Ethics and Information Technology (2024) 26:40
https://doi.org/10.1007/s10676-024-09769-3

ORIGINAL PAPER

Conceptualizing understanding in explainable artificial intelligence (XAI): an abilities-based approach

Timo Speith[1,2] · Barnaby Crook[1] · Sara Mann[3] · Astrid Schomäcker[1] · Markus Langer[4]

Accepted: 20 April 2024 / Published online: 15 June 2024

Abstract

A central goal of research in explainable artificial intelligence (XAI) is to facilitate human understanding. However, understanding is an elusive concept that is difficult to target. In this paper, we argue that a useful way to conceptualize understanding within the realm of XAI is via certain human *abilities*. We present four criteria for a useful conceptualization of understanding in XAI and show that these are fulfilled by an abilities-based approach: First, thinking about understanding in terms of specific abilities is motivated by research from numerous disciplines involved in XAI. Second, an abilities-based approach is highly versatile and can capture different forms of understanding important in XAI application contexts. Third, abilities can be operationalized for empirical studies. Fourth, abilities can be used to clarify the link between explainability, understanding, and societal desiderata concerning AI, like fairness and trustworthiness. Conceptualizing understanding as abilities can therefore support interdisciplinary collaboration among XAI researchers, provide practical benefit across diverse XAI application contexts, facilitate the development and evaluation of explainability approaches, and contribute to satisfying the societal desiderata of different stakeholders concerning AI systems.

Keywords Explainability · Explainable AI · XAI · Understanding · Abilities · Evaluation · Conceptualization

Introduction

Many artificial intelligence (AI) systems remain opaque to the stakeholders who interact with them (Burrell, 2016; Mann et al., 2023). This hinders the fulfillment of societal desiderata such as trust, fairness, and accountability (Langer et al., 2021c). For example, when companies use AI systems for hiring decisions, it can be problematic if those systems remain a black box to the human resource managers who use them or the applicants who are subject to their decisions. In light of these concerns, it is increasingly recognized that an ethically acceptable integration of AI into society will require humans to *understand* AI systems. To address this issue, the growing, interdisciplinary research field of explainable AI (XAI) tries to render AI systems understandable (Barredo Arrieta et al., 2020; Langer et al., 2021c; Páez, 2019).[1]

✉ Timo Speith
timo.speith@uni-bayreuth.de

Barnaby Crook
barnaby.crook@uni-bayreuth.de

Sara Mann
sara.mann@tu-dortmund.de

Astrid Schomäcker
astrid.schomaecker@uni-bayreuth.de

Markus Langer
markus.langer@psychologie.uni-freiburg.de

1 Chair for Philosophy, Computer Science and Artificial Intelligence, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Bavaria, Germany

2 Center for Perspicuous Computing, Saarland University, Campus, 66123 Saarbrücken, Saarland, Germany

3 Department of Philosophy and Political Sciences, TU Dortmund, Emil-Figge-Straße 50, 44227 Dortmund, North Rhine-Westphalia, Germany

4 Department of Psychology, University of Freiburg, Engelbergerstraße 41, 79085 Freiburg im Breisgau, Baden-Württemberg, Germany

[1] Note that we interpret XAI broadly, as the interdisciplinary research field addressing epistemic challenges arising from the opacity of AI systems. This contrasts with a narrow construal of XAI as the development of post-hoc explanation methods (see, e.g., Rudin, 2019).

However, a roadblock to progress in XAI is a lack of clarity about what it means to understand an AI system. This creates several problems. First, as an interdisciplinary endeavor (Langer et al., 2021a, 2021c; Páez, 2019; Miller et al., 2017; Lipton, 2018), XAI requires collaboration across disciplines. When ethicists and legal scholars demand that we need to understand AI systems to ensure fairness or legal liability (e.g., Vredenburgh, 2022; Deeks, 2019) or when computer scientists present a new method to enable better understanding of AI systems (e.g., Ribeiro et al., 2018; Lapuschkin et al., 2016), these parties need to agree on what they mean when they refer to "understanding" AI.

Second, appeals to "sufficient understanding" of AI-based systems (as in Article 14 of the proposal for a European AI Act[2]) do not reflect how requirements for understanding vary depending on the specific application context (e.g., on the *stakeholder* who aims to understand AI, the *system-related aspect*[3] they want to understand, and the *desideratum* which motivates their need for understanding). For example, in the case of an AI hiring system, an applicant may want to understand why their application was rejected (Wachter et al., 2017), while the developer may need to understand whether the system is fair (Hutchinson & Mitchell, 2019).

Third, understanding is often construed as a cognitive concept that can be analyzed in terms of an agent's internal states (see, e.g., Wilkenfeld, 2013). However, accounts that are not closely associated with empirically testable measures remain an elusive target for the design and evaluation of explainability approaches.

Fourth, although there is widespread agreement that a greater understanding of a system will be beneficial (Barredo Arrieta et al., 2020; Chazette et al., 2021; Langer et al., 2021a, 2021c; Speith, 2022a; Hoffman et al., 2018), it remains unclear how, exactly, the concept of understanding relates to the satisfaction of societal desiderata.

In this paper, we propose that advancing research on XAI requires a conceptualization of understanding that is tailored to the specific needs of the field. To this end, we argue for explicating different ways of understanding[4] an AI system in terms of specific understanding-related *abilities*[5] (e.g., *assessing, predicting*) a person may have with regard to that system. We believe that an abilities-based conceptualization of understanding has the potential to overcome the aforementioned problems.

The structure of this paper is as follows: Drawing on the specific needs of XAI outlined above, we begin by establishing four criteria that a useful conceptualization of understanding in XAI should fulfill: It must be (i) motivated by prior research across relevant disciplines, (ii) versatile, (iii) operationalizable, and (iv) establish a link to the satisfaction of relevant societal desiderata. The rest of the paper is devoted to showing that *abilities* fulfill these criteria to a high degree. First, we demonstrate that abilities unite interdisciplinary research by documenting their importance for accounts of understanding in different disciplines. Then, we argue that abilities are versatile by presenting six clusters of abilities that capture different ways of understanding required for diverse XAI application contexts. Subsequently, we show that our proposed ability clusters are operationalizable, as reflected by prior research using abilities to evaluate XAI methods. Next, we show how abilities can clarify the relationship between understanding and the satisfaction of societal desiderata. Finally, we address potential limitations of an abilities-based approach to understanding and conclude.

Criteria for a useful conceptualization of understanding in XAI

In philosophy, there is an extensive debate about what exactly constitutes understanding (Baumberger et al., 2017; Hannon, 2021). Similarly, different models of what it is to understand have been proposed across various disciplines (Pearl & Mackenzie, 2018; Endsley, 1995; Thórisson et al., 2016; Bloom et al., 1956). What is lacking, however, is a shared notion of understanding that is tailored to XAI. In this paper, we provide a conceptualization[6] of understanding that is useful specifically for the debate on explainability, and that allows for fruitful interdisciplinary research advancing the field. Based on our knowledge of XAI, we specify four criteria below which a conceptualization of understanding should fulfill in order to serve that purpose.[7]

[2] https://artificialintelligenceact.eu/the-act/

[3] In the XAI debate, there are many possibilities for the exact aspect that is supposed to be understood (for an overview, see Chazette et al., 2021). For our argument, however, it is only important that the understanding is concerned with some aspect of a system, that is, a *system-related aspect*.

[4] We speak of *ways* or *forms* of understanding in order to remain neutral regarding debates on different types and degrees of understanding. For discussion on these issues, see, e.g., Baumberger (2014) and Baumberger et al. (2017).

[5] Unless further qualified, we will use "abilities" in the following to mean specific abilities related to understanding.

[6] By *conceptualization* we mean a specific account of a concept that may not capture every aspect of it but serves a particular research purpose.

