KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-023-00806-9

TECHNICAL CONTRIBUTION

Increasing the Value of XAI for Users: A Psychological Perspective

Robert R. Hoffman[1] · **Timothy Miller**[2] · **Gary Klein**[3] · **Shane T. Mueller**[4] · **William J. Clancey**[1]

Received: 17 January 2023 / Accepted: 15 May 2023

Abstract
This paper summarizes the psychological insights and related design challenges that have emerged in the field of Explainable AI (XAI). This summary is organized as a set of principles, some of which have recently been instantiated in XAI research. The primary aspects of implementation to which the principles refer are the design and evaluation stages of XAI system development, that is, principles concerning the design of explanations and the design of experiments for evaluating the performance of XAI systems. The principles can serve as guidance, to ensure that AI systems are human-centered and effectively assist people in solving difficult problems.

Keywords Explanation · Explainable AI · Psychological principles · Self-explanation · Empirical evaluation

1 Introduction

The general objective of Explainable AI (XAI) is to develop methods that enable people to achieve a satisfying and useful understanding of an AI system's capabilities and vulnerabilities [see 70]. This objective pertains to stakeholders such as policy makers, system integrators, and trainers, but it pertains especially to end-users. Understanding how to use an AI system makes it possible for people to act appropriately, including cross-checking and complementing the automation to accomplish the intended function within a broader established activity [17]. The provision of a machine-generated explanation has been proposed as a way to assist people in gaining this expertise. When are explanations needed? What are the best ways to explain AI systems? How should the performance of XAI systems be evaluated?

1.1 Background

The literature on explanation and explanatory reasoning is vast, if not overwhelming. It spans centuries of scholarship in philosophy and research in psychology. Explanatory reasoning has clear linkages to the equally vast literatures on causation and causal reasoning. Key concepts of XAI were manifest in research spanning the 1970s-80s on Intelligent Tutoring Systems [16–18], on intelligent assistants [10, 11, 30] and on expert systems [16, 65]. The research revealed challenges for explanation systems, which will be familiar to developers of XAI systems:

- The need to provide both global or architectural explanations (*How it works)* and local or procedural explanations (*Why it did that*),
- The need to explain the reason or rationale for procedures and explanations,
- The need to match explanations to user goals, including their changing goals,

✉ Robert R. Hoffman
rhoffman@ihmc.us

Timothy Miller
tmiller@unimelb.edu.au

Gary Klein
gary@macrocognition.com

Shane T. Mueller
shanem@mtu.edu

William J. Clancey
wclancey@ihmc.us

1 Institute for Human and Machine Cognition, 40 S. Alcaniz St, Pensacola, FL, USA

2 Department of Philosophy, University of Melbourne, Parkvillle, VIC 3010, Australia

3 MacroCognition, LLC, 5335 Far Hills Ave., Suite #21, Dayton, OH, USA

4 Department of Cognitive and Learning Science, Michigan Technological University, Houghton, MI, USA

- The need to settle the matter of "fragile credibility" or trust and reliance (e.g., users often prefer to pursue their own goals rather than follow the given advice),
- The need to enable users to actively test the explanations and explore the boundary conditions of the AI,
- The need to fully integrate a help system with the application,
- The need to be able to deal with user misinterpretations of explanations, via a dialog,
- The need to deal with "tangled errors" when one mistake leads to another and the situation becomes uncorrectable,
- The need to reveal usefulness and usability challenges (e.g., help systems can trigger annoyance rather than help users).

The topic of XAI has ramped up quickly. Researchers have created Machine Learning systems for object recognition that "explain" their categorization of objects or images using saliency or "heat" maps showing what the AI system "is looking at." Researchers have created software agent systems that explain their choice of courses of action using such displays as histograms or matrices of probabilities [see 30]. There has been a wave of review articles and a wave of attempts to taxonomize machine-generated explanations (qualities, formats, data types, purposes, etc.) [7, 12, 62, 64, 72, 79].

Many issues have emerged in this research. For instance, there have been lamentations about the ambiguity of key concepts including transparency, interpretability and explainability [2, 6, 8, 15, 46, 47, 53, 61]. Interpretability tools are advertised as exposing machine learning interpretability algorithms, and thereby explain the output of machine learning systems. But the "explanations" that these systems provide are code-intensive representations of the results from statistical or game-theoretic modeling. They serve more to justify system architectures to other computer scientists [1] than to explain AI systems in ways that make sense to people generally [57]. Recent research has begun to address the issue of experimental adequacy and rigor [4, 9, 37, 40, 50, 54, 67, 73].

