frontiers | Frontiers in Computer Science

TYPE Review
PUBLISHED 06 February 2023
DOI 10.3389/fcomp.2023.1096257

OPEN ACCESS

EDITED BY
Chathurika S. Wickramasinghe,
Capital One, United States

REVIEWED BY
Mark Pfaff,
The MITRE Corporation, United States
Robert Hugh Thomson,
United States Military Academy, United States

*CORRESPONDENCE
Robert R. Hoffman
✉ rhoffman@ihmc.us

SPECIALTY SECTION
This article was submitted to Theoretical Computer Science, a section of the journal Frontiers in Computer Science

RECEIVED 11 November 2022
ACCEPTED 16 January 2023
PUBLISHED 06 February 2023

CITATION
Hoffman RR, Mueller ST, Klein G and Litman J (2023) Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Front. Comput. Sci.* 5:1096257. doi: 10.3389/fcomp.2023.1096257

Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance

Robert R. Hoffman[1]*, Shane T. Mueller[2], Gary Klein[3] and Jordan Litman[4]

[1]Institute for Human and Machine Cognition, Pensacola, FL, United States, [2]Department of Psychology, Michigan Technological University, Houghton, MI, United States, [3]MacroCognition, LLC, Dayton, OH, United States, [4]Department of Psychology, University of Maine at Machias, Machias, ME, United States

If a user is presented an AI system that portends to explain how it works, how do we know whether the explanation works and the user has achieved a pragmatic understanding of the AI? This question entails some key concepts of measurement such as explanation goodness and trust. We present methods for enabling developers and researchers to: (1) Assess the *a priori* goodness of explanations, (2) Assess users' satisfaction with explanations, (3) Reveal user's mental model of an AI system, (4) Assess user's curiosity or need for explanations, (5) Assess whether the user's trust and reliance on the AI are appropriate, and finally, (6) Assess how the human-XAI work system performs. The methods we present derive from our integration of extensive research literatures and our own psychometric evaluations. We point to the previous research that led to the measurement scales which we aggregated and tailored specifically for the XAI context. Scales are presented in sufficient detail to enable their use by XAI researchers. For Mental Model assessment and Work System Performance, XAI researchers have choices. We point to a number of methods, expressed in terms of methods' strengths and weaknesses, and pertinent measurement issues.

KEYWORDS

explanatory reasoning, machine-generated explanations, measurement, explanation goodness, mental models, trust, performance

1. Introduction

Explainability is an issue for decision makers who rely upon Artificial Intelligence, Machine Learning, Data Analytics, and related areas. If a computational system relies on a simple statistical model, decision makers can perhaps understand it and convince executives who have to sign off on a system that it is reasonable and that seems fair. They can justify the analytical results to shareholders, regulators, etc. But for Machine Learning and Deep Net systems, they can no longer do this. There is a need for ways to explain the computational system to the decision maker so that they know that the AI's process is going to be reasonable.

> ... current efforts face unprecedented difficulties: contemporary models are more complex and less interpretable than ever; [AI is] used for a wider array of tasks, and are more pervasive in everyday life than in the past; and [AI is] increasingly allowed to make (and take) more autonomous decisions (and actions). Justifying these decisions will only become more crucial, and there is little doubt that this field will continue to rise in prominence and produce exciting and much needed work in the future (Biran and Cotton, 2017; p. 4).

This brings into relief the importance of Explainable AI (XAI). A proposed regulation before the European Union (Goodman and Flaxman, 2016) prohibits "automatic processing" unless user's rights are safeguarded. Users have a "right to an explanation" concerning algorithm-created decisions that are based on personal information. Future laws may restrict AI, which represents a challenge to industry (see European Intellectual Property Office, 2022). The importance to the field of AI is made salient by the recent reviews of XAI systems, highlighting the goals of explanations for "responsible AI," spanning such issues as trustworthiness, informativeness, confidence, and fairness (e.g., Arrieta et al., 2020).

AI systems are receiving considerable attention in the recent popular press (Harford, 2014; Bornstein, 2016; Alang, 2017; Champlin et al., 2017; Hawkins, 2017; Kuang, 2017; Pavlus, 2017; Pinker, 2017; Schwiep, 2017; Voosen, 2017; Weinberger, 2017). Reporting and opinion pieces have discussed social justice, equity, and fairness issues that are implicated by AI (e.g., Felten, 2017).

The goals of explanation involve answering questions such as, *How does it work?* and *What mistakes can it make?* and *Why did it just do that? Why did it do x rather than y?* The issue addressed in this article is: If we present to a user an AI system that explains how it works, how do we go about measuring whether or not the explanation works, whether it works well, and whether the user has achieved a pragmatic understanding of the AI? Our focus is on the key concepts of measurement, and measurement methods for the evaluation of XAI systems and human-AI work system performance.

1.1. Key measurement concepts

The concept or process of explanation has been explored in one way or another by scholars and scientists of all schools and specializations, spanning literally all of human civilization. To say that the pertinent literature is enormous is an understatement. In modern times, the concept has been a focus in Philosophy of Science, Psychology (Cognitive, Developmental, Social, and Organizational), Education and Training, Team Science, and Human Factors. "While explainable AI is only now gaining widespread visibility, [there is a] continuous history of work on explanation and can provide a pool of ideas for researchers currently tackling the task of explanation (Biran and Cotton, 2017, p. 4)."

Expert Systems researchers implemented methods of explanation (Clancey, 1984, 1986; McKeown and Swartout, 1987; Moore and Swartout, 1990). Additionally, explanation is what Intelligent Tutoring Systems were (and are) all about (Sleeman and Brown, 1982; Polson and Richardson, 1988; Psotka et al., 1988; Ritter and Feurzeig, 1988; Anderson et al., 1990; Lesgold et al., 1992; Forbus and Feltovich, 2001).

Key concepts include causal reasoning, abductive (or hypothetic) inference, comprehension of complex systems, counterfactual reasoning (*Why didn't z happen instead of x?*), and contrastive reasoning (*What would have happened if q had been different?*). For reviews of the literature, see Byrne (2017), Miller (2017), Hoffman et al. (2018), and Mueller et al. (2019).

An early conceptual model of explaining is presented in Figure 1. This diagram was intended to call out the places in the XAI evaluation process where key factors would have to be measured. According to this model, initial instruction in how to use an AI system would enable the user to form an initial mental model of the task and the AI system. Subsequent experience, which would include machine-generated explanations, would enable to participant to refine their mental model, which in turn should lead to better performance and appropriate trust and reliance.

As a psychological or cognitive model this is deficient for a variety of reasons. But by this model, a number of types of evaluation measures are required (Miller, 2017). In this article we detail each of the four classes of measures and offer specific methodological guidance.

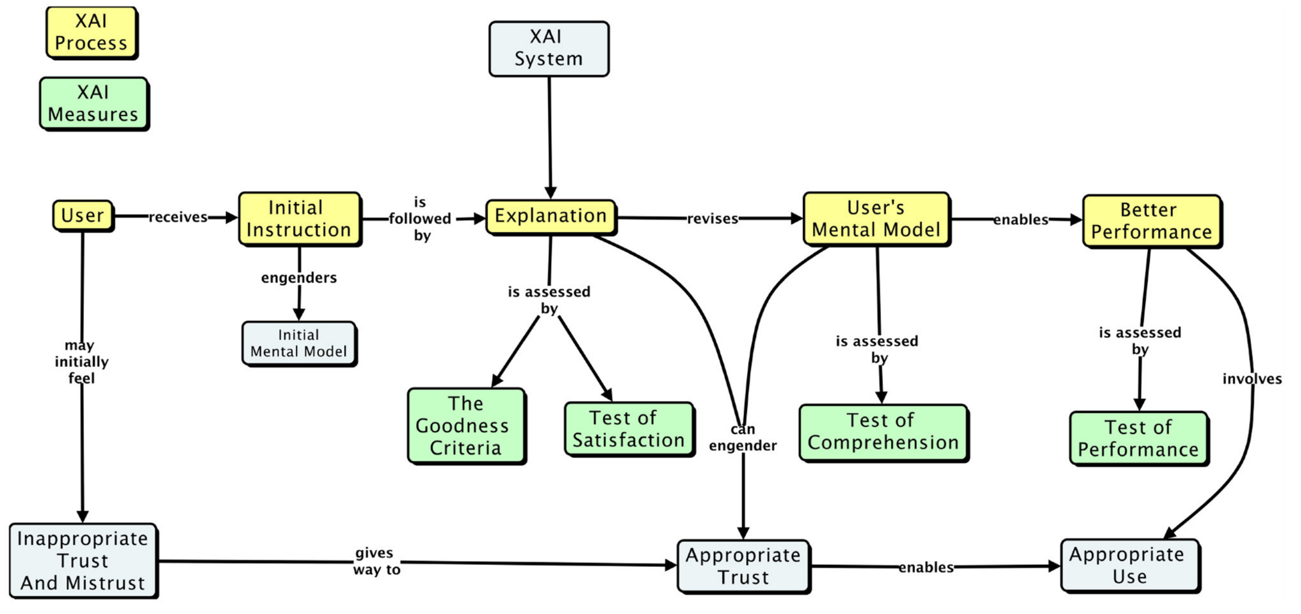

FIGURE 1
A conceptual model of the process of explaining, in the context of evaluating XAI systems.

TABLE 1 Triggers of the need for explanation, and corresponding learner's and goals.

Triggers	User/learner's goal
How do I use it?	Achieve the primary task goals.
How does it work?	Feeling of satisfaction at having achieved a "global" understanding of the AI.
What did it just do?	Feeling of satisfaction at having achieved a "local" understanding of how the AI made a particular decision.
What does it achieve?	Understanding of the AI's functions and uses.
What will it do next?	Feeling of trust based on the observability and predictability of the system.
How much effort will this take?	Feeling of effectiveness and achievement of the primary task goals.
What do I do if it gets it wrong?	Desire to avoid mistakes.
How do I avoid the failure modes?	Desire to mitigate errors.
What would it have done if x were different?	Resolution of curiosity at having achieved an understanding of the system.
Why didn't it do z?	Resolution of curiosity at having achieved an understanding of the local decision.

2. Explanation goodness and satisfaction

The property of "being an explanation" is not a property of statements, it is an interaction. What counts as an explanation depends on what the learner/user needs, what knowledge the user (learner) already has, and especially the user's goals. This leads to a consideration of the function and context of the AI system (software, algorithm, and tool), that is, why does a given user need an explanation? In the various pertinent literatures, this is expressed in terms of the different kinds of questions that a user might have. These "triggers" for explanation are listed in Table 1.