[7] Our criteria bear resemblance to those used in conceptual engineering research (v. Carnap, 1962; Brun, 2016). In conceptual engineering, a new term is introduced to serve a specific theoretical purpose. Well-known criteria for assessing the new term are similarity to the previous term (our Criterion 1) and fruitfulness for further research (our Criteria 2–4).

Criterion 1: interdisciplinary motivation

First, a useful conceptualization of understanding in XAI should be *motivated across disciplines*. The research field of XAI is highly interdisciplinary, stretching from computer science via psychology and law to philosophy and the social sciences. Such interdisciplinary endeavors come at the risk of producing misunderstandings as different disciplines use concepts according to their particular needs. However, the general development of research on XAI as well as the development and assessment of individual approaches requires a smooth collaboration between the fields (Langer et al., 2021a, 2021c; Speith, 2022b). Therefore, it is crucial to align the use of important concepts.

In light of the interdisciplinary nature of XAI, insisting on a conceptualization of understanding that is particular to any one discipline would be misguided (v. Miller et al., 2017; Miller, 2019). Instead, it is preferable to conceptualize understanding in a way that draws from multiple disciplines and, consequently, is plausible to most researchers in the XAI debate. We thus aim to find a conceptualization that unites research on the concept of understanding from different disciplines relevant for XAI.

Criterion 2: versatility

The next property of a useful conceptualization is *versatility*. We are looking for a conceptualization that is sensitive to the variety of understanding-related requirements of concrete XAI application contexts. We thus need a conceptualization of understanding that allows for differentiating the ways in which understanding can manifest.

To illustrate, we refer to the following example throughout the paper:

> **Hiring Case Example**
> An AI system is used to support hiring decisions (Hickman et al., 2022; Langer et al., 2021b; Baum et al., 2022). As input, it takes application data such as résumés or results from cognitive ability tests. It then outputs a ranking of applicants. A hiring manager plays the role of a human-in-the-loop and decides, with the help of this ranking, who should be invited to interview for a position. Explainability is relevant for such a system, for instance, to ensure the system's outputs do not discriminate unfairly against any protected groups (e.g., women, people of color, LGBTQIA*; Carvalho et al., 2019; Langer & König, 2023).

Based on this example, we can think of several XAI application contexts in which different forms of understanding might be important: The applicants might want to understand why their application was rejected (i.e., they need to understand a single output of the program). Alternatively, if we want to hold the hiring manager accountable for evaluating applicants based on the system's output, the manager might need to understand how the system uses information about applicants to determine hirability (Baum et al., 2022). Finally, a system developer, who wants to improve the system, needs to understand how the system's decision making is implemented in the code.

Clearly, then, various forms of understanding are required in different XAI application contexts. Therefore, it is important to rely on a conceptualization of understanding that is able to accommodate the distinctions between these forms. Making such distinctions allows for greater precision and specificity when thinking about what is needed in each case. This should help researchers and practitioners spell out concrete requirements for the evaluation and design of explainability approaches by focusing attention on the form of understanding relevant for the specific application context.

Criterion 3: operationalizability

Third, for the field of XAI to advance, it needs to adopt a conceptualization of understanding that can be empirically evaluated. An important objective within XAI research is to assess and compare the success of different explainability approaches at inducing understanding in different contexts. To that end, the effects of an approach on different stakeholders with varying background knowledge and in diverse contexts should be measured in empirical studies. This requires a conceptualization of understanding that allows for measurement of whether an explainability approach facilitates the required understanding of an AI system.

Not all conceptualizations of understanding immediately offer themselves to such an operationalization. For example, some philosophical accounts of understanding focus on purely cognitive aspects of understanding like the manipulation of mental representations (v. Wilkenfeld, 2013).[8] For pragmatic reasons, a conceptualization of understanding should be relatable to observable human behavior.

Criterion 4: link to desiderata

The last property that we consider important for a useful conceptualization of understanding in XAI is that it can be used to specify the relationship between explainability,

[8] As we are not looking to give a definition of understanding, we do not want to exclude that these accounts point to some important aspects of the nature of understanding in general.

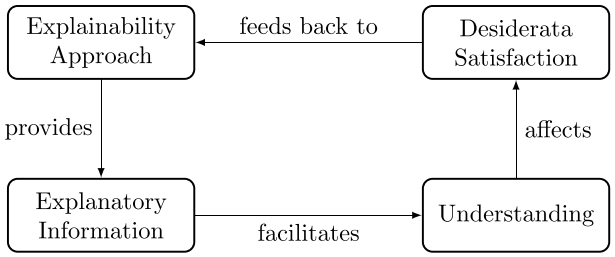

Fig. 1 A simplified version of the explainability models that Langer et al. (2021c) and Hoffman et al. (2018) have proposed

understanding, and desiderata. Understanding is usually not a goal in itself for XAI, but we rather want to understand AI to ensure, e.g., fairness, accountability, or human oversight (Krishnan, 2020; Barredo Arrieta et al., 2020; Chazette et al., 2021; Langer et al., 2021a, 2021c; Speith, 2022a, 2022b; Hoffman et al., 2018).

In particular, Hoffman et al. (2018) and Langer et al. (2021c) propose similar models that outline how key concepts in XAI relate to one another (see Fig. 1). According to their models, explainability approaches provide explanatory information with the aim of facilitating people's understanding. This understanding, in turn, affects the satisfaction of desiderata. Depending on the final status of desiderata satisfaction, the explainability approach may need to be adapted or a new one chosen.

Langer et al. (2021c) specify their model further by introducing the distinction between *epistemic* and *substantial* desiderata satisfaction. For substantial desiderata satisfaction, an AI system, or an element of the sociotechnical system in which it is embedded, has to have a certain property (e.g., the AI's outputs have to fulfill some mathematical fairness measure, or its user needs to be in a position to bear responsibility for its failure). For epistemic desiderata satisfaction, stakeholders require epistemic access to whether or not a desideratum is satisfied substantially (e.g., they have to know whether a system is fair). Understanding primarily leads to epistemic desiderata satisfaction. However, understanding can also support substantial desiderata satisfaction (e.g., when understanding the system supports debugging to make it fairer). Sometimes, understanding is even required for substantial desiderata satisfaction (e.g., human oversight of a system can only be ensured when certain people understand it).

In many cases, however, the exact relationship between explainability, understanding, and (epistemic or substantial) desiderata satisfaction is not clear (v. Kästner et al., 2021; Deck et al., 2024). Thus, a useful conceptualization of understanding should serve to clarify how, exactly, desiderata can be satisfied by explainability.

In the following sections, we will argue that conceptualizing understanding via abilities fulfills all of these criteria to a high degree.

Abilities are motivated by research across disciplines

Our first criterion requires that a useful conceptualization of understanding in XAI is both motivated by and fruitful for disciplines relevant to XAI. We demonstrate interdisciplinary plausibility of focusing on abilities by documenting convergence across various fields in utilizing abilities to analyze and operationalize understanding.

In *philosophy*, one of the central aspects of understanding is considered to be the *grasping* of relationships, e.g., between the elements of an explanation, a theory, or a scientific model (Kvanvig, 2009; Baumberger, 2014; Riggs, 2003; Grimm, 2011). Grasping, in turn, is usually spelled out in terms of distinct abilities someone who understands is taken to possess (Baumberger et al., 2017). For instance, Hills (2016) lists several abilities characteristic of *understanding-why* (e.g., understanding why a job applicant was rejected). These include the ability to follow an explanation, to explain in one's own words, or to draw conclusions about relevantly similar cases. Others have suggested further abilities, e.g., the ability to answer questions about counterfactual cases (Grimm, 2011), to make predictions (de Regt, 2015), to qualitatively solve problems (Newman, 2017), to construct (scientific) models (de Regt, 2015), or to evaluate competing explanations (Khalifa, 2013).

