5-9-2023

Evaluating machine-generated explanations: a “Scorecard” method for XAI measurement science

Robert R. Hoffman
Florida Institute for Human & Machine Cognition

Mohammadreza Jalaeian
The Ohio State University

Connor Tate
Florida Institute for Human & Machine Cognition

Gary Klein
LLC

Shane T. Mueller
Michigan Technological University, shanem@mtu.edu

Recommended Citation

Hoffman, R., Jalaeian, M., Tate, C., Klein, G., & Mueller, S. (2023). Evaluating machine-generated explanations: a “Scorecard” method for XAI measurement science. *Frontiers in Computer Science, 5*. http://doi.org/10.3389/fcomp.2023.1114806
Retrieved from: https://digitalcommons.mtu.edu/michigantech-p/17127

Follow this and additional works at: https://digitalcommons.mtu.edu/michigantech-p

 Part of the Cognitive Science Commons

frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 09 May 2023
DOI 10.3389/fcomp.2023.1114806

OPEN ACCESS

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

REVIEWED BY
Christos Troussas,
University of West Attica, Greece
Siming Chen,
Fudan University, China
Scott Cheng-Hsin Yang,
Rutgers University, Newark, United States

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

RECEIVED 02 December 2022
ACCEPTED 03 April 2023
PUBLISHED 09 May 2023

CITATION
Hoffman RR, Jalaeian M, Tate C, Klein G and Mueller ST (2023) Evaluating machine-generated explanations: a "Scorecard" method for XAI measurement science. *Front. Comput. Sci.* 5:1114806. doi: 10.3389/fcomp.2023.1114806

Evaluating machine-generated explanations: a "Scorecard" method for XAI measurement science

Robert R. Hoffman[1]*, Mohammadreza Jalaeian[2], Connor Tate[1], Gary Klein[3] and Shane T. Mueller[4]

[1]Institute for Human and Machine Cognition, Pensacola, FL, United States, [2]Department of Integrated Systems Engineering, Ohio State University, Columbus, OH, United States, [3]MacroCognition, LLC, Dayton, OH, United States, [4]Department of Cognitive and Learning Sciences, Michigan Technological University, Houghton, MI, United States

Introduction: Many Explainable AI (XAI) systems provide explanations that are just clues or hints about the computational models-Such things as feature lists, decision trees, or saliency images. However, a user might want answers to deeper questions such as *How does it work?, Why did it do that instead of something else? What things can it get wrong?* How might XAI system developers evaluate existing XAI systems with regard to the depth of support they provide for the user's sensemaking? How might XAI system developers shape new XAI systems so as to support the user's sensemaking? What might be a useful conceptual terminology to assist developers in approaching this challenge?

Method: Based on cognitive theory, a scale was developed reflecting depth of explanation, that is, the degree to which explanations support the user's sensemaking. The seven levels of this scale form the Explanation Scorecard.

Results and discussion: The Scorecard was utilized in an analysis of recent literature, showing that many systems still present low-level explanations. The Scorecard can be used by developers to conceptualize how they might extend their machine-generated explanations to support the user in developing a mental model that instills appropriate trust and reliance. The article concludes with recommendations for how XAI systems can be improved with regard to the cognitive considerations, and recommendations regarding the manner in which results on the evaluation of XAI systems are reported.

KEYWORDS

explainable AI, sensemaking, self-explanation, explanation scale, mental model

1. Introduction

There have appeared many discussions of types of explanations, described in terms of format, content and detail (e.g., Samek et al., 2017). There have been many discussions of the qualities of explanations, and some attempts to measure and empirically evaluate explanation quality and the effectiveness of explanations (e.g., Doshi-Velez and Kim, 2017; Miller, 2017; Lage et al., 2019; Mueller et al., 2019; Buçinca et al., 2020; Johs et al., 2020; Cabitza et al., 2023; Hoffman et al., 2023).

An individual who is using an AI system is struggling to form a mental model that explains how the AI works, how it fails, and how failures can be anticipated or surmounted (Mueller et al., 2019). Reviews of the literature on explanation (in diverse disciplines) and

studies of how people explain complex systems demonstrate the depth of reasoning that can be involved in sensemaking (e.g., Miller, 2017; Hind et al., 2019; Mueller et al., 2019).

On the other hand, reviews have shown that many machine-generated explanations are rather superficial, only providing the users with hints or clues about how the AI system works. The broader literature on explanation and explanatory reasoning, along with the literature on Intelligent Tutoring Systems (Troussas et al., 2019; Clancey and Hoffman, 2022), pedagogy (Chi and VanLehn, 1991) and Intelligent Help Systems (Carroll and Aaronson, 1988) and expert systems (Jingge, 1988) suggested that richer explanations and richer AI-user engagement are not only possible, but necessary. Recent work on intelligent help systems has shown how instructional agents can engage in personalized conversation (e.g., Troussas et al., 2017).

Compounding this shortfall has been divergence in how such terms as "explanation," "justification," and "interpretation" are used in the XAI community (e.g., Adadi and Berrada, 2018; Kaur et al., 2020). This has been compounded further by the assumption that an explanation that is good for system developers and computer scientists would be good for a general population of general users. For example, in the XAI system by Preece et al. (2018), the explanations are lists of features and the output of the interpretability models. The authors assert that the outputs of interpretable models are explanations for users.

Many papers refer to explanation and explainable AI but that focus on formal interpretability, ontologies, or transparency, in such modes as decision trees and logical expressions (e.g., Tomsett et al., 2018; Calegari et al., 2019; Felzmann et al., 2019; Kaur et al., 2020; Tjoa and Guan, 2020). For example, Chari et al. (2020) present an ontology system to aid in the design of explanatory systems that provides support for expressing the semantic relationships among explanation attributes. The researchers consider explanation types including those identified in prior literature: case-based, conceptual, contrastive, counterfactual, every-day, scientific, simulation-based, statistical, and trace-based. These are expressed using formal conceptual graphs. They also present "competency questions" that can be used by a systems designer for the development of user and context dependent explanations (e.g., "What AI model is capable of generating this explanation type?"). Vilone and Longo (2020) list 36 dimensions of explainability.

Although this and many other papers speak to the matter of the explanatory value of explanations, the matter of the extent to which machine-generated explanation support the user's sensemaking remains in the background of the formal analyses. At the same time, there is a consensus that user needs should be understood, prioritized and addressed (Liao et al., 2020).

Interviews with stakeholders of various stripes have shown that some users are quite interested in doing "deep dives" into how the AI works, and are quite capable of it (Hoffman et al., 2021). One might wonder how a chunk of code or a cryptic decision tree would have explanatory value to some users–as opposed to having explanatory value to system developers. Some users might prefer a digestible explanation of an AI's reasoning. But such an explanation may be misrepresentational to some degree. This might be said about explaining how an AI system works by using metaphor or analogy (e.g., "neural nets"). At one extreme, the explanation would be complete and correct. And likely it would have explanatory value primarily to developers or computer scientists. This entails the conundrum that a complex system can only be modeled by something that is as complex as itself. So, any "interpretation" of an AI model is bound to be reductive. The issue is, what does it explain? and "to whom does the explanation have explanatory value?"

An important distinction has been drawn between epistemological (formal/scientific) explanation and psychological (cognitive) explanation (Cabitza et al., 2023). It is this latter category that is the focus of the present work: An attempt to develop a scale for evaluating the explanatory value of the machine-generated explanations to people (users, stakeholders), vs. a formalist understanding of explainability, interpretability, or transparency. The Explanation Scorecard presented here does not classify explanations in terms of form or format (images, diagrams, text, etc.); nor does it classify explanations in terms of the widely-used distinction between "local" explanations (*Why it decided this)* vs. "global" explanations (*How it works*). Furthermore, the Scorecard is agnostic with regard to the explanation requirements of different stakeholder groups (see Klein et al., 2020; Mohseni et al., 2020). Rather, it ranks explanations in terms of cognitive depth of support for the user's sensemaking.

The Scorecard originated from discussions with XAI system developers, early in the Explainable AI Program initiated by the U.S. Defense Advanced Research Projects Agency (2018). Although the concept of explainable AI pre-dated that Program (see Samek et al., 2017), the DARPA Program was a major impetus for the field. The initial version of the scorecard (Klein et al., 2020) was applied to the explanation approaches being utilized in the early phase of the DARPA Program. The system developers generally found the Scorecard interesting; it allowed them to reflect on how they might enhance their systems' explanation capabilities. The discussions also called out some subtle distinctions that entailed refinements of that initial Scorecard. Subsequent empirical evaluations, reported in the present article, also resulted in refinements to the Scorecard.

The next section of this article discusses the background on the taxonomics of types of explanations and dimensions of explanation, which leads to a consideration of the cognitive dimension that is formative of the Scorecard. Following that is a discussion of the process by which the Scorecard was developed. This is followed by presentation of the results of an empirical validation that utilized the Scorecard in an analysis of recent XAI literature. The Conclusion discusses the key findings and some recommendations for how work in this area might be advanced.

2. Classifying explanations

Across the broad and trans-disciplinary literature on explanation there have appeared many distinctions on types and categories of explanations; traceable explanations, removable explanations, layers of explanation, and so forth. Variations emerge from the focus of the research, that is, whether the focus is on the philosophy of science, cognitive psychology, instructional design, or computer science. In the XAI context, explanations can be

described according to their form and content. An AI system that relies on mathematical algorithms such as regression might need to explain such things as co-variation. Such differences in explanation format and content are of course important to developers but may not be to other users or stakeholders.

Explanations of *How it works* have often been called global. These refer to mechanism, function, or architecture, often using analogies (e.g., neural network). Explanations of *Why it did that* refer to features and instances and are called local (Lipton, 1990; Miller, 2017). Many of the explanations presented in the XAI literature involve the analysis of features, classes, or instances (Mueller et al., 2019; Belle and Papantonis, 2020; Covert et al., 2021). The global-local distinction is indeed convenient, but it is seductive. When people try to explain complex systems to other people (Klein et al., 2014, 2021), explanations of *How it works* are typically accompanied by examples or instances. Conversely, explanations about *Why it did that* serve as hints about *How it works*. The distinction has seemed important because, for example, an understanding of the AI's mechanism or process might or might not be sufficient to a user who needs to appreciate why or when an AI system might make mistakes. In other words, global vs. local does not well describe actual explanations.

In the XAI work there has been a natural focus on the explanation requirements of developers and computer scientists. In many cases, the explanations are really justifications (*Why we designed it this way*) (Lundberg and Lee, 2017; Covert et al., 2021). The explanations are often model outputs either of the AI or of a simplified model of the AI (called a *post hoc* or white box model). The formal analysis of interpretations and justifications involves the need to determine which model or class of models is better, according to computational metrics.

Holzinger et al. (2020) presented a scale for evaluating explanations in terms of user judgements (understandability, consistency, completeness, etc.). As such, it can be applied to any type, format or content of explanation; in other words, it is a scale for evaluating user satisfaction (see Hoffman et al., 2023). It is not for the evaluation of the depth of reasoning that explanations support.