Thus, the seeking of an explanation can tacitly be an expression of a need for a certain kind of explanation, to satisfy certain user purposes of user goals.

A number of XAI developers have recognized the importance of measuring the qualities of explanations of AI systems (e.g., Ehsan et al., 2019). Holzinger et al. (2020) proposed a 10-item scale called the System Causability Scale. Some items referred to the measurement context, such as the timeliness of the explanations. Some items referenced such features as detail, completeness, understandability and learnability of the explanation. This scale can be understood as tapping into two separable things: (1) the intrinsic goodness of explanations and (2) the user's satisfaction with the explanations.

2.1. Explanation goodness (system designer's perspective)

Looking across the scholastic and research literatures on explanation, assertions are made about what makes for a good explanation, from the standpoint of explanations as statements. There is a general consensus on this; factors such as clarity and precision. Thus, one can look at a given explanation and make an *a priori* (or decontextualized) judgment as to whether or not it is good. Table 2 presents a Goodness Checklist of the features that make explanations good, according to the research literature. The reference is to the properties of explanations. The participant simply checks off "yes" or "no" in response to each question.

TABLE 2 The explanation goodness checklist.

The explanation helps me understand how the [software, algorithm, tool] works.
The explanation of how the [software, algorithm, tool] works is satisfying.
The explanation of the [software, algorithm, tool] sufficiently **detailed**.
The explanation of how the [software, algorithm, tool] works is sufficiently **complete**.
The explanation is **actionable**, that is, it helps me know how to use the [software, algorithm, tool].
The explanation lets me know how **accurate or reliable** the [software, algorithm] is.
The explanation lets me know how **trustworthy** the [software, algorithm, tool] is.

This Checklist can be used by researchers to build goodness into the explanations that their XAI system generates, or to evaluate the *a priori* goodness of the explanations that an XAI system generates. In a properly controlled experiment, the researchers who complete the checklist, with reference to some particular XAI-generated explanation, would not be the ones who created the XAI system under study.

2.2. Explanation satisfaction (user's perspective)

Are the researchers correct in claiming that their explanations are adequate, or good? While an explanation might be deemed good in the *a priori* manner described above, it may not be adequate or satisfying to users-in-context. Many AI researchers have recognized the importance of empirically evaluating explanations from the user's perspective. Ehsan et al. (2019) focused on the "human-likeness" of explanations, that is, the degree to which a machine-generated explanation seems like the sort of thing a human might say. The researchers had players in the game Frogger express their rationales for their game play actions. These rationales were then evaluated by expert game players. Another group of participants evaluated rationales that had been selected at random and rationales that had been deemed best by a game expert. The rationales were rated for human-likeness, and understandability. The results showed that participants wanted rationales that were understandable, reliable, and of sufficient detail. They preferred rationales that had implications for immediate and longer-term actions.

Explanation Satisfaction is defined here as the degree to which users feel that they sufficiently understand the AI system or process being explained to them. Compared to Goodness, as defined above, satisfaction is a contextualized, *a posteriori* judgment of explanations [And it must be noted that a person may say that they feel satisfied with an explanation when in fact their understanding is piecemeal or flawed; see diSessa (1993)].

Based on our review of the psychological literature, including theoretical and empirical work by Muir (1987, 1994) and by Cahour and Forzy (2009), we identified several key attributes of explanation

TABLE 3 The explanation satisfaction scale.

From the explanation, I know how the [software, algorithm, tool] works.
This explanation of how the [software, algorithm, tool] works is **satisfying**.
This explanation of how the [software, algorithm, tool] works has **sufficient detail**.
This explanation of how the [software, algorithm, tool] works seems **complete**.
This explanation of how the [software, algorithm, tool] works **tells me how to use** it.
This explanation of how the [software, algorithm, tool] works is **useful to my goals**.
This explanation of the [software, algorithm, tool] shows me how **accurate** the [software, algorithm, tool] is.

satisfaction: understandability, feeling of satisfaction, sufficiency of detail, completeness, usefulness, accuracy, and trustworthiness (Muir and Moray, 1996). These were aggregated and formed as a Likert scale. The sale items are listed in Table 3. For each item, the participant would provide a rating using a 1–5 scale (*I agree strongly, I agree somewhat, I'm neutral about it, I disagree somewhat, I disagree strongly*).

Like the Explanation Goodness Checklist, the Explanation Satisfaction Scale was based on the literatures in cognitive psychology, philosophy of science, and other pertinent disciplines regarding the features that make explanations good. Thus, the Explanation Satisfaction Scale is very similar to the Explanation Goodness Checklist. However, the application context for the two scales is quite different.

- The Explanation Goodness Checklist is intended for use by researchers who have created explanations. The Explanation Goodness Checklist is intended for use as an independent evaluation of explanations, by other researchers. The reference is to the properties of explanations.
- Explanation Satisfaction is an evaluation of explanations by users. The Explanation Satisfaction Scale is for collecting judgments by research participants after they have worked with the XAI system that is being explained, and have been the beneficiaries of one or more explanations.

3. Measuring mental models

In cognitive psychology, mental models are defined as representations or expressions of how a person understands some sort of event, process, or system (Klein and Hoffman, 2008). There is a large body of psychological research on mental models (Johnson-Laird, 1980, 1989; Gentner and Gentner, 1983; Greeno, 1983; Carroll, 1984; Kintsch et al., 1984). Praetorious and Duncan (1988) present an elegant treatment of methodology of eliciting mental models. Staggers and Norcio (1993) provide a good summary of the conceptual and theoretical issues.

There is a consensus that mental models can be inferred from empirical evidence (Johnson-Laird, 1983; Zhang and Wickens, 1987; Glenberg and Langston, 1992; Qin and Simon, 1992; Bogacz and Trafton, 2004; Clement, 2004; Heiser and Tversky, 2006; Klein and Hoffman, 2008). There is sufficient evidence to conclude that different methods for eliciting mental models can converge (Evans et al., 2001; van der Veer and Melguzio, 2003). People may not be able to tell you everything about their understanding, and they may not be able to tell it well. But with adequate scaffolding by some method of guided task reflection, people can tell you how they understand an event or system, they can express their knowledge of it, and the concepts and principles that are involved.

To be sure, people's mental models are rarely complete and coherent (diSessa, 1993), or even consistently correct. "Knowledge shields" are arguments that learners make that enable them to preserve their simplifying and misleading understandings (Feltovich et al., 2001). A focus for instructional design has been to develop methods to get people to recognize when they are employing a knowledge shield that prevents them from developing richer mental models (Hilton, 1996; Hilton and Erb, 1996; Prietula et al., 2000; Tullio et al., 2007; Schaffernicht and Groesser, 2011).

People sometimes overestimate how well they understand complex causal systems. This can be corrected by asking the learner to explicitly express their understanding or reasoning (Chi et al., 1989; Van Lehn et al., 1990; Chi and Van Lehn, 1991; O'Reilly et al., 1998; Rosenblit and Kein, 2002; Mills and Keil, 2004; Rittle-Johnson, 2006; Fernbach et al., 2012; Bjork et al., 2013).

These considerations pertain directly to how people understand machines, spanning process control systems, complex computational systems, decision aids, and intelligent systems (Bainbridge, 1979, 1988; de Kleer and Brown, 1983; Williams et al., 1983; Young, 1983; Harris and Helander, 1984; Rasmussen, 1986; Moray, 1987; Goodstein et al., 1988; Carberry, 1990; May et al., 1993; Staggers and Norcio, 1993; Rasmussen et al., 1994; Samurcay and Hoc, 1996; Doyle et al., 2002; Mueller and Klein, 2011).

In the XAI context, a mental model is a user's understanding of the AI system and its context. This involves more than their understanding of just the AI; it involves their task, goals, current situation, etc. (see Miller, 2017; Mueller et al., 2019). XAI system development requires a method for eliciting, representing, and analyzing users' mental models.

3.1. Overview of mental model elicitation methods

Table 4 lists some methods that have been used to elicit mental models, and exemplary studies. The methods' strengths and weaknesses presented in Table 5 should be useful to system developers who have to make choices about how to investigate user sensemaking about an AI system.

3.2. Application to the XAI context

Explanations can entail what will happen in the future (Mitchell et al., 1989; Koehler, 1991; Lombrozo and Carey, 2006). Thus, a prediction task has the user anticipate what the AI will do, such as the classification of an image by a Deep Net. A prediction task might be about *What do you think will happen next?* but it can also be about a counterfactual: *Why wouldn't something else happen?* A prediction task can serve as a method for peering into users' mental models, especially if its application is be accompanied by a confidence rating and a free response elaboration in which the users explain or justify their predictions, or respond to a probe about counterfactuals.

Diagramming can enable a user to convey the understanding to the researcher, and it can also support a process in which diagrams are analyzed in terms of their proposition content (Johnson-Laird,

TABLE 4 Some methods that can be used to elicit mental models.

Method	Illustrative references
Think-aloud problem solving task, in which participants think aloud during a task.	Reports by Williams et al. (1983) and Rasmussen et al. (1994) are good illustrations of the task used to elicit mental models specifically of devices (see also Gentner and Stevens, 1983; Greeno, 1983; Ericsson and Simon, 1984; Beach, 1992; Rasmussen et al., 1994; Ward et al., 2019).
Think-aloud task with concurrent question answering.	Gentner and Gentner, 1983; Williams et al., 1983; Dodge et al., 2021; Khanna et al., 2021.
Task reflection or retrospection task, in which participants describe their reasoning after conducting a task (e.g., fault diagnosis). The retrospection can be scaffolded, for example, by a replay of their task performance such as in a video.	Fryer (1939) provides a clear and succinct presentation of a method that combines retrospection with Likert scale questionnaire to quantify participants' reasoning. See also Praetorious and Duncan, 1988; Frederick, 2005; Lippa et al., 2008; Dodge et al., 2021; Khanna et al., 2021.
Structured interview, essentially retrospection task with question-answering.	Fryer, 1939; Friedman et al., 2017.
Card sorting task based on the semantic similarity among a set of domain concepts.	The review article by van der Veer and Melguzio (2003) highlights this method (see also Chi et al., 1981; Evans et al., 2001; St-Cyr and Burns, 2002; van der Veer and Melguzio, 2003).
Nearest neighbor task, in which participants select the explanation or conceptual diagram that best fits their beliefs.	Hardiman et al., 1989; Klein and Milltello, 2001.
Self-explanation task (also called teach-back), in which the user/learner expresses their own understanding. Similar to the Retrospection/Reflection Task.	Ford et al., 1993; Cañas et al., 2003; van der Veer and Melguzio, 2003; Molinaro and Garcia-Madruga, 2011; Fernbach et al., 2012.
Glitch detector task (also called accident-error analysis), in which people identify the things that seem wrong in an explanation.	Taylor, 1988; Hoffman et al., 2001.
Prediction task, in which users are presented test cases and are asked users to predict the results and then explain why they thought the predicted results would obtain.	Muramatsu and Pratt, 2001.
Diagramming task (also concept mapping), in which users create a conceptual diagram that lays out their knowledge of concepts, processes, events and their relations.	Novak and Gowin, 1984; Evans et al., 2001; Cañas et al., 2003; Moon et al., 2011; Hoffman and Hancock, 2017.
Shadow box task, in which learners compare their understandings to those of a domain expert (see Table 6, below).	Klein and Borders, 2016.

