A Two-Dimensional Explanation Framework to Classify AI as Incomprehensible, Interpretable, or Understandable

Ruben S. Verhagen[1(✉)], Mark A. Neerincx[1,2], and Myrthe L. Tielman[1]

[1] Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
{R.S.Verhagen,M.A.Neerincx,M.L.Tielman}@tudelft.nl
[2] TNO, Kampweg 55, 3769 DE Soesterberg, The Netherlands

Abstract. Because of recent and rapid developments in Artificial Intelligence (AI), humans and AI-systems increasingly work together in human-agent teams. However, in order to effectively leverage the capabilities of both, AI-systems need to be understandable to their human teammates. The branch of eXplainable AI (XAI) aspires to make AI-systems more understandable to humans, potentially improving human-agent teamwork. Unfortunately, XAI literature suffers from a lack of agreement regarding the definitions of and relations between the four key XAI-concepts: transparency, interpretability, explainability, and understandability. Inspired by both XAI and social sciences literature, we present a two-dimensional framework that defines and relates these concepts in a concise and coherent way, yielding a classification of three types of AI-systems: incomprehensible, interpretable, and understandable. We also discuss how the established relationships can be used to guide future research into XAI, and how the framework could be used during the development of AI-systems as part of human-AI teams.

Keywords: Explainable AI · Human-agent teaming · Transparency · Interpretability · Understandability · Explainability

1 Introduction

Rapid developments in the field of Artificial Intelligence (AI) have resulted in the design and adoption of intelligent systems/agents (A/IS) working together with humans. For such human-AI teams to work effectively and efficiently, it is crucial that AI-systems are understandable and predictable to their human teammates [22–24]. The eXplainable Artificial Intelligence (XAI) community aims to make AI more understandable, however, there is a lack of clear definitions and relationships between key concepts in XAI. The objective of this paper is to identify similarities, differences and inconsistencies in the description and usage

D. Calvaresi et al. (Eds.): EXTRAAMAS 2021, LNAI 12688, pp. 119–138, 2021.
https://doi.org/10.1007/978-3-030-82017-6_8

of these concepts, and to establish a framework in which the concepts can be unambiguously defined and related to each other.

Autonomous and intelligent systems/agents (A/IS) are characterized by their abilities to *sense* their environment, *reason* about their observations and goals, and consequently make *decisions* and *act* within their environment in a goal-driven manner [9]. Thanks to these capabilities, A/IS often outperform humans with respect to handling complex problems and rapid and rational decision-making. Consequently, the adoption domains of A/IS range from applications in healthcare to military defense. On the other hand, humans still surpass A/IS regarding the handling of uncertainty and unexpected situations. In an attempt to assemble their diversity in skills and leverage the unique abilities of both, A/IS and humans are increasingly paired to create human-agent teams (HATs).

Several factors are crucial for and determine the success of human-agent teams. Some of the most cited involve mutual trust and understanding; shared mental models and common ground; observability, predictability and directability; transparency and explainability; and teaming intelligence [22–24,33]. Unfortunately, many of these factors are lacking in contemporary human-agent teams. For example, most A/IS demonstrate extremely limited directability and often possess only rudimentary teaming intelligence (i.e., the knowledge, skills, and strategies necessary to effectively team) [23]. Furthermore, A/IS often demonstrate poor transparency and explainability, making it hard for human teammates to properly understand their inner workings, behavior, and decision-making [3,26,30]. This, in turn, negatively affects factors like mutual trust and understanding, eventually resulting in decreased global team performance [22,23].

To understand the behavior of A/IS, humans attribute A/IS behavior by assigning particular mental states (i.e., Theory of Mind) that explain the behavior [3,14,28–30]. Such mental states involve beliefs, desires/goals, emotions, and intentions. For example, humans trying to understand a robot entering a burning house can do so by attributing it to the goal to save a victim. A/IS capable of self-explaining their behavior and actions based on the reasons for the underlying intentions (e.g., beliefs, goals, emotions) help human teammates to build this ToM of the A/IS. This, in turn, will result in better understanding of the capabilities and limits of the A/IS and eventually better human-agent collaboration [3].

Explainable AI (XAI) methods, techniques, and research emerged as a means of making AI-systems more *understandable* to humans [16]. This relatively new community is characterized by the distinction between data-driven - and goal-driven XAI [3] (or perceptual vs. cognitive XAI [31]). Data-driven XAI is about *explaining* and *understanding* the decisions and inner workings of "black-box" machine learning algorithms given certain input data [3,15]. In contrast, goal-driven XAI/explainable agency refers to building goal-driven A/IS (e.g., robots) *explaining* their actions and reasons leading to their decisions to lay users [3,25].

Although fundamentally different branches, both data- and goal-driven XAI are characterized by the same fundamental issue: a lack of consensus with regards

to the definition of and relations between key XAI concepts. Furthermore, provided definitions often suffer from a high level of ambiguity because they frequently refer to related notions. For example, the concepts of *transparency*, *interpretability*, *explainability*, and *understandability* are all frequently used in XAI literature, but often interchangeably, differently, with recourse to each other, or without even being defined. Without establishing clear distinctions and relations between these notions, the resulting ambiguity significantly hampers the comprehensibility of research centered around these concepts. We argue that prior to implementing, manipulating, or investigating these key concepts it is fundamental to first define and relate them. Only in this way, we can truly know what exactly we are trying to develop and evaluate.

To address the lack of agreement concerning the definition of and relations between key XAI notions, we propose a two-dimensional explanation framework that establishes clear concept definitions and relationships between them. This framework is based on both XAI and social sciences literature, and focuses primarily on A/IS disclosing and clarifying causes underlying their behavior and reasoning to human teammates (i.e., goal-driven XAI). Our framework explicitly addresses the lack of consensus and ambiguity problem by establishing clear distinctions and relations between system *transparency*, *interpretability*, *explainability*, and *understandability*. More specifically, the framework discriminates between system *interpretability* and *understandability* as passive and subjective characteristics concerning user knowledge of the system, versus system *transparency* and *explainability* as active and objective characteristics involved with disclosing and clarifying relevant information. Ultimately, these definitions result in the classification of three types of AI-systems: *incomprehensible*, *interpretable*, and *understandable* systems. We argue *transparency* can make *incomprehensible* systems *interpretable*, and *explainability* can make *interpretable* systems *understandable*. Adopting our distinctive concept definitions and mutual relationships can benefit XAI community by clarifying what kind of systems can be developed, and how we can evaluate them.