A program established in 2018 by the US Defense Advanced Research Projects agency was a major impetus for the field (although the concept of XAI pre-dates it [see 69]). What has been accomplished in the field of XAI most recently, what assumptions have been revealed, and what remains to be done in order to ensure that AI systems are "human-centered"? XAI research has identified some principles that should be appreciated [12, 64]. These represent the findings of a broad range of research and scholarship at the nexus of AI, automation, and human-machine systems [31, 60]. We highlight the ways in which these principles have been invoked and in some cases have been integrated into AI research.

1.2 Organization

This article is organized as a set of twelve principles, some of which have recently been instantiated in XAI systems. The purpose of the principles is to express guidance for XAI system developers, not to solve immediate problems of the design of explanations, displays or experimental evaluations. The principles might be thought of as cautionary tales, or more strongly as necessary considerations. Details concerning the design of explanations and evaluations, and that are in accordance with the Principles, are presented in [42, 47, 66].

The Principles fall into these groupings:

- Cognition as a determiner of what it means to explain,
- Orchestration of the process of explaining,
- Designing explanations,
- Designing empirical evaluations.

2 The Cognitive Perspective: What does it Mean to Explain?

2.1 Principle 1: All Explanation Involves Self-Explanation

This principle is about how XAI system developers should think about the cognitive process of explanation. The principle might be deemed the Golden Rule of XAI systems: *Explain unto others in such a way as to help them explain to themselves.* From the moment the user receives their first instruction on how to use an AI system, there is an active process of self-explanation. It may be brief and superficial, resulting in knowledge that is fragmentary and inconsistent [22] or the self-explanation process may be drawn-out and deliberative. Sometimes users/learners simply do not care "how it works" and just want to get on with their task or job. But self-explanation is often a highly motivated desire to understand. This assertion derives from psychological experimentation [13] and empirical research on how people explain complex systems to other people [51, 52].

Explanations can help learners form and refine their mental model of the AI and the task. But the achievement of an understanding is not just the ingestion of information. Self-explaining involves changing preconceptions by both assimilation and accommodation, to use Piagetian terminology. Thus, it can be useful for the learner to see what happens when the AI system fails, or is at its boundary conditions.

Failure cases can be a signal of a model mismatch [38], especially if the AI system fails in ways that humans would never fail (e.g., confusing a turtle with a rifle) [14, 41].

The self-explanation process can be seen in research that has used cognitive interviews to assess explanatory efficacy [5]. A number of researchers have noted that surprising events or violations of expectations trigger a need for an explanation. This has been considered in psychology, especially in research on curiosity [58]. A model of self-explanation is emerging in XAI based on the psychological investigation of explanatory reasoning [52] and XAI research on levels of intelligibility [56].

This first principle entails a reconsideration of the basic conceptual model that motivated some early XAI research, including the DARPA Explainable AI Program, which was a major impetus for the field in 2018–2021. This model is depicted in Fig. 1. This model approached the topic from a programmatic, computer science point of view. The explanation is generated and then is delivered, (ideally) to good effect. The model highlights the things that would have to be measured in an evaluation of an XAI system (the shaded nodes in Fig. 1):

(1) Is the explanation good in the sense that it accords with the criteria that have been espoused in the literatures on explanatory reasoning (succinct, understandable, etc.)?
(2) Is the explanation satisfying in the judgment of the users?
(3) Do the users really understand the explanation?
(4) Does performance improve because of the explanation?

This model expressed the premise that explanation of how the AI system works is presented early, at the point where the user is being instructed on the task and the tool. A further assumption was that subsequent human-computer interaction would involve the generation and presentation of specific machine decisions [62]. These assumptions had a positive impact in that they highlighted the distinction between local (*Why did it decide that?*) and global (*How does it work?*) explanations, a distinction that many XAI researchers embraced [28, 77].

In parallel with the computer science work, empirical analysis was conducted of a large corpus of cases where a person was the recipient of an explanation of how some sort of complex system works [52]. Many of the cases illustrated how global explanations often include local information (e.g., instances that exemplify rules or principles), and local explanations often include some global information (e.g., a generalization over a class of instances). Thus, in actual explanation events, the global-local distinction gets blurry. A second revelation from the empirical research was that the "spoon feeding" paradigm expressed in Fig. 1 is blind to the fact that users engage in a motivated, deliberative attempt to make sense of the AI system and any explanatory material that may be presented. This is diagrammed in Fig. 2. In this psychological model, explanatory systems benefit by providing information that empowers users to self-explain, rather than just delivering some sort of representation of the output of an algorithm, a representation that is believed to be adequate as an explanation.

