explanation in terms of interacting entities is grounded in established work from social science, linguistics, and human-centred research. A notable early contribution is Hilton (1990)'s conversational model of explanation, which distinguishes between identifying why an event occurred and communicating this to another party. More broadly, work on conversational communication, such as Grice (1975), has emphasised that successful explanation depends not only on content but also on relevance, adequacy, and clarity of presentation.

These ideas have been carried into XAI in several ways. A number of authors explicitly treat explanations as interactive or conversational processes rather than static outputs (Adadi & Berrada, 2018; Miller, 2019; Miller, Howe, & Sonenberg, 2017). In this view, explanations involve an explainer and an explainee, may require follow-up questions or clarification, and should adapt to the goals, expertise, and needs of the user. Related work stresses the role of stakeholders, context, and user-specific understanding in shaping what counts as a satisfactory explanation (Bellucci, Delestre, Malandain, & Zanni-Merk, 2021; Chazette, Brunotte, & Speith, 2021; Langer et al., 2021; Nyrup & Robinson, 2022), while other contributions emphasise explanation as a support for broader human-AI coordination and alignment of mental models (Mucha, Robert, Breitschwerdt, & Fellmann, 2021; Picard, Mualla, Gechter, & Galland, 2023).

Overall, this literature supports an interactional view of XAI in which explanation is not reducible to a technical artefact extracted from a model, but instead emerges through relations among a system, a recipient, and a context of use. This provides an important foundation for our framework, as discussed in Section 3.

2.5 Research gap and research questions

The literature reviewed above shows not only that no commonly accepted definitions of interpretability and explainability exist, but also that the field lacks a structured account of how these concepts are composed and related. Their meanings vary across studies, their mutual relationship remains unclear, and existing work does not provide a sufficiently systematic analysis of definitions themselves. Nor does it offer a multidimensional framework capable of capturing the entities, relations, and properties that recur across them.

A further limitation is that definitions are rarely framed in explicitly interactional terms. Although many implicitly involve an AI system, a human recipient, and some explanatory or interpretative process between them, the literature does not yet provide a clear conceptual structure that represents these elements within a unified model. This motivates a systematic review focused specifically on definitions, with the aim of extracting their core semantic components and organising them within a more rigorous conceptual framework.

In response, this study addresses the open theoretical problem of defining interpretability and explainability through the following research questions:

- What entities, properties and relations recur across existing definitions of interpretability and explainability?
- How can these recurring elements be synthesised into a coherent conceptual framework?
- How do interpretability and explainability relate to one another within this framework, in terms of overlap, distinction, and possible unification?
- To what extent can an interactional perspective capture the conceptual structure of interpretability and explainability?

To address these questions, we conduct a systematic investigation of definitions of interpretability and explainability in the literature and propose a unified taxonomical framework grounded in their recurring conceptual components and relations.

3 Methodology

3.1 Overview

This study begins by conducting a systematic literature review of the concepts of interpretability and explainability, focusing specifically on published, peer-reviewed definitions proposed in the literature. The review follows an adapted version of the PRISMA methodology (Page et al., 2021), tailored to the conceptual and qualitative nature of the present study.

The extracted definitions are analysed by decomposing them into semantic components, which are then reorganised within a novel interaction-based framework. As a first step, we construct a descriptive table designed to capture the full range of semantic elements identified in the definitions and to organise them within a hierarchical conceptual structure. This structure is centred on four main entities: *AI system*, *Human user*, *Interaction*, and *Context*.

Building on this detailed representation, we then derive a more abstract conceptual framework that preserves the higher-level hierarchical relations while omitting lower-level detail. This conceptual framework serves as the basis for an integrated representation of interpretability and explainability. Finally, we propose a set of taxonomical dimensions for classifying existing definitions of these two concepts.

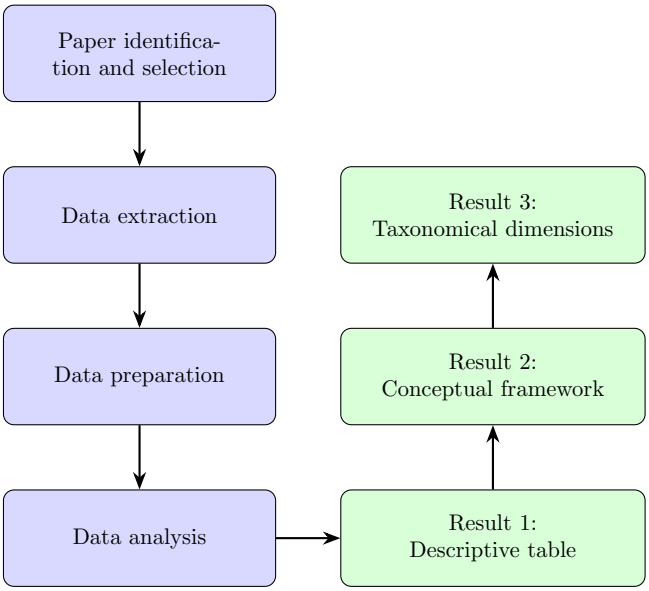


Fig. 1 Overview of the methodology

3.2 Systematic review process

Eligibility criteria

Papers were included in the review if they met the following criteria:

- they were published in peer-reviewed venues;
- they addressed artificial intelligence or explainable artificial intelligence;
- they included at least one explicit definition of either interpretability or explainability.

We excluded papers that referred to these concepts only through associated properties, desiderata, requirements, or constraints, or that defined them in a circular or overly elementary manner, without providing an explicit definition.

Information sources

The initial automatic search was conducted in the Scopus database on 15 January 2025 and was updated on 1 February 2026. Additional relevant papers were identified through backward reference searching of selected articles. We also included relevant papers retrieved from tabulated collections of definitions reported in previously identified studies.

Search strategy

The automatic search targeted documents whose abstracts contained the terms ("explainability" OR "interpretability") AND ("definition" OR "define").

To improve precision, the search was restricted to the subject area of Computer Science and to documents associated with the following keywords: *Interpretability*,

Explainability, *Explainable AI*, *Explainable Artificial Intelligence*, *Artificial Intelligence*, *Machine Learning*, and *Deep Learning*. No date restrictions were applied.

For papers identified through reference lists or tables in previously selected articles, we performed a title-based search and retained an accessible full-text version.

Selection process

The selection process was carried out iteratively. Papers returned by the automatic search were first screened at abstract level in order to exclude clearly irrelevant records. The remaining papers were then assessed through full-text review against the eligibility criteria described above. The same procedure was followed for papers identified through reference lists and tabulated collections of definitions. The review process was conducted by the lead author under supervision of the co-authors. No automation tools were used.

3.3 Data extraction and preprocessing

Data collection process

Data collection consisted of extracting all relevant definitions identified in the selected literature. Each definition was recorded in a separate spreadsheet to support subsequent analysis. In addition, relevant textual excerpts and visual material were noted where they contributed to the interpretation of the definitions. The extraction process was conducted by the lead author.

Data items

The primary data for this study consist of textual definitions of interpretability and explainability, understood as explicit textual descriptions of what these concepts are, represent, or refer to. We distinguish such definitions from discussions of associated requirements, desiderata, properties, or goals, which are not treated here as definitions in themselves.

Under restricted conditions, we also included definitions of interpretation and explanation. These were considered only when a paper was situated in an XAI context and did not provide an explicit definition of either interpretability or explainability. In such cases, interpretation and explanation were treated as closely related conceptual proxies and analysed in the same way as the other definitions. When both types of definition were present in the same paper, priority was given to interpretability and explainability. This decision is consistent with prior literature that has analysed these terms together within the same conceptual space (Cabitza et al., 2023).

Risk of bias assessment

Although data extraction was conducted by the lead author only, the subsequent analysis was reviewed by the two co-authors in order to reduce interpretative bias.

Preprocessing and preparation for synthesis

After extraction, each data item was assessed for relevance and appropriateness for inclusion. The retained definitions, together with relevant supporting passages, were compiled into a shared working document for analysis.