The remainder of the paper is structured as follows. In Sect. 2 we demonstrate the terminology problem by providing an overview of literature defining the key concepts. Next, we present our two-dimensional framework in Sect. 3. In Sect. 4 we discuss how the framework can be used to guide future XAI research, be applied in practice, and other relevant future directions. Finally, we conclude our paper in Sect. 5.

2 Background

Several works introduced or defined key XAI concepts such as *interpretability*, *explainability*, *transparency*, and *understandability*. However, the lack of consensus on the exact meanings and relations between these notions remains a prevalent issue. This section aims to highlight the problem and discuss relevant and significant prior contributions, before proposing our framework attempting to establish clear distinctions and relations between the concepts. First, we

Table 1. Several definitions for key XAI concepts, illustrating their ambiguity and relatedness.

Concept	Definition
Explainability	How well humans can understand AI-system decisions [30,37]
Interpretability	To explain or present in understandable terms to humans [4,11] How well humans can understand AI-system decisions [30,37]
Transparency	Representing system states in a way that is open to scrutiny, analysis, interpretation, and understanding by humans [1] Characteristic of model to be understandable for humans [4] Capacity of method to explain how a system works, even when behaving unexpectedly [37]
Understandability	To make a human understand how a model works, without any need for explaining its internal structure [4] Measuring how well humans understand model decisions [4] Capacity of a method of explainability to make a model understandable by end users [37]

demonstrate the lack of consensus problem and ambiguity of several proposed definitions. Next, we discuss some definitions, distinctions, and classifications that influenced our work. Finally, we discuss a framework that might help to unambiguously define and relate XAI concepts.

2.1 Problem

Unambiguously defining and relating XAI concepts is challenging. A small survey of available definitions in the literature demonstrates it is particularly hard to do so without recourse to related concepts (Table 1). Table 1 clearly demonstrates the ambiguity and relatedness of the defined concepts, and fails to provide any clear distinctions between them. For example, all of these concepts are defined at least once as *how understandable the AI-system is to humans*.

2.2 Transparency

Turilli and Floridi [36] introduce a clear definition for *transparency* which influenced our work. They suggest *transparency* refers to forms of *information visibility* and the possibility of *accessing* information, intentions, or behaviors that have been intentionally revealed through a process of *disclosure*. This disclosed information (i.e., made explicit and openly available) can then be exploited by potential users to support their own decision-making process.

Despite considering *transparency* and *explainability* as synonyms, Walmsley's [38] discussion of *transparency* influenced our work. Walmsley [38] divides the notion of *transparency* into two major categories: outward - vs. functional *transparency*. Outward *transparency* concerns the relationship between the AI-system and externals, such as developers and users. This includes *transparency*

about development reasons, design choices, values driving the system developers, and capabilities and limitations of the system. In contrast, functional *transparency* concerns the inner workings of the system. This includes *transparency* about how and why the system behaves in general (type functional *transparency*[1]), or came up with certain decisions or actions (token functional *transparency*[2]).

2.3 Related Work

Ciatto et al. [8] propose an abstract and formal framework for XAI that, in contrast to most work, introduces a clear distinction between *interpretation* and *explanation*. The framework stresses the objective nature of *explanation*, in contrast with the subjective nature of *interpretation*. The act of *interpreting* some object X is defined as the activity performed by an agent A assigning a subjective meaning to X. Furthermore, Ciatto et al. [8] argue an object X is interpretable for an agent A if it is easy for A to assign a subjective meaning to X (i.e., A requires low computational or cognitive effort to *understand* X). The authors stress the subjective nature of *interpretations*, as agents assign them to objects according to their State of Mind and background knowledge.

In contrast, *explaining* is defined as the epistemic and computational activity of producing a more *interpretable* object X' out of a less interpretable one X, performed by agent A. They argue this activity can be considered objective because it does not depend on the agent's perceptions and State of Mind. Consider, for example, decision tree extraction (the *explaining* activity) from a neural network (object X) to produce a decision tree (the *explanation*/object X'). In the end, the effectiveness of the explanations always remains a subjective aspect.

This framework differs from ours in a few ways. In particular, Ciatto et al. [8] provide a formal framework focused on data-driven XAI, whereas we provide more general definitions in a goal-driven XAI context. In contrast, the intentions of the paper and provided definitions are similar to our work. We also define *interpretability* as a subjective system characteristic reflecting user knowledge about a system, and *explainability* as an epistemic and computational activity aimed at increasing user knowledge about the system.

Barredo Arrieta et al. [4] provide a brief clarification of the distinctions and similarities between *transparency*, *interpretability*, *explainability*, and *understandability*. So this part of their work is very similar in its intents to our work, despite focusing on data-driven XAI instead of goal-driven XAI. However, we argue that their attempt at clarifying the distinctions and similarities between the concepts fails to resolve any ambiguity. For example, the authors first argue *interpretability* is a passive model characteristic referring to the level at which a given model makes sense for a human, but later as the ability to explain or provide the meaning in *understandable* terms to a human.

[1] Also referred to as global explanations in XAI literature.
[2] Also referred to as local explanations in XAI literature.

In summary, Barredo Arrieta et al. [4] define *interpretability* (i.e., their first definition), *understandability*, and *transparency* as passive model characteristics reflecting human knowledge and understanding of a model. In contrast, they define *explainability* as an active model characteristic, denoting any action taken by a model with the intent of clarifying or detailing its internal functions. Unlike Barredo Arrieta et al. [4], we consider *transparency* as an active system characteristic concerned with disclosing information to generate knowledge about system elements. Similar to them, we also define *interpretability* and *understandability* as passive characteristics reflecting system knowledge and understanding, and *explainability* as actively clarifying or detailing system elements.