Sheh and Monteath (2018) attempted to classify explanations by form, by content, and by the intended beneficiary of the explanations. Their "scope" dimension ranges from Teaching to Justification. Their "depth" dimension ranges from Attributes to Models. The "source" dimension refers to whether the explanation is from the AI or from a *post hoc* model of the AI. Explanations are further classified according attribute identity (e.g., saliency maps) vs. attribute use, which involves explaining how the AI arrived at a determination (p. 263) and the understanding of decision boundaries (p. 264). It is not clear how any of the dimensions qualify as actual dimensions, although it is possible to bin examples. For example, decision trees and rules satisfy all of the quadrants of their classification scheme, being "Introspective, Model-based, and Teaching explanations".

In brief, explanations can take many forms and formats. They can express very different kinds of content. They can be classified in many ways, such as according to the purpose or context of use (e.g., teaching). They can be evaluated by users or by the researchers themselves. Explanations can be tailored to a particular beneficiary. There is one thing, however, that all developers, users, and stakeholders have in common: the desire to make sense of the AI. It was this cognitive aspect that was formative of the Explanation Scorecard.

2.1. Explanation as sensemaking

Early in the DARPA XAI Program, a simplified model of the explanation process was used to idealize the process of explaining:

(1) The XAI system generates an explanation,
(2) The explanation is provided to the user,
(3) The user understands the explanation,
(4) Performance improves.

This model was created in order to inform system developers of the kinds of things that need to be measured at steps 2, 3, and 4: explanation goodness, user satisfaction, user mental model, performance, and reliance (Hoffman et al., 2023). From a cognitive perspective, this "spoon feeding" model glosses over the user's engagement in deliberative sensemaking.

The concept of "depth of processing" emerged in the psychology of verbal learning in the 1970s (Craik and Lockhart, 1972; Craik and Tulving, 1975). Numerous studies demonstrated that deeper processing (comprehension of meaning) resulted in better recall and recognition memory, compared to shallow processing of the surface features of the to-be-learned material. In a related vein, research has shown that deliberate and deliberative self-explanation plays a significant role in learners' understanding, in domains and contexts ranging from schoolhouse learning to learning about complex systems and devices (Chi and VanLehn, 1991; Calin-Jageman and Ratner, 2005; Rittle-Johnson, 2006; Lombrozo, 2016; Gajos and Mamykina, 2022). The purpose of self-explanation can be to satisfy curiosity, to enable the learner to develop a richer mental model, to make more accurate predictions of the AI's behavior, to develop appropriate trust in the AI, or to improve performance via appropriate reliance. Consistent with this is the key finding from the decades of research on Intelligent Tutoring Systems (Clancey and Hoffman, 2022). This research showed that to be an effective tutor, the AI must enable the user to make comparisons among cases. It must help the user integrate fragmentary knowledge into more general structural schemas. It must help the user reflect on experience and integrate general and situated knowledge.

These two extremes—"superficial hints" vs. "cognitive depth"—suggested that a scale might be formed that characterizes machine-generated explanations in terms of support for sensemaking. Differences in the content of the information and in how the information is presented can manifest as differences in the extent to which the explanations support the user's sensemaking. For example, a table showing the weights of surface features by itself does not help much in understanding how an object recognition system works, but in conjunction with examples of both positive cases and failures, information about features can be useful in figuring out how the system arrives at good answers but sometimes

gives bad answers. The Levels of Explanation presented in this article are an attempt to capture these differences.

3. The explanation Scorecard

The Levels present an ordinal variable reflecting the extent to which the information supports sensemaking to a greater depth. The ordination of the Levels 1–7 is not simply "less to more" explanation or "weaker to better" explanation or "sparser-to-richer" explanation. Moving up the Levels does not mean that the user needs to engage in less mental effort because the explanations are more complete. Nor does it mean that more mental effort is required because the explanations are more detailed and technical. Rather, it means that the explanation helps the user to engage in sensemaking to enrich their mental model. Moving up the Levels, somewhat greater sophistication to inference making is required of the user, and yet at the same time there is more support for the user who is trying to understand how to use the AI as a tool (e.g., how to anticipate confusions). Considering that the emphasis in the XAI literature is on explanations and justifications that are suited for programmers or system developers, going from Level 1 to Level 7 there is greater consideration of the *user's* needs to achieve a satisfactory and actionable understanding, rather than a purely technical understanding. In other words, the Levels are not just about explanation, they are about *self*-explanation. The sensemaking perspective differs significantly from the "spoon-feeding" perspective.

Following the presentation of the Scorecard in Table 1 is a consideration of how it covers some of the ideas in other classification schemes.

Based on analysis at either Level 1 or Level 2, a user might formulate multiple hypotheses about the rule or system that the AI is utilizing. Level 3 (Instances of Failures) is distinguished from Level 2 (Instances of Successes) because Level 3 supports the disconfirmation of hypotheses. As AI systems achieve higher levels of performance, instances of AI failures may become more salient. The discovery of what happens when the AI gets something wrong can push the user to a next level of understanding. Explanations at Levels 1 and 2 may promote users' feeling of trust or confidence in the AI by showing what the AI looks at. But such trust may not always be appropriate. Level 3 explanations adduce cautionary tales that may enable a sense of justified trust and contribute to appropriate reliance.

Level 4 (AI Reasoning) and Level 5 (Diagnosis of Failures) expand self-explanation yet further. Some XAI systems (during the time frame of the DARPA XAI Program) presented depictions of AI rules, logic, etc., and this motivated the attempt to develop the Scorecard, to provide developers with concepts and categories that may help them consider the importance of user sensemaking and self-explanation.

In the application of the Scorecard, the attribution for a given XAI system defaults to the highest level of the individual aspects of the explanation. For example, a heat map (Level 1) in conjunction with positive cases (Level 2) portraying choice logic (Level 4; AI Reasoning) would be scored as Level 4. If the XAI system included negative cases that violate the choice logic, then it would be scored

TABLE 1 The explanation Scorecard.

Null
No material is provided to support self-explaining. The user can only guess.
1. Surface Features *"Here's what it looked at."*
Level 1 explanations can be thought of as the "cues" to what the AI perceives. Surface features can be indicated by salience ("heat") maps, bounding boxes, linguistic features (text), and semantic bubbles, representing the outputs derived by the AI. The features typifying a class can be listed in text form, or in a matrix or histogram in which probabilities or other scalar variables are associated with individual features.
2. Instances of Success *"Here's examples of cases it got right."*
Level 2 explanations can be thought of as providing "hints." These reference instances or demonstrations of the AI generating correct categorizations, predictions or recommendations. The explanations might consist of clear cases or exemplars of a category; the results of various categorization analyses. Additionally, success cases might be scaled by a value on some machine-generated measure of correctness or likelihood. Examples of successes along with identifications of surface features or the values of classification attributes let the user make richer inferences about how the AI is working.
3. Instances of Failures *"Here's examples of cases it got wrong."*
While examples of successes (Level 2) are hints as to how the AI is making decisions, hints can also be in the form of examples of failures, which are often presented in contrast with exemplars or successes. Examples often include highlighting of features or differences. The comparison of failures to successes allows the reconsideration of hypotheses and the generation of alternative hypotheses. Example cases might be considered "failures" if they are accompanied by categorizations or analyses indicating a low probability or low machine confidence of their being correct.
4. AI Reasoning *"Here's how it decides."*
Level 4 explanations go beyond cues and hints, to reasons. These are decision rules: expressions of how the AI makes its determinations. These provide the user with a capability to think about when and why an AI decision was correct. These explanations can be in the form of categorization rules, choice logic, parse graphs or other symbolic forms. Goal stacks show the goals that are most activated when the AI made a decision about particular instances. These explanations are often formal or semi-formal, but they might include text or even be in the form of text. Decision rules can reference features or instances, to illustrate how the AI weights different features in order to make choices.
5. Diagnosis of Failures *"Here's why it got those things wrong."*
Level 5 explanations make the reasons for AI failures or mistakes explicit. These provide the user with a capability to anticipate failure, and determine how or why an AI decision was correct or incorrect. Explanations are diagnostic; they refer to violations of feature constraints, decision rules or choice logic. These explanations can be semi-formal, but they might include text or even be in the form of text.
6. Exploration *"Why did it get those things right? "What things can it get wrong?"*
At Level 6 there is a jump in machine capability to support self-explanation. The AI enables the user to explore contrasts in the variation of categories, features, concepts, or events. Machine- or user-generated contrasts show how the AI's determination would change or might not change if some feature of an instance were to be changed. Contrastives can be in the form of counterfactuals or semifactuals (Kenny and Keane, 2009; Wachter et al., 2017; Miller, 2018). Counterfactuals can involve categorizations (i.e., *Why did it decide X instead of Y?)* or features (i.e., *If q had changed to z would the outcome be different?*).

(Continued)

TABLE 1 (Continued)

Semi-factuals are of the form *If feature z were changed, would the instance still have been called a Q?* Manipulations can involve fuzzying, feature breaking, region deletion, inpainting, or other techniques. In one way or another, the user is able to create failure and success conditions and to make their own predictions by manipulating input features, weights, etc. in order to see the effects on the AI outputs.
7. Interactive Adaptation *"Here's how the XAI could improve; here's how the user's mental model can improve."*
At Level 7 there is another jump in machine capability: The user can provide the XAI with actionable feedback to augment either the AI models or the XAI component of the system. The interaction has to involve reciprocation, in which both the XAI and the user adapt. Specifically, the user provides feedback that enables the AI to improve its models and improve its explanations for the user. The engagement can be in the form of question-and-answer between the XAI and the user; it can be in the form of annotations or manipulations to cases. It can be about the adequacy of the XAI-generated explanations. The goal is to improve the XAI-generated explanations, which may themselves fall at any of the lower Levels. At Level 7 the distinction between an explanatory system and an Intelligent Tutor dissolves.

as Level 5 (Diagnosis of Failures). These possibilities, and others as well, are explicated in Section 4.

3.1. Inheritance

The Levels do not assume or impose inheritance constraints. It is easy to imagine such constraints. Many XAI systems present success or failure cases (Levels 2 and 3), but the cases are described in terms of the next lower level— features (Level 1). An explanation that qualifies as Level 4 (AI Reasoning) can describe reasoning by reference to features (Level 1), by reference to success cases (Level 2), or by reference to failure cases (Level 3). But the Scorecard Levels do not require that each Level inherit all of aspects of all of the lower Levels. An explanation in the form of success cases (Level 2) need not necessarily describe the cases by reference to features (Level 1), although they often do. An explanation at any Level might involve aspects of one or more of the lower Levels. For example, the explanations presented by Kenny and Keane (2009) and Dodge et al. (2019) qualify as Level 6 (Exploration) by virtue of the inclusion of counterfactuals. But the exploration is in terms of case features (Level 1), and success cases Level 2 (successes). Yet, there is no inheritance from the intermediate Levels 3 (Failures) or 4 (AI Reasoning).

3.2. About levels 6 (Exploration) and 7 (Interactive Adaptation)

Level 6 (Exploration) and Level 7 (Interactive Adaptation) may stand out from the lower levels because Levels 6 and 7 explicitly involve active exploration and thus explicitly involve agency on the part of the user. However, the assumption of the Scorecard as a whole is that users engage in an active, deliberative attempt to self-explain how the AI works. Agency on the part of the user is involved across all the levels, though it is certainly the case that some users simply take the AI outputs as givens and do not desire to sensemake to any depth.