1983 Ch. 2; Cañas et al., 2003). A diagramming task, along with analysis of concepts and relations (including causal connections or state transitions) has been noted in the field of education as a method for comparing the mental models of students to those of experts or their instructor (see Novak and Gowin, 1984). It has been noted in social studies as a method for comparing individuals' mental models of social groups [i.e., individuals and their inter-relations; Carley and Palmquist (1992)]. It has been noted in operations research as a method for comparing mental models of dynamical systems (see Schaffernicht and Groesser, 2011).

Another method that can accomplish these things this is Cued Retrospection. Probe questions are presented to participants about their reasoning just after the reasoning task has been performed (see Ward et al., 2019). For instance, they might be asked, *Can you describe the major components or steps in the [software system, algorithm]?* The probes can also reference metacognitive processes, for example by asking, Based on the explanation of the [software, algorithm], can you imagine circumstances or situations in which the [software, algorithm] might lead to error conditions, wrong answers, or bad decisions?

Explicit Self-explanation has been shown to improve learning and understanding. This holds for both deliberate self-motivated self-explanation and also self-explanation that is prompted or encouraged by the instructor. Self-explanation can have a significant and positive impact on understanding. Self-generated deductions and generalizations help learners refine their knowledge (Chi et al., 1989, 1994; Chi and Van Lehn, 1991; Rittle-Johnson, 2006; Molinaro and Garcia-Madruga, 2011; Lombrozo, 2016). Having learners explain the answers of experts also enhances the learner's understanding (Calin-Jageman and Ratner, 2005).

Understanding also accrues from a task that allows a learner to compare their reasoning to that of an expert. The ShadowBox Lite method [presented here by permission from Klein and Borders (2016)] is a self-explanation task that is applicable to the XAI context, and provides a quick window into user mental models: It avoids the necessity of eliciting and analyzing an extensive recounting of the user's reasoning. In the method, the user is presented a question such as *How does a car's cruise control work?* Accompanying the question is a proposed explanation. The task for the user is to identify one or more ways in which the explanation is good, and ways in which it is bad. After doing this, the participant is shown a Good-Bad list that was created by a domain expert (see Table 6, below). The participant's comparison of the lists can lead to insights.

Compelling evidence of the value of task reflection comes from a set of studies that used methods of knowledge elicitation for the evaluation of explainable AI systems (Anderson et al., 2020; Dodge et al., 2021; Khanna et al., 2021; Tabatabai et al., 2021). The researchers created a strategy game in which two software agents competed, and learned to adjust their strategies. Research participants were shown a game replay. After every few decision points, the game was paused and participants predicted what action an agent would take, then after the game resumed and was again paused, the participants were asked to described the action the agent did take, and finally, they explained why they thought the agent would take a particular action. In one of the experiments, participants were also asked to describe what they felt to be an agent's reasoning errors.

TABLE 5 Methods strengths and weaknesses.

Method	Strength	Weakness
Concurrent think-aloud problem solving	Can provide rich information about mental models.	Transcription and protocol analysis can be time consuming, result in a great deal of data, require much analysis.
Think-aloud task with concurrent question answering	Enables the researcher to present targeted probes to the user during task performance.	Highly dependent on the researcher/interviewer's skill at question design and interviewing.
Task reflection or retrospection	Can be conducted as a structured interview, as a questionnaire task, or as a post-task verbalization of a self-explanation. Can provide rich information about mental models, or a quick window into mental models. The process of self-explanation itself has learning value.	People can overestimate how well they understand complex systems. Transcription and protocol analysis can be time consuming, result in a great deal of data, require much analysis. Less effort would be required in a questionnaire method, though questionnaire design would be non-trivial.
Card sorting	Can provide information about domain concepts and their relations.	Can provide sparse data about concepts, events or processes. The data consist of similarity ratings (i.e., semantic nets).
Nearest neighbor	Can provide a quick window into mental models.	People can overestimate how well they understand complex systems.
Glitch detector	Can support users to discover and explain aspects of their mental model that are reductive or incorrect.	Glitches have to be built-in. Knowledge shields may inhibit the awareness of reductive tendencies.
Question-answering/structured interview	Enables researcher to probe selected aspects of a user's mental model.	Highly dependent on the researcher/interviewer's skill at question design and interviewing.
Prediction task	Can provide a quick window into mental models. The predictions should be accompanied by a confidence rating and a free response elaboration that explains or justifies the predictions.	The free responses require content analysis. Requires clear rationale for the choice of instances or cases to be the focus of the predictions.
Diagramming task	Can provide a rich and thorough representation of the user's mental model. Relations are not restricted to similarity (see Card Sorting Task, above).	Can take time to create, although user friendly software systems are readily available.
Box or "Shadowbox Lite" task (see Table 6, below).	Can provide a quick window into mental models.	May not result in a thorough expression of the mental model.

TABLE 6 An example of propositional coding using the ShadowBox Lite Task.

Expert's explanation	
The control unit detects the rotation of the drive shaft from a magnet mounted on the drive shaft, and from that can calculate how fast the car is going. The control unit controls an electric motor that is connected to the accelerator linkage. The cruise control adjusts the engine speed until it is disengaged.	
What is right and helpful about this explanation?	*What is problematic or wrong about this explanation?*
The cruise control unit has to know how fast the car is going.	It seems overly technical, with some concepts left unexplained.
The cruise control has to control the engine throttle or accelerator.	I do not think the cruise control detects the engine speed.

The task retrospections included the asking of counterfactual and contrastive explanations, to ask why a software agent did not decide in a particular way, or ask why decided one way when it should have decided another (see also Goyal et al., 2019). In other words, participants were expressing their reasoning about the AI.

This corpus of work using various methods of task reflection and self-explanation has established the value of revealing user's mental models. It illustrates the need for structure or "scaffolding" that supports the user in explaining their thoughts and reasoning. It shows that task retrospection enhances the understanding of learners, and thus has value as a learning or training tool for new users. The researchers cited here have provided detailed descriptions of their methods, including ways of making the process efficient for AI developers (to determine, for example, how an AI might be improved). This is in recognition of the fact that XAI developers need a method that can elicit mental models quickly, and can result in data that can be easily scored, categorized, or analyzed.

3.3. Design considerations

Referencing these kinds of knowledge elicitation tasks, and the other tasks listed in Tables 4, 5, it is recommended that the evaluation of user mental models within the XAI context should employ more than one method for eliciting mental models. Performance on any one type of task might not align with performance at some other task. For instance, in one study it was found that the adequacy of a user-generated diagram did not match to better performance at a prediction task (see St-Cyr and Burns, 2002). Performance on a simulated industrial process control task can be good and yet the operator's understanding can be limited and even incorrect (see Berry and Broadbent, 1988).

Not all of the participants in a study have to be presented with a mental model elicitation task. Indeed, a reasonably sized and representative set of 10 to 12 participants can be presented one or more mental model elicitation tasks. If the analysis of the goodness of the mental models aligns with measures of performance, then subsequent studies might use performance measures as a surrogate for mental model analysis.

3.4. Analysis of user mental models

An empirically-derived expression of the content or the ebbs and flows that compose a user's mental model must contribute to the evaluation of mental model goodness (i.e., correctness, comprehensiveness, coherence, and usefulness). Evaluation is usually based on proposition analysis, Results from most elicitation tasks will be sentence-like utterances that can be recast as propositions, broken out by the concepts and their relations (see Crandall et al., 2006). Carley and Palmquist (1992) provide illustrative examples of propositional coding for transcripts of interviews of students by their teachers. The user model developed by Friedman et al. (2017) is based on propositional encoding of interview transcripts. We illustrate propositional analysis using the framework of the ShadowBox Lite Task. Table 6 presents an expert explanation, although this might just as easily be thought of as an explanation generated by an XAI system. The bottom two cells present the propositions expressed by the participant.

The products from a diagramming task can also be recast as propositions. The explanation can be decomposed into the component concepts, relations and propositions. In the case of the cruise control example, the expert's explanation has ten concepts (drive shaft rotation, car speed, etc.) seven relations (mounted on, disengage, etc.), and six propositions (e.g., Magnet is mounted on the drive shaft, Control unit calculates car speed).

Concepts, relations and propositions can be counted and the counts aggregated and analyzed in a number of ways. For instance, one can calculate the percentage of concepts, relations, and propositions that are in the user's explanation that are also in the expert's explanation. This can suggest the completeness of the user's mental model.

4. Measuring curiosity

There are theoretical and empirical reasons why curiosity might be considered an important factor in Explainable AI. Fundamentally, the seeking of an explanation can be driven by curiosity. As referenced in Table 1 of the "triggers" for explanation, learners sometimes wonder about such things as *What would the AI have done if x were different?* and *Why didn't the AI do z?* Therefore, it is important that XAI systems harness the power of curiosity. Explanations may promote curiosity and set the stage for the achievement of insights and the development of better mental models.

On the other hand, explanations can actually suppress curiosity and reinforce flawed mental models. This can happen in a number of ways:

- An explanation might overwhelm people with details,
- The XAI system might not allow questions or might make it difficult for the user to pose questions,
- Explanations might make people feel reticent because of their lack of knowledge,
- Explanations may include too many open variables and loose ends, and curiosity decreases when confusion and complexity increase.

For these reasons, the assessment of users' feelings of curiosity might be informative in the evaluation of XAI systems.

Epistemic curiosity is the general desire for knowledge, a motive to learn new ideas, resolve knowledge gaps, and solve problems, even though this may entail effortful cognitive activity (Berlyne, 1960, 1978; Loewenstein, 1994; Litman and Lunsford, 2010). Stimulus novelty, surprisingness, or incongruity, can trigger curiosity. All of these features refer to circumstances when information is noted as being missing or incomplete.

