A Situation Awareness-Based Framework for Design and Evaluation of Explainable AI

Lindsay Sanneman(✉) and Julie A. Shah

Massachusetts Institute of Technology, Cambridge, MA 02139, USA
{lindsays,julie_a_shah}@csail.mit.edu

Abstract. Recent advances in artificial intelligence (AI) have drawn attention to the need for AI systems to be understandable to human users. The explainable AI (XAI) literature aims to enhance human understanding and human-AI team performance by providing users with necessary information about AI system behavior. Simultaneously, the human factors literature has long addressed important considerations that contribute to human performance, including how to determine human informational needs. Drawing from the human factors literature, we propose a three-level framework for the development and evaluation of explanations about AI system behavior. Our proposed levels of XAI are based on the informational needs of human users, which can be determined using the levels of situation awareness (SA) framework from the human factors literature. Based on our levels of XAI framework, we also propose a method for assessing the effectiveness of XAI systems.

Keywords: Explainable AI · Human-AI collaboration · Interpretability

1 Introduction

With the recent focus on explainable artificial intelligence (XAI) in the AI literature, defining which information XAI systems should communicate and how to measure their effectiveness is increasingly important. Gunning and Aha [21] define XAI as "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future." We adopt this definition of XAI and define explanations as the information necessary to support human inference of the above details about AI systems, including information about their inputs, models, and outputs. The motivation for development of XAI techniques is often stated as the need for transparency within increasingly complex AI systems [20,31] and the need to engender user trust in increasingly opaque systems [6,20,31]. Both increasing AI system transparency and accounting for human trust in these systems contribute to improved human-AI team performance; thus, supporting

D. Calvaresi et al. (Eds.): EXTRAAMAS 2020, LNAI 12175, pp. 94–110, 2020.
https://doi.org/10.1007/978-3-030-51924-7_6

human-AI team performance is one of the primary aims of XAI. Some literature argues that there is a performance-explainability trade off in that more explainable AI systems sacrifice algorithmic performance in some way [21,31]. However, if a lack of system explainability inhibits overall team performance, benefits provided by improved algorithmic performance might be lost. Therefore, we view optimization of human-AI team performance, enabled by explanations about the system's behavior, as the primary goal of XAI.

There exists a rich literature in human factors that explores scenarios in which humans interact with automated systems, as well as the various factors that influence human performance during task execution. The concept of situation awareness (SA), which has been studied within the field of human factors and in the context of human-automation teams [9,13], defines the informational needs for humans operating in any scenario [13]. XAI systems, as systems that provide information about AI behavior, can contribute to the subset of a human user's SA that is related to AI behavior. Human-AI team performance can be improved through information provided by XAI systems that support SA; however, overall SA, in addition to the subset of SA supported by XAI, are necessary for but not solely sufficient to support team performance [13].

The human factors literature has additionally introduced methods and metrics for assessment of a human's SA [37]. Just as SA supports but is not equivalent to performance, high-quality explanations provided by XAI systems support but are not equivalent to SA. Assessing XAI systems based on methods related to SA can contribute to an understanding of whether the provided explanations achieve the ultimate goal of enhancing human-AI team performance. Measuring SA as an intermediate aim of XAI can provide clarity as to the potential confounds that exist in performance assessment. The XAI literature currently lacks a comprehensive set of suitable methods and metrics for assessing explanation quality. While it may not be possible to explicitly and independently define an explanation's quality, explanations are only "good" insofar as they contribute to intermediate goals, such as SA, and the ultimate goal of improved performance. In this paper, we discuss how a human factors-based SA assessment method can be useful for evaluating XAI systems.

The remainder of the paper is organized as follows: in Sect. 2, we discuss the relevant situation awareness literature as it relates to XAI. In Sect. 3, we propose a framework for design and evaluation of XAI systems in light of findings within the human factors community. In this framework, we propose levels of XAI that define which information about AI algorithms and processes should be supported by XAI systems; these levels map closely to those of SA as proposed by Endsley [13] (discussed in Sect. 2). Our framework applies to XAI generally, including explainable machine learning (ML), explainable agents/robots, and multi-agent/multi-human teams. There exist other frameworks in the XAI literature that are primarily agent-centric in that they categorize systems based on agent attributes, such as stages of explanations [3,36], types of errors [42,43], or agent internal cognitive states [23]. The framework we propose is complementary to these in that ours is human-centric and focuses on human informational

needs. Other frameworks propose human-centric approaches [38,40], but these are largely human role-based, and our framework applies more generally and is role-agnostic. One other framework focuses on the theory of mind (ToM) of the robot and human [26]. The authors of that work discuss the need to define which information a robot should communicate, which our framework addresses.