Where necessary, definitions were normalised into a standard subject-predicate-complement form, such as "Interpretability is ..." or "Explainability is ...". This was done only when the original wording was syntactically irregular or difficult to analyse directly, and without altering the intended meaning.

The definitions were then decomposed into atomic semantic units, that is, minimal textual units expressing a single meaningful element. These units did not correspond to a fixed number of words: in some cases they consisted of a single adjective or noun phrase, while in others they involved longer expressions.

3.4 Synthesis process

Procedure

The extracted atomic semantic units were reorganised into a multilayer, tree-like hierarchical conceptual structure. In this structure, each node represents a concept and each child node specifies or refines the concept represented by its parent. The hierarchy includes only vertical relations of specification and abstraction; horizontal relations between concepts in different semantic areas were not modelled at this stage.

The synthesis was guided by an interactional perspective on explanation, discussed in Section 2.4, which motivated the use of the entities *AI system*, *Human user*, *Interaction*, and *Context* as the top-level elements of the framework.

This procedure was applied separately to each definition in order to preserve traceability between individual definitions and the semantic units they contained. To improve the readability of the hierarchy, intermediate conceptual layers were introduced where necessary to group related lower-level concepts.

The synthesis followed an iterative combination of top-down and bottom-up reasoning. A top-down approach was used to define the foundational entities of the framework and to impose a small number of higher-level interaction-related categories, such as those associated with communicative structure. A bottom-up approach was used for the remaining concepts by grouping atomic semantic units according to semantic similarity and associating them with the most appropriate higher-level entity. For example, units relating to a model's internal functioning were grouped under the *AI system* entity.

The hierarchical structure was refined iteratively as additional semantic units were processed and new conceptual distinctions became necessary. This allowed us to reconstruct the semantic scope of the studied terminology as it appears in the literature while organising it into a coherent abstract framework. Particular attention was given to minimising overlap between conceptual areas and to consolidating synonymous or near-synonymous expressions under shared conceptual categories.

A hierarchical representation was adopted because it allows broader conceptual categories to be decomposed into more specific sub-concepts while preserving their semantic relations. Similar hierarchical structuring has been used in related work

on interpretability, explainability requirements, and evaluation frameworks (Alonso, Castiello, & Mencar, 2015; Chen & Eickhoff, 2024; Nauta et al., 2023; Ploug & Holm, 2020). In the present study, however, hierarchy is used not to organise metrics or requirements, but to reconstruct the conceptual structure of published definitions.

Artefacts produced

The synthesis process produced three artefacts.

The first artefact is a *descriptive table*, designed to preserve the maximum level of detail extracted from the original definitions. In this table, each row corresponds to a definition and each column to a semantic unit. A mark indicates the presence of a given unit in a given definition, while semantically similar units are positioned in neighbouring columns. This artefact supports the management of synonyms, semantically equivalent expressions, and fine-grained conceptual distinctions. The descriptive table is supplied as a spreadsheet (Online Resource 1).

The second artefact is a more abstract *conceptual framework*. This was derived from the descriptive table by retaining the most meaningful conceptual layers while removing lower-level detail. The purpose of this step was to move from a detailed descriptive representation of the literature to a more general and usable conceptual structure.

The third artefact is a set of *taxonomical dimensions* of definitions of interpretability and explainability. These were derived from patterns identified during the synthesis of the first two artefacts and capture higher-level dimensions along which definitions can be compared and classified.

The synthesis was not purely descriptive in the sense of mechanically reproducing the lexical surface of each definition. As the corpus expanded, it became clear that similar semantic units could be classified in multiple ways depending on local phrasing, which would have led to redundancy, instability, and limited comparability across definitions. For this reason, the framework construction involved a partly prescriptive step, in which conceptually similar units were reorganised into a coherent higher-level structure. This was necessary to preserve meaning while enabling systematic comparison and abstraction.

For example, expressions such as "inner working", "inner mechanism", and "how the model produces its output" may differ lexically, but often point towards the same conceptual area. A purely surface-level treatment would therefore multiply categories unnecessarily, whereas the present framework groups such expressions under shared higher-level concepts.

Certainty assessment

To assess the completeness and plausibility of the proposed framework, we compare it with selected well-cited conceptual frameworks from the literature, examining areas of overlap, divergence, and representational coverage. This comparison highlights both similarities and differences, and allows us to examine the extent to which our framework captures or reorganises information already present in prior conceptual models.

A worked example of how individual definitions were decomposed and mapped into the framework is provided in A.

4 Findings and analysis

4.1 Corpus and study selection

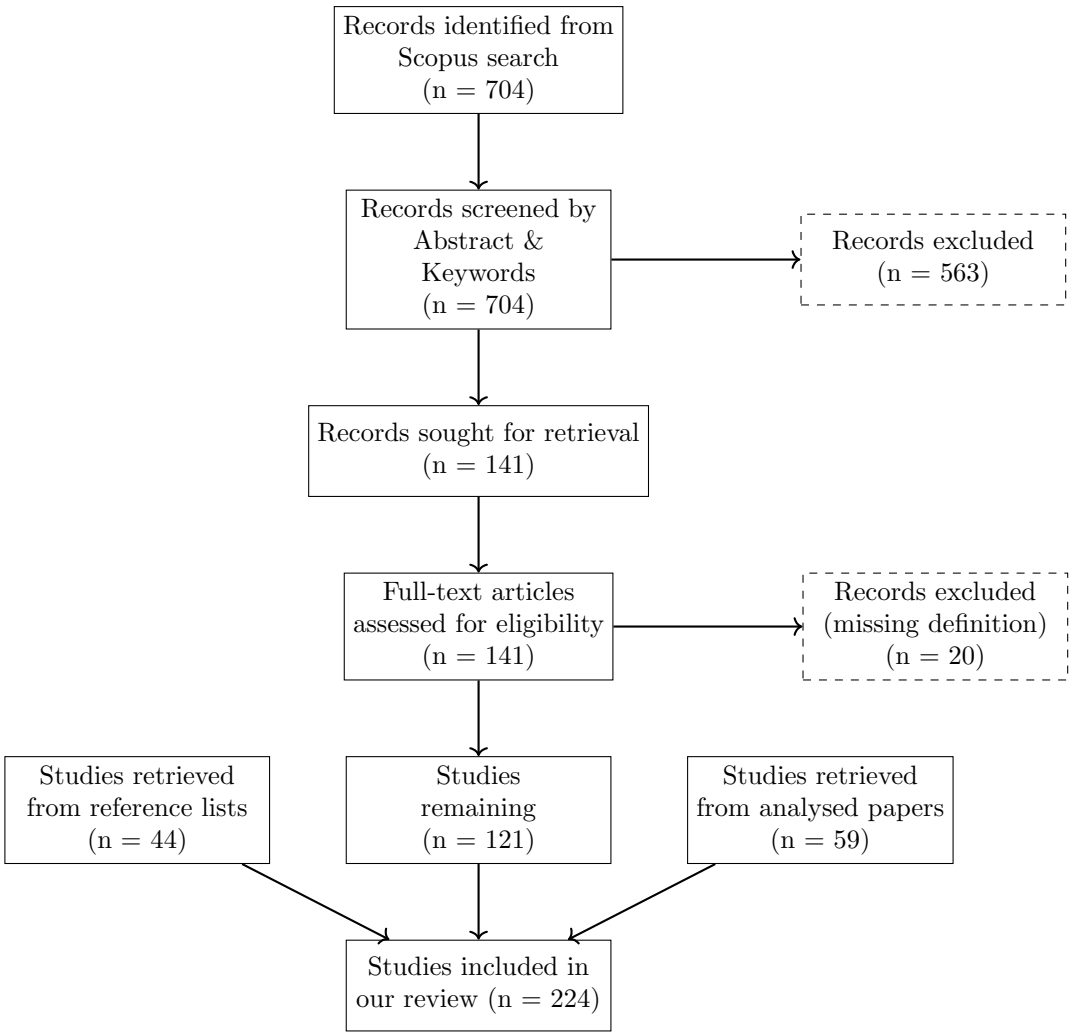


Fig. 2 PRISMA flow diagram of the literature selection process

As visualised in the PRISMA flow diagram of Fig. 2, the automatic search returned 704 papers. After abstract screening, 141 papers were retained, and after full-text

screening this number was reduced to 121. A further 44 papers were identified through backward reference searching, and 59 additional relevant papers were retrieved from the full text or tabulated material of selected articles. In total, 224 papers were analysed. A full list of these 224 papers is provided as a spreadsheet (Online Resource 2).