The Fig. 1 model is Programmatic, that is, it was a description of an approach to evaluation methodology (i.e., what to measure and when to measure it). The emphasis

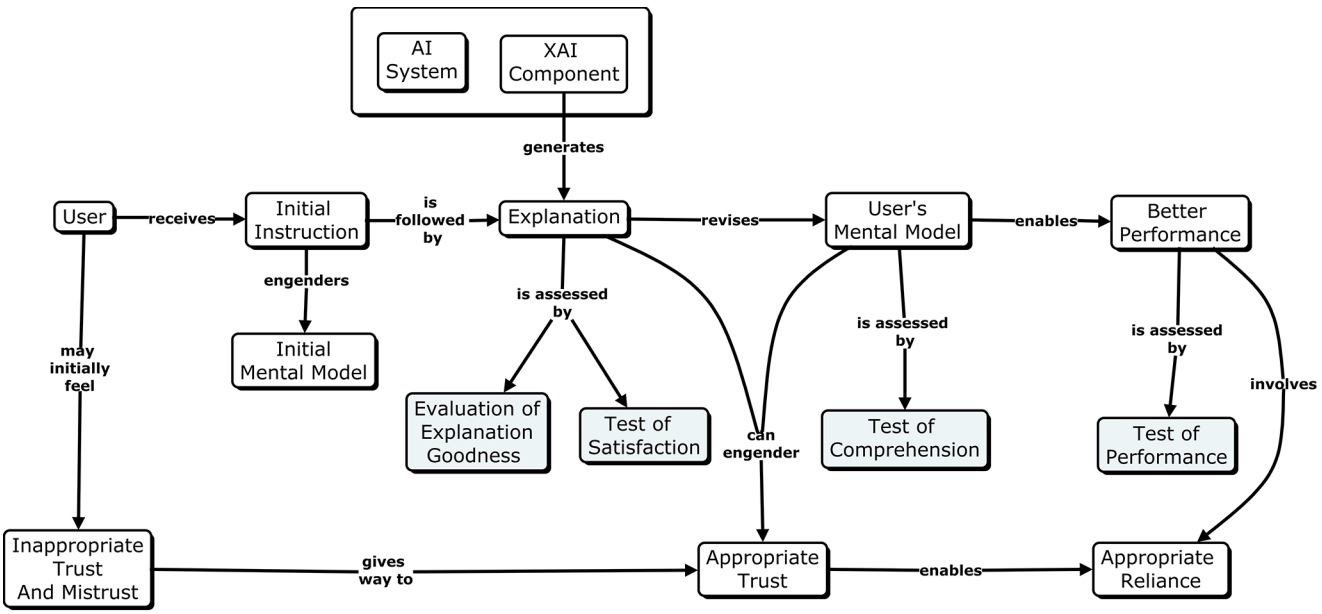

Fig. 1 2018 model of the XAI process, from the perspective of computer science

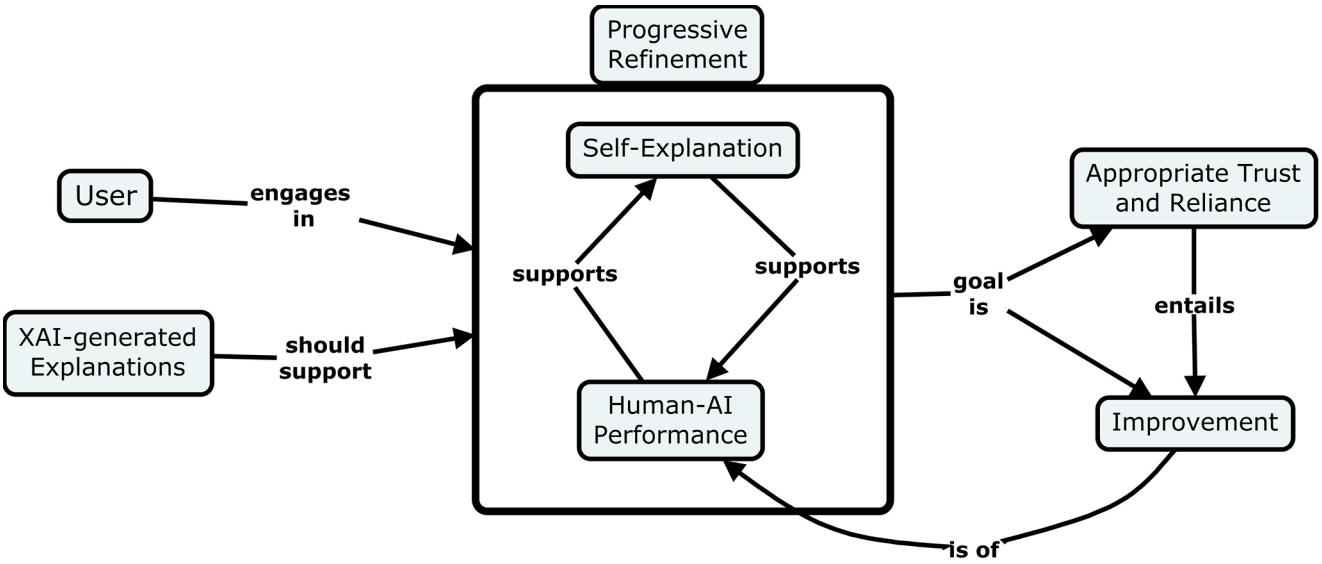

Fig. 2 Psychological Model of the XAI process

was on the performance of the user, given that the user benefits from the machine-generated explanations. As such, the model made limiting assumptions about the psychological aspects of explanation. While the measurement nodes of Fig. 1 might be ported into the Fig. 2 model, the Fig. 2 psychological model refers to the performance of the human-XAI system. Additionally, the model situates performance in a loop with the process of self-explanation, which can be thought of as the user's refinement of their mental model.

2.2 Principle No. 2: Explanation is an Exploratory Activity

Following from the notion of progressive refinement that is illustrated in Fig. 2, Principle 2 is also about how XAI system developers should think about the process of explaining. Most taxonomies of types of explanations or features of explanations are predicated on the view that the property of being an explanation is a property of text, visualizations, etc. The mantram *Explanation is Exploration* is another lesson from the XAI research. It harkens back to John Dewey's theory of inquiry [70]. Explanations should not force learners to adopt the AI's model, a model that is not their own, and is arguably not human, and can only be assimilated with difficulty. Explanations can actually impair learning if they do not encourage the learner to discover patterns. Explanations should enable people to diagnose situations, predict or anticipate the future and justify decisions or actions. Explanations should enable people to notice apparent inconsistencies, and the XAI system should enable the user to restore consistency between their mental model and empirical reality. Thus, the explanation process must involve the exchange of meaningful (and not just computationally formal) information. XAI researchers have recognized the value of narrative explanations that describe causal relations, and the value of narrative explanations [31].