In Sect. 4 we provide a non-comprehensive set of examples of how a set of existing XAI techniques fit into our framework in order to clarify how our framework might be applied. Section 5 discusses how to determine human informational needs at each of the three levels proposed in our framework. In Sect. 6, we discuss how methods used to evaluate existing XAI techniques map to assessments of SA from the human factors literature, and we propose one key SA-related method for the assessment of XAI systems. Section 7 provides a motivating example, which we use to clarify our discussion of the levels of XAI and the suggested SA-related assessment method. Finally, Sect. 8 suggests future directions for XAI research, and Sect. 9 concludes the paper.

2 Situation Awareness in the Human Factors Literature

The concept of situation awareness has been widely studied in the human factors literature, especially in the context of human-automation teams operating in complex environments [13]. The concept originally received attention in the study of aviation systems, particularly with the rise of cockpit automation and the need to support pilot awareness of aircraft behavior [46]. However, its applicability extends to any complex scenario in which humans have informational needs for achieving the tasks they are performing. Accordingly, it has additionally been studied in the context of many other domains including air traffic control, emergency management, health care, and space, among others [15].

Different definitions of situation awareness and corresponding frameworks have been proposed in the literature [5,13,44]. We adopt the three-level definition from Endsley [13]: "the perception of elements in the environment within a volume of time and space (level 1), the comprehension of their meaning (level 2), and the projection of their status in the near future (level 3)." This definition is the most widely cited and applied of the existing definitions [47]. It has direct value for designers of complex systems due to its relative simplicity and its division into three levels, which allow for easy definition of SA requirements for different scenarios and for effective measurement of a person's SA [41]. The SA construct has been empirically validated in various contexts [16,47], and connections between SA and other task-related measures such as performance and error frequency have been demonstrated in the literature [15]. SA has also been used to define a framework for agent transparency [9], which focuses primarily on information that interfaces should display about agent behavior. We apply a similar approach to that of Chen et al. [9], but we focus on XAI specifically and define our framework based on AI system behavior more generally.

Endsley [13] further defines an assessment technique for measuring a person's SA: the Situation Awareness Global Assessment Technique (SAGAT). Since SA

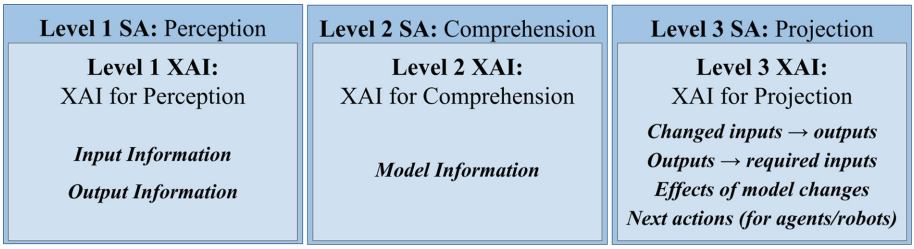

Fig. 1. Levels of XAI framework

represents the "diagnosis of the continuous state of a dynamic world", there exists a "ground truth" against which a person's SA can be measured [37]. The SAGAT test aims to measure the discrepancies between a human user's SA, or their knowledge of the state of the world, and this "ground truth" state of the world. We detail SAGAT and its applicability to XAI further in Sect. 6.4.

SA is relevant to the XAI community since it contributes to defining human informational needs, and XAI aims to meet them. In particular, XAI provides human users with the subset of their SA that relates to AI behavior. It is not equally valuable to provide just any information to human users via XAI, but only information that is relevant to them given their respective tasks and contexts. In fact, providing excessive or irrelevant information can be detrimental to human-AI team performance by causing confusion or unnecessarily increasing workload [37]. Therefore, it is important for XAI practitioners to consider which information is relevant to users and then to measure whether users have received and understood that information. Our proposed framework provides a guideline for determining which information XAI systems should communicate about AI system behavior, and our suggested use of the SAGAT method provides a way to measure how effectively this information is delivered.

3 Situation Awareness-Based Levels of XAI Framework

As AI systems become increasingly ubiquitous and humans interact with more complex AI systems, XAI support of adequate SA can benefit human-AI team performance. According to the definition of SA provided by Endsley [13], an individual working towards a goal requires all three levels of SA to support their decision-making processes, which can in turn improve performance of goal-oriented tasks. It is important to note the distinction between general SA (related to the situation as a whole) and SA related specifically to AI behavior: the latter is a subset of the former and is the focus of this paper. SA, in the most general sense, comprises user awareness of the environment, other situational factors, and other human teammates in addition to information about the AI's behavior.