Curiosity is also triggered in circumstances where one experiences a violated expectation (Maheswaran and Chaiken, 1991). Violated expectations essentially reflect the discovery that events that were anticipated to be comprehensible are instead confusing—more information and some sort of change to one's understanding is needed to make sense of the event and thus resolve the disparity between expectation and outcome. Such situations lead people to engage in effortful processing, and motivates them to seek out additional knowledge in order to gain an insight and resolve the incongruency (Loewenstein, 1994).

This is directly pertinent to XAI. Curiosity is stimulated when learner recognizes that there is a gap in their knowledge or understanding. Recognizing a knowledge gap, closing that gap, and achieving satisfaction from insight make the likelihood of success from explanations or self-explanations seem feasible. This leads to the question of how to assess or measure curiosity in the XAI context, and what to do with the measurements.

Figure 2 presents a simple conceptual model of self-explanation, that focuses on the role and place of curiosity, noting that some learners may not be curious.

Unlike in Figure 1, the conceptual model presented in Figure 2 does not describe the explanation process as a series of fixed stages. Sensemaking based on explanations is rarely a "one-and-done." Research has shown that there are many possible paths to reasoning about complex systems, in which people engage in deliberative self-explanation (Klein et al., 2023). Some users need to know how the AI works, but may not be entirely clear about what it is that they want or need to know. Explanatory information provided by the XAI system might help in this sense making process, or it might not.

Some users might be curious, but only occasionally. Realizing that there is a gap in their understanding, they may tell themselves a story about how the AI system works and evaluate that story for its plausibility.

4.1. Measuring curiosity in the XAI context

A number of psychometric instruments have titles that make them seem pertinent to XAI, such as the Cognitive Reflection Test (Frederick, 2005). But this taps numerical fluency and competence. Available scales of curiosity, such as the Curiosity Exploration Inventory (Kashdan et al., 2004), Cacioppo's Need for Cognition scale (Cacioppo et al., 1984), and the I-Type/D-Type Curiosity Scales (Litman and Jimerson, 2004) consider curiosity to be a pervasive style or personality trait. As such, the instruments ask questions such as: *I actively seek as much information as I can in a new situation; I feel stressed or worried about something I do not know*; *I like to discover new places to go.* Such items are barely applicable in the XAI context, in which curiosity is situation or task specific, and

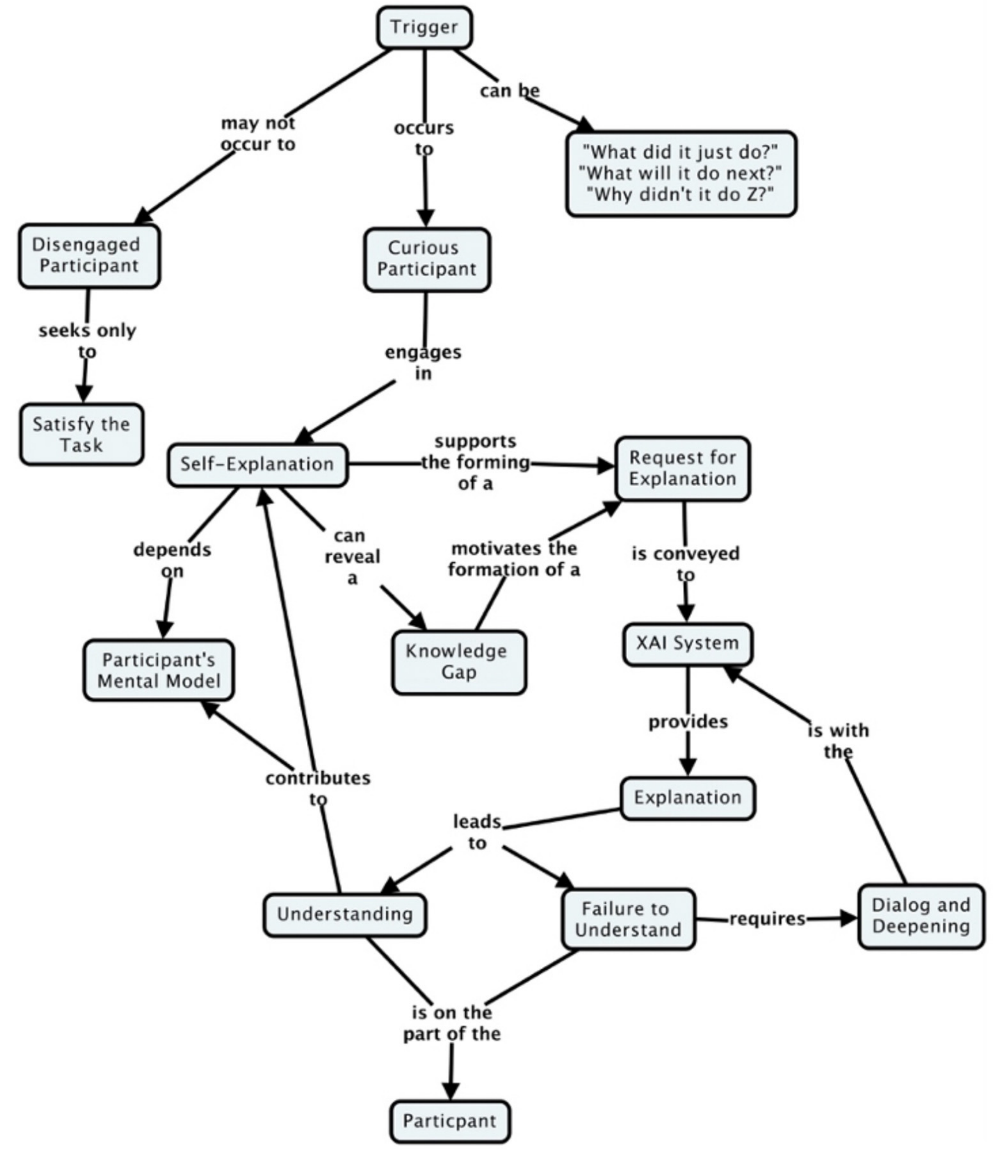

FIGURE 2
A model of the explanation process in the XAI context from the perspective of the learner, showing the role of curiosity.

refers to the workings of computational devices, rather than daily life or experiences.

The discussion of the knowledge gap model of curiosity (see, above) implies that the evaluation of XAI systems might benefit from asking users to identify the triggers that motivated them to ask for an explanation. As for the other measurement classes we have discussed (Figure 1, above), it is valuable for the curiosity measurement method to present a "quick window." Table 7 presents a simple questionnaire that can be administered to research participants whenever they ask for an explanation.

Responses will be informative with regard to these things:

- Responses may serve as parameters or constraints that the XAI system uses to generate explanations.
- Responses may provide a window into aspects of the AI system's operations that need explaining.
- Responses may reveal ways in which the AI's explanation method might suppress or inhibit curiosity, and
- Responses may make it possible to use depth of curiosity (i.e., more triggers are checked) as an independent variable.

TABLE 7 A curiosity checklist.

Why have you asked for an explanation? Check all that apply.
I want to know what the AI just did.
I want to know that I understand this AI system correctly.
I want to understand what the AI will do next.
I want to know why the AI did not make some other decision.
I want to know what the AI would have done if something had been different.
I was surprised by the AI's actions and want to know what I missed.

5. Measuring trust in the XAI context

Trust in automation is of concern in computer science and cognitive systems engineering, as well as in the popular media (e.g., Hoffman, 1989; Shadbolt, 2002; Lee and See, 2004; Huynh et al., 2006; Woods and Hollnagel, 2006; Merritt and Ilgen, 2008; Hoffman et al., 2009, 2013; Naone, 2009; Montague, 2010; Fitzhugh et al., 2011; Kucala, 2013; Merritt et al., 2013, 2015; Chancey et al., 2015; Hoff and Bashir, 2015; Pop et al., 2015; Wickens et al., 2015). Trust is implicated in many ways in AI system development. For example, stakeholders sometimes express a need to be able to trust vendors; they express a need to get explanations from trusted systems engineers. Users need to understand whether they can trust the data that were used to train an AI system (see Hoffman et al., 2022). The measurement scale presented in this article is focused specifically on the end-user's trust in machine-generated explanations.

Trust in computers is typically understood in terms of the metaphor of "calibration" (Lee and See, 2004). The metaphor assumes that trust is a state, a single state, and that it develops toward some single stable and ideal value. This does not fit the facts of the matter concerning the dynamics and complexities of trust (Hoffman, 2017). Some users may take the computer's assertions (data, claims) as valid and true because they come from a computer. But other users may require some sort of justification—empirical reasons to believe that the computer's presentations or assertions are valid and true. Just as there are varieties of trusting, there are varieties of negative trusting (such as mistrust and distrust). Trust in automation can rapidly break down under conditions of time pressure, or when there are conspicuous system faults or errors, or when there is a high false alarm rate (Dzindolet et al., 2003; Madhavan and Wiegmann, 2007). The trusting of machines can be hard to reestablish once lost.

However trust is measured, in the XAI context, the measurement method must be sensitive to the emergence of negative trusting states (such as mistrust and distrust). XAI systems should enable the user to know whether, when, and why to trust and rely upon the XAI system and know whether, when, or why to mistrust the XAI and either not rely upon it, or rely on it with caution. People always have some mixture of justified and unjustified trust, and justified and unjustified mistrust in computer systems. A user might feel positive trust toward an AI system with respect to certain tasks and goals, and simultaneously feel mistrusting or distrusting when other tasks, goals or situations are involved. Indeed, in complex human-machine work systems, this is undoubtedly the norm (Sarter et al., 1997; Hoffman et al., 2013). Only if this trusting relation is achieved can the user's reliance on the computer be confident (Riley, 1996).

Appropriate trust and reliance emerge from the user's experience, especially as the user encounters tough cases or cases that fall at the boundary of the work system's competence envelope. Appropriateness refers to the fact that mistrust, as well as trust, can be justified. Presumably, appropriate reliance (knowing when and when not to rely on the system's outputs) hinges on appropriate trust.

5.1. Trust measurement scales

The scientific literature on trust presents a number of scales for measuring trust. The majority of trust scales have been developed for the context of interpersonal trust. Scales have been proposed that specifically for measuring trust in robots 9 (e.g., Schaefer, 2013). We focus here son scales designed for use in the assessment of trust in automation. Minimally, a trust scale asks two basic questions: *Do you trust the machine's outputs?* (trust) and *Would you follow the machine's advice?* (reliance). Indeed, these two items comprise the scale developed by Adams et al. (2003).