Abilities are also central to how understanding is envisioned and studied in *cognitive psychology*. For example, Williams and Lombrozo (2010) treated the ability to reason and generalize about how object properties relate to object categories as a measure of understanding (see also Williams et al., 2010). Relatedly, Rozenblit and Keil (2002) used the ability to produce diagrammatic explanations of everyday objects as a measure of understanding, showing that people were systematically overconfident in their self-assessed understanding. Finally, research in cognitive psychology suggests that our intuitive judgments of explanation and understanding may be ability-centered (v. Vasilyeva et al., 2015, 2017; Lombrozo & Carey, 2006).

In *educational psychology*, assessing students' abilities is foundational to measuring their understanding of educational content. The locus classicus of this line of research is *Bloom's Taxonomy* (Bloom et al., 1956), an influential attempt to systematize the wide variety of abilities that capture student understanding. Krathwohl (2002)'s revision of the taxonomy consists of six hierarchical levels with

associated abilities: *remembering*, *understanding*, *applying*, *analyzing*, *evaluating*, and *creating*.

In research on *AI and causal inference*, Pearl and Mackenzie (2018)'s *Ladder of Causation* relates understanding to possessing certain abilities. The ladder has three rungs: *association*, *intervention*, and *counterfactuals*. In another contribution from the field of AI, Thórisson et al. (2016) provide a pragmatic way of testing the understanding of any potentially intelligent agent. In this case, the authors associate understanding with an agent's possession of the following abilities (with respect to a target phenomenon): *predicting*, *achieving goals*, *explaining*, and *re-creating*.

Abilities are also used to operationalize understanding in the area of *human factors* research. For example, *situation awareness*, which is understanding of one's environment, concerns "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley, 1995, p. 36). Situation awareness has three hierarchical levels; in ascending order: *perception*, *comprehension*, and *projection*.

Overall, numerous disciplines consider abilities to be a means of capturing understanding. Further, there is a significant overlap between the abilities referred to by different disciplines. For example, abilities such as *generalizing*, *explaining*, and *counterfactual reasoning* are part of many conceptualizations of understanding. Additionally, many disciplines propose some kind of hierarchy of abilities, corresponding to increased degrees of understanding, according with the intuition that certain abilities are more demanding than others (Bloom et al., 1956; Pearl & Mackenzie, 2018). This convergence suggests that an abilities-based account of understanding is well-suited to serve as a unified conceptualization for interdisciplinary research in the field of XAI.

Despite these similarities, however, the various disciplines we have discussed present different accounts of which abilities are relevant to understanding. We take this to indicate that *which* abilities best reflect understanding can vary depending on the context. We suggest, therefore, that XAI application contexts require a tailored account of understanding-related abilities. To this end, we next present a systematization of abilities suited to capturing the different forms of understanding relevant to XAI application contexts.

Abilities are versatile

Our second criterion for a useful conceptualization of understanding in XAI demands that it should be versatile enough to capture the variety of forms of understanding that may be required in XAI application contexts. To this end, in this section, we integrate and systematize the insights gleaned from our interdisciplinary review and propose six clusters[9] of abilities (see Fig. 2) that reflect differing ways of understanding: *recognizing*, *assessing*, *predicting*, *intervening*, *explaining*, and *designing*. We intend these six clusters to reflect existing work and, crucially, to capture the ways of understanding that are relevant across XAI application contexts.

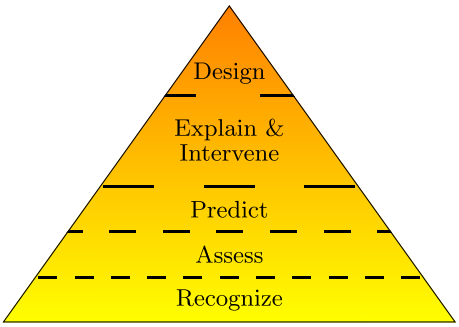

Fig. 2 Our proposed clusters of abilities relevant to characterizing different ways of understanding in XAI. The hierarchical structure reflects that some abilities tend to be more demanding than others. However, how demanding the acquisition of an ability is in a specific case is also influenced by contextual factors

Recognizing

A foundational ability cluster is *recognizing*. Recognizing is apprehending what is signified by a piece of information such as a text, an image, or a number. As such, in many cases, recognizing is an undemanding ability possessed, e.g., by any person reading a text consisting of familiar words in their own language. However, while recognizing may seem trivial, many representational formats common to the AI domain, such as vector representations, correlation matrices, and saliency maps, may be unrecognizable to laypeople, especially if they are presented without adequate labeling or accompanying description (Langer et al., 2021c; Franconeri et al., 2021; Speith, 2022a). In general, recognizing is crucial for basic use of AI systems and often serves as an important prerequisite for achieving more demanding forms of understanding.

For example, consider the hiring manager in our application scenario. To use the system in the first place, the manager needs to recognize the system's outputs as ascribing a certain rank to an applicant (e.g., that a woman named April is ranked 10th). This limited understanding is sufficient for a person to follow the recommendations of

[9] We call them "clusters" to reflect that they represent umbrella concepts whose concrete operationalization may change depending on the context. For the same reason, we entreat readers to focus on the substantive descriptions of the abilities, rather than their labels.

the decision support system, even if nothing else about the system is known. In this example, the output being presented in an accessible way would suffice to induce this ability. Insight into the inner workings of the system is not required.

Abilities similar to what we call recognizing can be found in many disciplines, always on a comparably low hierarchical level. For instance, *perceiving* in situational awareness (Endsley, 1995) and *remembering* in Bloom's Taxonomy (Krathwohl, 2002) are close to what we mean by recognizing.

Assessing

The next ability cluster we specify is *assessing*. Once one recognizes what a system-related aspect (e.g., an output) refers to, the next step is an (initial) assessment of the adequacy of that aspect. This ability is characterized, then, by accurately judging when a system-related aspect is inadequate, unreasonable, or has an unacceptable risk of being so.

In some cases, recognizing an output, in combination with basic background knowledge[10], will suffice for assessing. For instance, an AI system that classifies a human being as a gorilla is obviously flawed (Garcia, 2016). In other cases, AI systems behave inadequately in less straightforward ways, particularly when they are intended to satisfy complex and multifaceted desiderata such as fairness. In such cases, information about the reasoning processes of the AI may be needed to assess the adequacy of an output or some other system-related aspect (Baum et al., 2022).

Assessing, as we construe it, relates to the ability to *evaluate* that occupies the fifth level in the revision of Bloom's Taxonomy (Krathwohl, 2002) and constitutes one facet of *comprehension* in situation awareness (Endsley, 1995).

Being able to assess a system's output is a prerequisite for taking appropriate action in response. This could mean demanding further explanation, deciding not to use a system, or bringing a legal challenge against a system's operators. Assessing is typically more demanding than recognizing, as one can recognize what a system's output denotes without being able to judge whether it is appropriate or satisfactory.

Predicting

The third ability cluster we propose is *predicting*. We characterize predicting as the anticipation of a system's behavior. As such, this ability is exemplified by accurately estimating what a system will output given some input. For example, will a job applicant be ranked favorably or not? How will a particular image be classified? A person possesses this ability with respect to a system if they are able to reliably predict (roughly) what it will do.

Consider our hiring scenario. Assume that April contacts a consultant to find jobs for her. Drawing on their experience, the consultant can predict that April will be ranked favorably at a company employing the hiring system. Thus, they encourage her to apply there. Explainability approaches that highlight the relevance of particular features may be effective at supporting this ability of the consultant (e.g., LIME, see Ribeiro et al., 2016 or Anchors, see Ribeiro et al., 2018; see also Speith, 2022b).