The informational needs defined by SA can serve to dictate the information XAI systems should provide about AI behavior. For many scenarios in which XAI systems are useful and relevant, humans in the loop must know what the

AI system did or what decision it made (perception), understand why the system took the action or made the decision it did and how this relates to the AI's own sense of its goals (comprehension), and predict what the system might do next or in a similar scenario (projection). Thus, just as SA is divided into three levels, we introduce three levels of XAI systems. Our proposed framework is shown in Fig. 1. The three levels of XAI in our framework are defined as follows:

1. Level 1: XAI for Perception - explanations of what an AI system did or is doing and the decisions made by the system
2. Level 2: XAI for Comprehension - explanations of why an AI system acted in a certain way or made a particular decision and what this means in terms of the system's goals
3. Level 3: XAI for Projection - explanations of what an AI system will do next, what it would do in a similar scenario, or what would be required for an alternate outcome

Our framework generalizes to cover both explainable ML and explainable agents/robots. It can also be applied for both "black box" AI systems that are fundamentally uninterpretable to human users and high-complexity systems that may or may not be inherently interpretable/"white box" but that human users cannot grasp due to their complexity. Note that our focus is on the informational content of explanations rather than explanation modality (natural language, communicative actions, etc.), which is a separate but important consideration. The following sections further detail each of the levels of XAI in our framework.

3.1 Level 1: XAI for Perception

Level 1 XAI includes explanations about what an AI system did or is doing as well as the decisions made by the system. It covers information about both AI system inputs and outputs and aims to answer "what" questions as they are defined by Miller [34]. In the context of explainable ML, level 1 information might include inputted data or outputted classification, regression, or cluster information, for example. For explainable agents and robots, level 1 information could include inputted state information, a particular decision or action taken by the system, an outputted plan/schedule (sequence of decisions/actions) from a planning agent, a particular resource allocation, and others. While level 1 XAI might seem straightforward in many applications since it is simply information about a system's inputs or outputs, providing this information might be challenging when explaining a complex model that makes decisions over many different input factors and produces numerous outputs, only a subset of which are relevant to the user. The primary technical challenge for level 1 XAI is determining which specific information is relevant to users of complex systems.

3.2 Level 2: XAI for Comprehension

Level 2 XAI includes explanations about why an AI system acted in a particular way or made a certain decision and what this means in terms of the system's

goals. The primary aim of level 2 XAI is to provide information about causality in AI systems [22] as it relates to a specific instance or decision made by the system. Level 2 XAI answers "why" questions (as defined by Miller [34]) and typically provides information about a system's model. In the context of explainable ML, level 2 information might relate to sensitivities to inputs, semantic feature information, simplified feature or model representations, cluster information, or abstracted representations of model details. For explainable agents and robots, level 2 information could include details about system goals, objectives, constraints, pre-/post-conditions, rules, policies, costs, or rewards.

In identifying level 2 XAI informational requirements, it is important to identify which causal information is most relevant to a user attempting to understand the system. Miller [34] states that explanations are fundamentally contrastive and that when humans seek explanations, they often have a particular "foil" (defined by the author as a counterfactual case) in mind. Reasoning about the most likely foils users have in mind when interacting with a system can help determine which causal information to provide. Note that by our definition, level 2 XAI provides answers to "why" questions for specific instances or in relation to specific foils and might only involve some limited information about a system's model. Therefore, level 2 explanations alone do not necessarily enable users to make all necessary predictions; as such, information beyond level 2 XAI may be required for projection (level 3). We detail this distinction further in Sect. 3.3.

3.3 Level 3: XAI for Projection

Level 3 XAI includes explanations about what an AI system will nominally do next or would do in a different circumstance or context. Level 3 XAI provides answers to "what if"/"how" questions as they are defined by Miller [34]. It aims to explain what would happen if certain system inputs or parameters changed or what the system would do if human users took particular actions. Level 3 XAI incorporates counterfactual or other simulated information in order to provide explanations about a system's future behavior in the presence of changes to either inputs or system parameters, which might occur due to human actions.

While level 2 XAI provides information about why a decision was made based on model-related factors, level 3 XAI provides insight as to what degree of change to inputs, model parameters, or constraints would yield a different outcome. Further, while level 2 explanations provide information about a decision made in a specific instance, level 3 information helps users to reason about what will happen in different contexts and what exactly would need to change about the given circumstance in order to alter the system's output. In the context of explainable ML, level 3 information could include information about what effect a changed input would have on the output, which changes in the input would be required to achieve a given output, or what would change about the output if the model changed in some way. Similarly, for explainable agents and robots, level 3 information would provide information about changed inputs and outputs, changed models (such as the addition/removal of constraints or differently weighted objectives), or the nominal continued course of action.