It has been argued that counterfactuals lack explanatory value (White and Garcez, 2021). This claim seems counter-intuitive. But it hinges on a model that assumes that an explanation is a resolution: an explanation has served its purpose once it has been delivered and understood. However, the delivery of an explanation is not a terminus in the sensemaking process. Indeed, counterfactuals have a very important purpose: They show the user that further exploration is possible, and they show how exploration can be conducted. This is powerful, as it supports the exploration when the AI is operating at the boundaries of its competence envelope—instances that fall inside a classification but nearly do not, and instances that fall outside the classification but nearly do not. Such cases show when a small change makes a difference to the categorization.

Levels 6 and 7 move the XAI prospectus closer to Intelligent Tutoring Systems. The two areas of research have much in common, but most importantly, the work on Intelligent Tutoring Systems revealed many of the challenges that have confronted the field of XAI. To be an effective tutor, the AI system has to be able to form and rely upon a model of the task domain, a model of the task, a model of the user's mental model of the task, and a model of the AI system. Clancey and Hoffman (2022) review the relation of XAI and ITSs, illustrating ITSs that manifest the Levels 6 and 7 in the Scorecard. Thus, Levels 6 and 7 are at the cutting edge of current XAI capabilities. A number of XAI researchers have noted the value of interactive explanation—to allow the user to "peek inside a model's behavior" and a capability that lets users "toggle" the information in an explanation (see Bhatt et al., 2020, p. 5; also Amershi et al., 2014).

At Level 7, the user reciprocates by advising the program in some way (Goyal et al., 2019; Yeh et al., 2019; Kim et al., 2021). A number of XAI researchers have proposed systems that would fall at this level. For example, Dahan (2020) proposed a system for the analysis of trademark logos in which a user could use bounding boxes to highlight regions of a given test logo, to thereby adjust how the AI analyzed similar logos that it had identified. We know that it is possible to implement an XAI system that achieves Level 7. Two of the projects funded by the DARPA XAI Program culminated in systems that included an explanatory manipulation capability (see Stefik et al., 2021).

4. Related work

Although the Scorecard Levels are specifically for describing explanations of AI systems, one would expect them to be consistent with taxonomies developed in cognitive psychology and instructional design.

4.1. Related work in instructional design

Perhaps the most cited such scheme is that advanced by Benjamin Bloom (Bloom and Krathwohl, 1956; Anderson et al., 2001). It references categories that are types of knowledge (factual, conceptual, procedural, and meta), and cognitive processes

(remember, understand, apply, analyze, evaluate, and create). The scheme identifies types of knowledge applied by using particular cognitive processes, for example using factual knowledge in order to conduct an analysis. Discussions of the limitations of the Bloom taxonomy appear in Moore (1982) and Bereiter and Scardamalia (2005).

The Scorecard Levels embrace Bloom's types of knowledge by virtue reference to facts about why an AI made a particular determination and how the AI works (e.g., decision rules). The Scorecard Levels are formed on the basis of a cognitive dimension (sensemaking), and they directly reference the cognitive processes of understanding, applying (developing appropriate trust and reliance), analysis, and evaluation (exploration).

A number of other taxonomies combine categories of purpose (remembering vs. application), content (facts, concepts, principles) and procedures (such as analysis) (e.g., Moore, 1982; Sugrue, 2002). The Scorecard Levels are consistent with those, in a manner similar to their consistency with the Bloom taxonomy.

Given that these schemes list varieties of knowledge content and varieties of cognitive processes, it would be difficult for the Scorecard Levels to *not* be consistent with them in some respects. But the integration of schemes is handicapped when some refer to dimensions when in fact the orderings are categorial. Marzano and Kendall (2007) argued that it is impossible to unambiguously order learning content or processes by the degree of cognitive demand or mental difficulty. Complex procedures can sometimes be learned easily. It can be difficult to consistently order examination questions according to cognitive difficulty (Dempster and Kirby, 2018). Ordering by cognitive difficulty is tacit in some of the taxonomic schemes, including the Bloom scheme.

The Scorecard Levels do not attempt such an ordering; they are about the manner in which explanations support sensemaking, not the degree of difficulty of sensemaking. The emphasis on sensemaking is consistent with the emphasis of the learning sciences on metacognition, that is, deliberative reasoning about one's own understanding (Marzano and Kendall, 2007).

The Scorecard Levels also provide coverage of the types of explanation of AI systems.

4.2. Related work in AI

The Levels cover various distinctions that have been proposed with regard to types of explanations. For example, Level 6 covers what (Covert et al., 2021) call "removal-based explanations". Level 2 (examples as explanations) covers what some XAI researchers refer to as prototypes (as in Kim et al., 2016; Vong et al., 2018).

The Scorecard covers classifications of types of machine-generated explanations. For example, the Levels cover the "explainability categories" presented in the review by Belle and Papantonis (2020). Levels 1 and 4 correspond to what they call local explanations. The Scorecard Levels cover the types of *post-hoc* interpretations listed by Amershi et al. (2014); saliency maps (called "local explanations") are featural (Level 1), explanation by nearest neighbors is successes (Level 2). The Scorecard Levels map to the categories of explanations developed by Vilone and Longo (2020). Numerical explanations and visual explanations would be Level 1; selection of prototypes would be Level 2; rule-based explanations would be Level 4.

The Levels cover what Preece et al. (2018) call traceable explanations, justifications and assurances (Level 4). These reference features (Level 1), input-output relations (decision rules, Level 4) and success cases (Level 2). Preece et al. refer to these types as "layers". The layering involves a computational linkage (e.g., Layer 2 "links" to Layer 1) but the layering also involves the notion that different stakeholders need different types of explanation. The Scorecard is agnostic with respect to the intended beneficiary (developers, stakeholders, etc.). Any of the Levels can apply to explanations suited to any stakeholder group, although the form and content of those might differ from one group to another (see Klein et al., 2020). Throughout this article we use the term "user" to refer to the sensemaking activity of any individual.

The scaling of explanations in terms of their support for self-explaining involves merging some types that are distinguished for the purposes of computer science. Guidotti et al. (2018) distinguished decision trees and decision rules (both are Scorecard Level 4). They distinguished features of importance, saliency mapping, "sensitivity analysis" and "activation maximization" (all are Scorecard Level 1). This focus on distinctions of importance to computer science is clearly different from the consideration of explanatory value for individuals who are not computer scientists.

5. An application of the Levels

The Levels were used to score a set of reports on XAI system development. The method was one that has been utilized in protocol analysis (Hoffman et al., 1998; Crandall et al., 2006). Independent evaluators classify statements using a pre-defined scheme (in this case, the Levels of Explanation). Then there is a discussion of the disagreements, which can sometimes be resolved directly. This enables calculation of the rate of essential agreement. Remaining agreements are resolved by clarifications to the coding scheme.

We expected that many machine-generated explanations would not reach the higher levels, yet hoped to see instances of systems that fell at highest Level (Interactive Adaptation) (see Table 1).

5.1. Materials

The scope of the evaluation of the Levels of explanation had to be bounded. Within the field of computer science, it might be argued that all AI systems that have an interface for data visualization, data fusion, and visual analytics are a form of Explainable AI (for example, see Liu et al., 2017). However, the goal of such systems is to explain data, not explain how the AI works. A literature search was conducted to identify reports (publications, proceedings) using the following Boolean: ["Explainable AI" OR "XAI" OR "Explainable Artificial Intelligence" OR "explainability + AI"]. The search was also narrowed to the years of the DARPA XAI Program 2019–2021. The resulting set of 165 articles included journal publications, publication preprints, and conference presentations. Three of the

researchers read the abstracts and performed a cursory read of the articles. This enabled a down-select, retaining only those articles that explicitly related to explainable AI, methods of explanation, examples of explanation, or evaluation of explanation methods (many articles referenced explanation in their titles or key words lists but were actually about formal interpretability or *post hoc* models, for example). Next, review articles and opinion pieces were removed from the focus set. Most of those referenced articles that had been published prior to 2019, but more importantly, the example explanations that were provided were not always ones that had actually been machine-generated. In many cases, the example explanations were ones generated by the authors (of the various selected articles) to illustrate what they felt to be good explanations.

The final set consisted of 33 articles that reported on the development of XAI systems and that included examples of machine-generated explanations. One of the articles reported on the development of two XAI systems, so the total number of XAI systems that could be scored was 34.

5.2. Procedure

The articles were evaluated by three of the authors. One evaluator was a Ph.D. cognitive psychologist specializing in educational psychology, and one was a Ph.D. student in the field of Intelligent Systems and robotics. Both were familiar with the literature on Explainable AI and had conducted an extensive review of the literature. The third evaluator was a Ph.D. Cognitive psychologist who has conducted and developed methods of cognitive task analysis and protocol analysis.

The evaluation process requires is a qualitative analysis (see Corbin and Strauss, 1990; Hughes and Wood-Harper, 2000; Yardley et al., 2020). Previous research that has utilized the method of independent judgments in protocol analysis has found that only one or two scoring passes are required for evaluators to resolve differences in their judgments and achieve consensus (see Hoffman et al., 1998). The present analysis involved three waves of scoring. There was progressive refinement of the scoring categories (the Levels). Judgment benefitted from the review of multiple and different XAI systems, each bearing its own unique characteristics and context. This supported adjustment of the definitions of the Levels until a consensus was achieved.

In the first wave, two evaluators independently scored the XAI systems' explanations utilizing the initial version of the Scorecard, which had been developed early in the DARPA XAI Program (Klein et al., 2020). The first pass scoring was conducted during the final stages of the process of down-selecting the articles. It was clear when explanations involved reference to features (Level 1) and this held for a many of the articles. Overall, however, there was only about a 50 percent agreement between the first two raters.

Discussion revealed that it was not always clear what counted as a "success" or a "failure." For Level 1 explanations (e.g., saliency maps, graphs of feature weights, etc.) is it not clear what it would mean for an explanation to reference an example of a success (as in Abdollahi and Nasraoui, 2016; Anderson et al., 2019). For Level 2 and Level 3, an explanation might be presented as a Success case, but if the AI was merely reporting empirical data (e.g., the percentage of people who liked a particular movie), that could not be legitimately attributed as a success on the part of the AI system. Some scoring disagreements were easily resolved by discussion, and both raters changed some of their scorings. This resulted in 70% rate of essential agreement.

Next, a third researcher conducted a scoring. Discussion of the results mandated the rejection of the assumption that there would be strict inheritance as one progressed up the Levels (described above). This resolved most of the remaining disagreements. For example, explanations that involved Instances of Successes (Level 2) usually described successes in terms of features (Level 1). But sometimes they did not. Explanations that expressed AI Reasoning (Level 4) usually expressed rules in terms of Features (Level 1) and Success examples (Level 2), but sometimes they did not. As a third example, an explanation that was in the form of a saliency map (Level 1) might involve a comparison of multiple maps, which would qualify as Level 6 (Exploration). Two articles reported explanations that were scored at Level 6, but there was no inheritance from Levels 3 or 5.