From these papers, we extracted 310 definitions relevant to the present study, of which 130 concerned interpretability and 180 concerned explainability. This corpus constitutes, to our knowledge, one of the largest collections of published definitions of these concepts assembled for the purpose of conceptual analysis.

4.2 Emerging patterns in the definitions

Analysis of the corpus confirmed a high degree of conceptual heterogeneity. Definitions varied not only in wording, but also in scope, emphasis, and assumed structure. Some highlighted human understanding, others the provision of reasons, and others still the communicative or contextual conditions under which a system becomes understandable or explainable. The corpus also revealed variation in how the two concepts are related: some definitions treated them as overlapping or near-synonymous, while others sharply distinguished them or placed one within the scope of the other.

At the same time, several recurrent semantic areas appeared across the definitions. These included the AI system or model being examined, the human recipient or stakeholder, the interaction through which understanding or explanation is achieved, and the contextual conditions that shape what counts as satisfactory understanding. These recurring elements suggest that the definitional space is diverse but not unstructured.

4.3 From extracted components to conceptual structure

The descriptive table was used to preserve the maximum level of semantic detail extracted from the corpus and to support the grouping of synonymous or closely related expressions. On this basis, recurring semantic units were reorganised into broader conceptual categories and then into a hierarchical conceptual structure. This process made it possible to move from a dispersed body of lexical formulations to a more explicit representation of the conceptual architecture underlying the literature.

A key result of this synthesis was the identification of four top-level entities that recur across definitions and support the organisation of the definitional space: *AI system*, *Human user*, *Interaction*, and *Context*. These entities form the basis of the conceptual framework presented in the next section.

4.4 Implications for conceptual modelling

The findings show that the literature contains both substantial variability in terminology and identifiable recurrent structure. This makes a purely narrative comparison insufficient, while also indicating that the definitional space is amenable to systematic modelling. The conceptual framework introduced in the next section is intended to capture this recurring structure at a higher level of abstraction, while the taxonomical dimensions derived from it provide a way of classifying existing definitions.

5 A conceptual framework for interpretability and explainability

5.1 Framework overview

The central contribution of this section is a unified conceptual framework for interpretability and explainability organised around an interactional scenario.

The central feature of the framework is the *Interaction* between an *AI system* and a *Human user* within a particular *Context*. In the present work, this interaction is modelled primarily in communicative terms, drawing on established communication-based perspectives on explanation discussed earlier in the paper. This makes it possible to organise many of the extracted semantic units in relation to how information is conveyed, received, and interpreted.

The framework is based on the aforementioned four fundamental entities that recur across the definitions analysed in this study: *AI system*, *Human user*, *Interaction*, and *Context*.

The role of these entities can be summarised as follows:

- *AI system*: the system, model, or decision-making artefact to be interpreted or explained;
- *Human user*: the human recipient of explanatory information, any stakeholder of the AI system for whom interpretability or explainability is relevant;
- *Interaction*: the processes through which explanatory or interpretative information is exchanged, queried, or otherwise made available;
- *Context*: the situational, task-related, and user-related conditions that shape how the other entities and their relations should be understood.

The framework is intended to reorganise the conceptual content of published definitions, rather than to classify definitions directly. For this reason, it should be distinguished from the taxonomical dimensions introduced later in the paper. Whereas the framework captures the recurring entities and relations underlying the definitional space, the dimensions focus on higher-level comparison criteria, including formalism and other meta-level characteristics.

In constructing the framework, we prioritised specific conceptual distinctions over generic references. Broad labels without meaningful specification were omitted in order to keep the structure analytically useful and conceptually lean. As a result, lower-level concepts are treated as progressively more specific refinements of higher-level ones, rather than as loosely associated labels.

Figure 3 illustrates the two topmost levels of the conceptual framework hierarchy. The full hierarchy is provided as a diagram (Online Resource 3). The following subsections describe the conceptual components associated with each of its four entities.

5.2 AI system

The *AI system* entity captures the technical characteristics, internal organisation, and operational aspects of the system to be interpreted or explained. It includes those

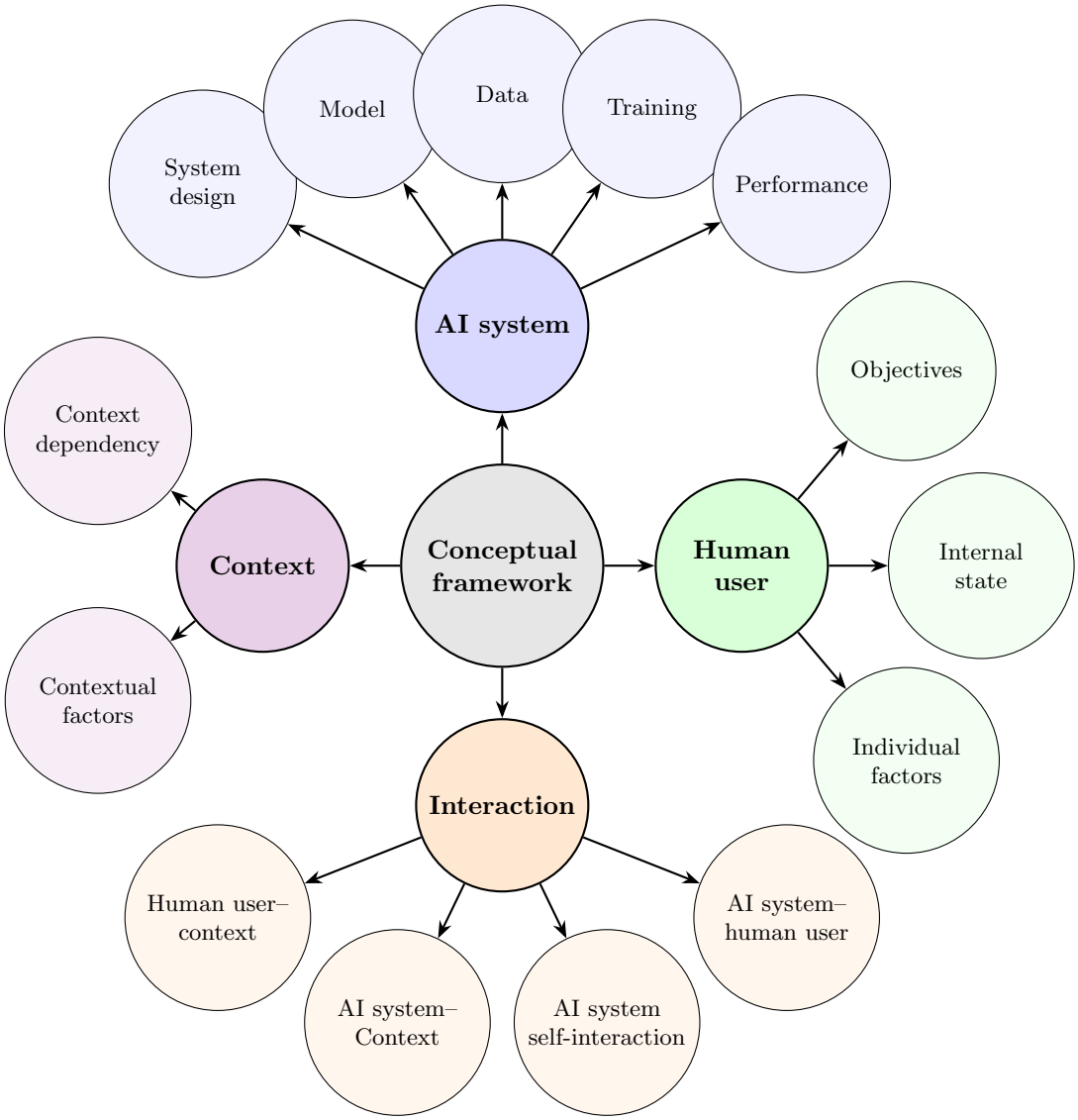


Fig. 3 Top two levels of the conceptual framework.

semantic components that refer primarily to the artefact under examination, rather than to the recipient of explanations, the interaction through which they are conveyed, or the surrounding context. This entity is organised into a set of major subcomponents that reflect the most recurrent technical areas identified across the analysed definitions.