We define two types of prediction that can be supported with level 3 explanations. First, backward reasoning helps a user start with a desired outcome and work backwards to determine what would be necessary to achieve that outcome. For example, consider a situation in which a user interacting with a neural network hopes to understand what type of input would be required for a particular classification. In such a case, successful level 3 XAI would help the user understand the ranges of inputs to the neural network that would result in the desired classification. Second, forward simulation helps a user understand what will happen given any changes in the inputs or model that occur. An example of forward simulation in a robot planning scenario might involve a user who hopes to add a constraint to the planning problem based on their own preferences about the robot's actions. Successful level 3 XAI would help such a user to understand the effect the new constraint would have on the outputted plan.

4 Example Approaches Achieving the Levels of XAI from the XAI Literature

The following sections discuss how a limited, non-comprehensive set of example XAI techniques fit into our framework.

4.1 Example Approaches Achieving Level 1 XAI

Level 1 XAI relates to AI system inputs and outputs. Whether a system has provided adequate level 1 explanations depends on whether a human user has sufficient information about these things. Many explainable ML techniques provide level 1 XAI implicitly through their inputs and outputted results. For example, Kim et al. [30] and Ribeiro et al. [39] provide users with the system's outputted classification (level 1) in addition to explanations about the reasons behind the outputs (level 2). In these cases, the entire system output is captured by a single or small number of classifications, and the human user can easily understand the entire set of outputs. In other cases, such as with some clustering techniques, the entire set of outputs (i.e. features that represent a cluster) contains extraneous information in addition to information that is directly relevant to the human user's understanding of the outputted clusters. Kim et al. [29] designate a set of clusters in a feature space, find the most quintessential prototype of each, and, for each prototype, down-select to a subset of features to present to the user.

In the explainable agents and robots literature, explainable Belief-Desire-Intention (BDI) agents explain their actions (intentions) based on their goals (desires) and their observations (beliefs) [7,23,24]. Belief-based explanations are level 1 explanations, since they provide information about inputs that agents use in their decision-making processes. Harbers et al. [24] implement a BDI agent that produces explanations of both its observations (inputs) and actions (outputs), which both constitute level 1. Beyond BDI agents, Floyd and Aha [19] implement an agent that explains when it changes its behavior (output) in order to increase transparency. Lomas et al. [32] propose a framework for explainable

robots which includes explanations about which actions a robot took (outputs) and what information it had about the world at the time (inputs). Finally, AI planning systems that provide users with a partial or entire plan [6,8,45] implicitly provide level 1 XAI through their outputted plans.

4.2 Example Approaches Achieving Level 2 XAI

Level 2 XAI is fundamentally related to supporting user comprehension of a system's behavior through the understanding of its model, including reasoning about objectives, constraints, features, or other model aspects. Successful level 2 XAI adequately explains the relevant aspects of why a system behaved the way it did. Much of the current XAI literature falls into the category of level 2 XAI.

Various XAI techniques for ML models aim to explain which features, parts of the model, or other feature abstractions have the greatest bearing on a system's decision making. Ribeiro et al. [39] introduce the LIME technique, which learns an approximation of a complex classifier over a human-understandable set of features in order to explain which of these features were most important in generating a classification for a given input. Kim et al. [30] propose a technique that allows users to define abstract concepts (which may be distinct from the original set of features used for classification) and learn about the significance of a concept's contribution to a given classification. Other approaches, such as saliency maps, highlight important aspects of inputs [1].

In the explainable agent and robot literature, explainable BDI agents that explain their actions based on their goals (desires) [7,23,24] contribute to level 2 XAI. The agent proposed by Floyd and Aha [19] provides explanations about why it changes its behavior (level 2) based on user feedback. Hayes and Shah [25] propose a policy explanation technique that can answer questions about why an agent did not take a given action by reasoning about predicates that constitute its state. The technique proposed by Dannenhauer et al. [10] explains agent behavior based on the agent's rationale and goal. Dragan et al. [12] discuss the distinction between legible and predictable robot motions. By their definition, legible robot motions support human inference of the robot's goal and would therefore be considered level 2 XAI. Work related to explainable planning has proposed explanations according to human-understandable aspects of AI models, such as predicates or system objectives. Sreedharan et al. [45] introduce a technique that explains model predicates to a user in order to fill perceived gaps in the user's understanding of the model based on foils they suggest. Finally, Borgo et al. [6] propose a set of techniques that explain system decisions by incorporating user-produced foils into planning and demonstrating that the modified plans are sub-optimal or infeasible.