Predicting is a broad and graded ability that is related to multiple notions from our interdisciplinary survey. For example, it corresponds closely to what Pearl and Mackenzie (2018) call *association*. In line with Pearl and Mackenzie (2018)'s *Ladder of Causation*, this ability does not necessarily require causal knowledge about *how* the input is transformed into the output (as also noted by Knüsel & Baumberger, 2020). In practice however, accurate prediction often requires augmenting observations of input-output pairs with insights into how the transformation takes place.

Further notions, such as *projection* in human factors or *generalizing* in cognitive psychology, are also kinds of prediction (see, e.g., Williams & Lombrozo, 2010). For example, generalizing is predicting the properties of an object, given knowledge of its class. What unifies predicting, despite its heterogeneity, is the mental simulation of a mapping from inputs to outputs.

Predicting is distinct from assessing. One can assess a system's behavior, that is, notice when it produces inadequate outputs, without being able to predict what those outputs will be. In this sense predicting is more intimately related to knowledge of and experience with a particular system than assessing.

Intervening

The next cluster of abilities we identify is *intervening*. Intervening is acting on a system so as to accomplish a goal. It therefore requires finding the means to achieve some specified ends.

Intervention can have different objects. On the one hand, intervention can be directed at the *information* a system processes, e.g., when tailoring input features to achieve a certain result. On the other hand, intervention can aim at altering how the system *processes* information, e.g., to achieve some desired system property such as monotonicity.

In our hiring case, the first type of intervention could be achieved by the applicant. Aiming to improve her ranking, April might tailor her application by including specific

[10] By "background knowledge" we mean any (propositional or procedural) knowledge a person possesses prior to their interaction with an AI system. Background knowledge plays a crucial role, as the explanations provided by explainability approaches always need to match the stakeholder's previous background knowledge in order to support the abilities we discuss.

information that is known to be influential. The second kind of intervention would largely be exclusive to developers, namely, changing the inner workings of the system in order to, e.g., improve the system's accuracy or fairness.

Notice that being able to intervene in the latter way requires knowledge about the properties of the system and how it implements the transformation of inputs into outputs. As such, this ability typically depends upon significant background knowledge. However, it may also be supported by model-specific explainability approaches that shed light on how different system properties and components contribute to overall behaviors (e.g., Lapuschkin et al., 2016; Stock & Cissé, 2018; see also Speith, 2022b).

Intervening is related to de Regt (2015)'s view that scientific understanding of phenomena is characterized by being able to use a model to *control* its behavior as well as Thórisson et al. (2016)'s notion of *achieving goals*.

Intervening can be distinguished from predicting, which requires forward simulation from a set of initial conditions to their likely consequences. Intervening, in contrast, depends on reasoning backwards from a chosen outcome to the changes in the conditions that would produce it. While learning to predict is possible from pure observation of input-output behavior without specific knowledge of how a system works at the level of components (Knüsel & Baumberger, 2020), intervention typically demands knowledge of which variables need to be changed to alter the system's behavior.

Explaining

Our fifth cluster is *explaining*. Explaining can mean providing causal information about an occurrence or giving a description of how something works in terms of its components and their interactions (Rozenblit and Keil, 2002; Halpern & Pearl, 2005).[11] As with other abilities, explaining comes in degrees. Partial explanations describe some of the causes, components, and interactions relevant to a phenomenon, while complete explanations describe them all.

To exemplify explaining, assume that the hiring system in our example is based on a decision tree classifier. One can partially explain the system's output by pointing to influential input features (e.g., educational attainment). A complete explanation, however, would also require describing the tree's nodes, how they are organized, and how they interact to produce the system's outputs. Providing detailed explanations is a demanding ability which typically requires significant expertise with respect to the target of explanation (Rozenblit and Keil, 2002).

Information provided by explainability approaches may be crucial in supporting the development of the ability to explain. For example, model-specific approaches to explainability often attempt to characterize how components contribute to the functioning of a system (e.g., Lapuschkin et al., 2016; Cammarata et al., 2020; see also Speith, 2022b).

Our notion of explaining coheres with Rozenblit and Keil (2002)'s operationalization of understanding and is close to accounts in the philosophy of science which require, e.g., being able to specify the causal components and relationships relevant to some phenomenon (Newman, 2017; Halpern & Pearl, 2005; Strevens, 2008).

Explaining closely relates to intervening, as both clusters presuppose some knowledge of inputs or processes in a system. However, while intervening does not require the ability to articulate this knowledge, explaining does. On the other hand, explaining may not require the technical know-how needed for altering inputs or changing the system. For this reason, we place these abilities on the same level in our hierarchy.

Designing

The final cluster of abilities we identify, constituting comprehensive understanding, is *designing*. Designing typically applies to whole systems and requires being able to specify which components are needed, how they are organized, and how they function to produce outcomes satisfying relevant requirements. As such, being able to design an AI system from scratch, such that it satisfies specific constraints on, say, accuracy and fairness, would constitute a deep understanding of that system and the problem it addresses.

In our hiring case, a model developer might be tasked with designing an appropriate system. This could require an iterative process of feature engineering to construct meaningful and robust features that support fair and accurate decisions (Rudin, 2019). Note that, since we take designing to require specifying *how* a system functions, training black-box models like deep neural networks does not suffice for this ability.

Designing is related to Thórisson et al. (2016)'s notion that fully understanding a phenomenon means being able to *re-create* it and de Regt (2015)'s view that scientific understanding involves the construction of models. Additionally, it resembles the final level of Bloom's revised taxonomy: *creating* (Krathwohl, 2002). This form of understanding requires being able to put elements of knowledge together to develop a novel product.

[11] This holds at a practically relevant level of abstraction. That is, explaining an artificial neural network might require specifying how input data and learned weights interact to produce outcomes, but not how those properties are implemented in machine language (Craver & Kaplan, 2020; Potochnik, 2010).

Notice the distinction between intervening and designing. The former requires altering a system in some way, perhaps by tweaking the values of its parameters, while the latter requires actually *specifying* the components of a system and how they are organized. Further, designing is related to, but more demanding than explaining. While partial explanations are often sufficient for the satisfaction of desiderata, incomplete designs are of limited value. However, a complete diagrammatic explanation of a system including all components and their interactions essentially *is* a design for that system (v. Cammarata et al., 2020).

Note that we depict our clusters in the form of a pyramid to reflect that some of them tend to be more demanding than others (see Fig. 2). However, how demanding it is to acquire each ability is also influenced by contextual factors. For instance, *assessing* whether a single prediction relied on protected attributes may be easier than assessing whether the system satisfies fairness desiderata at a global level. Furthermore, there are scenarios where someone might possess an ability higher up in the hierarchy without possessing the abilities below it. For example, in absence of relevant domain knowledge, the ability to *explain* does not necessarily entail the ability to *assess* whether an output was appropriate. Crucially, explainability approaches should not always aim to produce the abilities we place higher up in the pyramid. Rather, they ought to facilitate whatever form of understanding is required to satisfy downstream desiderata (see the "Abilities can be linked to desiderata" section).

Abilities are operationalizable

The next criterion of a useful conceptualization of understanding in XAI is that it should be operationalizable. While the interdisciplinary review of understanding in the "Abilities are motivated by research across disciplines" section has already demonstrated that abilities can be operationalized for empirical studies (e.g., in cognitive or educational psychology), we now show that they can also be used in XAI. More specifically, we show that the abilities we describe in our clusters have already been used in studies to evaluate explainability approaches; thus, the clusters we propose are operationalizable.