The creation of an XAI-appropriate scale requires the deconstruction of existing scales, and then cherry-picking of the appropriate scale items. The scale developed by Johnson (2007) asks only about reliance and the rareness of errors. Some scales for assessing trust in robots are not applicable to the XAI context, or to any generic trust-in-automation context.

The scale by Dzindolet et al. (2003) was created for the study of trust in a software system for evaluating terrain in aerial photographs, showing images in which there might be camouflaged soldiers. Thus, the hypothetical technology was referred to as a contrast detector. The experiment was one in which the error rate of the hypothetical detector was a primary independent variable. As a consequence, the scale items refer to trials e.g., *How well do you think you will perform during the 200 trials?* and How many errors do you think you will make during the 200 trials? Some of the scale items can be adapted to make them appropriate to the XAI context, but the result of this modification is just a few items, which are ones that are in the Cahour-Fourzy Scale (Cahour and Forzy, 2009) (e.g., *Can you trust the decisions the [system] will make?*).

Of those scales that have been subject to psychometric analysis, results suggest that trust in automation scales can be reliable. Of those scales that have been subject to validity analysis, high Cronbach alpha results have been obtained. The report by Jian et al. (2000) illustrates those psychometric analyses. From validated scales, we have distilled a set of items that are appropriate for use in XAI research. This XAI Trust Scale asks users directly whether they are confident in the XAI system, whether the XAI system is predictable, reliable, efficient, and believable. Most of the items are adapted from the Cahour-Fourzy scale, which has been shown to be reliable. The XAI Trust Scale incorporates items from other scales, indicated in Table 7. Items in the Trust Scale bear overall semantic similarity to items in the Madsen-Gregor-Scale, and that scale too was also shown to have high reliability. The XAI Trust Scale items are listed in Table 8. The items are presented to participants in a Likert format, using a 5–1 scale (*I agree strongly, I agree somewhat, I'm neutral about it, I disagree somewhat, I disagree strongly*).

TABLE 8 The trust scale for the XAI context.

1. I am confident in the [tool]. I feel that it works well.
2. The outputs of the [tool] are very predictable.
3. The tool is very reliable. I can count on it to be correct all the time.
4. I feel safe that when I rely on the [tool] I will get the right answers.
5. The [tool] is efficient in that it works very quickly.
6. I am wary of the [tool] (adopted from the Jian, et al. Scale and the Wang, et al. Scale).
7. The [tool] can perform the task better than a novice human user (adopted from the Schaefer Scale).
8. I like using the system for decision making (adapted from the Madsen-Gregor Scale).

6. Measuring performance

The goal of performance measurement is to determine the degree of success of the human-machine system at effectively conducting the tasks for which the technology (and the work system as a whole) is designed. Based on the model in Figure 1, hypotheses are:

- User performance (including measures of joint user-system performance) will improve as a result of being given satisfying explanations.
- User performance will be a function of the qualities of their mental model (e.g., correctness, completeness, etc.).
- User performance may be affected epistemic trust.
- User performance (that is, reliance) will be appropriate if the user has been able to explore the competence envelop of the AI system, that is, experience how, when and why the AI fails.

The evaluation of the performance of an XAI system cannot be entirely divorced from the evaluation of the performance of the user, or from the performance of the human-machine work system as a whole. Thus, there are considerations that go beyond those called out in Figure 1. User performance would be positively impacted if there is an opportunity for the user and the AI to engage in meaningful dialog, for instance.

6.1. Performance with regard to the primary task goals

The human-AI work system will have some primary task goal or goals. This might be to correctly categorize images, correctly identify certain kinds of actions in videos, or conduct an emergency rescue operation. Performance can be measured in terms of the number of trials on which the work system met with success, within some pre-specified time period. In a search-and-rescue use case, work system performance might be measured in terms of the number of trials that a user has to work in order to reach some pre-determined criterion. Basic measures of efficiency can be applied, expressing the ratio of the number of tasks or sub-tasks completed per some period of time.

6.2. Performance with regard to the user

Another aspect of performance measurement is the quality of the performance of the user, such as the correctness of the user's predictions of what the AI will do. For such aspects of performance, just like performance with regard to primary task goals, one can measure response speed and correctness of the user's predictions of the machine outputs (hits, errors, misses, and false alarms). Examination can be made for both typical and atypical cases/situations. Additionally, using a knowledge elicitation task in concert with a success measure, one can assess the completeness of the user's explanation of the machine's output for cases that are rare, unusual, or anomalous.

6.3. Performance with regard to the work system

With regard to performance at the work system level, many considerations other than raw efficiency come into play (Koopman and Hoffman, 2003; Hoffman and Hancock, 2017). For example, an XAI system that drives the work efficiency (say, to deal with data overload issues) may not make for a very contented workplace (Merritt, 2011). The complexity of performance at the work system level requires analysis based on this and other trade-offs (Woods and Hollnagel, 2006).

The analysis of work system level performance may employ some measure of controllability, that is, the extent to which the human can induce an intended outcome based on given inputs or conditions. Analysis may employ some measure of correctability. This is a measure of the extent or ease with which the user can correct the machine's activities so as to make the machine outputs better aligned with either objective states of affairs or the user's judgments of what the machine should be determining.

The analysis of work system level performance can involve comparing the work productivity of the work system to productivity in current practice (baseline). A related method is to examine learning curves. This involves establishing a metric on a productivity scale, a metric that identifies when performance is satisfactory. How many trials or test cases must a user work successfully in order to reach that learning criterion? Is the learning rapid? Why do some people take a long time to reach criterion? The advantage of a trials-to-criterion approach to measurement is that it could put different XAI systems on a "level playing field" by making the primary measure a derivation of task completion time. A variation on this method is to compare performance in this way with performance when the Explanation capability of the XAI is somehow hobbled.

Perhaps the most powerful and direct way of evaluating the performance of a work system that includes an XAI is to evaluate how easy or difficult it is to get prospective users (stakeholders) to adopt and use the XAI system. In discussing early medical diagnosis systems, van Lent et al. (2004) stated, "Early on the developers of these systems realized that doctors weren't willing to accept the expert system's diagnosis on faith" (p. 904), which led to development of the first explanation-based AI systems. Many users may be satisfied with shallow explanations in that they will be willing to adopt the system that has an XAI capability, or may prefer it over another non-XAI system when making adoption decisions. Measures of such choice behavior were advocated by Adams et al. (2003) as a way of measuring human trust in automation.

7. Limitations and prospects

Our discussion of measurement key concepts, measurement scales, and methods has not broached the subject of measurement methodology, that is, how scales are actually applied in context. For example, a response of "yes" to *I was surprised by the AI* (in an application of the Curiosity Checklist) would be an open invitation for the researcher to ask additional questions—using a mental model elicitation task. In other words, the actual administration of a psychometric scale or a performance evaluation method is not merely the collection of some numbers.

We look forward to research that utilizes the measurement scales and methods presented in this article. Our own research went only so far as to develop, and then empirically evaluate the measures in small-scale, targeted studies. Some of the measures presented in this article were adopted by Performer Teams in the out-years of the DARPA XAI Program. Some of the measures have been adopted by researchers outside of that Program (Schraagen et al., 2020). The findings include the validation of XAI metrics and also the interesting suggestion that it takes fewer than ten trials or cases for learners to develop reasonable mental models and begin to trust an AI system.

Based on the model in Figure 1, an assumption made in some XAI work has been that a measure of performance can be used simultaneously as a measure of the goodness of a user's mental model. This assumption should be empirically investigated. It may be especially revealing to compare results from participants who perform the best and participants who perform the worst. Comparison of their mental models, and theirs with those of an expert, would explore the assumption that a measure of performance can indeed be used simultaneously be a measure of the goodness of the user's mental model.

All measures get interpreted. An assumption that has been made in some XAI work that a measure of AI predictability can be used simultaneously as a measure of the goodness of the user's mental model. That is, if the user can predict what the AI will do, then the user must have a good mental model of the AI. While this may be true according to the XAI measurement scheme presented in Figure 1, it begs the question of whether the AI is needed at all. More to the point, it is important that all measures undergo appropriate psychometric evaluation to confirm their proper interpretation.

None of the existing trust scales, including the one presented here, really treat trust as a process; they treat it as a static quality or target state, which has typically been measured once, after the research participants have completed their experimental tasks. In contrast, XAI trust measurement might be a repeat measure. Selected scale items can be applied after individual trials or blocks of trials (e.g., after individual XAI categorizations or recommendations; after individual explanations are provided, etc.). The full scale could be completed part way through a series of experimental trials, and again at the conclusion of the final experimental trial. Multiple measures taken over time could be integrated for overall evaluations of human–machine performance, but episodic measures would be valuable in tracking such things as How do users maintain trust? and What is the trend for desirable movement toward appropriate trust?

A crucial consideration is methodological in nature. How often should users/trainees be asked to complete a scale? Should a curiosity scale be administered before or after experience with an AI/XAI system? Should measurement be longitudinal? While multiple measurements on multiple measures is desirable there is trade-off in that burdening the research participants is not desirable.

Measurement is the foundation of empirical inquiry. One of the purposes of making measurements is to improve the measures. Standard practice in psychometrics involves assuring that a scale is valid. Psychometric research pursuing this for XAI scales would be valuable.

We advocate for a multi-measure approach. Such an approach seems mandated by the fact that the human-XAI work system is a complex cognitive system. We do not regard the scales and methods that presented in this article as being final. The distinction between explanation goodness as an *a priori* evaluation by researchers vs. explanation satisfaction as an *a posteriori* evaluation by learners emerged from our own empirical inquiry. We look forward to refinements and extensions of all the ideas presented in this article.

An important measurement topic that we have not addressed is that of "metrics." In modern discourse, the word metric is used to mean both measure and metric. A metric is a threshold on a measurement scale that is used to make evaluations or decisions (e.g., performance is "acceptable," "poor," etc.). In this article we have discussed measures, not metrics. In the evaluation of, say, a machine learning system for recognizing objects that are depicted in photographs, at what level of performance does performance cross over from being unacceptable to being promising? Ideally, machine learning systems would be infallible, always manifesting a hit rate of 100 percent. Achieving that that seems unlikely. So, where lies the point of diminishing returns? A hit rate of 90 percent? Ninety-five percent? As another concrete example, a surgeon specializing in carpal tunnel syndrome who has a success rate less than 90 percent might well be in trouble. On the other hand, a specialist in spinal surgery with a 90% success rate would be considered a miracle worker. The metric that is laid on a measurement scale depends on the application context [For a fuller discussion, see Hoffman (2010)].

Metrics bring the notion of practical significance into focus. For example, suppose an XAI work system achieves a 95 percent level of performance. As the work system capability reaches that point does it actually become likely that those rare cases where errors are made are ones that are potentially more impactful?