4.3 Example Approaches Achieving Level 3 XAI

Fundamentally, level 3 XAI is about supporting user prediction of AI behavior through enabling understanding of what a system would do if its inputs changed or if the model were to change in any way. Successful level 3 XAI helps users to

predict what a system will do next or what it would do in a different context and answers "what if" questions about system behavior.

In the explainable ML literature, some approaches provide users with predictions of contexts in which an AI system will fail [4] or predictions of which changes in inputs would be required to amend misclassified examples [33]. Others that provide level 2 information could be extended to support level 3. For example, the SP-LIME algorithm [39] chooses a subset of local model approximations produced by the LIME algorithm (discussed in Sect. 4.2) in order to provide a more "global" explanation of the interpretable features that impact classification in different scenarios. Ideally, if these examples are chosen according to human informational needs for prediction, the human user would be able to predict the outcome of a new example. However, with very complex systems, adequately providing information in this manner might be intractable, and other ways of providing level 3 explanations might be necessary. Other methods, such as the one described by Kim et al. [30] (discussed in Sect. 4.2), could be augmented to provide combinations of relevant "concepts" or could be complemented with other contextual information in order to support prediction more fully.

In the explainable agent and robot literature, Amir and Amir [2] provide explanations of global agent behavior by selecting "important" states in the state space and providing traces of subsequent states and actions (determined by the agent's policy). These state-action pairs support human user prediction of future agent behavior. The policy explanation technique proposed by Hayes and Shah [25] can support both backward reasoning by answering questions about when (from which states) it will take certain actions and forward reasoning by answering questions about what the agent will do given different states. Some explainable agents provide more direct prediction-related information by explaining their next action(s), such as explainable BDI agents that provide sequence-based explanations [7,23] and others that provide their plans [19]. Note that providing users with plans that agents are executing online is level 3 XAI, while providing users with plans outputted by a planning agent is level 1 XAI. Finally, in the discussion of legibility versus predictability [12], predictability is related to human inference of a robot's actions based on a known goal, so we categorize predictable robot motions as level 3 XAI. As with explainable ML, information provided by level 2 XAI techniques can be combined and amended in order to produce level 3 XAI to support prediction of future robot or agent actions.

5 Determining Human Informational Needs

In designing XAI systems and measuring their effectiveness, defining human informational requirements according to the above framework is of value. This information depends upon the overall goal of the human-AI team and the individual roles of the autonomous agent(s) and human(s) within that team. Endsley [16] describes a process called goal-directed task analysis (GDTA) for determining SA requirements for a given context, both for individuals and for those operating in larger teams. In this process, the major goals of each human teammate's

task are identified along with their associated sub-goals. Then, required decisions associated with each sub-goal are enumerated. Finally, SA requirements at all three levels are defined for each of these decisions (i.e. the information required to support human decision-making). The GDTA process is detailed at length in [16] and can be applied by XAI practitioners to define which information human users need about AI system behavior in order to achieve their respective goals. The definition of informational requirements with GDTA also informs the assessment of XAI systems, which will be discussed further in Sect. 6.

In many scenarios, users do not require information about all of a system's behavior but only the aspects that are relevant to their specific tasks. Often, a human cannot possess information about the entirety of a complex system's behavior; therefore, defining the specific information that users require (through GDTA or a similar process) in order to support human-AI team goals is critical. This is especially relevant when considering teams of humans, who each have their own roles and corresponding goals. Informational requirements in these cases are user-specific, and consequently, XAI systems might need to be able to adapt to users playing different roles in the team, providing each with the specific information relevant to his or her own task and potentially at different levels of abstraction. An extended discussion of the definition and support of team SA is provided by Endsley and Jones [18].

One important aspect of team SA is the interdependence of individual team members. Johnson et al. [28] detail an "interdependence analysis" process for assessing individual team members' needs given different possible team configurations. This process results in the definition of observability (level 1) and predictability (level 3) requirements for each teammate in the context of their interdependence on each other. Since it defines information-sharing requirements in the team, it can also be useful for defining information requirements for XAI systems given different possible team configurations. We recommend using a modified version of this process that includes the definition of "comprehensibility" requirements (level 2) in order to define which role an XAI system should play in the context of a team. Once informational needs are identified, appropriate XAI techniques can be chosen to provide necessary information.