It should be noted that the *AI system* refers to the system, model, or decision-making artefact whose behaviour is to be interpreted or explained. Post hoc explainer

models are not treated here as the primary *AI system*, but as part of the *Interaction* entity, to the extent that they mediate the production of explanatory information.

A first subcomponent relates to the AI *system design*, which refers to the technical characteristics that guide development. A major subcomponent concerns the *model* itself. This includes references to architectural structure, operating principles, parameters and settings, system state and inference processes. Within the latter area, the analysed definitions frequently refer to the model's *inner working*, that is, the internal aspects, mechanisms, procedures, or rules through which the system produces its outputs. Closely related to this is the theme of *decision-making*, which includes references to decision criteria, reasoning processes, causal relations, and the underlying rationale or "why" behind predictions and recommendations. A further recurring theme is the *input-output mapping*, including the functional transformation of inputs into outputs, the formalism through which this mapping is expressed, and the internal representations used to realise it. Lastly, *output* refers recurrently to the outcome of the inference process. Collectively, these components show that many definitions of interpretability and explainability remain strongly centred on opening up the internal functioning of the model.

A further major subcomponent is *data*. This includes references to datasets, data preprocessing, feature engineering, training and test splits, adversarial or target distributions, and the variables or features involved in prediction. At this level, the analysed definitions do not focus only on data in a generic sense, but also on specific aspects such as which inputs are involved, how many features are considered, and what relations hold between domain variables. This shows that, in part of the literature, interpretability and explainability are linked not only to the model as an abstract mechanism, but also to the informational basis on which its behaviour depends.

The *training* subcomponent captures those references that concern the learning phase of the system, including the learning objective and the algorithmic process through which the model is fitted. These aspects occur less frequently than model- and data-related components, but they remain relevant because some definitions associate understanding a system with understanding how it was trained or optimised.

Finally, the *performance* subcomponent covers references to empirical behaviour and evaluation, including the assessment objective, such as output quality or error, and explicit metrics such as accuracy, stability, simplicity, or predictivity scores. Although performance is not always treated as part of the definition itself, some accounts do connect interpretability or explainability to measurable aspects of how well a model behaves.

Overall, the *AI system* entity shows that the technical side of interpretability and explainability extends well beyond the model in a narrow sense. It encompasses architecture, data, training, inference, outputs, and evaluation, thereby capturing the broader technical structure within which understanding, explanation, and interpretation are situated.

5.3 Human user

The *Human user* entity captures those semantic components that relate to the person, stakeholder, or recipient for whom interpretability or explainability is relevant.

Across the analysed definitions, this entity is organised primarily around three main themes: *objectives*, *internal state* and *individual factors*. This reflects the fact that interpretability and explainability are often defined not only in terms of what a system reveals, but also in terms of what users seek, need, or are able to do with the information provided.

The most prominent subcomponent is *objectives*, which is consistent with prior work showing that explanatory information is often valued in relation to the goals it helps users achieve in practice (Theis et al., 2023). This subcomponent is organised into five broad categories. The first consists of *epistemic objectives*, centred on knowledge acquisition and understanding. These include understanding information provided by the system, acquiring knowledge about its behaviour, being guided towards discovery, reasoning on the basis of the information received, and reducing uncertainty. This category is especially prominent in the corpus and reflects the fact that many definitions treat interpretability and explainability as means of making AI outputs cognitively accessible.

A second group consists of *operational objectives*, which concern the practical use of the information provided. These include utility, usability, and operationalisation, that is, enabling the user to apply the received information in decision-making, analysis, or action. In this sense, explanations are not valued only for their descriptive content, but also for their capacity to support effective use.

A third group comprises *ethical objectives*, including trustworthiness, safety, fairness, transparency and ethical permissibility. These objectives show that, for many authors, interpretability and explainability are linked to certain expectations about how AI systems should behave and how their behaviour should be judged by users.

A fourth category concerns *governance and management objectives*. This includes references to data governance, system purpose and intentions, monitoring and control, debugging and improvement, validation and verification, evaluation, impact assessment, and the management of conflicts or disagreements. These objectives indicate that the human user is often conceived not merely as a passive recipient of explanations, but also as an evaluator, overseer, or modifier of the system, consistent with a human-in-the-loop perspective.

Finally, *alignment objectives* concern the relation between user expectations and system behaviour. These include the alignment of model outputs with user predictions, the conformity of algorithmic behaviour with specifications, the alignment of systems with human values, and consistency with users' mental models. This category highlights that interpretability and explainability are often tied to whether a system behaves in ways that users can anticipate, relate to, or endorse.

The other two main subcomponents are *internal state* and *individual factors*. The former captures the user's internal cognitive state, such as mental models or representations. The latter includes user preferences, explanatory desires, subjective interpretative differences, user capabilities, prior knowledge, and other psychological or cognitive factors that shape how explanations are received and understood. These elements capture the heterogeneity of interpretability and explainability from the perspective of human factors: different users may require different forms, levels, or styles of explanation depending on their background, abilities, and expectations.

As a whole, these components show that the *Human user* entity is not limited to a generic notion of comprehension. Instead, it encompasses cognitive, practical, ethical, managerial, and individual dimensions, thereby capturing the many ways in which interpretability and explainability are oriented towards human needs, expectations, and limitations.

5.4 Interaction

The *Interaction* entity captures the relational processes through which interpretability and explainability are realised. Across the analysed definitions, this entity is dominated by the interaction between the *AI system* and the *Human user*, while a smaller number of references concern other forms of interaction. This confirms that many definitions do not treat interpretability or explainability as isolated properties of a system, but as phenomena emerging through exchanges between parties.

5.4.1 AI system–human user interaction

The most frequently occurring branch of the *Interaction* entity concerns the interaction between the *AI system* and the *Human user*. Within this branch, the extracted notions cluster around two broad dimensions: *communicational dimensions* and *activities*.

Communicational dimensions

A first major dimension concerns the communicational structure of the interaction. At this level, the framework captures both general communication-related concepts and more specifically XAI-related ones.

The first group includes standard communicational aspects such as *encoding* and *decoding*, that is, the translation of information from an initial form into a final form that is understandable to the user. In the analysed definitions, this often involves movement from abstract, low-level, formal, or otherwise inaccessible representations towards human-understandable, higher-level, or expert-usable forms. The communicational dimension also includes the *content* of the interaction, specifically explanatory primitives, constructs, approaches (perturbation, contrastivity, feature attribution), properties of explanatory content (e.g. fidelity, completeness, correctness), and information properties (e.g. comprehensibility, transferability).

Other communicational aspects include the representational *format*, such as linguistic, visual, or formal forms, the medium through which information is conveyed, the possibility of *noise* or distortion in transmission, and the degree of *interactivity*, understood as the possibility of back-and-forth exchange, clarification, or follow-up.

A second group consists of XAI-specific communicational aspects. These include the *scope* of the explanation, particularly the distinction between local and global forms; the question of whether interpretability or explainability is treated as an *intrinsic* property of the system or as something requiring an intermediate method; and the *timeliness* of the interaction, e.g. whether explanation occurs post hoc, after prediction, or after training. Together, these elements show that the communicational side of explainability is not exhausted by the mere presence of information, but also depends on how, when, and in what form that information is made available.