As was discovered in the research on Intelligent Tutoring Systems (ITS), an XAI system's explanation capabilities need to be designed on the basis of a pedagogical model—an instantiated model of the instructional process of structured interaction, which in turn is based on a cognitive model of learning [18]. XAI systems might attempt to identify the user's current understanding (so that it can better predict how to transform this knowledge), and support the information that will help make these transformations. This may be asking a lot of XAI system developers who might not think of XAI as a form of ITS. But for XAI to work as intended, it needs to accord with the key findings from ITS research [see 18]. Recent research has demonstrated that XAI systems can support the user in a process of active exploration [37].

2.3 Principle 3: Explanation Relates to Trust, but Trust must be Considered a Dynamic Process, not a State

This Principle is about how XAI system developers should conceive of the relation between explanation and trust/reliance in an AI system. The notional model of the XAI process (Fig. 1, above) assumed that good explanation will result in trust. But trust is not a single state that develops to some level of calibration. In work contexts in which humans rely on computational technologies, people always feel some mixture of justified and unjustified trust *and* justified and unjustified mistrust. These attitudes are in constant flux and rarely develop in a progression to some ideal and stable point [4, 35, 39, 55]. Trust can come and go in a flash. When

the AI fails in a way that a human would never fail, reliance can collapse.

3 Orchestrating of the Process of Explaining

3.1 Principle 4: Exploration is Interaction; it is never a "One-Off"

This Principle is an extension of Principle 2. For many XAI systems, explaining has to be thought of as a process that extends over multiple uses and interactions. Explaining could extend over multiple sessions if the user wants to check back on a previous case. Or, the user might want to test the coverage of the AI by presenting it various counterfactual test cases.

Explanation cannot be a one-off if only because the AI system is a machine learning system which itself can continually evolve. Especially for AI systems that learn or are applied in dynamic contexts, users often need repeated explanations and re-explanations. How has the algorithm changed? How are AI system's decisions affected by the new data, and if the new data themselves are not valid, is the AI decision invalid? XAI systems might benefit from considering the long-term interaction with users, even in simple ways like recognizing that once learned, an explanation may not need to be given again unless something important has changed. Such design decisions are contextual (see Principle 7) and specific to individuals (see Principle 10).