6 Evaluating Explanation Quality: A Method for Situation Awareness-Based XAI Assessment

In the following sections, we discuss a selection of existing human-based metrics for XAI from the literature. We then suggest the use of the SAGAT method from human factors for the assessment of the effectiveness of XAI systems.

6.1 Existing Level 1 XAI Methods and Metrics

Since providing a user with a system's outputs is inherent to many existing XAI techniques, most literature does not aim to assess whether the human properly understood these outputs upon receiving them. Kim et al. [29] do this in part by

assessing whether users are able to appropriately assign outputted prototypes to clusters based on the subset of features presented. As the XAI community moves towards explaining higher-complexity systems with multiple inputs and outputs, it will be increasingly important to measure whether users understand the correct inputs/outputs in the contexts of their intended goals. Section 6.4 outlines one approach that could be applied for such assessments.

6.2 Existing Level 2 XAI Methods and Metrics

Metrics for level 2 explanations should indicate whether users understand the meaning of a given system's actions or decisions and what these actions or decisions imply in terms of progress towards team goals. Some of the literature has proposed survey-like questions for assessing explanation quality as it relates to user understanding. For example, Hoffman et al. [27] propose a "goodness" scale that includes a question about whether the user understands how the given algorithm works. They also detail a set of questions related to the perceived understandability of a system from the Madesen-Gregor scale for trust. Doshi-Velez and Kim [11] suggest human experiments requiring users to choose which of two possible system outputs is of higher quality, which necessitates understanding of the system. While these questions and metrics represent a step towards measuring whether adequate level 2 explanations have been provided to users, a more comprehensive way of defining comprehension-related informational needs and assessing whether they have been met through XAI is needed. As mentioned previously, we outline one possible approach to this in Sect. 6.4.

6.3 Existing Level 3 XAI Methods and Metrics

Metrics for level 3 explanations should indicate whether human users can predict what a system will do next or what it would do given an alternate context or input. To this end, Doshi-Velez and Kim [11] suggest running human experiments in which human users perform forward simulation, prediction, and counterfactual simulation of system behavior given different inputs for XAI assessment. Hoffman et al. [27] discuss the use of prediction tasks to measure explanation quality and further detail a Likert-scale survey for trust measurement that includes a question about predictability of system actions. Questions and experiments such as these can be used to assess the quality of level 3 explanations provided by XAI systems. Beyond these assessment techniques, a comprehensive way of assessing whether level 3 informational needs are met is discussed in Sect. 6.4.

6.4 The SAGAT Test and Its Applicability for Assessment of XAI

In assessing the quality of XAI techniques, it is important to determine whether human users receive the information they need in order to perform their roles. Miller et al. [35], in particular, stress the need for human evaluations of XAI systems. As discussed in Sects. 6.1–6.3, existing XAI literature includes some

human-based evaluation metrics; however, to our knowledge, none have comprehensively assessed whether human informational needs are met by XAI systems.

Endsley [14] proposes the situation awareness-based global assessment technique (SAGAT) for SA measurement. SAGAT is a widely-used objective measure of SA that has been empirically shown to have a high degree of sensitivity, reliability, and validity in terms of predicting human performance [16]. It has been applied for measurement of SA in a variety of domains [16], has been extensively used to measure team SA [17], and has been shown to outperform other SA measures in terms of sensitivity, intrusivity, and bias, among other factors [17]. In the SAGAT test, simulations of representative tasks are frozen at randomly selected times, and users are asked questions about their current perceptions of the situation [17]. The questions asked are based directly on the human informational needs defined according to a process such as GDTA (discussed in Sect. 5) and therefore directly measure whether humans have the information required. More complete discussions of SAGAT are provided in [13,14,16], and implementation recommendations for the test are discussed by Endsley [16].

Since the SAGAT test measures whether human informational needs are met, we propose that a SAGAT-like test can be applied to assess XAI systems. Situational information needs related to AI behavior should be thoroughly defined, and in the assessment of an XAI system, user knowledge of this information can be evaluated through a SAGAT-like test focused on information related to specifically AI behavior. Such a test could more adequately determine whether XAI systems achieve the purpose of communicating relevant information about system behavior to human users than current assessment techniques allow.

7 Example Application of the Framework

Here we introduce a simple planetary rover example to demonstrate the application of our framework and the use of the SAGAT test for assessment. The example touches on aspects of explainable ML, explainable agents/robots, and XAI for human teams. In our example, a rover on another planet is executing a learned exploration policy. Its objective is to search for water, which is more likely to be found in areas with certain types of rocks. There are costs associated with navigation time and science task duration, and there are differing rewards associated with performing science tasks on the different types of rocks, some of which are more valuable. The rover has a camera onboard and an ML-based image processing system that allows it to classify rocks. There are constraints associated with the rover power requirements, and some terrain is not traversable. Human users include one engineer who monitors rover health and one scientist who monitors science activities. The scientist and engineer can also request new rover actions during the mission. Below are examples of information constituting levels 1–3 XAI for the engineer and scientist and the types of information they represent (in parentheses).