Certainly AI measurement science needs metrics to accompany its measures. When is performance superior, acceptable, or poor? When is an explanation sufficient or not? When is a mental model rich or impoverished? But metrics for these sorts of decisions do not emerge directly (or easily) from the theoretical concepts that are being measured, or from the operationalized measures that are being used. The operational definition of a measure tells you how to make measurements, not how to interpret them. Metrics come from policy (see Hoffman, 2010). Resolution of the metrics challenge can only emerge as more XAI projects are carried through all the way to rigorous performance evaluation.

Author contributions

RH integrated material composed by each of the co-authors on the topics of mental models and curiosity, contributed the material on trust, performance, and methodology. SM conducted most of the statistical tests. All authors contributed to the article and approved the submitted version.

Funding

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8650-17-2-7711. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

Acknowledgments

The authors would like to acknowledge William J. Clancey of the Florida Institute for Human and Machine Cognition for his contributions to the conceptual foundations of XAI. The authors thank Timothy Cullen (Col., US Army, Ret.) for his consultation on this project. The authors thank the two reviewers, who provided comments that were insightful, challenging, and uniformly helpful. This material is approved for public release. Distribution is unlimited.

Conflict of interest

GK was employed by MacroCognition, LLC.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of AFRL or the U.S. Government.

References

Adams, B. D., Bruyn, L. E., and Houde, S. (2003). *Trust in automated systems*. Report, Ministry of National Defence, United Kingdom.

Alang, N. (2017). *Turns out algorithms are racist. The New Republic*. Available online at: https://newrepublic.com/article/144644/turns-algorithms-racist?utm_content=buffer7f3ea (accessed August 31, 2017).

Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., et al. (2020). Mental models of mere mortals with explanations of reinforcement learning. *ACM Trans. Inter. Intell. Syst.* 10, 1–37. doi: 10.1145/3366485

Anderson, J., Boyle, C. F., Corbett, A. T., and Lewis, M. W. (1990). "Cognitive modeling and intelligent tutoring," in *Artificial intelligence and learning environments,* eds. W. J. Clancey and E. Soloway (Cambridge, MA: Bradford Books) 7–49. doi: 10.1016/0004-3702(90)90093-F

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Bainbridge, L. (1979). Verbal reports as evidence of process operator's knowledge. *Int. J. Man-Mach. Stud.* 11, 411–436. doi: 10.1016/S0020-7373(79)80035-8

Bainbridge, L. (1988). "Types of representation," in *Tasks, errors and mental models,* eds. I.P. Goodstein, H.B. Andersen, and S.E. Olsen (New York: Taylor and Francis) 70–91.

Beach, L. R. (1992). "Epistemic strategies on causal thinking in expert and nonexpert judgment," in *Expertise and decision support,* eds. G. Wright and F. Bolger (New York: Plenum) 107–127. doi: 10.1007/978-0-585-34290-0_6

Berlyne, D. E. (1960). *Conflict, Arousal, and Curiosity.* New York: McGraw-Hill. doi: 10.1037/11164-000

Berlyne, D. E. (1978). Curiosity and learning. *Motiv. Emot.* 2, 97–175. doi: 10.1007/BF00993037

Berry, D. C., and Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *Br. J. Psychol.* 79, 251–272. doi: 10.1111/j.2044-8295.1988.tb02286.x

Biran, O., and Cotton, C. (2017). "Explanation and Justification in Machine Learning: A Survey," in *IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI).* Available online at: http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf (accessed January 24, 2023).

Bjork, R. A., Dunlosky, J., and Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Ann. Rev. Psychol.* 64, 417–444. doi: 10.1146/annurev-psych-113011-143823

Bogacz, S., and Trafton, J. G. (2004). Understanding dynamic and static displays: Using images to reason dynamically. *Cogn. Syst. Res.* 6, 312–319. doi: 10.1016/j.cogsys.2004.11.007

Bornstein, A. M. (2016). *Is Artificial Intelligence Permanently Inscrutable?* Available online at: http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable (accessed August 29, 2017).

Byrne, R. M. (2017). Counterfactual thinking: from logic to morality. *Curr. Direct. Psychol. Sci.* 26, 314–322. doi: 10.1177/0963721417695617

Cacioppo, J. T., Petty, R. E., and Kao, C. F. (1984). The efficient assessment of need for cognition. *J. Person. Assess.* 48, 306–307. doi: 10.1207/s15327752jpa4803_13

Cahour, B., and Forzy, J. F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Sci.* 47, 1260–1270. doi: 10.1016/j.ssci.2009.03.015

Calin-Jageman, R. J., and Ratner, H. H. (2005). The role of encoding in the self-explanation effect. *Cogn. Instr.* 23, 523–543. doi: 10.1207/s1532690xci2304_4

Cañas, A. J., Coffey, J. W., Carnot, M. J., Feltovich, P., Hoffman, R., Feltovich, J., et al. (2003). *A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support.* Report to The Chief of Naval Education and Training, prepared by the Institute for Human and Machine Cognition, Pensacola, FL.

Carberry, S. (1990). Second international workshop on user modeling. *AI Magaz.* 11, 57–60.

Carley, K., and Palmquist, M. (1992). Extracting, representing and analyzing mental models. *Soc. Forces* 70, 601–636. doi: 10.2307/2579746

Carroll, J. M. (1984). Minimalist training. *Datamation* 1, 125–126.

Champlin, C., Bell, D., and Schocken, C. (2017). AI medicine comes to Africa's rural clinics. *IEEE Spectrum* 54, 42–48. doi: 10.1109/MSPEC.2017.7906899

Chancey, E. T., Bliss, J. P., Proaps, A. B., and Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. *Human Factors* 57, 947–958. doi: 10.1177/0018720815582261

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cogn. Sci.* 13, 145–182. doi: 10.1207/s15516709cog1302_1

Chi, M. T., Leeuw, N., Chiu, M.-H., and LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cogn. Sci.* 18, 439–477. doi: 10.1207/s15516709cog1803_3

Chi, M. T., and Van Lehn, K. A. (1991). The content of physics self-explanations. *J. Learn. Sci.* 1, 69–105. doi: 10.1207/s15327809jls0101_4

Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cogn. Sci.* 5, 121–152. doi: 10.1207/s15516709cog0502_2

Clancey, W. J. (1984). "Methodology for building an intelligent tutoring system," in *Method and tactics in Cognitive Science,* eds. W., Kintsch, J. R., Miller, and P. G. Polson (Hillsdale, NJ: Lawrence Erlbaum Associates) 51–83.

Clancey, W. J. (1986). From GUIDON to NEOMYCIN and HERACLES in twenty short lessons. *AI Magazine* 7, 40–60.

Clement, J. (2004). "Imagistic simulation and physical intuition in expert problem solving," in *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (Hillsdale, NJ: Lawrence Erlbaum) 201–205. doi: 10.4324/9781315789354-35

Crandall, B., Klein, G., and Hoffman, R. R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis.* Cambridge, MA: MIT Press. doi: 10.7551/mitpress/7304.001.0001 de Kleer, J., and Brown, J. S. (1983). “Assumptions and ambiguities in mechanistic mental models,”. in *Mental models,* eds. D. Gentner and A.L. Stevens (New York: Psychology Press) 155–190.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cogn. Instr.* 10, 105–225. doi: 10.1080/07370008.1985.9649008

Dodge, J., Khanna, R., Irvine, J., Lam, K.-H., Mai, T., Lin, Z., et al. (2021). After-action review for AI (AAR/AI). *ACM Trans. Intell. Syst.* 11, 29. doi: 10.1145/3453173

Doyle, J. K., Ford, D. N., Radzicki, M. J., and Trees, W. S. (2002). “Mental models of dynamic systems,” in *System Dynamics and Integrated Modeling, Encyclopedia of Life Support Systems*, eds. Y. Barlas (Oxford: EOLSS Publishers).

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* 58, 697–718. doi: 10.1016/S1071-5819(03)00038-7

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). “Automated rationale generation: a technique for explainable AI and its effects on human perceptions,” *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 263–274. doi: 10.1145/3301275.3302316

Ericsson, K. A., and Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data.* Cambridge, MA: MIT Press.

European Intellectual Property Office (2022). *AI-powered Trademark Dispute Resolution - Expert Opinion Commissioned by the European Union Intellectual Property Office (EUIPO).* doi: 10.2814/062663 (accessed January 24, 2023).

Evans, A. W., Jentsch, F., Hitt, J. M., Bowers, C., and Salas, E. (2001). “Mental model assessments: Is there convergence among different methods?,” in *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (Santa Monica, CA: Human Factors and Ergonomics Society) 293–298. doi: 10.1177/154193120104500406

Felten, E. (2017). What does it mean to ask for an “explainable” algorithm? Available online at: https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/ (accessed August 29, 2017).

Feltovich, P. J., Coulson, R. L., and Spiro, R. J. (2001). “Learners’(mis) understanding of important and difficult concepts: A challenge to smart machines in education,” in *Smart Machines in Education,* eds. K. D. Forbus and P. J. K. Feltovich (Menlo Park, CA: AAAI/MIT Press) 349–375.

Fernbach, P. M., Sloman, S. A., Louis, R. S., and Shube, J. N. (2012). Explanation friends and foes: How mechanistic detail determines understanding and preference. *J. Cons. Res.* 39, 1115–1131. doi: 10.1086/667782

Fitzhugh, E. W., Hoffman, R. R., and Miller, J. E. (2011). “Active trust management,” in *Trust in Military Teams,* eds. N. Stanton (London: Ashgate) 197–218.

Forbus, K. D., and Feltovich, P. J. (2001). *Smart Machines in Education.* Menlo Park, CA: AAAI/MIT Press.

Ford, K. M., Cañas, A. J., and Coffey, J. (1993). “Participatory explanation,” in *Presented at the FLAIRS 93: Sixth Florida Artificial Intelligence Research Symposium (FLAIRS)* (Pensacola, FL: Institute for Human and Machine Cognition) 111–115.

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect.* 19, 25–42. doi: 10.1257/089533005775196732

Friedman, S., Forbus, K., and Sherin, B. (2017). Representing, running and revising mental models: A computational theory. *Cogn. Sci.* 42, 1110–1145. doi: 10.1111/cogs.12574

Fryer, D. (1939). Post quantification of introspective data. *Am. J. Psychol.* 52, 367–371. doi: 10.2307/1416744

Gentner, D., and Gentner, D. R. (1983). “Flowing waters or teem crowds: Mental models of electricity,” in *Mental models,* eds. D. Gentner and A.L. Stevens (New York: Psychology Press) 99–129.