Perhaps the best example of scoring subtleties is an article that presented explanations that consisted of saliency maps plus rewards that were assigned for a particular course of action. This made it appear as if the XAI explanations revealed AI Reasoning (Level 4), but the explanations were actually just a number of different kinds of cues, all of which would be Level 1. And yet, later in the article it was pointed out that the AI agent was programmed to sometimes fail, enabling the user to see the failure and the accompanying proportional decrease in the reward for the chosen course of action. Thus, the highest Level achieved by this system was Level 3 (Instances of Failures).

The third round of scoring resulted in a refinement of the definitions of the Levels 4 through 7. For example, the description of Level 6 was adjusted to be more definitive about the use of contrastives (counterfactuals and semi-factuals).

In the discussion of the results of the scoring by the three researchers, a number of changes were made to the scoring assignments, in which each of the three raters changed between one and three of their scorings. This second round resulted in pairwise agreement rates between 70 percent and 87 percent. A final discussion of the few remaining disagreements led to convergence and 100 percent agreement. These final adjustments were incidental. For example, one report presented additional material about the machine-generated explanations late in the paper, rather than at the point where the machine-generated explanations were described. The final version appears in Table 1.

5.3. Results

The down-select procedure resulted in reports that can be categorized as: classification systems (objects, events, situations, handwriting, text, patients) ($n = 15$), Decision making systems (health care, finance, criminal justice, selection of interpretability tools, selection of classification models) ($n = 13$), Planning systems (recommenders, game strategy analysis, course of action evaluation) ($n = 2$), Data analytic systems, ($n = 1$) and Vehicle control systems ($n = 2$). We now present summaries of the selected

reports, describing their AI application, the application domain, the nature of the machine-generated explanations, and their Levels scorings.

Abdollahi and Nasraoui (2016)
AI application: planning system

The application domain is movie recommendation. Explanations were empirically derived from ratings of movies on a 1–5 scale. Text expressed the proportion of people who rated a given movie as "x or greater out of 5". A histogram showed the proportion of people who rated the given movie as "1–2" vs. "3" vs. "4–5". The method by which the explanations were generated was (apparently) not described to the participants. The explanations are of the recommendations.

Score: Level 1 (Features)

Adebayo et al. (2018)
AI application: classification system

The application domain is the classification of objects. The outputs of image processing algorithms were regarded as explanations. The presented example consists of a figure composed of multiple images including three original photographs (of a bird, a dog, and a person eating an ear of corn), each of which is accompanied by a series of seven images showing the outputs of various saliency algorithms and an edge detector. In the presented example, the instances are all correct classifications, and so can considered successes.

Score: Level 2 (Instances of Successes)

Akula et al. (2020)
AI application: classification system

The application domain is classification of objects. The XAI system generates counterfactuals about the classification of objects by identifying the minimal ("semantic") features that would have to be changed in order to change the AI's determination. The explanations take the form of an exemplar (e.g., a toad) accompanied by images of semantic features (e.g., bumpiness of the skin) that would have to be changed to alter the determination (e.g., a frog).

Score: Level 6 (Exploration)

Anderson et al. (2019)
AI application: decision making system

The application domain is analysis of courses of action by a machine agent in video game play. Explanations took the form of views of the game board along with the player's score. Saliency maps were said to show what the machine agent was seeing. The XAI system would remove game elements, and the participant's task was to anticipate what the agent would do at the decision points. After making a prediction, the XAI system showed either a saliency map or a histogram showing the rewards for each of the agent's alternative courses of action. The researchers assert that "By comparing the bars of two actions, a human can gain insight into the trade-offs responsible for the agent's preference" (p. 2).

Score: Level 3 (Instances of Failures)

Anguita-Ruiz et al. (2020)
AI application: classification system

The application domain is the identification of gene expression patterns related to obesity. The explanations were an interface showing the results of the processing of gene expression data, along with annotations about gene functions from gene data bases and encyclopedias. The processing system identified rules that described transcription pathways. The goal of the research was to evaluate the "biological quality" of the rules (i.e., spurious associations vs. true causal relationships). Experts in the genetics of obesity evaluated the rules by relying on a display showing the associations among genes along with values on a number of measures (e.g., gene expression change). About 100 genes (identified by codes names such as "8127A/ME1") were arrayed along the perimeter of a circle. Running among them were connecting lines showing the rules (i.e., gene associations). These were sometimes isolates and sometimes formed clusters of convergence and divergence. The display enabled the experts to determine the "biological meaningfulness" of rules. The XAI system did not explain how the rules were discovered by the AI, but the explanations that the AI had discovered.

Score: Level 4 (AI Reasoning)

Cheng et al. (2021)
AI application: decision making system

The application domain is clinical/surgical decision making. The Visual Analytics XAI system presents a complex, multi-window display of a number of data types and algorithm outputs. For example, a two-bar histogram indicates the pre-surgery and in-surgery risk level for a patient. Timelines show changes in patient status. Displays of numerical and symbolic data show patient pre-surgical information (e.g., oxygen saturation level). These various data fields provide contextual information that is intended to help clinicians. The system tool presents a detailed visualization of case features (Level 1) with a capability that enables users to investigate the contribution of features to the treatment recommendation and treatment outcome through a "what-if" tool. This would qualify the system as Level 6 (Exploration). The qualification is that little information is provided to the users about how the variation of features impacts the treatment outcome prediction, leaving the reader to assume functionality of the provided features.

Score: Level 6 (Exploration)

Choo and Liu (2018)
AI application: classification system

The application domain is handwriting analysis. MNIST are described using notional network-like diagrams to illustrate the "inner workings" of deep net models, and color-coded arrays to depict the activation patterns of nodes. Icon colors and color clusters depict proximity in a feature space (Level 1). The system includes a capability whereby users can change configurations of deep nets to see the effects on the machine outputs. This would

qualify as Level 6. There is also a capability of "interactive model refinement," in which expert knowledge can be used to refine a deep learning method.

Score: Level 7 (Interactive Exploration)

Dahan (2020)
AI application: decision making system

The application domain is the choice of classification models by legal professionals. The XAI system is intended to assist in determining legal standing of IP and trademarks, based on case law and an Intellectual Property registry. A number of classification models were compared, including SHAP, Logistic Regression, Decision Trees, XGBoost, Unsupervised Saliency Maps, and feature detection algorithms. The trademark analysis utilized saliency maps to highlight similarities of the given trademark logo with other similar trademark logos. The system classifies the input based on visual, phonetic and conceptual similarity to existing textual and image data. The XAI provides explanations of the results in the form of references to the specific laws used to arrive at the decisions. The explanations include highlighted words in pieces of legal text, intended to explain how the AI classified a given case. These explanations referred to visual similarity of trademarks (e.g., "The difference between the marks at issue cannot counteract the similarities..."), phonetic similarity (e.g., "It must be concluded that the Board of Appeal did not err in taking the view that the two signs at issue are phonetically similar"), or conceptual similarity (e.g., "The Board of Appeal did not err in taking the view that the two signs at issue are conceptually similar"). Each analysis was accompanied by a "confidence score".

Score: Level 2 (Instances of Successes)

David et al. (2012)
AI application: decision making system

The application domain is financial decision making but the experiment's task involved an interactive game in which participants could earn coins by selling lemonade. Participants had to decide how many cups to make per day, based on the price of lemons and a weather forecast. After some practice at the game, participants could take recommendations from an "advisor." The advisor's recommendation could be a good one or a flawed one. Participants had the option to accept the recommendation without seeing the details of it. Participants were told that the recommendations came either form a human expert or a computer algorithm.

The Recommendations (explanations) all took the form of text, which could be in one of three forms:

- Global Explanation–Information about the type and extent of data it uses (sales data from many lemonade stands over several years).
- Feature-based Explanation–The features that were used to generate the prediction (e.g., "Based on data from lemonade stands over several years, your previous sales, and market demand, the algorithm recommendation is to make six cups of lemonade").
- Performance-Based Explanation–A confidence rating for the advice (e.g., "90 percent certainty").

The explanations were presented in two windows. One showed the weather prediction and lemonade cost, and a field for the participant to the number of cups to be made. The second window showed the actual weather, the actual demand for lemonade and the number of cups sold. Below the two windows was a progress bar showing the main events across the game session, starting from the point where advice became available, to the point where the advice failed, to the point where the advice came with a cost. How the explanatory text was incorporated into the display is not shown. The explanations had a clear effect in that when the algorithm/advice performed well the adoption rate increased. There was no difference between being told the advisor was a human vs. an algorithm.

Score: Level 1 (Features)

Dodge et al. (2019)
AI application: decision making system

The application domain is fairness judgments concerning predictions of recidivism. Explanations took the form of text. Case Explanations asserted such things as *The training set contained 10 individuals identical to Illana and 60% of them re-offended.* Demographic explanations referenced the likelihoods of reoffending for individuals of the same race and age as the given case. The Case Explanations and the Demographic Explanations are arguably Level 2 (Successes). "Input-Influence explanations" expressed the likelihood that an individual with certain features would be more/less likely to reoffend (e.g., race, age, number of prior convictions) (e.g., Iliana's age is 18-2; if it had been older than 39, she would have been predicted as NOT likely to reoffend). These are counterfactuals.

Score: Level 6 (Exploration)

Ehsan et al. (2019)
AI application: planning system

The application domain is the evaluation of machine agent game strategy. The explanations consisted of text that provided the "rationale" for the actions of an "agent" which was playing a video game (Frogger). The rationales described what each action was and the machine agent's reason for the action. The text was accompanied by a view of the game board, showing the game grid and the sequence of game piece moves made by the agent/game player. The participants in an evaluation study were told that the game moves and rationale statements were made by a computer agent. Participants rated the rationale statements for confidence, human-likeness, adequacy and understandability. The rationale statements did not describe how the agent works or how the explanations were generated. The actual instructions given to the participants are not provided in the article.

Score: Level 4 (AI Reasoning)

Harbone et al. (2018) and Preece et al. (2018)

AI application: classification system

The application domain is the identification, monitoring, and prediction of traffic congestion in images and video. A large data set of images were annotated for ground truth. A subset was then used to train Deep Net classifiers. One explanation form was a saliency map, with green coloring indicating evidence toward "Not congested" and red coloring indicating evidence toward "congested" (Level 1). Images were also annotated with arrows, derived based on a neural network, indicating cars that contributed to the "congested" classification. The annotations were referred to as "*Post-hoc* explanation via salient semantic object identification." The researchers refer to the presentation of a rule that states if the ratio between the average car pixel velocity and the speed limit is lower than a threshold, the road is classified as congested. Such a rule could be presented to the decision maker to provide an explanation of how the system used the component data to infer the final classification (Level 4). However, this seems to have been just a prospectus for a form of explanation.