- **Level 1 XAI:**
  Engineer - terrain information, current battery level (inputs); current path plan and next stopping point/time (plan); next science action (decision/action)
  Scientist - next science action (action/decision); inputted image of rock for science analysis (input); rock classification (output)
- **Level 2 XAI:**
  Engineer - terrain map with rover path costs including untraverseable areas with infinite cost (policy information - costs); battery usage for current path (constraints); list of possible science actions and associated rewards (policy information - rewards); battery usage for each science action (constraints)
  Scientist - list of possible science actions and associated rewards (policy information - rewards); list of semantic features, such as color, contributing to the rock classification (feature information); sensitivity to light given inputs (sensitivity information)
- **Level 3 XAI:**
  Engineer - map of maximum traverseable distance given current battery level (continued action); remaining battery level after each possible science activity (continued action)
  Scientist - predicted rock classification under different lighting conditions (changed inputs)

The scientist and engineer have individual informational requirements in addition to some shared requirements, such as which science activities are planned. Each is only provided with necessary information in order to avoid a cognitive overload from excess information, which poses a risk to task performance.

Measuring SA Through SAGAT. In order to apply SAGAT to this example, specific informational requirements can be enumerated from the high-level informational needs listed above. A list of questions regarding this specific information at all three levels can be specified, and a simulated mission can be run with the scientist and engineer. At various randomly-selected points during the simulated mission, the experiment should be frozen, and the scientist and engineer would then be asked a subset of the specified questions for each level. For example, the following questions might be asked of the engineer regarding the battery during rover traversal between two science activity locations: What is the current battery level of the rover? (Level 1); How much power is required to get to the next location? (Level 2); Does the rover have enough battery to get to the next location and perform the science task? (Level 3).

8 Future Directions

One natural future direction for this work would be to implement a system that addresses the three levels of XAI in a goal- or performance-oriented context and to perform human experiments to assess whether improved SA, enabled through XAI, correlates with improved performance of the human-AI team.

Such a system could be a combination of existing techniques addressing each of the three levels or a single system that can address all three levels of XAI. To our knowledge, no system exists that can, on its own, address all three levels. While explainable BDI agents have addressed aspects of each of the three levels [7,23], additional techniques beyond these solutions will be needed to fully address levels 2 and 3 XAI. In general, development of an XAI system that can independently address all three levels of XAI would be a valuable next direction. Such a system will also require the development of techniques that provide user-tailored explanations in a way that goes beyond what exists in the literature. While there is some existing literature that considers user needs or context in a limited way [8,19,45], producing explanations that fully consider user contexts and tasks remains an understudied area. To this end, another possible future direction would be to perform inference of human models in order to inform explanation generation.

9 Conclusion

In this paper, we propose a three-level framework for the design of XAI systems based on human user informational needs. This framework is based on the situation awareness framework in the human factors literature, which has been studied in relation to performance of human-autonomy teams. We further propose a method for assessment of explanations with respect to the three levels of information that XAI systems should provide. Finally, we propose future directions for XAI research.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems, pp. 9505–9515 (2018)
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. International Foundation for Autonomous Agents and Multiagent Systems (2018)
3. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019)
4. Bansal, A., Farhadi, A., Parikh, D.: Towards transparent systems: semantic characterization of failure modes. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 366–381. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_24
5. Bedny, G., Meister, D.: Theory of activity and situation awareness. Int. J. Cogn. Ergon. **3**(1), 63–72 (1999)
6. Borgo, R., Cashmore, M., Magazzeni, D.: Towards providing explanations for AI planner decisions. arXiv preprint arXiv:1810.06338 (2018)