Gentner, D., and Stevens, A. L. (1983). *Mental Models.* New York: Psychology Press.

Glenberg, A. M., and Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *J. Memory Lang.* 31, 129–151. doi: 10.1016/0749-596X(92)90008-L

Goodman, B., and Flaxman, S. (2016). “European Union regulations on algorithmic decision-making and a ‘right to explanation.’,” in *Presented at the ICML Workshop on Human Interpretability in Machine Learning*, New York, NY.

Goodstein, I.P., Andersen, H.B., and Olsen, S. E. (1988). *Tasks, errors and mental models.* New York: Taylor and Francis.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual visual explanations. *arXiv:1904.07451.*

Greeno, J. G. (1983). “Conceptual entities,” in *Mental Models* eds. D. Gentner and A. L. Stevens (New York: Psychology Press) 227–252. doi: 10.21236/ADA123387

Hardiman, P. T., Dufresne, R., and Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory Cogn.* 17, 627–638. doi: 10.3758/BF03197085

Harford, T. (2014). Big data: Are we making a big mistake? *Significance.* 11, 14–19. doi: 10.1111/j.1740-9713.2014.00778.x

Harris, S. D., and Helander, M. G. (1984). “Machine intelligence in real systems: Some ergonomic issues,” in *Human-Computer Interaction,* eds. G. Salvendy (Amsterdam: Elsevier Science Publishers) 267–277.

Hawkins, J. (2017). Can we copy the brain? What intelligent machines need to learn from the Neocortex. *IEEE Spectrum* 54, 34–71. doi: 10.1109/MSPEC.2017.7934229

Heiser, J., and Tversky, B. (2006). Arrows comprehending and producing mechanical diagrams, *Cogn. Sci.* 30, 581–592. doi: 10.1207/s15516709cog0000_70

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Think. Reason.* 2, 273–308. doi: 10.1080/135467896394447

Hilton, D. J., and Erb, H.-P. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking Reason.* 2, 273–308.

Hoff, K. A., and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 407–434. doi: 10.1177/0018720814547570

Hoffman, R. R. (1989). “Whom (or what) do you trust: Historical reflections on the psychology and sociology of information technology,” in *Proceedings of the Fourth Annual Symposium on Human Interaction with Complex Systems* (New York: IEEE Computer Society) 28–36.

Hoffman, R. R. (2010). “Theory → Concepts → Measures but Policies → Metrics,” in *Macrocognition metrics and scenarios: Design and evaluation for real-world teams,* eds. E. Patterson and J. Miller (London: Ashgate) 3–10. doi: 10.1201/9781315593173-2

Hoffman, R. R. (2017). “A Taxonomy of Emergent Trusting in the Human–Machine Relationship,” in *Cognitive Systems Engineering: The Future for a Changing World,* eds. P. J. Smith and R.R. Hoffman (Boca Raton, FL: Taylor and Francis) 137–163. doi: 10.1201/9781315572529-8

Hoffman, R. R., Coffey, J. W., Ford, K. M., and Carnot, M. J. (2001). “STORM-LK: A human-centered knowledge model for weather forecasting,” in *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society,* eds. J. M. Flach (Santa Monica, CA: Human Factors and Ergonomics Society) 752. doi: 10.1177/154193120104500807

Hoffman, R. R., and Hancock, P. A. (2017). Measuring resilience. *Human Factors* 59, 564–581. doi: 10.1177/0018720816686248

Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A. (2013). Trust in automation. *IEEE: Intell. Syst.* 28, 84–88. doi: 10.1109/MIS.2013.24

Hoffman, R. R., Klein, G., Jalaeian, M., Tate, C., and Mueller, S. T. (2022). The Stakeholder Playbook for explaining AI systems. *Frontiers in AI*. Available online at: https://www.ihmc.us/technical-reports-onexplainable-ai/ (accessed January 24, 2023).

Hoffman, R. R., Klein, G., and Mueller, S. T. (2018). *Literature Review and Integration of Key Ideas for Explainable AI.* Report to the DARPA XAI Program.

Hoffman, R. R., Lee, J. D., Woods, D. D., Shadbolt, N., Miller, J., and Bradshaw, J. M. (2009). The dynamics of trust in cyberdomains. *IEEE Intell. Syst.* 24, 5–11. doi: 10.1109/MIS.2009.124

Holzinger, A., Carrington, A., and Mueller, H. (2020). Measuring the quality of explanations: the System Causability Scale. *Kunstliche Intell.* 34, 193–198. doi: 10.1007/s13218-020-00636-z

Huynh, T. D., Jennings, N. R., and Shadbolt, N. R. (2006). An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents Multi-Agent Syst.* 13, 119–154. doi: 10.1007/s10458-005-6825-4

Jian, J. Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* 4, 53–71. doi: 10.1207/S15327566IJCE0401_04

Johnson, D. S. (2007). Achieving customer value from electronic channels through identity commitment, calculative commitment, and trust in technology. *J. Inter. Market.* 21, 2–22. doi: 10.1002/dir.20091

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cogn. Sci.* 4, 71–115. doi: 10.1207/s15516709cog0401_4

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Cambridge, MA: Harvard University Press.

Johnson-Laird, P. N. (1989). “Mental models,” in *Foundations of Cognitive Science,* eds. M. I. Posner (Cambridge, MA: MIT Press) 469–499.

Kashdan, T. B., Rose, P., and Fincham, F. D. (2004). Curiosity and exploration: Facilitating positive subjective experiences and personal growth opportunities. *J. Person. Assess.* 82, 291–305. doi: 10.1207/s15327752jpa8203_05

Khanna, R., Dodge, A., Anderson, R., Dikkala, J., Shureih, K.-H., Lan, C. R., et al. (2021). Finding AI’s faults with AAR/AI: an empirical study. *IEEE Trans. Inter. Intell. Syst.* 12, 1–33. doi: 10.1145/3487065

Kintsch, W., Miller, J. R., and Polson, P. G. (1984). *Methods and Tactics in Cognitive Science.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Klein, G., and Borders, J. (2016). The ShadowBox approach to cognitive skills training: an empirical evaluation. *J. Cogn. Eng. Decis. Mak.* 10, 268–280. doi: 10.1177/1555343416636515

Klein, G., and Hoffman, R. R. (2008). “Macrocognition, mental models, and cognitive task analysis methodology,” in *Naturalistic decision making and macrocognition,* eds. J.M. Schraagen, L.G. Millitello, T. Ormerod and R. Lipshitz (Aldershot, England: Ashgate) 57–80.

Klein, G., Jalaeian, M., Hoffman, R. R., and Mueller, S. T. (2023). The Plausibility Gap: A model for sensemaking, in *Frontiers in Psychology: Special Issue on Naturalistic Decision Making (NDM)*. Available online at: https://www.ihmc.us/technical-reports-onexplainable-ai/ (accessed January 24, 2023).

Klein, G., and Milltello, L. G. (2001). "Some guidelines for conducting a cognitive task analysis," in *Human/technology interaction in complex systems,* eds. E. Sales (Oxford: Elsevier Science Ltd) 161–197.

Koehler, D. (1991). Explanation, imagination, and confidence in judgement. *Psychol. Bull,* 110, 499–519. doi: 10.1037/0033-2909.110.3.499

Koopman, P., and Hoffman, R. R. (2003). Work-arounds, make-work, and kludges. *IEEE: Intell. Syst.* 18, 70–75. doi: 10.1109/MIS.2003.1249172

Kuang, C. (2017). Can A.I. be taught to explain itself? *The New York Times*. Available online at: https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html (accessed November 21, 2017).

Kucala, D. (2013). *The truthiness of trustworthiness.* Chief Learning Officer 57–59.

Lee, J. D., and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Lesgold, A. M., Lajoie, S. P., Bunzo, M., and Eggan, G. (1992). "SHERLOCK: A coached practice environment for an electronics troubleshooting job," in *Computer Assisted Instruction and Intelligent Tutoring Systems: Shared Issues and Complementary Approaches,* eds. J. Larkin and R. Chabay (Hillsdale, NJ: Lawrence Erlbaum Associates) 201–238. doi: 10.4324/9781315044361-8

Lippa, K., Klein, H. A., and Shalin, V. L. (2008). Everyday expertise: Cognitive demands in diabetes self-management. *Human Factors* 50, 112–120. doi: 10.1518/001872008X250601

Litman, J. A., and Jimerson, T. L. (2004). The measurement of curiosity as a feeling-of-deprivation. *J. Person. Assess.* 82, 147–157. doi: 10.1207/s15327752jpa8202_3

Litman, J. A., and Lunsford, G. D. (2010). Incurious motives to seek information about potential threats. *Eur. J. Person.* 24, 1–17. doi: 10.1002/per.766

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychol. Bull.* 116, 75–98. doi: 10.1037/0033-2909.116.1.75

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends Cogn. Sci.* 20, 748–759. doi: 10.1016/j.tics.2016.08.001

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

Madhavan, P., and Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors* 49, 773–785. doi: 10.1518/001872007X230154

Maheswaran, D., and Chaiken, S. (1991). Promoting systematic processing in low-motivation settings: Effect of incongruent information on processing and judgment. *J. Person. Soc. Psychol.* 61, 13–25. doi: 10.1037/0022-3514.61.1.13

May, J., Barnard, P. J., and Blandford, A. (1993). Using structural descriptions of interfaces to automate the modeling of user cognition. *User Model. User-Adapted Inter.* 3, 27–64. doi: 10.1007/BF01099424

McKeown, K. R., and Swartout, W. R. (1987). Language generation and explanation. *Ann. Rev. Comput. Sci.* 2, 401–449. doi: 10.1146/annurev.cs.02.060187.002153

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors.* 53, 356–370. doi: 10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., LaChapell, J., and Lee, D. (2013). I trust it, but don't know why: Effects of implicit attitudes toward automation in trust in an automated system. *Human Factors* 55, 520–534. doi: 10.1177/0018720812465081

Merritt, S. M., and Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human–automation interactions. *Human Factors* 50, 194–201. doi: 10.1518/001872008X288574

Merritt, S. M., Lee, D., Unnerstall, J. L., and Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation- aided task. *Human Factors* 57, 34–47. doi: 10.1177/0018720814561675

Miller, T. (2017). Explanation in artificial intelligence: insights from the social sciences. ArXiv:1706.07269 [Cs]. Available online at: http://arxiv.org/abs/1706.07269

Mills, C. M., and Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *J. Exper. Child Psychol.* 87, 1–32. doi: 10.1016/j.jecp.2003.09.003

Mitchell, D., Russo, E., and Rennington, N. (1989). Back to the future: Temporal perspective in the explanation of events. *J. Behav. Decis. Making* 2, 25–38. doi: 10.1002/bdm.3960020103

Molinaro, R. I., and Garcia-Madruga, J. A. (2011). Knowledge and question asking. *Psicothema* 23, 26–30. Available online at: https://www.psicothema.com/pdf/3845.pdf

Montague, E. (2010). Validation of a trust in medical technology instrument. *Appl. Ergon.* 41, 812–821. doi: 10.1016/j.apergo.2010.01.009

Moon, B. M., Hoffman, R. R., Cañas, A. J., and Novak, J. D. (2011). *Applied Concept Mapping: Capturing, Analyzing and Organizing Knowledge*. Boca Raton, FL: Taylor and Francis. doi: 10.1201/b10716

Moore, J. D., and Swartout, W. R. (1990). "Pointing: A Way Toward Explanation Dialogue," in *Proceedings of AAAI* (Menlo Park, CA: AAAI) 90, 457–464.