Activities

The second major dimension concerns the activities that take place during the explanatory or interpretative interaction. These include both the requirement that some entity perform an action and the nature of the action itself.

The extracted actions can be grouped into three broad categories. The first comprises *interaction management* actions, such as initiating, guiding, regulating, or querying the interaction. The second comprises *informative actions*, including the acquisition, extraction, inspection, provision, reporting, presentation, and transmission of information. The third comprises *cognitive actions*, which involve the transformation or interpretation of information, including justification, argumentation, assigning meaning, highlighting salient elements, summarisation, analysis, simulation, clarification, and explanation generation. This branch of the framework is particularly important because it shows that interaction in XAI is not merely a passive transfer of information, but often involves active processes of organisation, interpretation, and sense-making.

5.4.2 Other interactions

A smaller set of semantic components concerns other forms of interaction. In particular, some definitions imply forms of *AI system self-interaction*, including autonomous actions such as detecting the need for explanation, adapting explanatory messages, or reflecting internal structural awareness. Others refer to interactions between the context and the AI system or human user.

Overall, the *Interaction* entity shows that interpretability and explainability are often structured not simply around information or system properties, but around dynamic exchanges involving communication, action, and reciprocal influence. This makes interaction one of the key organising entities of the framework.

5.5 Context

The *Context* entity captures the meta-level information that specifies how the other entities and their relations should be understood. In the analysed definitions, context does not usually appear as an isolated component, but rather as a set of conditions that shape the meaning, adequacy, and applicability of interpretability and explainability. For this reason, the *Context* entity functions as a meta-informative layer over the interaction between the *AI system* and the *Human user*.

Context can be understood as the set of situational and specifying information necessary to characterise the explanatory interaction between the other entities of the framework. This use of context is consistent with prior work in context-aware computing and XAI, where context is treated as the situational information that characterises entities, tasks, and interactions in a given setting (Chazette et al., 2021; Köhl et al., 2019; Longo et al., 2024).

5.5.1 Context dependence

A first distinction concerns whether a definition includes context-dependent elements. Some definitions do not explicitly consider contextual factors, while others make their

meaning explicitly dependent on the circumstances of application. This distinction is important because a part of the reviewed literature suggests that what counts as a satisfactory explanation depends on who the user is, what system is being considered, what task is being performed, and what broader environment or purpose frames the interaction. In this sense, context is not merely background information, but a condition that affects the explanatory interaction.

5.5.2 Contextual factors

Several recurring contextual factors can be identified. A first group concerns the *situational setting* of the interaction, that is, the general setup within which the system is deployed or used. This includes the domain of application, the problem being addressed, the task at hand, and the application context of the AI system. Notably, this group also includes the environment in which explanatory or interpretative processes occur, whether physical or virtual. Such factors matter because they can shape all relevant elements of the explanatory interaction.

A second group concerns the *specification of the entities involved*. On the side of the *AI system*, this includes the type of system under consideration and whether interpretability or explainability is assumed to depend on the specific model or to be model-agnostic. On the side of the *Human user*, contextual specification includes the audience group, such as end users, developers, stakeholders, or domain experts, as well as the distinction between technical and non-technical recipients. These distinctions are significant because they show that the same explanatory content may not be equally suitable across audiences, levels of expertise, or application domains.

The second group also relates to the *conditions of interaction* between the *AI system* and the *Human user*. These include the adaptation of explanations to different audiences, the accessibility of explanatory content, the presence of a shared conversational ground, and the purpose of the interaction itself. Context therefore shapes not only who receives an explanation, but also how that explanation should be framed, what level of detail it should include, and what function it is intended to serve, whether justificatory, diagnostic, instructional, or otherwise.

Finally, some contextual factors concern broader *operational conditions*. These include the need for external knowledge, such as domain knowledge, common knowledge, or machine learning expertise, as well as the constraints and operational conditions under which explanatory interaction is considered appropriate, reliable, or valid. We include here practical constraints such as time pressure and the stakes involved, together with operational limitations of AI systems, such as functioning only under predefined conditions or only when sufficiently confident.

Taken together, these components show that context is not an optional background variable, but a constitutive element in many definitions of interpretability and explainability. It determines how the system, the user, and their interaction are specified, and therefore plays a central role in the conceptual structure captured by the framework. By including context as a distinct entity, the framework is able to represent not only what is explained and to whom, but also the circumstances under which explanation becomes meaningful, useful, or appropriate.

5.6 Comparing interpretability and explainability within the framework

A central question raised by the present study is whether interpretability and explainability can be clearly separated within the proposed framework, or whether they are better understood as overlapping concepts with different emphases. To explore this issue, we analysed how definitions associated with each concept are distributed across the four entities of the framework and their main sub-dimensions. Because the corpus contains more definitions of explainability than of interpretability, we normalised the counts of semantic units within each dimension in order to compare the relative density of semantic content across the two concepts.

At the level of the four entities as a whole, interpretability is more strongly associated with the AI system, the Human user, and the Context, whereas explainability is more strongly associated with the Interaction entity. This is consistent with a broader pattern in the literature: interpretability is linked more often to model properties and understanding, while explainability is linked more often to communication and explanatory action, such as provision of reasons.

Within the AI system entity, interpretability is more strongly associated with architecture and data. With regard to the model's inference process, interpretability is more often connected to inner working and input-output mapping, whereas explainability is more often associated with decision-making and output. This suggests that interpretability is more frequently framed in terms of understanding internal structure and functional behaviour, while explainability is more frequently framed in terms of making decisions and outcomes intelligible.

Within the Human user entity, interpretability is more closely associated with epistemic and alignment objectives, whereas explainability is more closely associated with governance objectives and, to a lesser extent, ethical objectives. This indicates that interpretability is more often linked to understanding and cognitive alignment, while explainability is more often linked to oversight, accountability, and practical scrutiny.

Within the Interaction entity, and especially within the AI system-human user interaction branch, explainability is more strongly represented in the actions sub-dimension, including both informative and cognitive actions. This supports the view that explainability is more explicitly tied to the active production, communication, and use of explanatory content.

At the same time, the analysis does not support a sharp separation between the two concepts. Semantic units associated with interpretability and explainability overlap across all four entities and across most major sub-dimensions. The framework therefore suggests that the two concepts are not cleanly separable in the literature as it currently stands, but occupy a shared conceptual space within which different authors place emphasis on different aspects.

This comparison also helps explain the persistence of fragmentation in the literature. Many definitions capture only partial aspects of a broader multidimensional conceptual structure. Some focus on model-centred aspects, others on user-centred goals, communicative processes, or contextual conditions. The proposed framework

does not eliminate these differences, but provides a common structure within which they can be located and understood.

5.7 Taxonomical dimensions of definitions

In addition to the conceptual framework presented above, the analysis also supports taxonomical dimensions of interpretability and explainability. Whereas the framework organises the recurring conceptual content of the definitions, the taxonomical dimensions can be used to classify the definitions themselves. These dimensions capture not only what definitions refer to, but also how they are formulated and under what assumptions they operate.