Recognizing

Recognizing is a fundamental ability that is often taken for granted. As such, most studies presuppose this ability by choosing participants that are familiar with the representational format used, or are primed beforehand. However, Vilone and Longo (2021)'s review of explainability approaches summarizes the strengths and weaknesses of different representational formats (e.g., numeric, rule-based, visual, textual), incorporating concerns about whether certain stakeholders can really recognize some representational formats (e.g., laypeople may struggle with mathematical formulae as explanations; see also Speith, 2022a, 2022b).

Assessing

Ribeiro et al. (2016, 2018) have proposed two explainability approaches, LIME and Anchors, both of which they evaluated via measuring abilities. In one of their studies evaluating LIME, participants were asked to choose between several equally accurate classifiers which relied on different classification strategies. This procedure tests for participants' ability to *assess* the quality of different systems. Another example can be found with Piltaver et al. (2014), who also used an *assessing* task to test the comprehensibility of classification trees. In particular, they had participants validate (parts of) the trees. Furthermore, Lapuschkin et al. (2016) demonstrated the efficacy of their Layer-wise Relevance Propagation (LRP) approach by using it to support *assessing* whether image classifiers were detecting meaningful features or artifacts. See Stock and Cissé (2018) and Allahyari and Lavesson (2011) for further tasks that can be attributed to our *assessing* cluster.

Predicting

Piltaver et al. (2014)'s study also measured *prediction* abilities. In particular, participants in their study had to derive the output of a decision tree given an input. Similarly, Huysmans et al. (2011) evaluated understanding of various classification models by having participants *predict* these models' classifications on previously unseen instances. See Alqaraawi et al. (2020), Poursabzi-Sangdeh et al. (2021), and Ribeiro et al. (2018) for further tasks from our *prediction* cluster.

Intervening and explaining

In the evaluation of LIME, Ribeiro et al. (2016) tasked participants to a) find a classifier's source of failure and b) improve a classifier after receiving explanations. The former task assessed the participants' ability to *explain* the failure of a system, while the latter evaluated their ability to *intervene* on the classifier. Piltaver et al. (2014) also conducted a study belonging to the intervening cluster. They tasked participants with identifying which attribute values in a decision tree had to be changed to alter a classification. Another explaining task can be found in Tullio et al. (2007).

Designing

Designing is usually so challenging that it is not addressed with empirical studies. Pragmatically, however, designing is the goal for model developers tasked with constructing inherently understandable models. This is exactly what Rudin (2019) calls for: She argues that in most problem domains, given sufficient time and developer expertise, such models can be *designed* (see also Crook et al., 2023). While designing, to the best of our knowledge, has not been tested empirically, it is evaluated by the process of constructing inherently understandable models, e.g., when checking whether a model satisfies certain functional requirements. Further, designing is sometimes a goal of scientific investigations. For example, Cammarata et al. (2020) attempt to reverse engineer a deep neural network so that it can be re-implemented from scratch (i.e., by setting the connection weights between neurons in the network manually). In principle, one could imagine using this task in a study.

Overall, we believe that the work cited above confirms not only that abilities can be operationalized (i.e., that they can be used to design and carry out empirical studies to measure understanding and thereby evaluate explainability approaches), but also that our proposed clusters are useful to this end.

Abilities can be linked to desiderata

Turning to the final criterion, a useful conceptualization of understanding in XAI should illuminate the intricate relationship between explainability, understanding, and desiderata. As the models of Langer et al. (2021c) and Hoffman et al. (2018) illustrate, it is commonly assumed in XAI that explainability approaches, by facilitating understanding, are supposed to satisfy desiderata (e.g., fairness, trust, or safety; see Fig. 1), either epistemically or substantially. However, it is often unclear how, exactly, this is supposed to work (see, e.g., Kästner et al., 2021; Deck et al., 2024).

We claim that by using the ability clusters we propose in the "Abilities are versatile" section, one can describe more precisely the relationship between explainability, understanding, and desiderata. In particular, abilities can be used to specify the requirements for epistemic or substantial desiderata satisfaction in a given situation.

Example 1: fairness

To exemplify the connection between abilities and desiderata, let us first consider a desideratum prevalent in the hiring case: *fairness*. In the XAI literature, it is often assumed that explainability is important for fairness (for a review, see Deck et al., 2024); however, as of yet, there is little agreement on how, exactly, a better understanding of the system can contribute to fairness. Therefore, anyone who wants to improve fairness via XAI needs to be clear about how they expect a better understanding of the system to affect fairness. In cases like this, our account proves useful.

Consider April again, whose application was rejected. In this situation, April may want to know whether she was treated unfairly. Thus, she requires epistemic access to whether the system's decision was fair. For this kind of epistemic desideratum satisfaction, April would need to be able to *assess* what led the system to the specific output "rejection," specifically concerning the influence of protected attributes.

Now consider an external auditor, Sam, whose role is to check whether the hiring support system adheres to a particular notion of fairness (e.g., subgroup parity, Hutchinson & Mitchell, 2019). This is another case that aims for epistemic desideratum satisfaction, which, however, requires different abilities. For Sam to gain epistemic access to whether the system is fair overall, he would not only need the ability to assess certain statistical properties of the system's behavior, but also to reliably *predict* that the system's outputs, overall, will continue to adhere to the chosen notion of fairness.

Other ability clusters can specify how understanding can contribute to the substantial satisfaction of the fairness desideratum: If the system under inspection turns out not to be fair in the desired sense, this may prompt a requirement for further abilities. For instance, to rectify an unfair system, a developer would need the ability to *intervene* on the system. This would require changing its internal processing in a way that ensures fair behavior in the future.

One can imagine even more demanding abilities being required to ensure system fairness. For instance, Rudin (2019) rejects so-called black-box models in high-stakes scenarios because it is hard to determine whether their decision making is fair. Instead, she prefers ante-hoc explainable models (e.g., logistic regression or decision lists). If ante-hoc explainability was the only way to ensure system fairness, this would require a complete description or explanation of the system's decision-making processes, as well as demonstration that those processes adhere to the chosen notion of fairness. As noted in the "Abilities are versatile" section, the ability to give such a complete explanation is tantamount to the ability to *design* the system.

The presented examples demonstrate that the relationship between our ability clusters and desiderata is so close that, in some cases, specifying which ability a stakeholder (or several stakeholders) needs to gain also describes the requirements for desiderata satisfaction. For both April and Sam, gaining the required abilities suffices for epistemic

desiderata satisfaction. In the other cases, our account allows us to specify which understanding-related abilities support substantial desiderata satisfaction. This is a clear advantage of our account over less versatile approaches to understanding.

Example 2: informed self-advocacy

Our ability clusters can also be used to specify the need for understanding described in existing work on desiderata in XAI. One example of that is our rephrasing of Rudin's position as a requirement for the ability to design in the previous section. For another, more complex example in this vein, we can look at Vredenburgh (2022)'s account of "informed self-advocacy;" a desideratum that even *requires* understanding, among other things, for substantial desiderata satisfaction. Her account relates to our proposal because she spells out informed self-advocacy in terms of abilities. Though these abilities do not readily fit with our clusters, our account allows us to specify exactly which understanding-related abilities are required for informed self-advocacy.

Vredenburgh argues that informed self-advocacy relies on three abilities, two of which require insight into systems:[12] The ability to "navigate systems of rules to achieve one's goals" (Vredenburgh, 2022, p. 213), which she calls *agency*, and the ability to hold decision-makers accountable if certain standards are not met, which she calls *accountability*. These abilities are specified at a rather high level. We think the forms of understanding required to enable agency and accountability in specific contexts can be specified concretely via our abilities-based conceptualization of understanding.