Rosenfeld and Richardson [32] formally define *explainability* and its relationship to *interpretability* and *transparency*, in the case of a ML-based classification algorithm. The authors define *explainability* as the ability for the human user to *understand* the algorithm's logic. This ability to *understand* is achieved from the *explanation*, which they define as the human-centric objective for the user to *understand* the algorithm, using an *interpretation*. *Interpretation/interpretability* is defined as a function mapping data, data schemes, outputs, and algorithms to some representation of the algorithm's internal logic. Furthermore, the authors argue an *interpretation* is *transparent* when the connection between the *interpretation* and algorithm is *understandable* to the human, and when the logic within the *interpretation* is similar to that of the algorithm.

All in all, the work of Rosenfeld and Richardson [32] differs from our work in several ways. First of all, they focus on data-driven XAI and provide formal definitions, whereas our work focuses on goal-driven XAI and provides more general definitions. More importantly, the provided definitions differ from our view. Rosenfeld and Richardson [32] consider *explainability* as passive and subjective, defining it as the ability to *understand*. In contrast, we consider *explainability* as an active system characteristic, and argue their definition of *explainability* reflects *understandability* instead. In addition, the authors consider *interpretability* as active and objective, defining it as providing representations of an algorithm's internal logic. However, we consider *interpretability* as passive and subjective, reflecting user knowledge and understanding of a system/algorithm, and argue their definition of *interpretability* reflects *explainability* instead.

Sanneman and Shah [34] propose an interesting situation awareness-based levels of XAI framework. This framework argues AI-systems part of human-AI teams should explain what the system did or decided (XAI for Perception), why the system did this (XAI for Comprehension), and what the system might do next (XAI for Projection). The authors argue XAI for Comprehension should provide information about causality in the system, aimed at supporting user comprehension of the system's behavior. Examples include explanations linking behavior to the system's goals, constraints, or rules.

This framework broadly aligns with ours, but includes a few differences as well. First of all, we agree with their distinction between providing information for perception and comprehension. However, whereas Sanneman and Shah [34] define both of them as explanations, we refer to XAI for Perception as

transparency/disclosing information, and XAI for Comprehension as *explainability*/clarifying disclosed information. We argue XAI for Projection can be defined as both *transparency* and *explainability*, depending on whether the system discloses next actions (i.e., *transparency*) or also clarifies them (i.e., *explainability*). Furthermore, the framework only focuses on *explaining* AI-system behavior like actions or decisions. However, we argue it is also possible and sometimes even necessary to explain system elements like goals, knowledge, development reasons, or design choices. By doing so, human users can build more complete mental models of the AI-system. Therefore, our framework also incorporates disclosing and clarifying other relevant system elements like goals or knowledge.

Doran et al. [10] introduce an interesting distinction between *opaque*, *interpretable*, and *comprehensible* AI-systems that influenced our work. They define *opaque* AI-systems as systems where the mechanisms mapping inputs to outputs are invisible to users. Consequently, the reasoning of the system is not observable or understandable for users. In contrast, *interpretable* AI-systems are characterized as systems where users cannot only *see*, but also *study* and *understand* how inputs are mapped to outputs. The authors argue that *interpretable* systems imply *transparency* about the underlying system mechanisms. Finally, they define *comprehensible* AI-systems as systems emitting symbols (e.g., words or visualizations) along with their output to allow users to *relate* properties of the input to their corresponding output. According to this classification, *interpretable* systems can be inspected to be understood (i.e., letting users draw *explanations* by themselves), while *comprehensible* systems explicitly provide a symbolic *explanation* of their functioning [8].

This classification of AI-systems is quite similar to the one provided in our work. However, whereas Doran et al. [10] focus on data-driven XAI and argue the notions of *interpretation* and *comprehension* are separate, we focus on goal-driven XAI and argue *understanding/comprehension* implies *interpretation*. More specifically, we claim *transparency* can make *incomprehensible* systems *interpretable*, and *explainability* can make these *interpretable* systems *understandable*. We will explain our definitions, relationships, and classification in detail in the next section.

3 A Two-Dimensional Framework to Classify AI

In this section we present and discuss our two-dimensional explanation framework providing clear distinctions and relations between key XAI concepts (Fig. 1). In short, our framework makes a distinction between *incomprehensible*, *interpretable*, and *understandable* AI-systems, and argues system *transparency* can make *incomprehensible* systems *interpretable*, whereas *explainability* can make *interpretable* systems *understandable*. In the following sections, we will explain and illustrate our framework by introducing our definitions of the concepts *transparency* and *explainability* (Sect. 3.1), and *interpretability* and *understandability* (Sect. 3.2). After that, we illustrate and discuss our framework based on the example of a search and rescue human-agent teaming scenario where a

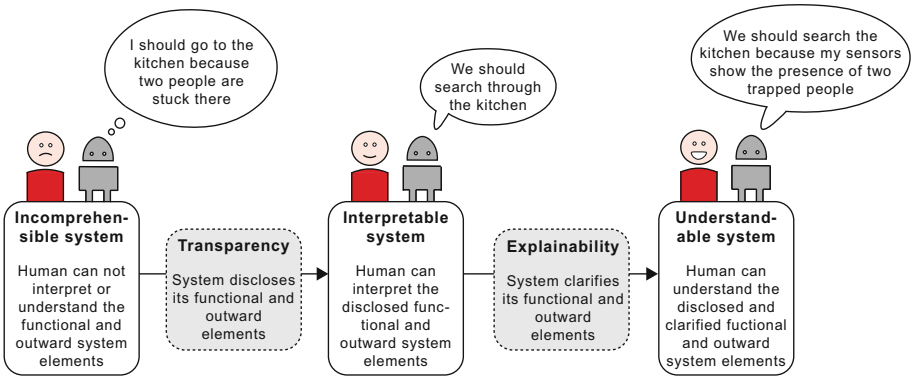

Fig. 1. Two-dimensional explanation framework providing distinctive definitions and relationships between key XAI concepts.

human collaborates with a goal-driven A/IS (Sect. 3.3). Finally, we extend our framework to include some other relevant factors enabled by system *transparency* and *explainability* in Sect. 3.4.