As was said above, XAI research initially assumed that the property of being an explanation is a property of text, images, etc. XAI research initially assumed that explanation only involves providing an explanation to the user, on the assumption that the explanation is good and sufficient. XAI researchers are escaping these assumptions. Explaining depends on the ability of the human and the machine to interact. Explanation as collaboration is another prime lesson from XAI research [25, 32, 64]. Some XAI systems have come to regard explaining as a two-way process in which the XAI program provides explanatory information, and the user advises the program in one way or another [28, 44, 49, 75, 82]. Thus, explanations allowing for user input might lead to improvements of the computer model. More work has to be done to extend this basic idea. Explanations work better when people can interact with them.

Table 1 "Triggers" of the need for an explanation, mapped onto the user's goals

Triggers	User/Learner's Goal	
	Need to Understand	*Need to Accomplish*
How do I use it?	Use	Primary Task Goals
How does it work?	Mechanism, Understandability	Feeling of Satisfaction
How will it help me do a better job?	Usefulness	Primary Task Goals
What did it just do?	Mechanism, Understandability	Feeling of Sufficiency
What does it achieve?	Function	Usefulness
What will it do next?	Observability	Trust, Reliance
How much effort will this take?	Usability	Primary Task Goals
What can't it do? What are its limitations?	Function	Trust, Reliance
What do I do if it gets it wrong?	Anticipation	Control
How do I avoid the failure modes?	Surprise	Adaptation, Work-around
What would it have done if x was different?	Mechanism, Understandability	Curiosity, Trust, Reliance
Why didn't it do z?	Anticipation, Surprise	Curiosity, Trust, Reliance
What do I do if I mistrust or distrust the tool?	Observability	Trust, Reliance
How can I repair the tool or help it work better?	Usefulness	Primary Task Goals, Feeling of Satisfaction, Adaptation

3.2 Principle 5: Explanations are most Needed when there is a "Trigger."

Explanations are not needed all the time, but in some early XAI research it was assumed that explanations should always be presented. The displays created for some early XAI always included a field showing "reasons" why the AI system made its determination (e.g., a list of key features, or a "saliency map" for classifier systems). It remains unclear as to whether the persistence of an explanation display detracts from performance by virtue of the dedication of display real estate to information that is not always needed. This is often because users have seen something interesting or surprising or they have some sort of goal that they strain to achieve. Some triggers have been mentioned above, but Table 1 lists those that have been referenced in work on XAI systems [see 62]. Furthermore, the triggers are manifest in interviews with stakeholders about their explanation requirements [38]. That said, not all XAI researchers may see it as the goal of their XAI systems to provide explanatory answers to some of these trigger questions.

Advances in AI (and XAI) will come when systems begin to understand and anticipate situations that are likely to engender surprise and violate user expectations [69].

3.3 Principle 6: Explanation Occurs in a Context, Carrying with it the Needs of the User

Some people prefer explanations that refer to single, necessary, or "focal" causes and that are both simple and yet broadly applicable. Some people will believe explanations to be good even when they contain flaws or gaps in reasoning. Some people are not satisfied with simple, superficial explanations. Some people are more deliberative and reflective in their explanatory reasoning. Some people are eager to learn how the system works. Expert users can prefer less detail, presumably because they have enough knowledge about causal relationships in the domain that they can link together the focal causes. Non-expert users may need more detailed information for their sensemaking.

Explanations can be desired for a number of reasons, that depend on the goals of the user [36, 75]. This includes the need to self-explain, discussed above, but also pragmatic, ask-relevant goals: *How will this AI system improve my performance? How much effort does it take to use this AI system? How can I trust it? How do I recover from the AI system's mistakes? How will I know when there has been an anomaly or a failure? How can I repair the AI system?* For want of satisfactory answers to these sorts of questions, the user might resort to trying a workaround.

3.4 Principle 7: People do not Necessarily Engage with Explainability Tools

It should not be taken for granted that people will engage with XAI. As outlined in Fig. 1 above, there was an assumption of the sequence: 'make decision' → 'explain decision' → 'understand' that people will use information from XAI tools to understand the AI's decisions. However, recent research suggests that this assumption should be questioned. For example, in a study involving judgments about nutrition [27], explanations were about the AI system's recommendations, but hid the actual recommendations themselves. The research showed that this resulted in better decisions and better incidental learning on their participants compared to giving just recommendations, and compared to giving both recommendations and explanations. The conclusion was that withholding decisions leads to higher cognitive engagement and therefore better decisions.

4 The Design of Explanations

4.1 Principle 8: Explanation Occurs in a Context, Carrying with it the Goals of the Work System

Explanation is an interaction among the user, the XAI system, and their activity in a task context. Different explanations support different information needs for different tasks within a broader set of work goals. Explanation needs of the user are also related to role of the user within that broader work context [38]. Stakeholders sometimes need access to others (e.g., trusted engineers, trusted vendors) in order for them to be able to develop satisfying mental models of AI systems. Trainers need to know how the AI system fails and how it misleads as much as they need to know how it works. Some stakeholders need to develop an understanding that enables them to explain the AI to someone else and not just satisfy their own sensemaking requirements.