Recall our illustrative hiring scenario. To ensure her agency, it may be sufficient for April to possess the ability to *predict* whether she will be invited to an interview, that is, to anticipate the system's behavior given her application data. However, if April gets rejected, securing her agency would require the ability to *intervene* on the system's input such that she will reach the interview in the next hiring round (e.g., by acquiring additional skills and updating her application accordingly). This could be supported by (ideally actionable, v. Karimi et al., 2021) counterfactual statements describing what would have been needed to be invited to an interview in the first hiring round (v. Wachter et al., 2017). The abilities of predicting and intervening enable April to navigate the system of rules that guides hiring decisions and therefore suffice to ensure April's agency in this context.

[12] The third ability is the ability to represent one's interests to decision-makers. As this ability is unrelated to the development and evaluation of an AI system (Vredenburgh, 2022, p. 213), we will not discuss it here.

Let us move to *accountability*, which requires access to normative reasons for a decision and appropriate recourse should those reasons be unjust. If April gets rejected surprisingly (e.g., despite her excellent credentials), she might suspect that her rejection was based on discrimination. In order to ensure accountability, April needs the ability to *assess* whether the outcome "rejection" was appropriate. This could be supported by feature attribution methods revealing which information in April's application led to her rejection (v. Speith, 2022b). If the revealed decision criteria are not self-explanatory, however, the explanation might need to be combined with normative reasons for their adequacy (Vredenburgh, 2022, p. 217). This could require the hiring manager to be able to *explain* to April why she was rejected (Baum et al., 2022). With this combination of abilities in place, April can both detect flawed decisions and hold someone at the hiring company accountable for them. Overall, then, the understanding required to support April's informed self-advocacy can be spelled out in terms of the ability clusters we propose.

In sum, there is a close relationship between the understanding-related abilities stakeholders might possess and whether and how societal desiderata can be fulfilled. Specifying which abilities are necessary to fulfill a certain desideratum allows academics, advocates for ethical AI, and policy makers to be more precise about how, exactly, explainability can contribute to desiderata satisfaction. This can aid communication and collaboration between researchers, support the progress of the debate on XAI, and improve the development and evaluation of individual explainability approaches. Furthermore, we saw that abilities fit into existing work on the subject: We have shown that both Rudin's rejection of black-box models as well as Vredenburgh's account of informed self-advocacy can be translated into demands for certain understanding-related abilities. Thus, we think that understanding-related abilities also meet the final criterion of a useful conceptualization of understanding in XAI: abilities can be used to clarify the link between explainability, understanding, and the satisfaction of relevant societal desiderata.

Limitations and objections

We have argued that abilities can serve as a useful conceptualization of understanding as they meet the criteria we identified in the "Criteria for a useful conceptualization of understanding in XAI" section: First, understanding-related abilities are theoretically motivated across relevant disciplines. Second, abilities are versatile enough to differentiate the various forms of understanding relevant to different XAI application contexts. Third, the abilities we propose are operationalizable. Finally, abilities can be used to identify

necessary conditions for desiderata satisfaction and serve as a concrete description of the forms of understanding that are needed to meet a given desideratum in a specific context. As such, abilities can establish the often assumed link between explainability, understanding, and satisfaction of stakeholder desiderata. Before concluding, we address three possible criticisms of our position.

Objection 1: abilities are not necessary for understanding

Writing about *interpretability*, Krishnan (2020) rejects an ability-based view as "more plausibly a consequence of interpretability than it is a statement of that in which interpretability itself consists." Thus, she judges pragmatic accounts of interpretability to be inadequate for contributing towards the objective of ethical AI. A similar objection could also be applied to understanding. Since contingent factors can prevent one's cognitive state manifesting in practical abilities (Wilkenfeld, 2013), someone might possess understanding but not the abilities we describe. For instance, the ability to *explain* will normally require linguistic capabilities, and the ability to *intervene* presupposes (usually minimal) physical capacities. Similar requirements likely hold for other understanding-related abilities, too.

While abilities may not capture the concept of understanding completely, we think the more relevant question is whether they can specify the concept well enough to be useful. To that end, the previous sections have tried to show that a conceptualization of understanding as abilities can have pragmatic value for XAI by fulfilling our four criteria.

Furthermore, we believe that focusing on abilities can be of particular importance in XAI contexts, e.g., when morally weighty desiderata make it indispensable to ensure that the appropriate way of understanding was indeed afforded by an explainability approach; or when the collaboration of different stakeholders requires interaction with other agents (e.g., by *explaining*) or with artifacts (e.g., by *intervening*). In such cases, these contingent factors that might not be constitutive of understanding are nevertheless required for desiderata satisfaction.

Objection 2: abilities are not sufficient for understanding

Another worry is that abilities may be present in the absence of understanding (v. Wilkenfeld, 2013). For instance, someone could demonstrate an ability to *explain* by parroting an explanation without truly grasping it. This scenario could occur in contexts where a stakeholder offloads the cognitive work involved in understanding to a reliable explainability approach. Similarly, for systems that can only provide a small range of possible outputs, someone might be able to *predict* the system's behavior by mere luck, without possessing any understanding. Similar scenarios are conceivable for the other abilities we suggest.

However, it is doubtful whether the abilities as we have described them in the "Abilities are versatile" section are indeed present in the above cases. Mere parroting arguably does not count as explaining, even if it appears otherwise; and lucky guessing seems to be fundamentally different from predicting. Instead of rejecting abilities as unreliable, researchers interested in measuring understanding need to be careful when operationalizing it. For instance, parroting and lucky guessing could be detected by demanding participants to reformulate an explanation or to make predictions on a range of counterfactual cases. A requirement for valid and robust empirical evaluations should ensure that stakeholders reliably possess the abilities in question. Such a requirement should also rule out subjective measures (e.g., self-reports) which cannot discriminate between a stakeholder's sense of understanding and the practical abilities associated with it.

Robust assessment of abilities is especially important in high-stakes contexts, where misjudgments could lead to harmful outcomes, and in what Bordt et al. (2022) call *adversarial explanation contexts*, where the parties involved in building, explaining, and using an AI system have conflicting interests. In such contexts, the explanation provider may be incentivized to furnish the explanation recipient with misleading explanations that induce false confidence in one's abilities or a false sense of understanding (i.e., one not accompanied by the relevant abilities; Bordt et al., 2022).

However, even if a reliable method indicates that a person possesses relevant abilities, this may still be insufficient for a demanding, philosophical account of understanding (for an overview of possibilities, see Baumberger et al., 2017; Baumberger, 2019). To someone who holds this kind of position, specific measures of abilities can only be fallible proxies for understanding; they suggest that it is present, but do not guarantee it. Though we grant that this may be the case, our goal was not to provide a metaphysical definition of understanding, but a useful conceptualization that is able to fulfill the role that is commonly attributed to understanding in XAI. We think abilities are the most suitable candidate for this role.

Objection 3: characterizing understanding is futile

In light of the difficulties that arise when one tries to specify the relationship between understanding and abilities, one might question whether these efforts are worthwhile. Krishnan (2020) argues that research addressing ethical issues arising from opaque AI algorithms need not focus on psychological constructs to achieve its goal of satisfying societal desiderata. In Krishnan's view, the challenges inherent to specifying nebulous terms like explainability

and understanding render them more problematic than useful. Instead of trying to characterize these terms, one should directly address the actual aim, that is, desiderata satisfaction.

We believe this argument is premature in rejecting a pragmatic approach to dealing with concepts like understanding. We agree with Krishnan that there is no *one* way in which systems can be explainable nor *one* form of understanding which will help satisfy all relevant desiderata. We disagree, however, that this is grounds to focus solely on desiderata. Instead, we believe our arguments have shown that conceptualizing understanding via abilities is a fruitful way to specify forms of understanding which can build a bridge between explainability approaches and desiderata. We believe that this approach is preferable to looking directly at desiderata for two reasons.