3.1 Transparency vs. Explainability

Whereas most prior work strongly ties or even equates system *explainability* to *interpretability* (e.g., [30,37]), we consider them fundamentally different. Instead, we strongly tie system *transparency* to *explainability*. However, we also argue for a major distinction between these two notions. Inspired by [1] and [36], we define system *transparency* as "*disclosing* the relevant outward and functional system elements to users, enabling them to access, analyze, and exploit this disclosed information". Here, functional system elements concern elements like goals, knowledge, beliefs, decisions, and actions. In contrast, outward elements concern aspects like development reasons, intended users, and design choices.

System *transparency* can answer "*what*"-questions [30] requiring descriptive answers concerning the system elements. Consider, for example, a goal-driven autonomous and intelligent agent collaborating with a human teammate to save victims after an earthquake. According to our definition, system *transparency* is both an *active* [4] and *objective* [8] system characteristic achieved by, for example, disclosing the goal to save all injured children first by collaborating with trained firefighters. By doing so, the human teammate can gain knowledge about these system elements (here a goal and intended users respectively), without necessarily always knowing the relations between them.

The disclosure of relevant elements can be considered *active* in the sense that it is an epistemic and computational *activity* aimed at increasing user knowledge, and *objective* because this activity itself does *not depend on the human's perceptions or State of Mind*. Put differently, the computational implementation of *transparency* is independent of the human user's perceptions and State of Mind,

and thus reproducible in principle [8]. However, the exact effectiveness and content of the disclosed information is a subjective aspect, reflected by measures of *interpretability* and *understandability*.

Inspired by [4,8], and [34] we define system *explainability* as "*clarifying* disclosed system elements by providing information about causality and establishing relations with other system elements, making it easier for users to *understand*, analyze, and exploit this information". *Explainability* can answer "*how*"- and "*why*"-questions [30] requiring clarifying answers concerning the system elements and how they relate and depend on each other. For example, system *explainability* can involve clarifying the disclosed goal to save all children first by linking it to the norm that children are most vulnerable, or that it will not give safety instructions because it assumes the user is a firefighter and familiar with these. Just as *transparency*, we characterize system *explainability* as an *active* [4] and *objective* [8] system characteristic aimed at increasing user knowledge and where the epistemic and computational activity itself does not depend on the human's perceptions or State of Mind.

In summary, the main difference between system *transparency* and *explainability* boils down to *disclosing* vs. *clarifying*. *Transparency* aims to provide descriptive answers providing knowledge about system elements. In contrast, *explainability* aims to ease understanding by clarifying the relations between system elements. Both are considered *active* and *objective* system characteristics, since they are epistemic and computational activities aimed at increasing user knowledge without depending on user's perceptions or State of Mind. We define *transparency* and *explainability* from a system-centric point of view as methods for sharing information, hence the categorization as active and objective/independent from the user. However, we argue that the subjective aspect concerning the effectiveness and content of the shared information also plays a crucial role, as reflected by measures of *interpretability* and *understandability*.

3.2 Interpretability vs. Understandability

In contrast to *transparency* and *explainability*, we define system *interpretability* and *understandability* as *passive* and *subjective* characteristics reflecting user knowledge of the system and depending on the user's State of Mind and background knowledge. In addition, we argue *transparency* makes system *interpretable*, whereas *explainability* makes *interpretable* systems *understandable*. Although we strongly tie *interpretability* to *understandability*, we argue for a major distinction between these two notions as well.

Inspired by [4,8,10] and [36], we define system *interpretability* as "the level at which the system's users can assign subjective meanings, draw explanations, and gain knowledge by accessing, analyzing, and exploiting disclosed outward and functional system elements". Our definition implies *interpretability* is both a *passive* [4] and *subjective* [8] system characteristic. *Passive* in the sense that *interpretability* reflects a degree of user knowledge about system elements, opposite to actively sharing information to generate knowledge (i.e., *transparency*).

Furthermore, *interpretability* can be considered *subjective* in the sense that it is highly dependent on the user's State of Mind and background knowledge [8].

Consider, again, the example of the goal-driven A/IS collaborating with a human to save victims after an earthquake. Disclosing its goal to save all children first enables human users to gain knowledge and assign subjective meanings or draw explanations by themselves (i.e., *interpret*). However, without clarifying the disclosed goal and relating it to other system elements (i.e., *explainability*), these interpretations can vary considerably. For example, the human could draw the conclusion that the system knows/beliefs the area contains a lot of children but only few elderly or adults.

On the other hand, we define system *understandability* as "the level at which the system's users have knowledge of disclosed and clarified outward and functional system elements, and the relationships and dependencies between them". *Understandability* involves knowing *how* and *why* the system reasons and functions, based on *explanations* clarifying and relating disclosed system elements. For example, clarifying the goal to save all children first because they are most vulnerable provides the user with knowledge about the relationship between the goal and a specific norm.

In summary, the main difference between system *interpretability* and *understandability* boils down to a difference in cognitive effort required to have knowledge of the system elements [8]. More specifically, we argue *interpretability* requires more cognitive effort because it implies inferring the meaning of and relations between disclosed information without explicit knowledge of this meaning and relations themselves. In contrast, *understandability* requires less cognitive effort because it implies knowing the meaning of and relations between disclosed and clarified information (facilitated by *explanations*). Both are considered *passive* [4] and *subjective* [8] system characteristics, since they reflect a degree of *user knowledge* about the system *depending on the user's State of Mind and background knowledge*. So we define *interpretability* and *understandability* from a user-centric point of view reflecting the subjective effectiveness of the *transparency* and *explainability* content. Here, *transparency* and *explainability* will be most effective when their content is tailored to the user's State of Mind and background knowledge.

3.3 Two-Dimensional Framework to Classify AI

Our framework (Fig. 1) distinguishes between three types of AI-systems (*incomprehensible*, *interpretable* and *understandable*) and establishes relations between them by integrating the defined concepts of Sect. 3.1 and Sect. 3.2. We will illustrate our framework in the context of a search and rescue human-agent teaming scenario, where a human collaborates with a goal-driven A/IS.