4.2 Principle 9: Explanations Tend to Work Better if they Include Demonstrations of Differences and Contrasts

This Principle is about the design of machine-generated explanations. A central lesson of XAI research is the utility of contrastive and counterfactual explanations in understanding the boundary conditions of a system [5, 25, 28, 60]. XAI researchers have begun to consider the importance of interactive explanations and counterfactual or contrastive explanations. These goals can reflect curiosity, or a need to know: *What did the AI system just do? Why didn't the AI system do z? What would the AI system have done if x had been different?* Explanation of "why something is what it is" entails an explanation of "why it is not something else." In other words, explanation and contrastive reasoning are co-implicative.

The importance of counterfactuals to explainability and to the automatic generation of explanations has been noted by XAI researchers [3, 19, 67, 79] (see [48] for a review). However, contrasts and counterfactuals might only be useful if the user already has in mind a mental model of typical behavior. That is, the user needs to understand what "right" is before they can understand what "not-right" is or what "wrong" is. Pedagogical use of examples requires understanding a failure in order to produce benefit—otherwise examples may end up simply frustrating the user and fostering distrust because of the salience of failures.

It has been argued that counterfactuals lack explanatory value because they do not provide a causal model [80]. This claim hinges on the assumption 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 exploration is possible, and they show how exploration can be conducted. This is powerful, as it supports exploration of situations when the AI is operating at the boundaries of its competence envelope—instances that fall inside a class but nearly do not, and instances that do not fall outside the class but nearly do. Such cases show when a small change to a case makes a difference to the categorization.

5 The Empirical Evaluation of XAI Systems

5.1 Principle 10: The Evaluation of Human-XAI Task Performance should Rely on Ecologically Valid Tasks

It might be taken for granted that XAI systems need to support task-essential capabilities that people cannot exercise well, or at all, without computational aids [71, 75]. Recent XAI research has shown that an XAI system is not interesting for people to use, or try to understand, if the user is always more skilled than the AI [74, 75, 77]. The task that is set before a human-AI work system has to be one that presents a genuine challenge. But the evaluation of XAI systems often involves human judgment of the acceptability or usefulness of algorithms. Such bench-testing may provide useful early guidance, but the benefits may not occur in situ. By implication from Principle 8, XAI evaluation should maintain the context of actual work. It is inadvisable to develop an explanation system by just looking at a convenient proxy task (e.g., bird classification) [9]. While proxy tasks may arguably be ecologically relevant, explanations have to relate the AI tool to the context of actual tasks and goals tasks [4, 9, 42].

Most real-world applications of AI systems involve work situations bearing multiple tasks and multiple goals. Explanations have to relate the AI tool to the user's knowledge in the context of those goals and tasks. Much XAI research has assumed a one-person, one-task problem situation. Failure to build ecological relevance into an XAI system makes research easy because it sidesteps the complexity of actual work settings, in which a group of people using an AI tool may have different roles and tasks, with conflicting constraints to satisfy. Additionally, the ecologically sparse settings and tasks used in XAI system design and evaluation assume that the user of the AI will only be interacting with one program. In most real-world settings, workers juggle multiple computational systems. Furthermore, the user and the AI system act interdependently [43]. Thus, the evaluation of the performance of an XAI system is to be thought of as an evaluation of the interactive performance of the user and the XAI system [64].

This said, some XAI research has focused on "real world" challenge cases, such as image analysis and activity recognition [64], air platform identification in aerial photographs [28], self-driving automobiles [49], dealing with cyberattacks and fake news [79, 82], controlling search and rescue drones [75], human-machine teaming [24, 25], and radiological pneumothorax diagnosis [70].

5.2 Principle 11: The Evaluation of Human-XAI Task Performance should Rely on Representative Users and Domain Experts

Also by implication from Principle 8, XAI evaluation should involve representative users as research participants. Work-centered design principles suggest involving representative intended beneficiaries of the AI system early and throughout the system development process [17, 20, 21, 29, 36]. Early evaluation of explainability interventions may rely upon non-experts on simplified tasks, but the findings may not transfer to the real context. Furthermore, a work-centered approach necessarily involves domain experts in the system design and development activity, from the beginning and throughout the development process—not just a one-off at the initial design stage or some sort of culminating evaluation. Recent XAI research attests to this [27, 75, 78].