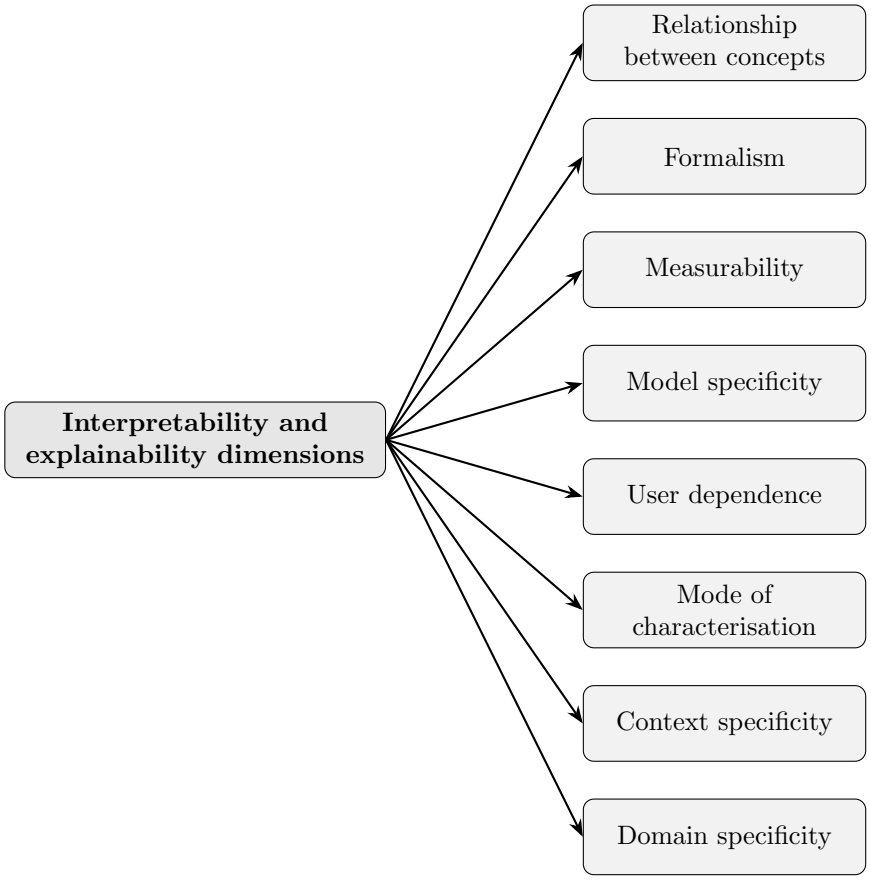


Fig. 4 Taxonomical dimensions of interpretability and explainability

The main taxonomical dimensions identified in this study are summarised below and visualised in Fig. 4.

- **Relationship between concepts.** This dimension captures how a definition positions interpretability and explainability with respect to one another. Definitions may treat them as synonyms, as related but distinct concepts, or as hierarchical concepts in which one subsumes the other.
- **Formalism.** This dimension concerns the language or representational form in which the definition is expressed. Definitions may be stated in natural language, or may adopt more formal mathematical, statistical, logical, or causal formulations.
- **Measurability.** This dimension captures whether a definition is associated with explicit quantification, a metric, or a set of measurable criteria.
- **Model specificity.** This dimension concerns whether the definition is tied to a particular model class or family, or whether it is intended to be model-agnostic.
- **User dependence.** This dimension captures whether the definition is presented as objective and general, or whether it depends on subjective, user-related, or audience-dependent factors.
- **Mode of characterisation.** This dimension concerns whether interpretability or explainability is treated as an intrinsic and passive property of the system, or as something actively produced, mediated, or achieved through additional means.
- **Context specificity.** This dimension captures whether the definition is intended to hold independently of context, or whether it is explicitly tied to a context of use.
- **Domain specificity.** This dimension concerns whether the definition is presented as general across domains, or as dependent on a particular application area.

These dimensions show that definitions of interpretability and explainability differ not only in conceptual content, but also in their degree of generality, formalisation, measurability, and dependence on models, users, contexts, and domains. The taxonomical dimensions therefore complement the framework by providing a structured basis for comparing definitions at a meta-conceptual level.

6 Discussion

6.1 Interpreting the framework

The proposed framework abstracts the main conceptual patterns emerging from the literature rather than attempting to enumerate every possible aspect of interpretability and explainability. Its value lies in making recurring dimensions visible and comparable despite the heterogeneity and incompleteness of the field.

Several broad patterns stand out. Within the *AI system* entity, the strongest concentration of semantic material lies in the *model*, especially the *inference process*. This is consistent with the historical development of both interpretability and explainability as responses to the black-box problem, and confirms that much of the literature still associates these concepts with understanding how outputs are produced and decisions are reached.

Within the *Human user* entity, the most prominent subcomponent is *objectives*, particularly *epistemic objectives* and, above all, *understanding*. This suggests that the literature continues to frame interpretability and explainability primarily in relation to human comprehension, even where operational, ethical, or governance-oriented

concerns are also present. At the same time, the prominence of individual capabilities shows increasing recognition that users differ in background, abilities, and expectations.

The *Interaction* and *Context* entities extend this picture beyond model-centred accounts. The former highlights that interpretability and explainability are often articulated through communicative and explanatory processes between an *AI system* and a *Human user*, while the latter shows the importance of audience adaptation, accessibility, and situational adequacy. Together, these dimensions reinforce the view that interpretability and explainability cannot be treated as self-contained properties of systems alone, but are shaped by the circumstances in which they are requested, delivered, and used.

A further result is the asymmetry of information flow. Although the framework allows for interactivity as a two-way process, the analysed literature mostly represents explanation as information moving from the AI system to the human user. This suggests that, despite frequent rhetorical emphasis on interaction, much of the field still conceptualises interpretability and explainability in predominantly one-directional terms. More broadly, the uneven distribution of semantic material across the framework is itself informative, as it reveals which aspects of interpretability and explainability have been most strongly developed in the literature and which remain comparatively underexplored.

6.2 Implications for human-centred XAI design

One important implication of the framework is that it supports a more human-centred and goal-oriented approach to XAI design. The *Context* entity shows that explanatory design should begin from the setting in which explanation is required, including the domain, task, audience, and operational conditions. The *Human user* entity highlights that explanations should be shaped by the objectives, capabilities, prior knowledge, and needs of the intended users. The *AI system* entity indicates that different explanatory settings may require attention to different aspects of the system, depending on what is most relevant to users and purposes. The *Interaction* entity shows that explanation should also be designed as a communicative process, involving appropriate representational formats, scope, and forms of exchange rather than being treated as a generic output.

Taken together, these four components support a view of XAI design in which explanation is shaped jointly by system properties, user needs, interactional form, and contextual constraints. This is consistent with growing emphasis in the literature on contextual adaptation, stakeholder sensitivity, and user-oriented explainability (Chakraborty et al., 2017; Chazette et al., 2021; Langer et al., 2021; Rudin et al., 2022; Tomsett et al., 2018).

6.3 Comparison with selected existing frameworks

To assess the scope of the proposed framework, it is useful to compare it with selected conceptual models from the literature. The aim is not to claim strict equivalence, but to examine whether the framework is expressive enough to accommodate important

existing perspectives and whether it provides a broader structure within which they can be situated. Table 1 summarises this comparison.

Table 1 Comparison of the proposed framework with selected existing conceptual models

Framework	Main focus	Where it maps in our framework	How our framework extends it
Alonso et al. (2015)	Hierarchical account of interpretability for fuzzy systems, with quantitative indices for different system components	Mainly within *AI system*, especially *Model*, including architectural structure and operating principles	Generalises beyond fuzzy systems and provides a broader conceptual structure rather than a domain-specific measured hierarchy
Cabitza et al. (2023)	Explanation typology based on a tripartite structure of *explanandum*, *explanans*, and *explanatory relation*	Distributed across *AI system*, *Interaction*, and *Human user*, especially explanatory content, communicational dimensions, decision-making, output, and user-oriented objectives	Situates structured accounts of explanation within a wider interactional and contextual framework for interpretability and explainability
Tomsett et al. (2018)	Role-based account of interpretability, linking different stakeholder roles to different interpretability needs	Mainly within *Human user* and *Context*, especially audience groups, objectives, individual factors, and audience adaptation	Embeds role-based variation within a broader framework that also includes the *AI system*, explanatory *Interaction*, and surrounding *Context*

These comparisons suggest that the framework is broad enough to accommodate hierarchical, role-based, and explanation-centred models while also providing a higher-level architecture within which they can be related. In this sense, it does not replace existing models, but offers a more integrated conceptual structure for situating them alongside one another.