When collaborating with *incomprehensible* systems, humans can not *interpret* or *understand* the system elements because they are not disclosed and clarified. For example, without disclosing and clarifying its decision to search through the kitchen because it perceived stuck people, a human will not be able to interpret or understand the system's behavior. Our framework argues *transparency*

can turn *incomprehensible* systems into *interpretable* ones. By disclosing its relevant functional and outward system elements (i.e., *transparency*), the human can access and exploit this information to assign subjective meanings and gain knowledge (i.e., *interpret*). Consider, for example, an A/IS disclosing the decision to search through the kitchen of a collapsed house to its human teammate. By doing so, the human can utilize this information to interpret that the A/IS perceived something urgent in the kitchen. Furthermore, we argue *explainability* can turn *interpretable* systems into *understandable* systems. By clarifying the disclosed system elements and relations between them (i.e., *explainability*), the human can more easily exploit this information to gain knowledge and build a mental model of the system (i.e., *understandability*). Consider, for example, an A/IS disclosing the decision to search through the kitchen, because it perceived two trapped children there. By providing a belief-based *explanation* for the decision, the system clarifies this decision and how it relates to other system elements like perceptions.

Our proposed framework has several implications. First of all, pursuing system *understandability* should be the ultimate goal, since it can improve collaboration and team performance in human-agent teams [3]. Furthermore, the framework implies that system *transparency* and *explainability* are *active* and *objective* characteristics which can be manipulated by designers to bring about the desired effects. In contrast, system *interpretability* and *understandability* are considered *passive* and *subjective* characteristics which can be measured to validate the effects of *transparency* and *explainability*.

3.4 Extended Framework

We extend our two-dimensional framework to include several often encountered XAI notions. This framework (Fig. 2) mainly illustrates the opportunities system *transparency* and *explainability* can provide to human teammates. Again, we discuss the framework in the context of a search and rescue human-agent teaming scenario where a human collaborates with a goal-driven A/IS.

The extended framework argues that when a system is *interpretable*, it is already both *controllable* and *directable*. Here, we define system *controllability* as "the extent to which human users can change or overrule functional system elements". For example, when the A/IS discloses the decision to search through the kitchen, its human teammate can overrule this decision by changing it to searching the basement instead (i.e., the system is *controllable*).

Next, we define system *directability* as "the extent to which human users can guide the actions of the system". This is different from system *controllability* in the sense that *directability* does not involve changing or overruling system elements, but rather accepting them and guiding the corresponding actions or dividing the work. For example, the human teammate could also accept the disclosed decision to search the kitchen but direct the action of the A/IS by giving the order to enter the kitchen first to assess its safety (i.e., the system is *directable*). Even though system *interpretability* already enables system *controllability* and *directability*, we argue system *understandability* will further improve

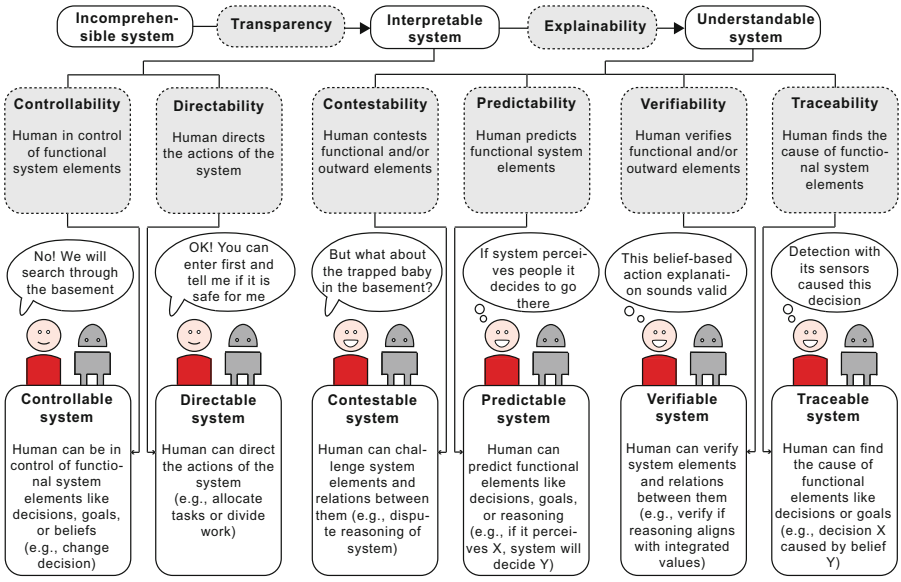

Fig. 2. Extended two-dimensional explanation framework providing distinctive definitions and relationships between key XAI concepts.

these two characteristics. For example, when the human teammate has more knowledge of the system, it can more effectively control and direct its functional elements such as actions or goals.

Furthermore, we argue that system *understandability* enables several other important notions such as system *contestability*, *predictability*, *verifiability*, and *traceability*. We define system *contestability* as "the extent to which human users can challenge or dispute system elements and the relations between them". Again, consider the example of the A/IS disclosing the decision to search through the kitchen, by clarifying it perceived two trapped people there. By doing so, the human teammate can contest this decision and dispute the underlying reason, for example by asking why they should search through the kitchen when there is a trapped baby in another room (i.e., the system is *contestable*).

We argue system *understandability* also enables system *predictability*. We define system *predictability* as "the extent to which human users can estimate future or other functional system elements". Consider the example of a system disclosing its goal to save all children first because of the norm that children are more vulnerable than adults. The human could use this explanation to predict that the agent's next actions will be focused on searching children rather than adults.

The extended framework also argues system *understandability* enables system *verifiability*. Here, we define system *verifiability* as "the extent to which human users can check that the system elements and relations between them make sense and sound valid". We do not refer to formal verification of systems using

formal methods involving mathematical models of systems and analyzing them using proof-based methods. Rather, we refer to a more informal verification of the plausibility of system elements and relations between them. Again, consider the example of a system disclosing its goal to save all children first because of the norm that children are more vulnerable than adults. Based on the provided explanation the human could informally verify that the reasoning aligns with the decision and sounds valid (i.e., the system is *verifiable*).