5.3 Principle 12: Research Evaluating XAI Systems Needs to "Get Inside the Heads" of the Users

This Principle is about methodology in the evaluation of XAI systems. XAI researchers have come to recognize the value of cognitive task analysis to reveal learners' mental models [23, 24, 28, 33, 45, 73, 75, 78]. Valuable findings have come from post-experimental structured interviews, asking the participants such questions as *What problems did you find?* and *Why do you think the AI system did that?* [5, 59]. Results are the most informative about how machine-generated explanations influence user reasoning, that is, the interview data help the developers make sense of the performance data. Presentations on XAI projects that include quotations from users in post-experimental inquiry demonstrate the kind of awareness that XAI evaluation research requires—awareness of the user's reasoning as they actively try to understand the AI system.

6 Discussion

The Principles are an attempt to capture the key ideas and challenges that have recently emerged, and in particular in the work that was conducted with the support of the DARPA XAI Program. Of course, more research, development and evaluation activity needs to be done to deepen and extend the application of the Principles.

The Principles speak to the value of cross-disciplinary synthesis and collaboration. The evaluation of an XAI system, or any AI system, is essentially a large-scale psychological experiment. Rigor is necessary to justify the investment of time and effort required to develop AI systems. The development and deployment of useful and useable XAI systems requires:

- A multi-disciplinary collaboration (philosophy, psychology, and education as well as computer science);
- A design process that embeds the human-XAI system in a realistic and relevant task context;
- A view that the good XAI system needs to benefit the machine as well as the human.

The Principles and their applications in XAI express the perspectives of experimental psychology and empirical philosophy. XAI research has certainly also resulted in insights from the perspective of computer science. For example, there has been a recognition of the idea that feedback from the user might be used to improve the computational model, that is, the benefits of explanatory reasoning work in both directions [37]. User feedback to an XAI system can improve the Machine Learning model that needs to be explained/understood. As XAI research matures there will no doubt be additional achievements and insights.

7 Acknowledgement and Disclaimer

The Authors would like to thank Timothy Cullen (Col., US Army, Ret.) for his consultation on this project.

The Authors would like to thank the anonymous Reviewers for their many helpful comments on the submission.

This material is approved for public release. Distribution is unlimited. This material is based on research sponsored by the Air Force Research Lab (AFRL) under agreement number FA8650-17-2-7711 and by Australian Research Council (ARC) Discovery Grant DP190103414: *Explanation in Artificial Intelligence: A Human-Centered Approach.* The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. 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

1. Abdollahi B, Nasraoui O (2016) Explainable restricted Boltzmann machines for collaborative filtering. [arXiv:1606.07129v1]
2. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence. IEEE Access 6:52138–52160 [https://doi.org/10.1109/ACCESS.2018.2870052]
3. Akula A, Wang S, Zhu S-C (2020) CoCoX: Generating conceptual and counterfactual Explanations via Fault-Lines. Proc AAAI Conf Artif Intell *34*(3):2594–2601
4. Amarasinghe K, Rodolfa KT, Jesus S, Chen V, Balayan V, Saleiro P, Bizarro P, Talwalkar A, Ghani R (2022) On the importance of application-grounded experimental design for evaluating explainable ML methods. [downloaded 29 January 2023 from arXiv:2206.13503].
5. Anderson A, Dodge J, Sadarangani A, Juozapaitis Z, Newman E, Irvine J, Chattopadhyay S, Fern A, Burnett M (2020) Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems* (TiiS). [https://doi.org/10.1145/3366485]
6. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik M, Barbado A, Garcia S, Gil-Lopez, Molina D (2020) Explainable artificial itelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inform Fusion *58*:82–115
7. Arya V, Bellamy RKE, and 18 others (2019) One explanation does not fit all: a toolkittaxonomy of AI explainability techniques. [arXiv:1909.03012v2]
8. Bojarski M, Yeres P, Choromanska A, Choromanski K, Firner B, Jackel LD, Muller U (2017) Explaining how a deep neural network trained with end-to-end learning steers a car. [arXiv:1704.07911]
9. Buçinca Z, Lin P, Gajos ZJ, Glassman EL (2020) Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. Association for Computing Machinery, New York, NY. [downloaded 29 March 2023 at [https://doi.org/10.1145/3377325.3377498]]
10. Carroll JM, Aaronson P (1988) Learning by doing with simulated intelligent help. Commun Assoc Comput Mach *31*(9):1046–1079
11. Carroll JMN, McKendree J (1987) Interface design issues for advice-giving expert systems. Commun Assoc Comput Mach *30*(1):14–31
12. Chari S, Gruen DM, Seneviratne O, McGuiness DL (2020) Foundations of knowledge-enabled systems [downloaded 29 March 202 at arXiv:2003.07520v1]