7. 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
8. Chakraborti, T., Sreedharan, S., Grover, S., Kambhampati, S.: Plan explanations as model reconciliation. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 258–266. IEEE (2019)
9. Chen, J.Y., Procci, K., Boyce, M., Wright, J., Garcia, A., Barnes, M.: Situation awareness-based agent transparency. Technical report, Army Research Lab Aberdeen Proving Ground MD Human Research and Engineering (2014)
10. Dannenhauer, D., Floyd, M.W., Molineaux, M., Aha, D.W.: Learning from exploration: towards an explainable goal reasoning agent (2018)
11. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
12. Dragan, A.D., Lee, K.C., Srinivasa, S.S.: Legibility and predictability of robot motion. In: 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 301–308. IEEE (2013)
13. Endsley, M.: Measurement of situation awareness in dynamic systems. Hum. Factors **37**, 65–84 (1995). https://doi.org/10.1518/001872095779049499
14. Endsley, M.R.: Situation awareness global assessment technique (SAGAT). In: Proceedings of the IEEE 1988 National Aerospace and Electronics Conference, pp. 789–795. IEEE (1988)
15. Endsley, M.R.: Situation awareness misconceptions and misunderstandings. J. Cogn. Eng. Decis. Mak. **9**(1), 4–32 (2015)
16. Endsley, M.R.: Direct measurement of situation awareness: validity and use of SAGAT. In: Situational Awareness, pp. 129–156. Routledge (2017)
17. Endsley, M.R.: A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and spam. Hum. Factors 0018720819875376 (2019)
18. Endsley, M., Jones, W.: A Model of Inter-and Intrateam Situation Awareness: Implications for Design. New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments. M. McNeese, E. Salas and M. Endsley. Human Factors and Ergonomics Society, Santa Monica (2001)
19. Floyd, M.W., Aha, D.W.: Incorporating transparency during trust-guided behavior adaptation. In: Goel, A., Díaz-Agudo, M.B., Roth-Berghofer, T. (eds.) ICCBR 2016. LNCS (LNAI), vol. 9969, pp. 124–138. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47096-2_9
20. Fox, M., Long, D., Magazzeni, D.: Explainable planning. arXiv preprint arXiv:1709.10256 (2017)
21. Gunning, D., Aha, D.W.: Darpa's explainable artificial intelligence program. AI Mag. **40**(2), 44–58 (2019)
22. Halpern, J.Y., Pearl, J.: Causes and explanations: a structural-model approach. Part I: causes. Br. J. Philos. Sci. **56**(4), 843–887 (2005)
23. Harbers, M., van den Bosch, K., Meyer, J.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. IEEE (2010)
24. 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

25. 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. IEEE (2017)
26. Hellström, T., Bensch, S.: Understandable robots-what, why, and how. Paladyn J. Behav. Robot. **9**(1), 110–123 (2018)
27. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
28. 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)
29. Kim, B., Rudin, C., Shah, J.A.: The Bayesian case model: a generative approach for case-based reasoning and prototype classification. In: Advances in Neural Information Processing Systems, pp. 1952–1960 (2014)
30. Kim, B., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). arXiv preprint arXiv:1711.11279 (2017)
31. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
32. 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, pp. 187–188 (2012)
33. Marino, D.L., Wickramasinghe, C.S., Manic, M.: An adversarial approach for explainable AI in intrusion detection systems. In: IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society, pp. 3237–3243. IEEE (2018)
34. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2018)
35. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum or: how i learnt to stop worrying and love the social and behavioural sciences. arXiv preprint arXiv:1712.00547 (2017)
36. 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
37. Parasuraman, R., Sheridan, T.B., Wickens, C.D.: Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Mak. **2**(2), 140–160 (2008)
38. Preece, A., Harborne, D., Braines, D., Tomsett, R., Chakraborty, S.: Stakeholders in explainable AI. arXiv preprint arXiv:1810.00184 (2018)
39. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
40. Ribera, M., Lapedriza, A.: Can we do better explanations? A proposal of user-centered explainable AI. In: IUI Workshops (2019)
41. Salmon, P.M., et al.: What really is going on? Review of situation awareness models for individuals and teams. Theor. Issues Ergon. Sci. **9**(4), 297–323 (2008)
42. Sheh, R., Monteath, I.: Introspectively assessing failures through explainable artificial intelligence. In: IROS Workshop on Introspective Methods for Reliable Autonomy (2017)
43. Sheh, R.K.: Different XAI for different HRI. In: 2017 AAAI Fall Symposium Series (2017)

44. Smith, K., Hancock, P.A.: Situation awareness is adaptive, externally directed consciousness. Hum. Factors **37**(1), 137–148 (1995)
45. Sreedharan, S., Srivastava, S., Kambhampati, S.: Hierarchical expertise level modeling for user specific contrastive explanations. In: IJCAI, pp. 4829–4836 (2018)
46. Stanton, N.A., Chambers, P.R., Piggott, J.: Situational awareness and safety. Saf. Sci. **39**(3), 189–204 (2001)
47. Wickens, C.D.: Multiple resources and mental workload. Hum. Factors **50**(3), 449–455 (2008)