First, clusters of abilities allow for the identification of commonalities between the epistemic requirements[13] for different desiderata. Langer et al. (2021c) and Chazette et al. (2021) identify numerous desiderata that are discussed within the XAI literature. For every desideratum, the lack of understanding of AI systems plays a role. While this cannot be addressed in the same way for every desideratum, there are recurring problems that our pragmatic conceptualization of understanding can capture.

Second, an independent characterization of understanding allows for greater specificity about the role of explainability for desiderata satisfaction. In some cases, we rely on explainability to guarantee epistemic desiderata satisfaction, that is, to gain epistemic access to the properties of a system, even when technical requirements on AI systems are fully satisfied (v. Langer et al., 2021c). In other cases, explainability is expected to contribute to substantial desiderata satisfaction, but it remains unclear *how* exactly. For example, it is still unclear whether and how explanations can calibrate trust in a system (Kästner et al., 2021), and it is ethically preferable only to use explanations to increase trust when it routes through an adequate ability to assess the trustworthiness of the system (Schlicker & Langer, 2021). For reasoning about both kinds of cases, a concrete conceptualization of understanding that is independent of desiderata satisfaction is required. Therefore, we consider our approach to be valuable even for those primarily interested in desiderata satisfaction.

[13] One could argue that our position is closer to Krishnan's than it might superficially appear, as we identify only the epistemic requirements of desiderata satisfaction rather than describing the metaphysical nature of understanding. However, given the omnipresence of "understanding" in the XAI debate, we consider it to be more useful to clarify different uses of understanding as opposed to eliminating talk of understanding. We thank an anonymous reviewer for pressing this point.

Conclusion

In this paper, we proposed a conceptualization of understanding useful for XAI endeavors: *abilities*. More specifically, we suggested that someone who understands some aspect(s) of an AI system typically possesses one or several understanding-related abilities (viz., the abilities to recognize, to assess, to predict, to explain, to intervene, and to design). Ceteris paribus, a person who possesses an ability to a higher degree than another person also has a higher degree of understanding. Furthermore, as indicated by the hierarchical organization of our ability clusters, the respective abilities tend to correspond to different degrees of understanding. To support our proposal, we showed that abilities are motivated by interdisciplinary research, versatile, operationalizable, and can be used to clarify the link between understanding and desiderata satisfaction. As such, our approach can benefit various stakeholders involved in XAI, both inside and outside computer science. For example, abilities can be used to precisely specify understanding-related requirements for the use and deployment of AI systems. In turn, having concrete requirements spelled out can help developers design new explainability approaches with a clear view of what they need to achieve.

Overall, the abilities-based conceptualization of understanding we propose can serve as a common language that facilitates communication and collaboration between different stakeholders involved in XAI. However, as a next step for such a conceptualization to effectively address challenges in XAI, research needs to investigate which experimental paradigms best operationalize understanding-related abilities. Furthermore, work exploring how particular abilities relate to specific desiderata would strengthen the theoretical basis on which to deploy explainability approaches in real-world contexts. Finally, our conceptualization of understanding in XAI contexts could inform a formal definition of this concept that takes additional philosophical and psychological debates on understanding into account. We believe that the clusters we have proposed will prove useful for these and related projects.

Acknowledgements Work on this paper was funded by the Volkswagen Foundation grants AZ 9B830, 98510, 98513, and 98514 "Explainable Intelligent Systems" (EIS) and by the DFG grant 389792660 as part of TRR 248 (in particular project A6). The Volkswagen Foundation and the DFG had no role in preparation, review, or approval of the manuscript; or the decision to submit the manuscript for publication. We thank two anonymous reviewers and the participants in the EIS colloquium for their valuable feedback on earlier versions of the paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors declare no other competing or financial interests.

References

Allahyari, H., & Lavesson, N. (2011). User-oriented assessment of classification model understandability. In Kofod-Petersen, A., Heintz, F., & Langseth, H. (Eds.), Proceedings of the 11th Scandinavian Conference on Artificial Intelligence (SCAI 2011), Frontiers in Artificial Intelligence and Applications, Vol. 227. IOS Press, pp. 11–19, https://doi.org/10.3233/978-1-60750-754-3-11

Alqaraawi, A., Schuessler, M., Weiß, P., et al. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In Paternò, F., Oliver, N., Conati, C., et al. (Eds.), Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI 2020). Association for Computing Machinery, pp. 275–285, https://doi.org/10.1145/3377325.3377519

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., et al. (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

Baum, K., Mantel, S., Schmidt, E., et al. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology, 35*(1), 12. https://doi.org/10.1007/s13347-022-00510-w

Baumberger, C. (2014). Types of understanding: Their nature and their relation to knowledge. *Conceptus, 40*(98), 67–88. https://doi.org/10.1515/cpt-2014-0002

Baumberger, C. (2019). Explicating objectual understanding: Taking degrees seriously. *Journal for General Philosophy of Science, 50*(3), 367–388. https://doi.org/10.1007/s10838-019-09474-6

Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 1–34). Routledge.

Bloom, B. S., Engelhart, M. D., Furst, E. J., et al. (1965). *Taxonomy of educational objectives. Handbook 1: Cognitive domain*. McKay.

Bordt, S., Finck, M., Raidl, E., et al. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. In Isbell, C., Lazar, S., Oh, A., et al. (Eds.), Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022). Association for Computing Machinery, pp. 891–905, https://doi.org/10.1145/3531146.3533153

Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis, 81*(6), 1211–1241. https://doi.org/10.1007/s10670-015-9791-5

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Cammarata, N., Carter, S., Goh, G., et al. (2020). Thread: Circuits. https://doi.org/10.23915/distill.00024.

Carnap, R. (1962). *Logical foundations of probability*. University of Chicago Press.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 1–34. https://doi.org/10.3390/electronics8080832

Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring explainability: A definition, a model, and a knowledge catalogue. In Cleland-Huang, J., Moreira, A., Schneider, K., et al. (Eds.), Proceedings of the 29th IEEE International Requirements Engineering Conference (RE 2021). IEEE, pp. 197–208, https://doi.org/10.1109/RE51729.2021.00025

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science, 71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Crook, B., Schlüter, M., & Speith, T. (2023). Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In Schneider, K., Dalpiaz, F., & Horkoff, J. (Eds.), Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops (REW 2023). IEEE, pp. 316–324, https://doi.org/10.1109/REW57809.2023.00060

Deck, L., Schoeffer, J., De-Arteaga, M., et al. (2024). A critical survey on fairness benefits of XAI. In F. Steibel, M. Young & R. Baeza-Yates (Eds.), Proceedings of the 7th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2024). Association for Computing Machinery. http://arxiv.org/abs/2310.13007

Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review, 119*(7), 1829–1850.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Franconeri, S. L., Padilla, L. M., Shah, P., et al. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest, 22*(3), 110–161. https://doi.org/10.1177/15291006211051956

Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal, 33*(4), 111–117. https://doi.org/10.1215/07402775-3813015

Grimm, S. R. (2011). Understanding. In S. Bernecker & D. Pritchard (Eds.), *The Routledge companion to epistemology* (pp. 84–94). https://doi.org/10.4324/9780203839065.ch9

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science, 56*(4), 889–911. https://doi.org/10.1093/bjps/axi148

Hannon, M. (2021). Recent work in the epistemology of understanding. *American Philosophical Quarterly, 58*(3), 269–290. https://doi.org/10.2307/48616060

Hickman, L., Bosch, N., Ng, V., et al. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, 107*(8), 1323–1351. https://doi.org/10.1037/apl0000695

Hills, A. (2016). Understanding why. *Noûs, 50*(4), 661–688. https://doi.org/10.1111/nous.12092

Hoffman, R. R., Mueller, S. T., Klein, G., et al. (2018). Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608