Score: Level 2 (Instances of Success)

Hind et al. (2019)

AI application: classification system

One application domain is the identification of "good" loan applications, based on such features as salary and risk (Scorecard Level 1). Another cited application is of a mapping of patients to treatments based on a list of pairs of patient features (symptoms) and treatments. A Cartesian product of a mapping of features to classes is the full set of all ordered pairs derivable from an n x n matrix. From that, an explanation is regarded as a justification of why a feature vector was mapped to a particular class (Scorecard Level 4). Based on a data set, the ML system finds the right number of classes. This would be an expression of decision rules. A given example is:

> For a loan application to be good:
> Number of satisfactory trades $\geq$ 23 AND
> External Risk Estimate $\geq$ 70 AND
> Net Fraction of Revolving Burden $<$ 63

Score: Level 4 (Reasoning)

Hohman et al. (2019)

AI application: data analytics

The intended beneficiary is the data scientist or system developer. That is, the goal was to make machine systems more interpretable, in the formal sense of that term. Explanations were in the form of data histograms ("visualizations") and text ("verbalizations"). The histograms described the contributions of features to the categorization of an instance. The text described the contribution of features to the system's determination. The histograms were regarded as an "overview" while the "verbalizations" highlighted particular features or trends. The display included a slider, with which the user could adjust the simplicity vs. detail of the machine-generated descriptions. When set to Brief, the explanation was just text that describes the difference between the two instance's predictions. An example is: "Predictions vary potentially due to some features." Dragging the slider to the second position, the sentence updates and displays the exact number of features that distinguish the two predictions. When set to Detailed position, the explanation lists the prediction values, the number of differing features, and the percent of the number of total features that distinguish the instances. An example is: "Overall predictions for instances 126,024 and 312,129 vary potentially due to nine features (i.e., 25%)." The researchers refer to this style of explanation as "interactive verbalization," but the interactivity is limited since the user can only adjust the detail/specificity of the explanations. The system supports the exploration of the features used to classify an instance and adjustment of the machine-generated explanations.

Score: Level 6 (Exploration)

Jesus et al. (2021)

AI application: decision making system

The application domain was fraud detection in financial transaction records. Three explainers were used (LIME, TreeInterpreter, and SHAP) to calculate feature distributions and determine which of the 112 possible features were used to build decision trees. The outputs were used to generate explanations of the model scores (i.e., which features contributed to the decision). Information about an ML's decision was provided to users cumulatively, starting with just the data (transaction features), then a ML "model score," then data and score plus an explanation. The model scores were referred to as an expression of the impact of features, but the quantification method was not explained and just a single model score was provided for one test case. The explanations consisted of a list of features, each of which was accompanied by a score expressing the degree to which each feature contributed to the classification, highlighted by a red vs. green square icon indicating whether a feature indicated risk or legitimacy.

Score: Level 1 (Features)

Kaur et al. (2020)

AI application: decision making system

The application domain was the choice of interpretability tools by data analysts. Interpretability tools (*post hoc* explanations) create mathematical descriptions of the input-output relations generated by black box systems, which are themselves said to be opaque or of low intrinsic interpretability. Looking in detail at two interpretability tools (GAMS and SHAP), the visualizations of the outputs of machine learning systems included a histogram showing the impact of features on model outputs. A numerical scale showing scores indicated how higher and lower values were thresholded for different categories (in the illustrated case, marital status and income). Kaur et al. referred to histograms of the important scores for individual cases as "local" explanations, which they were. Kaur et al. also referred importance of features as "global" explanations, which is a debatable appellation.

Score: Level 1 (Features)

Kenny and Keane (2009)
AI application: classification system

The application domain was the classification of handwriting using the MNIST data set. The XAI system focused on contrastive reasoning of two kinds: (1) counterfactuals (*If you had applied for a smaller loan, it would have been approved*) and (2) plausible semi-factuals (*Even if you had applied for a smaller loan, it still would have been refused*). Instances were paired with a variation that had a feature change that did not change the classification, a counterfactual variation where a feature change did result in a classification change, and a variation in which a subtle (hence plausible) feature change would lead to a different classification. The researchers conducted a formal evaluation of their system with other systems that generated counterfactuals. A number of mathematical comparisons were made based on a variety of measures, such as the distance of an instance from the nearest instance that was in the training set. This measure was interpreted as a measure of plausibility.

Score: Level 6 (Exploration)

Kim et al. (2021)
AI application: vehicle control system

The application domain is self-driving automobiles. An AI system learned vehicle control with the benefit of human advice. The XAI conducted an analysis of visual salience, to show the developers and the research participants which portions of an image the AI "looked at". The vehicle would summarize what it had observed and segmented with a semantic association. It expressed this in natural language along with a statement of what it intended to do (e.g., I see a pedestrian crossing, so I stop). This included justifications relating an action to a reason (e.g., The vehicle slows because it is approaching an intersection and the light is red). The XAI was pre-populated with action-justification statements for all of the events that occurred in the training data set. Although the user could provide the XAI with actionable feedback, the feedback was not about the machine-generated explanations (the XAI), but was about the reasoning of the AI (the automation's decision rules). Thus, the explanations do not satisfy Level 7. The justifications that the AI presented were decision rules.

Score: Level 4 (AI Reasoning)

Krause et al. (2016)
AI application: classification system

The application domain is the assignment of patients to risk categories of treatment risk on data on the relative effectiveness of various medications. However, data analysts are strongly implicated as the primary intended beneficiary of the explanations. The example explanation that was provided consisted of a graphical portrayal of nine scales, each of which consists of a color bar and a numerical scale. The color bar indicates the distribution of values. Each scale references a particular diagnostic having its unique measurement scale (e.g., blood glucose level, body mass index). The user (perhaps a diagnostician) can set thresholds for defining "high risk patients." Indeed, one use of the manipulation capability is in the conduct a diagnostic task, and not to develop an understanding how the AI works. However, "[the system] can be used for model interpretation, with a focus on visualizing input/output behavior rather than the model itself. Users can change the sort order of the partial dependence bars by using the buttons at the top. In addition to sorting by the feature weight and relevance as determined by the predictive model if available, users can also sort according to local feature importance and impactful changes. If impactful changes are chosen as the sort order, the suggested changes to each feature are indicated" (p. 109).

Score: Level 6 (Exploration)

Neerincx et al. (2018)
AI application: decision making system

The application domain is monitoring ocean-going vessels. The decision of a recommender system was accompanied by a sentence frame providing the reason for the AI's decision (e.g., "I expect changes to the weather condition and am reasonably certain of this estimate. I advise you to..."). These were generated by reference to an ontology and preformulated sentence frames. The key element to the explanations is the expression of the AI's confidence level. However, "... the particular design pattern behind the message also allows extra information to be provided on request of the user" (p. 207).

Score: Level 6 (Exploration)

Neerincx et al. (2018)
AI application: decision making system

The application domain is diabetes control by young patients. The explanations were in the form of text (accompanied by a cartoon robot who asks the questions of the patient). The system presents multiple choice decision problems, which list alternative courses of action. These serve to induce contrastive reasoning on the part of the user. "Therefore, the generic method for automated explanation extraction includes the identification of the foil for a contrastive explanation" (p. 207).

Score: Level 6 (Exploration)

Petsiuk et al. (2018)
AI application: classification system

The application domain is object classification based on photographs. "Importance maps" use a rainbow code to show the impact each pixel has in determining a class assignment. Examples show: (1) A photograph depicting an object (e.g., a bird, a cow), (2) Two importance maps showing the importance of pixels for each of two possible categories (e.g., a bird versus a person), and (3) The calculated likelihood (percentage) for each of the two categorizations. The presentation of a low and high likelihood categorizations can be regarded as a comparison of successes and failures. These are hints; they show what the AI "confused" but do not say why. Thus, there is no diagnosis (Level 4).

Score: Level 3 (Instances of Failures)

Pierrard et al. (2018)
AI application: classification system

The application domain is the classification of geometric shapes utilizing an artificial data set consisting of black and white images showing shapes (circles, ellipses, etc.) of varying degrees of fuzziness and located in various locations in an image field. The explanations consisted of text that stated the class assigned to an image and listed the features that typified the class (e.g., "Class 4 because object A is a disc and object B is a square, and a disc is above the ellipse").

Score: Level 1 (Features)

Pierrard et al. (2019)
AI application: classification system

The application domain was the identification of body organs in radiographs. An algorithm was trained to identify closed sets of relations (e.g., "is close to", "to the left of") associated with organs. The explanations consisted of the images annotated identifications of organs, and text expressing the reasons for the identifications (e.g., Organ 9 is the bladder because it is stretched and it is below the right kidney).

Score: Level 1 (Features)

Poursabzi-Sangdeh et al. (2021)
AI application: decision making system

The application domain is the prediction of selling prices for apartments based on various features (e.g., number of bedrooms, days on the market). The study presented participants with the results from two different interpretable (*post hoc*) models. The example explanations that are presented are lists of features and the output of the interpretability models. While these involved the formal notion of "interpretability," the authors assert that the outputs of interpretable models are explanations for users.

Score: Level 1 (Features)

Russell (2019)
AI application: decision making system

The application domain was financial (loan) decision making. The XAI system generates counterfactuals that describe how the data would have to be changed so as to result in a different outcome. This is achieved by finding minimal changes in the data that would lead to a different response by the algorithm (e.g., You have an income of $30,000 but if you had an income of $45,000 you would have been offered the loan). The XAI generated counterfactuals by using sentence frames (e.g., "You could have gotten a score_____ if _____ took the value of _____ rather than_____").

Score: Level 6 (Exploration)

Samek et al. (2017)
AI application: classification system

The application domain was object recognition in images (e.g., volcano, coffee cup) and activity recognition (e.g., "sitting up") in video clips. Explanations consisted of heat maps showing pixel salience. These were described as what the AI was "seeing". For a text classifier, text pieces were annotated by highlighting keywords that mapped to classifications (e.g., "sickness" and "discomfort" mapped to a medicine category).

Score: Level 1 (Features)

Singh et al. (2016)
AI application: decision making system

The application task is the prediction of the likelihood of hospital readmission based on various patient features. The explanations are "program snippets" that are a transformation of a decision tree into a simple if-then programs (two or three lines of code). The snippets were derived from black box models. An example snippet is:

```
if Diag: Other and not Tolbutamide:
Discharged: Home
else: Diag: Other
```

Score: Level 4 (AI Reasoning)

Turner (2016)
AI application: classification system

The application domain was face recognition. An "explanation face" was created as the product of the original image and a salience map. A model-derived description (saliency regions) was subtracted from the original image. This had the effect of showing the face with the regions of high salience (represented using white). Bounding boxes were added to highlight what the authors call "cues." From the provided example, it is not clear what the cues are cues of. Why, for example, should the person's nostrils or the corner of one eye be important in person identification?

Score: Level 1 (Features)

Vasu et al. (2021)
AI application: classification system

The application domain is content-based image retrieval and classification. The explanations consisted of an array of six images that all received a given classification (e.g., cats, tables, carrots), and salience maps for each of those images. The retrieval system was designed to incorporate user feedback. The user can indicate which of the retrieved images are, and which are not instances of the class. This can be regarded as training of the AI (the user effectively tells the AI that it was wrong about something.)

Score: Level 6 (Exploration)

Vilone and Longo (2020)
AI application: classification system

The application domain was object classification (e.g., cats, tables, carrots). The explanation for a given image was a saliency map (Level 1). These were said to "help the user understand

what it is paying attention to" (p. 1). The system also presents a set of comparison cases (Level 2). The system was designed to incorporate user feedback. The user could indicate which images in a retrieved set were and which were not instances of the class. This can be regarded as training of the AI: The user effectively tells the AI that it was wrong about something (Level 7). The researchers asserted that "[The] user helps the system understand their goal through saliency-guided relevance feedback" (p. 1).