6.4 Implications for research and practice

The present study contributes to the conceptual clarification of interpretability and explainability by reconstructing their definitional landscape into a more explicit and structured form. Methodologically, it shows that semantic analysis of published definitions can be used to recover conceptual structure in a fragmented research area. Conceptually, it introduces a framework organised around the entities AI system, Human user, Interaction, and Context, together with a taxonomical layer that captures higher-level differences in how definitions are formulated. The framework further supports the view that interpretability and explainability can be meaningfully unified through an interactive scenario that captures and generalises the main definitional patterns identified in the literature.

For researchers, this provides a clearer basis for comparing definitions, locating disagreements, and understanding why fragmentation has persisted in the field. It

also suggests that interpretability and explainability are best analysed through the elements and relations that constitute this unified interactive perspective. For practitioners, the framework offers a more structured way of reasoning about XAI design by drawing attention not only to model-centred aspects, but also to user objectives, communicative form, audience adaptation, and contextual constraints. More broadly, the same methodological approach could be extended to neighbouring concepts such as transparency, intelligibility, explicability, and trustworthiness.

6.5 Limitations of the current study

The present study has several limitations. First, the boundary between definitions and related notions such as requirements, desiderata, properties, and constraints is not always clear-cut in the literature. Although the review applied explicit inclusion criteria, some borderline cases still required judgement during selection and interpretation. Second, the construction of the framework involved interpretive choices in the decomposition and mapping of semantic units. Some terms were ambiguous, overlapped conceptually, or could plausibly have been assigned to more than one part of the framework. In such cases, conservative choices were made to preserve coherence and limit redundancy, but some degree of arbitrariness cannot be excluded entirely.

A further limitation concerns scope. The study focuses specifically on interpretability and explainability, which are central to XAI but do not exhaust the wider conceptual space that also includes transparency, explicability, intelligibility, understandability, and trustworthiness. Likewise, although the corpus used here is substantial, it is not claimed to be exhaustive. The framework should therefore be understood as a structured synthesis of a large and representative body of definitional material rather than as the final possible account of the domain.

Finally, the contribution remains conceptual rather than operational. The framework is intentionally semi-formal: it is designed to support communication, comparison, and conceptual clarification rather than to function as a fully formal ontology or measurement system. It does not yet include explicit operationalisation procedures, quality criteria for definitions, or associated metrics, and some modelling choices remain provisional. In particular, while the treatment of explanatory interaction as broadly communicative and interactional is theoretically well motivated, the paper does not claim that any single communication model has been conclusively established as definitive for XAI. Further refinement and comparative validation against additional conceptual models therefore remain important directions for future work.

7 Conclusion

This paper examined how interpretability and explainability are defined in AI and argued that the field still lacks a stable conceptual foundation. Through a systematic review of published definitions and a semantic decomposition of their content, we developed a framework organised around four recurring entities: AI system, Human user, Interaction, and Context. The framework provides a unified conceptual structure for interpretability and explainability.

The framework shows that the literature does not support a clean separation between interpretability and explainability. Instead, the two concepts overlap within a shared conceptual space while differing in the aspects they emphasise, especially system properties, user goals, interactional processes, and contextual dependence. The taxonomical dimensions derived from the review complement this framework by making it easier to compare how definitions are formulated. The value of the study is not that it imposes a final universal definition, but that it provides a clearer structure for comparison, critique, and future refinement.

Several directions for further research follow from the present work. First, developing quality criteria for definitions would allow future research to move beyond reconstruction towards comparative evaluation. Second, the analysis could be extended to related concepts such as transparency, explicability, intelligibility, understandability, and trustworthiness, to assess whether a broader unified structure for XAI terminology is possible. Third, requirements, desiderata, properties, and constraints, which were excluded as definitions in this work, should be incorporated to produce a fuller theoretical account. Finally, future work could explore how the framework might be operationalised through metrics and an automated process for evaluating AI systems (Ahmad, Baryannis, & Hill, 2024) against its dimensions.

Supplementary information. The following supplementary material are provided:

- Online Resource 1: Systematic literature review corpus
- Online Resource 2: Descriptive table linking papers, definitions and framework components
- Online Resource 3: Complete conceptual framework hierarchy

Declarations

7.1 Funding

No funding was received to assist with the preparation of this manuscript.

7.2 Competing interests

The authors have no relevant financial or non-financial interests to disclose.

7.3 Ethics approval and consent to participate

Not applicable.

7.4 Consent for publication

Not applicable.

7.5 Data availability

The complete systematic literature review corpus that underpins the research in this paper is available as a spreadsheet (Online Resource 1). The complete descriptive table

linking papers, definitions and framework components is also available as a spreadsheet (Online Resource 2).

7.6 Materials availability

Not applicable.

7.7 Code availability

Not applicable.

Appendix A Illustrative example of semantic mapping

To illustrate the reasoning process behind the construction of the framework, we consider the definitions of *interpretability* and *explainability* reported by Arrieta et al. (2020), which are presented there as summaries of positions found in the literature.

Interpretability is described as the "ability to explain or to provide the meaning in understandable terms to a human" and as something that "comes from the design of the model itself". In our framework, expressions such as "to explain" and "to provide meaning" are treated as interactive actions and mapped to the *Interaction* entity, specifically to cognitive explanatory actions. The reference to information provided "in understandable terms" is linked to communicational accessibility and to user-oriented epistemic goals, since it concerns the human comprehension of the explanatory content. The claim that interpretability arises from the design of the model itself is also relevant to the taxonomical dimension concerning the mode of characterisation, since it treats interpretability as an intrinsic or passive property of the system.

Explainability is defined by Arrieta et al. (2020) as the property of an AI system that, "given an audience", produces "details or reasons" to make its functioning clear or easy to understand. In our framework, the production of details or reasons and the clarification are again mapped to explanatory actions within the *Interaction* entity. The system's "functioning" is associated with the *AI system* entity, especially the subdimension of inner working. The emphasis on making the system easy to understand relates to epistemic objectives within the *Human user* entity and the accessibility within the *Context*. Finally, the explicit mention of an audience introduces a contextual specification of the recipient of the explanation, which is captured in the *Context* entity under audience identification.

This example shows how individual definitions were decomposed into semantic units and then reorganised within the framework. The goal of the procedure is not to reproduce surface wording mechanically, but to identify the recurring conceptual elements that allow definitions to be compared, grouped, and analysed within a common structure.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable

artificial intelligence (xai). *IEEE access*, *6*, 52138–52160, https://doi.org/10.1109/access.2018.2870052

Ahmad, M.A., Baryannis, G., Hill, R. (2024). Defining complex adaptive systems: An algorithmic approach. *Systems*, *12*(2), , https://doi.org/10.3390/systems12020045

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., ... Herrera, F. (2023). Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information fusion*, *99*, 101805, https://doi.org/10.1016/j.inffus.2023.101805

Alonso, J.M., Castiello, C., Mencar, C. (2015). Interpretability of fuzzy systems: Current research trends and prospects. *Springer handbook of computational intelligence*, 219–237, https://doi.org/10.1007/978-3-662-43505-2_14

Antoniou, G., Papadakis, E., Baryannis, G. (2022). Mental health diagnosis: A case for explainable artificial intelligence. *International Journal on Artificial Intelligence Tools*, *31*(03), 2241003, https://doi.org/10.1142/S0218213022410032

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... others (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, *58*, 82–115, https://doi.org/10.1016/j.inffus.2019.12.012

Baryannis, G., Tachmazidis, I., Batsakis, S., Antoniou, G., Alviano, M., Sellis, T., Tsai, P.-W. (2018). A Trajectory Calculus for Qualitative Spatial Reasoning Using Answer Set Programming. *Theory and Practice of Logic Programming*, *18*(3-4), 355-–371, https://doi.org/10.1017/S147106841800011X

Bellucci, M., Delestre, N., Malandain, N., Zanni-Merk, C. (2021). Towards a terminology for a fully contextualized xai. *Procedia Computer Science*, *192*, 241–250, https://doi.org/10.1016/j.procs.2021.08.025

Bibal, A., Lognoul, M., De Streel, A., Frénay, B. (2021). Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, *29*(2), 149–169, https://doi.org/10.1007/s10506-020-09270-4

Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., Holzinger, A. (2023). Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable ai. *Expert systems with Applications*, *213*, 118888, https://doi.org/10.1016/j.eswa.2022.118888

Carvalho, D.V., Pereira, E.M., Cardoso, J.S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832, https://doi.org/10.3390/electronics8080832

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature News*, *538*(7623), 20, https://doi.org/10.1038/538020a

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., ... others (2017). Interpretability of deep learning models: A survey of results. *2017 ieee smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (smartworld/scalcom/uic/atc/cbdcom/iop/sci)* (pp. 1–6).