Hutchinson, B., & Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In Boyd, D., & Morgenstern, J. H. (Eds.), Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019). Association

for Computing Machinery, pp. 49–58, https://doi.org/10.1145/3287560.3287600

Huysmans, J., Dejaeger, K., Mues, C., et al. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems, 51*(1), 141–154. https://doi.org/10.1016/j.dss.2010.12.003

Karimi, A. H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. In Elish, M. C., Isaac, W., & Zemel, R. S. (Eds.), Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021). Association for Computing Machinery, New York, NY, USA, pp 353–362, https://doi.org/10.1145/3442188.3445899

Kästner, L., Langer, M., Lazar, V., et al. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. In Yue, T., & Mirakhorli, M. (Eds.), Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops (REW 2021). IEEE, pp. 169–175, https://doi.org/10.1109/REW53955.2021.00031

Khalifa, K. (2013). Understanding, grasping and luck. *Episteme, 10*(1), 1–17. https://doi.org/10.1017/epi.2013.6

Knüsel, B., & Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A, 84*, 46–56. https://doi.org/10.1016/j.shpsa.2020.08.003

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology, 33*(3), 487–502. https://doi.org/10.1007/s13347-019-00372-9

Kvanvig, J. (2009). The value of understanding. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value* (pp. 95–111). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199231188.003.0005

Langer, M., & König, C. J. (2023). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Human Resource Management Review, 33*(1), 100881. https://doi.org/10.1016/j.hrmr.2021.100881

Langer, M., Baum, K., Hartmann, K., et al. (2021a). Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In Yue, T., & Mirakhorli, M. (Eds.), Proceedings of the 29th IEEE International Requirements Engineering Conference Workshops (REW 2021). IEEE, pp. 164–168, https://doi.org/10.1109/REW53955.2021.00030

Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology, 36*(5), 751–769. https://doi.org/10.1007/s10869-020-09711-6

Langer, M. (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

Lapuschkin, S., Binder, A., Montavon, G., et al. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. In Tuytelaars, T., Li, F. F., Bajcsy, R., et al. (Eds.), Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2016). IEEE, pp. 2912–2920, https://doi.org/10.1109/CVPR.2016.318

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

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99*(2), 167–204. https://doi.org/10.1016/j.cognition.2004.12.009

Mann, S., Crook, B., Kästner, L., et al. (2023). Sources of opacity in computer systems: Towards a comprehensive taxonomy. In Dalpiaz, F., Horkoff, J., & Schneider, K. (Eds.), Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops (REW 2023). IEEE, pp. 337–342, https://doi.org/10.1109/REW57809.2023.00063

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. In Aha, D. W., Darrell, T., Pazzani, M., et al. (Eds.), Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI 2017). IJCAI, pp. 36–42, arXiv:1712.00547

Newman, M. (2017). An evidentialist account of explanatory understanding. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 190–211). Routledge.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines, 29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Piltaver, R., Luštreka, M., Gams, M., et al. (2014). Comprehensibility of classification trees-Survey design. In Proceedings of 17th International Multiconference Information Society (IS 2014). Information Society, pp. 70–73

Potochnik, A. (2010). Levels of explanation reconceived. *Philosophy of Science, 77*(1), 59–72. https://doi.org/10.1086/650208

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., et al. (2021). Manipulating and measuring model interpretability. In Kitamura, Y., Quigley, A., Isbister, K., et al (Eds.), Proceedings of the 39th ACM Conference on Human Factors in Computing Systems (CHI 2021). Association for Computing Machinery, pp. 237:1–237:52. https://doi.org/10.1145/3411764.3445315

de Regt, H. W. (2015). Scientific understanding: Truth or dare? *Synthese, 192*(12), 3781–3797. https://doi.org/10.1007/s11229-014-0538-7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram, B., Shah, M., Smola, A. J., et al. (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016). Association for Computing Machinery, pp 1135–1144, https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In McIlraith, S. A., & Weinberger, K. Q. (Eds.), Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI 2018), the 30th Innovative Applications of Artificial Intelligence Conference (IAAI 2018), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI 2018). AAAI Press, pp. 1527–1535, https://doi.org/10.1609/aaai.v32i1.11491

Riggs, W. D. (2003). Intellectual virtue: Perspectives from ethics and epistemology. In M. DePaul & L. Zagzebski (Eds.), *Understanding 'virtue' and the virtue of understanding* (pp. 203–226). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199252732.003.0010

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

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

Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In Schneegass, S., Pfleging, B., & Kern, D. (Eds.), Proceedings of Mensch und Computer (MuC 2021). Association for Computing Machinery, pp. 325–329, https://doi.org/10.1145/3473856.3474018.

Speith, T. (2022a). How to evaluate explainability: A case for three criteria. In Knauss, E., Mussbacher, G., Arora, C., et al. (Eds.), Proceedings of the 30th IEEE International Requirements Engineering Conference Workshops (REW 2022). IEEE, pp. 92–97, https://doi.org/10.1109/REW56159.2022.00024

Speith, T. (2022b). A review of taxonomies of explainable artificial intelligence (XAI) methods. In Isbell, C., Lazar, S., Oh, A., et al. (Eds.), Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022). Association for Computing Machinery, pp. 2239–2250, https://doi.org/10.1145/3531146.3534639

Stock, P., Cissé, M. (2018). ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases. In Ferrari, V., Hebert, M., Sminchisescu, C., et al. (Eds.), Proceedings of the 15th European Conference on Computer Vision (ECCV 2018), Lecture Notes in Computer Science, Vol. 11210. Springer International Publishing, pp. 504–519, https://doi.org/10.1007/978-3-030-01231-1_31

Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press. https://doi.org/10.2307/j.ctv1dv0tnw

Thórisson, K. R., Kremelberg, D., Steunebrink, B. R., et al. (2016). About understanding. In Steunebrink, B., Wang, P., & Goertzel, B. (Eds.), Proceedings of the 9th International Conference on Artificial General Intelligence (AGI 2016), Lecture Notes in Computer Science, Vol. 9782. Springer International Publishing, pp. 106–117, https://doi.org/10.1007/978-3-319-41649-6_11.

Tullio, J., Dey, A. K., Chalecki, J., et al. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In Rosson, M. B., & Gilmore, D. J. (Eds.), Proceedings of the 25th ACM Conference on Human Factors in Computing Systems (CHI 2007). Association for Computing Machinery, pp. 31–40, https://doi.org/10.1145/1240624.1240630.

Vasilyeva, N., Wilkenfeld, D. A., & Lombrozo, T. (2015). Goals affect the perceived quality of explanations. In Noelle DC, Dale R, Warlaumont AS, et al (Eds.), Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci 2015). Cognitive Science Society, pp. 2469–2474, https://cogsci.mindmodeling.org/2015/papers/0424/paper0424.pdf

Vasilyeva, N., Wilkenfeld, D. A., & Lombrozo, T. (2017). Contextual utility affects the perceived quality of explanations. *Psychonomic Bulletin & Review, 24*(5), 1436–1450. https://doi.org/10.3758/s13423-017-1275-y

Vilone, G., & Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction, 3*(3), 615–661. https://doi.org/10.3390/make3030032

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy, 30*(2), 209–229. https://doi.org/10.1111/jopp.12262

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887. https://doi.org/10.2139/ssrn.3063289

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese, 190*(6), 997–1016. https://doi.org/10.1007/s11229-011-0055-x

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science, 34*(5), 776–806. https://doi.org/10.1111/j.1551-6709.2010.01113.x

Williams, J. J., Lombrozo, T., & Rehder, B. (2010). Why does explaining help learning? Insight from an explanation impairment effect. In Ohlsson S, Catrambone R (Eds.), Proceedings of the 32th Annual Meeting of the Cognitive Science Society (CogSci 2010). Cognitive Science Society, pp. 2906–2911