Score: Level 7 (Interactive Exploration)

Wang et al. (2019)
AI application: decision making system

The application domain is phenotyping for intensive patient care. From data on vital signs (features) interpretable models generate counterfactuals. The "Explanation sketches" consisted of a number of graphical components: A graph showing the values for 12 vital signs over time, a histogram showing vital signs' importance for each of five alternative diagnoses, and a text box showing the counterfactual rule for each prediction (e.g., The system would predict SHOCK if any of the following conditions were satisfied...). These were intended to suggest why the AI predicted a given condition, referencing the features (the vital signs).

Score: Level 4 (AI Reasoning)

Wang et al. (2022)
AI application: vehicle control

The application domain is the analysis of the behavior of autonomous vehicles. The information presented to the user is composed of spatio-temporal features (Scorecard Level 1): objects, object locations, movements, timelines, etc. The ability of the user to filter/select features allows the user to review the vehicle AI model, see examples of when it was successful or when it failed, as well as explore the connection between events, features and surrounding data (Levels 2 and 3). A display of object detection results and features shows green bounding boxes to indicate ground truth and red bounding boxes to indicate predicted positions. The darker the color, the lower the activation value (Scorecard Level 4). Like the work of Kim et al. (2021), Wang et al. focus on explaining why the AI (the autonomous vehicle) failed, i.e., the reasons for accidents (Level 5, Diagnosis of failures). The system offers the user an opportunity for users to explore what would happen if input data (scenario features or parameters) were changed. The user can also change data connections to see how the output would change.

Score: Level 6 (Exploration)

White and Garcez (2021)
AI application: classification system

The application domain, based on the Adult Data Set, is predicting whether an individual earns more than $50K per year. Individuals are described in terms of socioeconomic class, age, occupation, and other factors (Scorecard Level 1). The explanations consisted of a number of graphical components: A graph showing the values for the variables and a histogram showing importance of each measure for each alternative classification (Scorecard Level 2). Additionally, the XAI system was able to present counterfactuals, such as by changing marital status or education level. The results of counterfactuals are presented in a table showing the degree of change in the coefficient in the regression model. This would fall at Scorecard Level 6, but it is noted that the counterfactuals were generated by the system, and not posed as queries by the user.

Score: Level 6 (Exploration)

5.4. Summary of the findings

For the two largest applications categories (Decision Making and Classification Systems) systems, explanations were scored at Levels 1 through 4, and Level 6. Levels 1 and 6 were the most frequent (four systems at Level 1 and five systems at Level 6). For the other three applications categories (Planning, Data Analytic, and Vehicle Control Systems), scorings were predominately Levels 4 and 6.

Explanations at all Levels are ultimately expressed in terms of features or the manipulation of features. Explanations at Level 4 (AI Reasoning) often describe decision rules in terms of features (e.g., Hind et al., 2019; Jesus et al., 2021; exceptions are Singh et al., 2016 and Ehsan et al., 2019). Explanations at Level 6 describe explorations in terms of feature manipulations. An analysis of XAI systems that had been conducted early in the DARPA XAI Program showed that five out of the eleven XAI systems developed by the Performer Teams generated explanations that were exclusively at Level 1. In regard to this, the most intriguing findings of the present analysis are the relatively high proportion of systems that were Scored at Level 6 (Exploration), and the fact that two Classification systems achieved Level 7 (Interactive Exploration).

6. Limitation of the methodology

The Levels as defined are not free of all possible ambiguities or subtleties of interpretation, if only because the key concepts that are involved are intrinsically complex and subtle. Qualitative analysis requires judgment (see Corbin and Strauss, 1990; Hughes and Wood-Harper, 2000; Yardley et al., 2020). Because of this, it is possible to regard the Levels scoring process as limited. The Scorecard Levels form an ordinal scale, not an interval scale. Thus, for example, a system scored at Level 6 is not twice as explanatory as a system scored at Level 3. It would be desirable to have quantitative methods for evaluating and comparing machine-generated explanations. But such methods would themselves necessarily rely on categorical, conceptually-dependent judgments. For a deconstruction of the subjective-objective distinction see Mitroff (1974), Muckler and Seven (1992), or Annett (2002). Developing quantitative scales is certainly a challenge for XAI research, and its cognitive science complement.

Previous research that has utilized the method of independent judgments in protocol analysis has found that only one or two scoring passes are required for evaluators to resolve differences

in their judgments and achieve consensus (see Hoffman et al., 1998). The present analysis involved three waves of scoring. There was progressive refinement of the scoring categories (the Levels). Judgment benefitted from the review of multiple and different XAI systems, each bearing its own unique characteristics and context. This supported adjustment of the definitions of the Levels until a consensus was achieved.

The present analysis was complicated by the need to coordinate the work across multiple sessions and in a distance modality. The discussion of the scoring and negotiation of any disagreements should be conducted in person, by the entire group of evaluators.

The primary limitation of the present work is the restricted sample. The down-selection procedure revealed just those reports that included actual machine-generated explanations, published or posted up to the date of the analysis. None of the systems was scored at Level 5 (Diagnosis of failures). The system described by Wang et al. (2022) achieved Level 5 but was scored at Level 6 by default, that is, because it achieved that higher Level. We suspect that there are more XAI systems now that would be scorable at Level 5.

7. Use of the Scorecard

The success of XAI systems development hinges on an ability to empirically evaluate the quality and effectiveness of machine-generated explanations. Although early XAI research involved the development of systems that did not support the user's sensemaking to much cognitive depth, researchers have been aware of the need to do just that.

Scorecard Levels present a judgment scale whereby researchers and system developers can assess machine-generated explanations in term of the degree to which they support the user's sensemaking. This analysis can be prospective, in the sense of deciding early in system design the level of explanatory depth that is desired and they building the XAI system to meet the selected Level. The assessment can also be retrospective, in the sense that the researchers or independent evaluators can assess machine-generated explanations in the manner reported in our analysis.

The Scorecard scale represents a first attempt to do this. The scale can guide the creation, as well as the evaluation of XAI systems. The Scorecard is not interpreted as a measure of the explainability (of AI systems). The Scorecard levels are not interpreted with reference to explainability in the formal sense usually meant by computer scientists. Explainability is generally regarded as a formal property of AI systems (e.g., Koh and Liang, 2017; Adadi and Berrada, 2018; Kaur et al., 2020; Mohseni et al., 2020; Jesus et al., 2021).

The Explanation Scorecard is intended to be useful for identifying alternative ways to present information that might help users by supporting their self-explanation. The Scorecard provides the conceptual terminology needed to consider the cognitive value of machine-generated explanations, and how those explanations might be enriched. Developers can assess their system early in the design process, and determine whether (sometimes simple) interface or algorithm changes might support higher levels of self-explanation.

The Scorecard was composed so as to be agnostic with regard to the intended beneficiary of the explanations. The Levels can be directly applied to scale the explanations that are tailored and intended for use by individuals, or by one or another stakeholder group, such as developers. Explanations that are expressed formally or are intended to convey justifications can be evaluated in terms of the Levels, with respect to the depth of understanding the explanations convey for system developers.

8. Recommendations for the field of XAI

Clarity about the difference between explanation, interpretation, and justification seems achievable. But care is in order with regard to the differences between formal meanings of such terms as interpretability and understandability vs. their meanings in ordinary discourse (see Cabitza et al., 2023). Consistency can be achieved by disambiguating, e.g., not referring to explainability when the actual reference is to formal interpretability.

Out of our starting set of 165 reports on XAI system development, only 11 presented a discussion of methods and results from experiments involving human research participants and the empirical evaluation of the Human-XAI work system. This is a major shortcoming of the field. The mere demonstration of an XAI system, or mathematical proof of its computational capabilities, is not a substitute for empirical evaluation. The development of AI systems represents a significant investment, and assessment is necessary in order to realize the promise of that investment. Empirical evaluation of Human-AI work systems must adduce convincing empirical evidence that the work system is understandable and learnable; that the technology is usable and useful. Recent research have begun to address this matter of experimental adequacy and rigor (e.g., Hernandez-Orallo, 2017; Lage et al., 2019; Buçinca et al., 2020; Amarasinghe et al., 2022).

Reports on the development and evaluation of AI systems have one foot in computer science and one foot in psychology, hence the notion of empirical AI (Hoffman, 1992). The Levels of Explanation described in the Scorecard can be used as an independent variable in research that attempts to evaluate the effectiveness of explanations.

The evaluation of the performance of a human-XAI work system is essentially a large-scale psychological experiment. This brings with it the desirability of utilizing a reporting structure that is patterned after reports from the experimental psychology laboratory. Reports should have distinct and complete sections on Method, Procedure, and Results. This may seem obvious, but in practice it is important. For example, for many of the articles that were found in the literature search it was left unclear whether the presented examples of explanations were machine- or author-generated, or whether the intended beneficiary was the end user, or other developers, or general readers of the article. The method used in evaluation was often incomplete and scattered across the subsections of a report.

None of the reports in our focus set and none of the reports in the first phase of the DARPA XAI Program included full descriptions of the actual task instructions presented to the participants. Those might, even should, include some global explanation and not just be about the buttonology. It would be

assumed that as a part of a well-formed research method, some *How it works* explanatory information would always be presented in the initial instructions and training. Typically this form of explanation is presented as text, but may include example instances using other forms (e.g., diagrams, salience maps, etc.). It is uncertain whether *How it works* instructional materials have been well-formed across the broader field of in XAI.

9. Conclusions

Using the Levels of Explanation, we have evaluated the explanations generated by selected XAI systems with respect to their explanatory value to the user, that is, the degree to which they support the user's sensemaking and effort at self-explanation. Our results suggest that progress is being made. Achieving higher levels of explanation has been possible. Of the reports that underwent our analysis, 11 fell at a higher level (Levels 6 and 7), with five being dated 2020 or later. The black box might be opening up, but who is looking in? The field of XAI is still characterized, at least to some extent, by the development of "explanations" that are designer-centric. They rely on interfaces that are laden with alpha-numerics, data graphs, feature trees, color-coded matrices, all of which are supposed to illustrate the inner workings of Deep Nets or Machine Learning systems. These interfaces are cryptic, except to those who designed them. It not clear how these serve as explanations to users. Cognitive evaluation is one of a number of important opportunities fer advancing the field (see Liu et al., 2017). Do explanations across the Levels map onto user judgments of explanation goodness or explanation satisfaction? Do explanation across the Levels influence user judgments of trust in the AI? Do explanations at higher levels help users develop richer mental models of AI systems? It is possible to empirically evaluate the Scorecard levels using methods and metrics that have been tailored to XAI evaluation (see Hoffman et al., 2023). The work of Wang et al. (2022) (see above) included an evaluation by experts, indicating that the generation of explorable explanations was successful and the visual elements included with the spatial temporal feature selection and querying had explanatory value.