Finally, we argue system *understandability* enables system *traceability*. We define system *traceability* as "the extent to which human users are able to find the cause of functional system elements like decisions, goals, or beliefs". Again, consider the example of the A/IS disclosing the decision to search through the kitchen, by clarifying it perceived two trapped people there. The human teammate could use the provided explanation to infer that the decision to search the kitchen was caused by the detection of two trapped people.

In summary, the extended framework argues system *interpretability* and *understandability* enable important factors such as system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability*. These factors are crucial for and determine the success of human-agent teams [22–24,33]. Therefore, pursuing system *understandability* should be the main goal when developing AI-systems part of human-agent teams.

4 Discussion

In this work we have presented a two-dimensional explanation framework providing distinctive XAI concept definitions and relationships between them. In this section, we will discuss how our presented relationships between the concepts can be used to guide future research into these relationships. Additionally, we will describe how we believe this framework can be applied in practice.

4.1 Evaluation of Main Framework

Several assumptions arise from the proposed relationships in our presented framework. Below, we introduce these assumptions as claims and describe their corresponding requirements. Next, we discuss whether these assumptions can be evaluated, and how they offer a road map for future research.

- **Claim 1** - System *explainability* results in more knowledge/complete mental models of the system than *transparency*
  Requirement 1 - Manipulating/implementing system *transparency* and *explainability*
  Requirement 2 - Measuring user knowledge of a system
- **Claim 2** - Increased user knowledge of a system results in improved human-agent collaboration and eventually team performance
  Requirement 1 - Subjective and objective measurements of human-agent collaboration

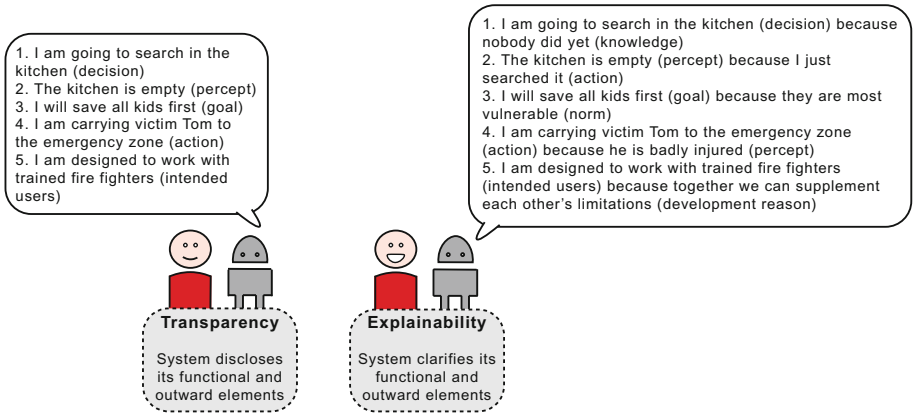

Fig. 3. Examples of system transparency and explainability in the context of a (simulated) search and rescue mission.

We will illustrate how these claims can be evaluated using the example of a simulated search and rescue mission where a human operator and self-explaining A/IS collaborate to search and rescue victims. To validate Claim 1, implementing system *transparency* and *explainability* would be required. Examples of implementing system *transparency* involve disclosure of the system's goals, decisions, and intended users. Figure 3 shows several examples of system *transparency* in the context of the search and rescue mission.

Implementing system *explainability* can be achieved in many different ways. However, a fundamental requirement is providing information about causality in the system and establishing relations between system elements. Existing approaches from the XAI literature include explanations of actions based on state information [2,19,27]; explanations of actions based on goals [5,17,18]; explanations of decisions based on demonstrating that alternative decisions would be sub-optimal [35]; and sequence-based explanations clarifying the next action(s) [5,17]. Figure 3 shows several concrete examples of system *explainability* in the context of the search and rescue operation.

Validating Claim 1 would also require the measurement of human user knowledge and understanding of the system, which can be done both subjectively and objectively. Subjective examples from the XAI literature include asking questions related to perceived understandability of the system and its model [20], and asking users to choose which of two possible system outputs is of higher quality (implicitly measures understanding) [11]. However, objectively measuring user knowledge and understanding of the system would be a more robust indicator than the subjective alternatives.

Currently, objective methods and metrics for measuring user knowledge and understanding of systems are lacking. Nevertheless, Sanneman and Shah [34] propose a relevant method based on the widely-used and empirically validated Situation Awareness Global Assessment Technique (SAGAT) [12,13]. In short,

their proposed technique involves freezing simulations of representative tasks at random time points, followed by asking questions measuring user knowledge about information related to system behavior. It is crucial to first define the human informational needs related to system behavior. Accordingly, a list of questions regarding the informational needs can be specified and used to measure user knowledge of the system.

Whereas Sanneman and Shah [34] focus solely on measuring user knowledge related to AI-system behavior, the test/technique can also be extended to include information related to other relevant system elements like goals, knowledge, decisions, or even development reasons. Some example questions based on the information in Fig. 3 include "Which room will the agent search next?"; "What is the current action of the agent?"; "Why is the agent going to search in the kitchen?"; and "Why will the agent save all kids first?".

Validating Claim 2 would require the subjective and objective measurement of human-agent collaboration and team performance. Subjective measures could include user satisfaction [7] or system usability [6], whereas objective measures could include aspects like the number of victims rescued or seconds required to finish tasks. The outlined example experiment, discussed example implementations of *transparency* and *explainability*, and suggested metrics for measuring user knowledge, human-agent collaboration, and team performance can be used as a road map for future work aimed at validating the assumptions arising from our framework.

4.2 Evaluation of Extended Framework

Several assumptions arise from the proposed relationships in our extended framework as well. Below, we introduce these assumptions as claims and describe their corresponding requirements. Next, we discuss whether these assumptions can be evaluated, and how they offer a road map for future research.