Moray, N. (1987). Intelligent aids, mental models, and the theory of machines. *Int. J. Man-Mach. Stud.* 7, 619–629. doi: 10.1016/S0020-7373(87)80020-2

Mueller, S. T., Hoffman, R. R., Clancey, W. J., Emrey, A., and Klein, G. (2019). *Explanation in human-AI Systems: A Literature Meta-Review, synopsis of Key Ideas, and Publications, and Bibliography for Explainable AI.* Technical Report, DARPA Explainable AI Program. arXiv:1902.01876 [pdf]

Mueller, S. T., and Klein, G. (2011). Improving users' mental models of intelligent software tools. *IEEE Intell. Syst.* 77–83. doi: 10.1109/MIS.2011.32

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *Int. J. Man–Mach. Stud.* 27, 527–539. doi: 10.1016/S0020-7373(87)80013-5

Muir, B. M. (1994). Trust in automation Part 1: Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 1905–1922. doi: 10.1080/00140139408964957

Muir, B. M., and Moray, N. (1996). Trust in automation. Part II Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 429–460. doi: 10.1080/00140139608964474

Muramatsu, J., and Pratt, W. (2001). "Transparent queries: Investigation users' mental models of search engines," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York: Association for Computing Machinery) 217–224. doi: 10.1145/383952.383991

Naone, E. (2009). *Adding trust to Wikipedia, and beyond. Technology Review* Available online at: https://www.technologyreview.com/2009/09/04/29900/adding-trust-to-wikipedia-and-beyond/ (accessed September 04, 2009).

Novak, J. D., and Gowin, D. B. (1984). *Learning How to Learn.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139173469

O'Reilly, T., Symons, S., and MacLatchy-Gaudet, H. (1998). A comparison of self-explanation and elaborative interrogation. *Contemp. Educ. Psychol.* 23, 434–445. doi: 10.1006/ceps.1997.0977

Pavlus, J. (2017). *Stop pretending you really know what AI is and read this instead.* Available online at: https://qz.com/1067123/stop-pretending-you-really-know-what-ai-is-and-read-this-instead/ (accessed September 06, 2017).

Pinker, S. (2017). *Uncommon insights into common knowledge. APS Observer, 30.* Available online at: https://www.psychologicalscience.org/observer/uncommon-insights-into-common-knowledge (accessed January 31, 2017).

Polson, M. C., and Richardson, J. J. (1988). *Foundations of Intelligent Tutoring Systems.* Hillsdale, NJ: Erlbaum.

Pop, V. L., Shrewsbury, A., and Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human Factors* 57, 545–556. doi: 10.1177/0018720814564422

Praetorious, N., and Duncan, K. D. (1988). "Verbal reports: A problem in research design," in *Tasks, Errors and Mental Models,* eds. I.P. Goodstein, H. B. Andersen, and S. E. Olsen (New York: Taylor and Francis) 293–324.

Prietula, M., Feltovich, P., and Marchak, F. (2000). Factors influencing analysis of complex cognitive tasks: A framework and example from industrial process control. *Human Factors* 42, 54–74. doi: 10.1518/001872000779656589

Psotka, J., Massey, L. D., and Mutter, S. A. (1988). *Intelligent Tutoring Systems: Lessons Learned*. Hillsdale, NJ: Erlbaum.

Qin, Y., and Simon, H. A. (1992). "Imagery as a process representation in problem solving," in *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (Hillsdale, NJ: Lawrence Erlbaum) 1050–1055.

Rasmussen, J. (1986). *Information Processing and Human-Machine Interaction.* Amsterdam, North-Holland.

Rasmussen, J., Pejtersen, A. M., and Goodstein, L. P. (1994). *Cognitive Systems Engineering.* New York: John Wiley.

Riley, V. (1996). "Operator reliance on automation: Theory and data," in *Automation Theory and Applications,* eds. R. Parasuraman and M. Mouloua (Mahwah, NJ: Erlbaum) 19–35.

Ritter, F., and Feurzeig, W. (1988). "Teaching real-time tactical thinking," in *Intelligent Tutoring Systems: Lessons Learned* (Hillsdale, NJ: Lawrence Erlbaum) 285–301.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development* 77, 1–15. doi: 10.1111/j.1467-8624.2006.00852.x

Rosenblit, L., and Kein, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cogn. Sci.* 26, 521–562. doi: 10.1207/s15516709cog2605_1

Samurcay, R., and Hoc, J.-M. (1996). Causal versus topographical support for diagnosis in a dynamic situation. *Le Travail Humain* 59, 45–68.

Sarter, N., Woods, D. D., and Billings, C. E. (1997). "Automation surprises," in *Handbook of Human Factors/Ergonomics*, ed. G. Salvendy (New York, NY: Wiley) 1926–1943.

Schaefer, K. E. (2013). *The perception and measurement of human-robot trust.* Doctoral dissertation, University of Central Florida, Orlando, Florida.

Schaffernicht, M., and Groesser, S. N. (2011). A comprehensive method for comparing mental models of dynamical systems. *Eur. J. Oper. Res.* 210, 57–67. doi: 10.1016/j.ejor.2010.09.003

Schraagen, J. M., Elasser, P., Fricke, H., and Ragalmuto, F. (2020). "Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models," in *Proceedings of the 2020 Human Factors and Ergonomics Society 64th International Annual Meeting.* doi: 10.1177/1071181320641077

Schwiep, J. (2017). *The state of explainable AI.* Available online at: https://medium.com/@jschwiep/the-state-of-explainable-ai-e252207dc46b (accessed May 3, 2017).

Shadbolt, N. (2002). A matter of trust. *IEEE Intell. Syst.* 20, 30–33. doi: 10.1109/MIS.2002.988440

Sleeman, D., and Brown, J. S. (1982). *Intelligent Tutoring Systems*. London: Academic Press.

Staggers, N., and Norcio, A. F. (1993). Mental models: concepts for human-computer interaction research. *Int. J. Man-Mach. Stud.* 38, 587–605. doi: 10.1006/imms.1993.1028

St-Cyr, O., and Burns, C. M. (2002). "Mental models and ecological interface design: An experimental investigation," in *Proceedings of the Human Factors and Ergonomic Society Annual Meeting* (Santa Monica, CA: Human Factors and Ergonomics Society) 270–274. doi: 10.1177/154193120204600311

Tabatabai, D., Ruangrotsakun, A., Irvine, J., Dodge, J., Sureih, Z., Lam, K.-H., et al. (2021). ""Why did my AI agent lose?": Visual analytics for scaling up After-Action Review," in *Proceedings of the 2021 IEEE Visualization Conference*, 16–20. doi: 10.1109/VIS49827.2021.9623268

Taylor, J. R. (1988). "Using cognitive models to make plants safer: Experimental and practical approaches," in *Tasks, errors and mental models,* eds. I.P. Goodstein, H.B. Andersen, and S.E. Olsen (New York: Taylor and Francis) 233–239.

Tullio, J., Dey, A. K., Chalecki, J., and Fogarty, J. (2007). "How it works: a field study of non-technical users interacting with an intelligent system," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery) 31–40. doi: 10.1145/1240624.1240630

van der Veer, G. C., and Melguzio, M. (2003). "Mental models," in *The human-computer interaction handbook,* eds. J.A. Jacko and A. Sears (Hillsdale, NJ: Lawrence Erlbaum) 52–80.

Van Lehn, K., Ball, W., and Kowalski, B. (1990). *Explanation-based learning of correctness: Towards a model of the self-explanation effect.* Report, Department of Psychology, Carnegie Mellon University. doi: 10.21236/ADA225644

van Lent, M., Fisher, W., and Mancuso, M. (2004). "An explainable artificial intelligence system for small-unit tactical behavior," in *Proceedings of the 19th National Conference on Artificial Intelligence* (Menlo Park, CA: AAAI Press) 900–907.

Voosen, P. (2017). How AI detectives are cracking open the black box of deep learning. *Science* 357, 22–27. doi: 10.1126/science.357.6346.22

Ward, P., Wilson, K., Suss, J., Woody, W. D., and Hoffman, R. R. (2019). "An historical perspective on introspection: Implications and guidelines for eliciting verbal and introspective-type reports," in *The Oxford Handbook of Expertise,* eds. P. Ward, J.M. Schraagen, T.C. Ormerod, and E. Roth (Oxford: Oxford University Press) 377–407. doi: 10.1093/oxfordhb/9780198795872.013.17

Weinberger, D. (2017). *Our machines now have knowledge we'll never understand.* Available online at: https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/ (accessed April 18, 2017).

Wickens, C. D., Clegg, B. A., Vieane, A. Z., and Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors* 57, 728–739. doi: 10.1177/0018720815581940

Williams, M. D., Hollan, J. D., and Stevens, A. L. (1983). "Human reasoning about a simple physical system," in *Mental Models,* eds. D. Gentner andA.L. Stevens (New York: Psychology Press) 131–154.

Woods, D. D., and Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Boca Raton, FL: CRC Press. doi: 10.1201/9781420005684

Young, R. M. (1983). "Surrogates and mappings: Two kinds of conceptual models of interactive devices," in *Mental Models,* eds. D. Gentner and A. L. Stevens (New York: Psychology Press) 35–52.

Zhang, K. L., and Wickens, C. D. (1987). "A study of the mental model of a complex dynamic system: The effect of display aiding and contextual system training," in *Proceedings of the Human Factors and Ergonomics Society 31st Annual Meeting* (Santa Monica, CA: Human Factors and Ergonomics Society) 102–107. doi: 10.1177/154193128703100123