The down-selection procedure that was utilized revealed just those reports that included actual machine-generated explanations, published or posted up to the date of the analysis. More reports that do this are appearing in the literature, and thus present an opportunity to not only apply the Levels and to refine them as well. That may result from applications of the scheme in either the system development or system evaluation context. Feedback from developers who apply the Levels would also be useful, in affirming or disconfirming the value of the Scoresheet for developers, and also in refining the Scoresheet. Thus, use of the Scorecard might enable the XAI community to track progress in the field.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MJ and CT assisted in the construction of the scale and the analysis of the selected articles. GK and SM created the initial version of the scale. RH directed the effort and participated in all of the activities of scale design, construction, validation, and application. 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. This material is approved for public release. Distribution is unlimited. 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 thank the three reviewers for their helpful comments on the submission. The authors would like to acknowledge Lamia Alam of the Michigan Technological University, for assisting in the data analysis, and William J. Clancey of the Florida Institute for Human and Machine Cognition, for his contributions to the conceptual foundations of XAI. The authors also thank Timothy Cullen (Col., US Army, Ret.) for his consultation on this project.

Conflict of interest

GK is 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

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

Abdollahi, B., and Nasraoui, O. (2016). *Explainable Restricted Boltzmann Machines for Collaborative Filtering*. Available online at: https://arxiv.org/pdf/1606.07129.pdf (accessed April 24, 2023).

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence. *IEEE Access*. 6, 52138–60. doi: 10.1109/ACCESS.2018.2870052

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). "Sanity checks for saliency maps," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. p. 9525–9536. Available online at: https://papers.nips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html (accessed 2 February 2023).

Akula, A., Wang, S., and Zhu, S. -C. (2020). "CoCoX: Generating conceptual and counterfactual explanations via fault-Lines," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New York, NY: Association for Computing Machinery), 2594–2601. doi: 10.1609/aaai.v34i03.5643

Amarasinghe, K., Rodolfa, K. T., Jesus, S., Chen, V., Balayan, V., Saleiro, P., et al. (2022). *On the Importance of Application-Grounded Experimental Design for Evaluating Explainable AL methods*. Available online at: https://arxiv.org/pdf/2206.13503.pdf (accessed April 24, 2023).

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014). Power to the People: the role of humans in interactive machine learning. *AI Magazine*. 35, 105–120. doi: 10.1609/aimag.v35i4.2513

Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., et al. (2019). Explaining reinforcement learning to mere mortals: an empirical study. *arXiv*. doi: 10.24963/ijcai.2019/184

Anderson, L. W., Krathwhol, D. R., Airasoian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (2001). *A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. New York: Longmans.

Anguita-Ruiz, A., Segura-Delgado, A., Alcala, R., Aguilera, C. M., and Alcala, J. (2020). eX-plainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research. *PLoS Comput. Biol.* 16, e1007792. doi: 10.1371/journal.pcbi.1007792

Annett, J. (2002). Subjective rating scales: science or art? *Ergonomics*. 45, 966–987. doi: 10.1080/00140130210166951

Belle, V., and Papantonis, I. (2020). Principles and practice of explainable machine learning. *arXiv*. doi: 10.3389/fdata.2021.688969

Bereiter, C., and Scardamalia, M. (2005). "Beyond Bloom's Taxonomy: Rethinking Knowledge for the Knowledge Age," in *International Handbook of Educational Change: Fundamental Change,* Fullan, M. (ed.) Cham, Switzerland: Springer Nature. p. 5–22. doi: 10.1007/1-4020-4454-2_2

Bhatt, U., Andrus, M., Weller, A., and Xiang, A. (2020). *Machine Learning Explainability for External Stakeholders*. Available online at: https://arxiv.org/abs/2007.05408 (accessed April 24, 2023).

Bloom, B. S., and Krathwohl, D. R. (1956). "Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners," in *Handbook 1: Cognitive Domain*. New York: Longmans.

Buçinca, Z., Lin, P., Gajos, K. Z., and Glassman, E. L. (2020). "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems," in *Proceedings of the ACM International Conference on Intelligent User Interfaces*. New York: Association for Computing Machinery. p. 454–464.

Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Scheenberger, D., Stoeger, K., et al. (2023). "Quod erat demonstrandum? - Toward a typology of the concept of explanation for the design of explainable AI," in *Expert systems with Applications, 313, Part 118888*. Available online at: https://www.sciencedirect.com/science/article/pii/S0957417422019066 (accessed 29 January, 2023).

Calegari, R., Ciatto, G., Dellaluce, J., and Omicini, A. (2019). "Proceedings of the 20th Workshop "From Objects to Agents," *Volume 2404 of the CEUR Workshop Proceedings*. p. 105–112. Available online at: https://apice.unibo.it/xwiki/bin/view/Events/Woa2019 (accessed 1 February, 2019).

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

Carroll, J. M., and Aaronson, A. (1988). Learning by doing with simulated intelligent help. *Commun. ACM*. 31, 1064–1079. doi: 10.1145/48529.48531

Chari, S., Seneviratne, O., Gruen, D. M., Morgan, A., Das, A. K., and McGuiness, D. L. (2020). "Explanation ontology: a model of explanations for user-centered AI," in *The Semantic Web – ISWC 2020,* Pan, J. Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., et al. (eds.). Berlin, Germany: Springer International Publishing.

Cheng, F., Liu, D., Du, F., Lin, Y., Zytek, A., Li, H., et al. (2021). Vbridge: Connecting the dots between features and data to explain healthcare models. *IEEE Trans. Vis. Comput. Graph.* 28, 378–388. doi: 10.1109/TVCG.2021.3114836

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

Choo, J., and Liu, S. (2018). Visual analytics for explainable deep learning. *IEEE Comput. Graph. Appl.* 38, 84–92. doi: 10.1109/MCG.2018.04273166

Clancey, W. J., and Hoffman, R. R. (2022). Methods and standards for research on explainable artificial intelligence: lessons from Intelligent Tutoring Systems. *Applied AI Letters downloaded* 2 February 2023 doi: 10.1002/ail2.53

Corbin, J. M., and Strauss, A. (1990). Grounded theory research: procedures, canons, and evaluative criteria. *Qual. Sociol.* 13, 3–21. doi: 10.1007/BF00988593

Covert, I. C., Lundberg, S., and Lee, S.-I. (2021). Explaining by removing: a unified framework for model explanations. *J. Machine Learn. Res.* 22, 1–90.

Craik, F. I., and Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *J. Exp. Psychol: General.* 104, 268–294. doi: 10.1037/0096-3445.104.3.268

Craik, F. I. M., and Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *J. Verbal Learning Verbal Behav.* 11, 671–684. doi: 10.1016/S0022-5371(72)80001-X

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

Dahan, S. (2020). "Analytics and the EU courts: the case of trademark disputes," in *The changing European Union: A critical view on the role of the courts,* Capeta, T., and Lang, I. G. (eds.). Hart Publishing. Available online at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3786069 (accessed 2 February, 2023).

David, D., Resheff, Y. S., and Tron, T. (2012). *Explainable AI and Adoption of Algorithmic Advisors: An Experimental Study*. Available online at: *https://arxiv.org/abs/2101.02555* (accessed April 24, 2023).

Defense Advanced Research Projects Agency. (2018). *Explainable AI Program. DARPA-BAA-16-53, Dr. Matt Turek, Program Manager*. Arlington, VA: U.S. Defense Advanced Research Projects Agency. Available online at: https://www.darpa.mil/program/explainableartificial-intelligence (accessed April 24, 2023).

Dempster, E. R., and Kirby, N. F. (2018). Inter-rater agreement in assigning cognitive demand to life sciences examination questions. *Persp. Educ.* 36, 94–110. doi: 10.18820/2519593X/pie.v36i1.7

Dodge, J., Liao, Q. V., Zhang, Y., and Bellamy, R. K. E. (2019). "Explaining models: an empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th International Conference on Intelligent User Interfaces* p. 275–285. New York: Association for Computing Machinery.

Doshi-Velez, F., and Kim, B. (2017). *Towards a Rigorous Science of Interpretable Machine Learning*. Available online at: https://arxiv.org/abs/1702.08608 (accessed April 24, 2023).

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," in *Proceedings of the 24th International Conference on Intelligent User Interfaces* (New York, NY: Association for Computing Machinery), 263–274. doi: 10.1145/3301275.3302316

Felzmann, H., Villaronga, E. F., Lutz, C., and Tamo-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc.* 6, 1–14. doi: 10.177/2053951719860542

Gajos, K. Z., and Mamykina, L. (2022). *Do People Engage Cognitively With AI? Impact of AI Assistance on Incidental Learning*. Available online at: https://arxiv.org/abs/2202.05402 (accessed April 24, 2023).

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

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., and Gianotti, F. (2018). A survey of methods for explaining black box models. *ACM Computing Surv.* 51, 1–93. doi: 10.1145/3236009

Harbone, D., Willis, C., Tomsett, R., and Preece, A. (2018). "Integrating learning and reasoning services for explainable information fusion," in *Proceedings of the 1st International Conference on Pattern Recognition and Artificial Intelligence* (ICPRAI). Available online at: https://dais-legacy.org/doc-2675/ (accessed 2 February, 2018).

Hernandez-Orallo, J. (2017). Evaluation in artificial intelligence: From task-oriented to ability-oriented measurement. *Artif. Intell. Rev.* 48, 397–447. doi: 10.1007/s10462-016-9505-7

Hind, M., Wei, D., Campbell, M., Codella, N. C. F., Dhurandhar, A., Mojsilovic, A., et al. (2019). *TED: Teaching AI to Explain its Decisions*. Available online at: https://arxiv.org/abs/1811.04896 (accessed April 24, 2023).

Hoffman, R. H., Klein, G., Mueller, S. T., Jalaeian, M., and Tate, C. (2021). *The Stakeholder Playbook. Technical Report, DARPA Explainable AI Program*. Available online at: https://psyarxiv.com/9pqez/ (accessed 17 March, 2023).

Hoffman, R. R. (1992). *The Psychology of Expertise: Cognitive Research and Empirical AI*. Mahwah, NJ: Erlbaum. doi: 10.1007/978-1-4613-9733-5

Hoffman, R. R., Crandall, B., and Shadbolt, N. (1998). A case study in cognitive task analysis methodology: the Critical Decision Method for the elicitation of expert knowledge. *Hum. Factors* 40, 254–276. doi: 10.1518/001872098779480442

Hoffman, R. R., Mueller, S. T., 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

Hohman, F., Srinivasan, A., and Drucker, S. M. (2019). "TeleGam: Combining visualization and verbalization for interpretable machine learning," in *Presentation at the 2019 IEEE Visualization Conference* (New York, NY: IEEE), 151–155. doi: 10.1109/VISUAL.2019.8933695

Holzinger, A., Carrington, A., and Müller, H. (2020). Measuring the quality of explanations. *KI-Künstliche Intell.* 34, 193–198. doi: 10.1007/s13218-020-00636-z

Hughes, J., and Wood-Harper, T. (2000). An empirical model of the information systems development process: a case study of an automotive manufacturer. *Accounting Forum.* 24, 391–406.

Jesus, S., Belem, C., Balayan, V., Bento, J., Saliero, P., Bizarro, P., et al. (2021). "How can I choose an explainer? An application-grounded evaluation of post-hoc explanations.," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery.

Jingge, W. (1988). "Reasoning explanation capability of expert system a new framework for explanation," in *Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics*. p. 836–838. Available online at: https://ieeexplore.ieee.org/document/712821 (accessed 1 February, 2023).