- **Claim 3** - System *transparency* already enables system *controllability* and *directability*, but not system *contestability*, *predictability*, *verifiability*, and *traceability*
  Requirement 1 - Implementing system *transparency*
  Requirement 2 - Measuring system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability*
- **Claim 4** - System *explainability* enables system *contestability*, *predictability*, *verifiability*, and *traceability*
  Requirement 1 - Implementing system *explainability*
  Requirement 2 - Measuring system *contestability*, *predictability*, *verifiability*, and *traceability*

Validating Claims 3 and 4 would require implementing system *transparency* and *explainability*, and measuring system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability*. An example of subjectively measuring these system characteristics could be freezing the simulated experiment at random points, followed by measuring perceived system *controllability*,

directability, *contestability*, *predictability*, *verifiability*, and *traceability*. One approach involves Likert-scale questions[3] asked to the human users. Table 2 shows example questions for each of these variables, though full questionnaires would require more research and validation of the exact scales. The outlined example experiment, discussed example implementations of *transparency* and *explainability*, and suggested metrics for measuring system *controllability*, *directability*, *contestability*, *predictability*, *verifiability*, and *traceability* can be used as a road map for future work aimed at validating the assumptions arising from our extended framework.

Table 2. Example questions for subjectively measuring the system variables in the extended framework.

Variable	Example question
Controllability	"I feel like I can change the system's decision"
Directability	"I feel like I can guide the system's behavior"
Contestability	"I feel like I can challenge the system's decision"
Predictability	"I feel like I can predict the system's next action"
Verifiability	"I feel like I can check that the system's behavior makes sense"
Traceability	"I feel like I can find the cause of the system's decision"

4.3 Application of Framework

Here we briefly address how our framework can be used in practice. Specifically, what difference can the framework make when developing systems part of human-agent teams? Consider the example of developing an autonomous and intelligent drone which should collaborate with a human operator (e.g., a firefighter) during the aftermath of an earthquake. The goal of the team is to search and rescue trapped victims as soon as possible. Our framework can be particularly helpful by mapping specific types of context and informational needs onto requiring either system *transparency* or *explainability*. For example, the drone can be developed/implemented in such a way that when the workload or time pressure is high, the drone displays *transparency* only. Similarly, contextual factors that could be mapped onto system *explainability* include low time pressure and operator workload, or when the user has an imprecise mental model of the system. In this way, the framework can contribute to developing adaptive systems able to tailor their communication of relevant information to the needs and requirements of both users and situations.

4.4 Future Work

Based on the work presented in this paper, we identify a few key ideas for future work. A possible first direction could be to conduct experiments aimed at vali-

[3] For example ranging from "Totally Disagree" to "Totally Agree" on a 7-point scale.

dating the assumptions arising from our framework. Some ideas, requirements, and examples concerning this validation have been discussed in more detail in Sect. 4.1 and Sect. 4.2.

For now, our framework focuses on sharing information regarding mental constructs like decisions or goals. A relevant suggestion for future work would be to extend the framework with a more physical domain as well by including literature/perspectives from explainable and understandable robots. For example, the role of visual and body cues could be incorporated in the framework. Furthermore, our provided framework is rather broad/general and informally defined. Therefore, another suggestion would be to formalize it and make it more concrete by providing examples in terms of different computational frameworks/architectures (e.g., transparency vs. explainability differences between agents using BDI vs. PDDL models). In addition, we currently do not consider situations where the user may be under the false impression of understanding the system, but only consider cases where their understanding actually matches the system's models/elements. We also do not consider different roles taken by human and agent, such as commander or supervisor. In future work, it would be interesting to extend the framework by including these two aspects, and see how it affects our proposed definitions.

Another future direction for this work would be to extend the framework to include context- and user-awareness required for tailoring system *transparency* and *explainability* to specific needs and requirements. The need for personalized and context-dependent system *transparency* and *explainability* is one of the main goals within XAI community and research [3]. However, the actual implementation and investigation is still somewhat in its infant stages. Currently, our proposed framework does not address context- and user-dependent system *transparency* and *explainability*, so this would be a relevant suggestion for future work. Ideas involve mapping specific types of context or user knowledge to requiring either system *transparency* or *explainability*. Furthermore, these aspects could also be mapped onto *transparency* and *explanation* modality/presentation instead of just content. Examples include mapping high workload to system *transparency*, rudimentary system knowledge to *explainability*, or visual thinkers to receiving visual *explanations* and verbal thinkers receiving textual *explanations*. Another idea involves adapting system *transparency* and *explainability* based on the interdependence relationship between human and system. For example, the system could adapt its communication based on whether joint activity is required (i.e., hard interdependence) or when joint activity is optional (i.e., soft interdependence) [21,22].

5 Conclusion

In this paper, we propose a two-dimensional explanation framework introducing clear distinctions and relationships between the key XAI notions *transparency*, *interpretability*, *explainability*, and *understandability*. This concise and comprehensive framework explicitly addresses the lack of consensus and ambiguity problem surrounding these concepts. We argue that adopting our distinctive concept

definitions and mutual relations can greatly benefit XAI community, as clearly defining concepts and relationships between them is a pre-requisite for both the implementation and evaluation of these concepts. Furthermore, the framework yields a classification of AI-systems as *incomprehensible*, *interpretable*, or *understandable*, guiding the research and development to establish understandable AI (e.g., by setting requirements for *contestability*, *predictability*, *verifiability* and *traceability*).

Acknowledgements. This work is part of the research lab AI*MAN of Delft University of Technology.