Chazette, L., Brunotte, W., Speith, T. (2021). Exploring explainability: a definition, a model, and a knowledge catalogue. *2021 ieee 29th international requirements engineering conference (re)* (pp. 197–208).

Chen, C., & Eickhoff, C. (2024). Evaluating search system explainability with psychometrics and crowdsourcing. *Proceedings of the 47th international acm sigir conference on research and development in information retrieval* (pp. 1051–1061).

Clinciu, M.A., & Hastie, H.F. (2019). A survey of explainable ai terminology. *1st workshop on interactive natural language technology for explainable artificial intelligence 2019* (pp. 8–13).

De Bruijn, H., Warnier, M., Janssen, M. (2022). The perils and pitfalls of explainable ai: Strategies for explaining algorithmic decision-making. *Government information quarterly*, *39*(2), 101666, https://doi.org/10.1016/j.giq.2021.101666

Dhurandhar, A., Iyengar, V., Luss, R., Shanmugam, K. (2017). A formal framework to characterize interpretability of procedures. *arXiv preprint arXiv:1707.03886*, , https://doi.org/10.48550/arXiv.1707.03886

Dib, L. (2024). Formal definition of interpretability and explainability in xai. *Intelligent systems conference* (pp. 133–151).

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, , https://doi.org/10.48550/arXiv.1702.08608

Fox, M., Long, D., Magazzeni, D. (2017). Explainable planning. *arXiv preprint arXiv:1709.10256*, , https://doi.org/10.48550/arXiv.1709.10256

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 ieee 5th international conference on data science and advanced analytics (dsaa)* (pp. 80–89).

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J.P., Yordanova, K., Vered, M., ... others (2023). A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. *Artificial intelligence review*, *56*(4), 3473–3504, https://doi.org/10.1007/s10462-022-10256-8

Grice, H.P. (1975). Logic and conversation. *Speech acts* (pp. 41–58). Brill.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42, https://doi.org/10.1145/3236009

Habiba, U.-e., Habib, M.K., Bogner, J., Fritzsch, J., Wagner, S. (2025). How do ml practitioners perceive explainability? an interview study of practices and challenges. *Empirical Software Engineering*, *30*(1), 18, https://doi.org/10.1007/s10664-024-10565-2

Hilton, D.J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, *107*(1), 65, https://doi.org/10.1037//0033-2909.107.1.65

Holzinger, A., Zatloukal, K., Müller, H. (2024). *Is human oversight to ai systems still possible?* Elsevier.

Jaakkola, R., Janhunen, T., Kuusisto, A., Rankooh, M.F., Vilander, M. (2025). Explainability via short formulas: the case of propositional logic with implementation. *Journal of Artificial Intelligence Research*, *83*, , https://doi.org/10.1613/jair.1.17422

Kaminski, M.E., & Malgieri, G. (2025). The right to explanation in the ai act. *Available at SSRN 5194301*, ,

Köhl, M.A., Baum, K., Langer, M., Oster, D., Speith, T., Bohlender, D. (2019). Explainability as a non-functional requirement. *2019 ieee 27th international requirements engineering conference (re)* (pp. 363–368).

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... Baum, K. (2021). What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial intelligence*, *296*, 103473, https://doi.org/10.1016/j.artint.2021.103473

Lipton, Z.C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57, https://doi.org/10.1145/3236386.3241340

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., ... others (2024). Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, *106*, 102301, https://doi.org/10.1016/j.inffus.2024.102301

Lundberg, S.M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st international conference on neural information processing systems* (p. 4768–4777). Red Hook, NY, USA: Curran Associates Inc.

Malik, S., Islam, S.M.R., Akram, T., Naqvi, S.R., Alghamdi, N.S., Baryannis, G. (2022). A novel hybrid meta-heuristic contrast stretching technique for improved skin lesion segmentation. *Computers in Biology and Medicine*, *151*, 106222, https://doi.org/10.1016/j.compbiomed.2022.106222

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1–38, https://doi.org/10.1016/j.artint.2018.07.007

Miller, T., Howe, P., Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, , https://doi.org/10.48550/arXiv.1712.00547

Montavon, G., Samek, W., Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital signal processing*, *73*, 1–15, https://doi.org/10.1016/j.dsp.2017.10.011

Mucha, H., Robert, S., Breitschwerdt, R., Fellmann, M. (2021). Interfaces for explanations in human-ai interaction: proposing a design evaluation approach. *Extended abstracts of the 2021 chi conference on human factors in computing systems* (pp. 1–6).

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., ... Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, *55*(13s), 1–42, https://doi.org/10.1145/3583558

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: a context-sensitive framework for explainable medical ai. *Ethics and information technology*, *24*(1), 13, https://doi.org/10.1007/s10676-022-09632-3

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*, , https://doi.org/10.1136/bmj.n71

Palacio, S., Lucieri, A., Munir, M., Ahmed, S., Hees, J., Dengel, A. (2021). Xai handbook: towards a unified framework for explainable ai. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 3766–3775).

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60, https://doi.org/10.1145/3241036

Picard, A., Mualla, Y., Gechter, F., Galland, S. (2023). Human-computer interaction and explainability: Intersection and terminology. *World conference on explainable artificial intelligence* (pp. 214–236).

Ploug, T., & Holm, S. (2020). The four dimensions of contestable ai diagnostics-a patient-centric approach to explainable ai. *Artificial intelligence in medicine*, *107*, 101901, https://doi.org/10.1016/j.artmed.2020.101901

Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "why should i trust you?" explaining the predictions of any classifier. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribera, M., & Lapedriza, À. (2019). Can we do better explanations? A proposal of user-centered explainable AI. C. Trattner, D. Parra, & N. Riche (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-WS.org. Retrieved from https://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf

Roundtree, A.K. (2023). Ai explainability, interpretability, fairness, and privacy: an integrative review of reviews. *International conference on human-computer interaction* (pp. 305–317).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, *1*(5), 206–215, https://doi.org/10.1038/s42256-019-0048-x

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, *16*, 1–85, https://doi.org/10.1214/21-ss133

Schneeberger, D., Röttger, R., Cabitza, F., Campagner, A., Plass, M., Müller, H., Holzinger, A. (2023). The tower of babel in explainable artificial intelligence (xai). *International cross-domain conference for machine learning and knowledge extraction* (pp. 65–81).

Swartout, W.R., & Moore, J.D. (1993). Explanation in second generation expert systems. *Second generation expert systems* (pp. 543–585). Springer.

Theis, S., Jentzsch, S., Deligiannaki, F., Berro, C., Raulf, A.P., Bruder, C. (2023). Requirements for explainability and acceptance of artificial intelligence in collaborative work. *International conference on human-computer interaction* (pp. 355–380).

Tomsett, R., Braines, D., Harborne, D., Preece, A., Chakraborty, S. (2018). Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, , https://doi.org/10.48550/arXiv.1806.07552

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106, https://doi.org/10.1016/j.inffus.2021.05.009

Von Eschenbach, W.J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & technology*, *34*(4), 1607–1622, https://doi.org/10.1007/s13347-021-00477-0

Zablocki, É., Ben-Younes, H., Pérez, P., Cord, M. (2022). Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision*, *130*(10), 2425–2452, https://doi.org/10.1007/s11263-022-01657-x

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & technology*, *34*(2), 265–288, https://doi.org/10.1007/s13347-019-00382-7

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.xlsx
- ESM2.xlsx
- ESM3.pdf