Johs, A. J., Agosto, D. E., and Weber, R. O. (2020). Qualitative investigation in Explainable Artificial Intelligence: a bit more insight from social science. *arXiv*. doi: 10.22541/au.163284810.09140868/v1

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman, V. J. (2020). "Interpreting Interpretability: understanding Data Scientists' Use of Interpretability Tools for Machine Learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. p. 1–14. Available online at: http://www-personal.umich.edu/~harmank/Papers/CHI2020_Interpretability.pdf (accessed 2 February, 2020).

Kenny, E. M., and Keane, M. T. (2009). "On generating plausible counterfactuals and semi-factual explanations for deep learning," in *Proceedings of The Thirty-Fifth AAAI Conference on Artificial Intelligence* (AAAI-21) (Menlo Park, CA: Association for the Advancement of Artificial Intelligence), 11575–11585.

Kim, B., Khanna, R., and Koyejo, O. O. (2016). "Examples are not enough, learn to criticize! criticism for," in *Proceedings of the 30th Annual Conference on Neural Information Processing Systems, NIPS 2016* (New York, NY: Association for Computing Machinery), 2288–2296. Available online at: https://papers.nips.cc/paper_files/paper/2016 (accessed April 24, 2023).

Kim, J., Rohrbach, A., Akata, Z., Moon, S., Misu, T., Chen, Y. -T., et al. (2021). Toward explainable and advisable model for self-driving cars. *Applied AI Letters*. doi: 10.1002/ail2.56

Klein, G., Hoffman, R. R., and Mueller, S. T. (2020). "The Scorecard for Self-Explaining Capabilities of AI Systems," in *Technical Report from Task Area 2, DARPA XAI Program* Available online at https://www.ihmc.us/technical-reports-on-explainable-ai/ (accessed 2 February, 2023).

Klein, G., Hoffman, R. R., Mueller, S. T., and Newsome, E. (2021). Modeling the process by which people try to explain complex things to others. *J. Cognitive Eng. Decis. Making.* 15, 213–232. doi: 10.1177/15553434211045154

Klein, G., Rasmussen, L., Mei-Hua, L., Hoffman, R. R., and Case, J. (2014). Influencing preferences for different types of causal explanation for complex events. *Hum. Factors.* 56, 1380–1400. doi: 10.1177/0018720814530427

Koh, O. W., and Liang, P. (2017). "Understanding black-box predictions via influence functions," in *Proceedings of ICML 17: The 34th International Conference on Machine Learning, Volume 70* (New York, NY: Association for Computing Machinery), 1885–1894. doi: 10.5555/3305381.3305576

Krause, J., Perer, A., and Bertini, E. (2016). "Using visual analytics to interpret predictive machine learning models," in *Presentation at the International Conference on Machine Learning: Workshop on Human Interpretability in Machine Learning*. Available online at: https://arxiv.org/abs/1606.05685 (accessed April 24, 2023).

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., et al. (2019). "An evaluation of the human-interpretability of explanation," in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018)* (New York, NY: Association for Computing Machinery).

Liao, Q. V., Gruen, D., and Miller, S. (2020). "Questioning the AI: Informing design practices for explainable Ai user experiences," in *Proceedings of the ACM/CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery).

Lipton, P. (1990). Contrastive explanation. *R. Inst. Philos. Suppl.* 27, 247–266. doi: 10.1017/S1358246100005130

Liu, S., Wang, X., Liu, M., and Zhu, J. (2017). Towards better analysis of machine learning models: a visual analytics perspective. *Visual Infor.* 1, 48–56. doi: 10.1016/j.visinf.2017.01.006

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

Lundberg, S. M., and Lee, S. -I. (2017). "A unified approach to interpreting model predictions," in *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)* (New York, NY: Association for Computing Machinery), 4768–4777. doi: 10.5555/3295222.3295230

Marzano, R. J., and Kendall, J. S. (2007). *The New Taxonomy of Educational Objectives*. Thousand Oaks, CA: Corwin Press.

Miller, T. (2017). *Explanation in Artificial Intelligence: Insights From the Social Sciences*. Available online at: http://arxiv.org/abs/1706.07269 (accessed 30 January, 2023).

Miller, T. (2018). Contrastive explanation: a structural-model approach. *arXiv*.

Mitroff, I. I. (1974). *The Subjective Side of Science*. Amsterdam: Elsevier.

Mohseni, S., Zarel, N., and Ragan, D. E. (2020). A multidisciplinary survey and framework for design and evaluation of explainable AI Systems. *arXiv*.

Moore, D. S. (1982). Reconsidering bloom's taxonomy of educational objectives: cognitive domain. *Educ. Theory* 32, 29–34. doi: 10.1111/j.1741-5446.1982.tb00981.x

Muckler, F. A., and Seven, S. A. (1992). Selecting performance measures: "objective" versus "subjective" measurement. *Hum. Factors.* 34, 441–455. doi: 10.1177/001872089203400406

Mueller, S. T., Hoffman, R. R., Clancey, W., 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*. Pensacola, FL: Institute for Human and Machine Cognition.

Neerincx, M. A., van der Waa, J., Kaptein, F., and van Diggelen, J. (2018). "Using perceptual and cognitive explanations for enhanced human-agent team performance," in *Proceedings of the International Conference on Engineering Psychology and Cognitive Ergonomics*. Cham, Switzerland: Springer. p. 204–214.

Petsiuk, V., Das, A., and Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. Available online at: https://arxiv.org/abs/1806.07421 (accessed April 24, 2023).

Pierrard, P., Poli, J. P., and Hudelot, C. (2018). "Learning fuzzy relations and properties for explainable artificial intelligence," in *Proceedings of 2018 IEEE International Conference on Fuzzy Systems* (FUZZ-IEEE). New York: IEEE. p. 1–8.

Pierrard, R., Poli, J. P., and Hudelot, C. (2019). "A new approach for explainable multiple organ annotation with few data," in *Proceedings of the IJCAI 2019 Workshop on Explainable AI* (Somerset, NJ: International Joint Conferences on Artificial Intelligence), 101–107. Available online at: https://arxiv.org/abs/1912.12932 (accessed April 24, 2023).

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. (2021). "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI conference on Human Factors in Computing Systems* (New York: Association for Computing Machinery), 1–52. Available online at: https://arxiv.org/abs/1802.07810 (accessed April 24, 2023).

Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018). *Stakeholders in Explainable AI*. Available online at: https:arxiv.org/abs/1810.00184v1 (accessed 2 February, 2023).

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

Russell, C. (2019). *Efficient Search for Diverse Coherent Explanations*. Available online at: arxiv.org/abs/1901.04909 (accessed 2 February, 2023).

Samek, W., Wiegand, T., and Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *Int. Telecommun. J. ICT Discover.* 1, 39–48. Available online at: https://www.itu.int/dms_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-P05-PDF-E.pdf (accessed April 24, 2023).

Sheh, R., and Monteath, I. (2018). Defining explainable AI for requirements analysis. *KI - Künstliche Intelligenz.* 32, 61–66. doi: 10.1007/s13218-018-0559-3

Singh, S., Ribeiro, M. T., and Guestrin, C. (2016). "Programs as black-box explanations," in *Presentation at the Conference on Neural Information Processing Systems, 1st Workshop on Neural Abstract Machines & Program Induction (NAMPI)* (New York, NY: Association for Computing Machinery). Available online at: https://arxiv.org/abs/1611.07579 (accessed April 24, 2023).

Stefik, M., Youngblood, G. M., Pirolli, P, Lebiere, C., Thomson, R., Price, R., et al. (2021). Explaining Autonomous Drones: An XAI Journey. *Appl. AI Lett.* 2, e54. doi: 10.1002/ail2.15

Sugrue, B. (2002). *Problems with Bloom's Taxonomy*. Performance Express. Available online at: https://eppicinc.files.wordpress.com/2011/08/sugrue_bloom_critique_perfxprs.pdf (accessed April 24, 2023).

Tjoa, E., and Guan, C. (2020). "A survey on explainable Artificial Intelligence (XAI): Toward medical XAI," in *IEEE Transactions on Neural Networks and Learning Systems*. Available online at: https://www.researchgate.net/publication/346017792_A_

Survey_on_Explainable_Artificial_Intelligence_XAI_Toward_Medical_XAI (accessed 2 February, 2023).

Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018). "Interpretable to whom? A role-based model for analyzing interpretable machine learning systems," in *Proceedings of the 2018 ICML workshop on Human Interpretability in Machine Learning (WHI 2018)*. Stockholm, Sweden.

Troussas, C., Krouska, A., and Virvou, M. (2017). "Integrating an Adjusted Conversational Agent into a Mobile-Assisted Language Learning Application," in *Proceedings of the IEEE 29th International Conference on Tools with Artificial Intelligence*. p. 1153–1157. Available online at: https://researchr.org/publication/ictai-2017 (accessed 1 February, 2023).

Troussas, C., Krouska, A., and Virvou, M. (2019). MACE: mobile artificial conversational entity for adapting domain knowledge and generating personalized advice. *Int. J Artif. Intell. Tools*. 28, 1940005. doi: 10.1142/S0218213019400050

Turner, R. (2016). A model explanation system: Latest updates and extensions. *arXiv*. doi: 10.1109/MLSP.2016.7738872

Vasu, B. Hu, B., Dong, B., Collins, R., and Hoogs, A. (2021). Explainable, interactive content-based image retrieval. *Appl. AI Lett.* 2, e41. doi: 10.1002/ail2.41

Vilone, G., and Longo, L. (2020). *Explainable Artificial Intelligence: A Systematic Review*. doi: 10.48550/arXiv.2006.00093

Vong, W. K., Sojitra, R. B., Reyes, A., Yang, S. C., and Shafto, P. (2018). "Bayesian teaching of image categories," in *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. p. 2677–2632. Available online at: http://scottchenghsinyang.com/paper/Vong-2018.pdf (accessed 31 March, 2023).

Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counter-factual explanations without opening the black box: automated decisions and the GDPR. *Harvard J Law Technol.* 31, 841–887. doi: 10.2139/ssrn.3063289

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. p. 1–15. New York: Association for Computing Machinery.

Wang, J., Li, Y., Zhou, Z., Wang, C., Hou, Y., Zhang, L., et al. (2022). "When, where and how does it fail? A Spatial-temporal visual analytics approach for interpretable object detection in autonomous driving," in *IEEE Transactions on Visualization and Computer Graphics*. Available online at: https://pubmed.ncbi.nlm.nih.gov/36040948 (accessed 7 March, 2023).

White, A., and Garcez, A.d'A. (2021). *Counterfactual Instances Explain Little*. Available online at: https://arxiv.org/abs/2109.09809 (accessed April 24, 2023).

Yardley, S., Mattick, K., and Dornbaum, T. (2020). "Close-To-Practice: Qualitative research methods," in *The Oxford Handbook of Expertise*, eds P. Ward, J. M. Schraagen, J. Ore, and E. Roth (Oxford: Oxford University Press), 409–428. doi: 10.1093/oxfordhb/9780198795872.013.18

Yeh, C. -K., Heish, C. -Y., Suggala, A. S., Inoyue, D. I., and Ravikumar, P. (2019). "On the (in)fidelity and sensitivity of explanations," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (New York, NY: Association for Computing Machinery), 10935–10946.