References

1. Alvarado, R., Humphreys, P.: Big data, thick mediation, and representational opacity. New Lit. Hist. **48**(4), 729–749 (2017). https://doi.org/10.1353/nlh.2017.0037
2. Amir, D., Amir, O.: Highlights: summarizing agent behavior to people. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1168–1176 (2018)
3. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, 13–17 May 2019, pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019)
4. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012. http://www.sciencedirect.com/science/article/pii/S1566253519308103
5. Broekens, J., Harbers, M., Hindriks, K., van den Bosch, K., Jonker, C., Meyer, J.-J.: Do you get it? User-evaluated explainable BDI agents. In: Dix, J., Witteveen, C. (eds.) MATES 2010. LNCS (LNAI), vol. 6251, pp. 28–39. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16178-0_5
6. Brooke, J.: SUS: a quick and dirty usability. In: Usability Evaluation in Industry, p. 189 (1996)
7. Chin, J.P., Diehl, V.A., Norman, K.L.: Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 1988, pp. 213–218. Association for Computing Machinery, New York (1988). https://doi.org/10.1145/57167.57203
8. Ciatto, G., Schumacher, M.I., Omicini, A., Calvaresi, D.: Agent-based explanations in AI: towards an abstract framework. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) EXTRAAMAS 2020. LNCS (LNAI), vol. 12175, pp. 3–20. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51924-7_1
9. van Diggelen, J., et al.: Pluggable social artificial intelligence for enabling human-agent teaming. arXiv preprint arXiv:1909.04492 (2019)
10. Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. CoRR abs/1710.00794 (2017). http://arxiv.org/abs/1710.00794
11. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

12. Endsley, M.R.: Situation awareness global assessment technique (SAGAT). In: Proceedings of the IEEE 1988 National Aerospace and Electronics Conference, vol. 3, pp. 789–795 (1988). https://doi.org/10.1109/NAECON.1988.195097
13. Endsley, M.R.: A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM. Hum. Factors **63**(1), 124–150 (2021). https://doi.org/10.1177/0018720819875376. pMID: 31560575
14. Goldman, A.I., et al.: Theory of mind. In: The Oxford Handbook of Philosophy of Cognitive Science, vol. 1 (2012)
15. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5) (2018). https://doi.org/10.1145/3236009
16. Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), nd Web **2**(2) (2017)
17. Harbers, M., van den Bosch, K., Meyer, J.: Design and evaluation of explainable BDI agents. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 2, pp. 125–132 (2010). https://doi.org/10.1109/WI-IAT.2010.115
18. Harbers, M., Bradshaw, J.M., Johnson, M., Feltovich, P., van den Bosch, K., Meyer, J.-J.: Explanation in human-agent teamwork. In: Cranefield, S., van Riemsdijk, M.B., Vázquez-Salceda, J., Noriega, P. (eds.) COIN -2011. LNCS (LNAI), vol. 7254, pp. 21–37. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35545-5_2
19. Hayes, B., Shah, J.A.: Improving robot controller transparency through autonomous policy explanation. In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction, HRI, pp. 303–312 (2017)
20. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects (2019)
21. Johnson, M., Bradshaw, J.M., Feltovich, P.J., Jonker, C.M., van Riemsdijk, B., Sierhuis, M.: The fundamental principle of coactive design: interdependence must shape autonomy. In: De Vos, M., Fornara, N., Pitt, J.V., Vouros, G. (eds.) COIN -2010. LNCS (LNAI), vol. 6541, pp. 172–191. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21268-0_10
22. Johnson, M., Bradshaw, J.M., Feltovich, P.J., Jonker, C.M., van Riemsdijk, M.B., Sierhuis, M.: Coactive design: designing support for interdependence in joint activity. J. Hum.-Robot Interact. **3**(1), 43–69 (2014). https://doi.org/10.5898/JHRI.3.1.Johnson
23. Johnson, M., Vera, A.: No AI is an Island: the case for teaming intelligence. AI Mag. **40**(1), 16–28 (2019). https://doi.org/10.1609/aimag.v40i1.2842. https://ojs.aaai.org/index.php/aimagazine/article/view/2842
24. Klien, G., Woods, D.D., Bradshaw, J.M., Hoffman, R.R., Feltovich, P.J.: Ten challenges for making automation a "team player" in joint human-agent activity. IEEE Intell. Syst. **19**(6), 91–95 (2004). https://doi.org/10.1109/MIS.2004.74
25. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable agency for intelligent autonomous systems. In: AAAI 2017, pp. 4762–4763 (2017)
26. Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
27. Lomas, M., Chevalier, R., Cross, E.V., Garrett, R.C., Hoare, J., Kopack, M.: Explaining robot actions. In: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI 2012, pp. 187–188. Association for Computing Machinery, New York (2012). https://doi.org/10.1145/2157689.2157748

28. Malle, B.F.: How the Mind Explains Behavior. Folk Explanation, Meaning and Social Interaction. MIT-Press, Cambridge (2004)
29. Malle, B.F.: Attribution theories: how people make sense of behavior. Theor. Soc. Psychol. **23**, 72–95 (2011)
30. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007. https://www.sciencedirect.com/science/article/pii/S0004370218305988
31. Neerincx, M.A., van der Waa, J., Kaptein, F., van Diggelen, J.: Using perceptual and cognitive explanations for enhanced human-agent team performance. In: Harris, D. (ed.) EPCE 2018. LNCS (LNAI), vol. 10906, pp. 204–214. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91122-9_18
32. Rosenfeld, A., Richardson, A.: Explainability in human-agent systems. Auton. Agent. Multi-Agent Syst. **33**(6), 673–705 (2019). https://doi.org/10.1007/s10458-019-09408-y
33. Salas, E., Sims, D.E., Burke, C.S.: Is there a "big five" in teamwork? Small Group Res. **36**(5), 555–599 (2005). https://doi.org/10.1177/1046496405277134
34. Sanneman, L., Shah, J.A.: A situation awareness-based framework for design and evaluation of explainable AI. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) EXTRAAMAS 2020. LNCS (LNAI), vol. 12175, pp. 94–110. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-51924-7_6
35. Sreedharan, S., Srivastava, S., Kambhampati, S.: Hierarchical expertise level modeling for user specific contrastive explanations. In: IJCAI, pp. 4829–4836 (2018)
36. Turilli, M., Floridi, L.: The ethics of information transparency. Ethics Inf. Technol. **11**(2), 105–112 (2009). https://doi.org/10.1007/s10676-009-9187-9
37. Vilone, G., Longo, L.: Explainable artificial intelligence: a systematic review. arXiv preprint arXiv:2006.00093 (2020)
38. Walmsley, Joel: Artificial intelligence and the value of transparency. AI Soc. 1–11 (2020). https://doi.org/10.1007/s00146-020-01066-z