13. Chi MTH, Van Lehn KA (1991) The content of physics self-explanations. J Learn Sci *1*(1):69–105
14. Choi CQ (2021) 7 revealing ways AIs fail: neural networks can be disastrously brittle, forgetful, and surprisingly bad at math. IEEE Spectr *58*(10):42–47 [https://doi.org/10.1109/MSPEC.2021.9563958]
15. Chromik M, Schuessler M (2020) A taxonomy for human subject evaluation of black-box explanations in XAI. In *Proceedings of the IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies* (ExSS-ATEC'20) [arXiv:2011.07130v2]
16. Clancey WJ (1986) From GUIDON to NEOMYCIN and HERACLES in twenty short lessons: ONR Final Report 1979–1985. The AI Magazine *7*(3):40–60
17. Clancey WJ (2020) *Designing agents for people: Case studies of the Brahms Work Practice Simulation Framework* Kindle Print Replica e-Book. [https://www.researchgate.net/publication/343224286_Designing_Agents_for_People_Case_Studies_of_the_Brahms_Work_Practice_Simulation_Framework_Excerpt_Contents_Preface_Reader%27s_Guide_Index]
18. Clancey WJ, Hoffman RR (2022) Methods and standards for research on explainable artificial intelligence: Lessons from Intelligent Tutoring Systems. Appl AI Lett. [https://doi.org/10.1002/ail2.53]
19. Covert IS, Lundberg S, Lee S-I (2021) Explaining by removing: a unified framework for model explanation. J Mach Learn Res *22*:1–30
20. Deal SV, Hoffman RR (2010), September/October The Practitioner's Cycles part 3: Implementation problems. *IEEE Intelligent Systems*, pp. 77–81
21. Deal SV, Hoffman RR (2010), March/April The Practitioner's Cycles, Part 1: The Actual World Problem. *IEEE Intelligent Systems*, pp. 4–9
22. diSessa AA (1993) Toward an epistemology of physics. Cognition and Instruction *10*:105–225. [https://doi.org/10.1080/07370008.1985.9649008]
23. Dodge J, Anderson A, Khanna R, Irvine J, Dikkala R, Lam HK-H, Tababai D, Ruangrotsakun A, Shureih Z, Khang M, Fern A, Burnett M (2021) From "no clear winner" to an effective explainable Artificial Intelligence process: an empirical journey. Appl AI Lett 2. [https://doi.org/10.1002/ail2.36]
24. Dodge J (2021) (with 13 others). After-Action Review for AI. ACM *Transactions on Interactive Intelligent Systems, 11*(3–4), Article 29, 1–35
25. Druce J, Niehaus M, Moody V, Harradon M, Daniels-Koch O, Voshell M (2021) "XAI Final Evaluation Reporting Request." Technical Report, Task Area 1, DARPA Explainable AI Program. Arlington, VA: DARPA
26. Ebrahimi S, Petryk S, Gokul A, Gan J, Gonzalez JE, Rohrbach M, Darrell T (2021) Remembering for the right reasons: explanations reduce catastrophic forgetting. Appl AI Lett *2*(4):e44. [https://doi.org/10.1002/ail2.44]
27. Gajos KZ, Mamykina L (2022) March. Do people engage cognitively with AI? Impact of AI assistance on incidental learning. In *27th International Conference on Intelligent User Interfaces* (pp. 794–806). [https://arxiv.org/pdf/2202.05402.pdf]
28. Goyal Y, Wu Z, Ernst J, Batra D, Parikh D, Lee S (2019) Counterfactual visual explanations. [arXiv:1904.07451]
29. Greenbaum J, Kyng M (eds) (1991) *Design at work: Cooperative design of computer systems*. Erlbaum, Mahwah, NJ
30. Grosz BJ (1975) Establishing context in task-oriented dialogs. In *Proceedings of the Proceedings of the 13th Annual ACL Meeting on Computational linguistics. American Journal of Computational Linguistics* (T.C. Diller, ed.), pp. 4–18. New York: Association for Computing Machinery
31. Gunning D, Vorm E, Wang JY, Turek M (2021) DARPA's explainable AI program: a retrospective. Appl AI Lett [https://doi.org/10.1002/ail2.61]
32. Hamidi-Haines M, Qi Z, Fern A, Li F, Tadepalli P (2019) Interactive naming for explaining deep neural networks: A Formative Study. IUI Workshop on EXplainable Smart Systems (EXSS). [arXiv:2006/00093v4]
33. Hilton DJ, Erb H-P (1996) Mental models and causal explanation: judgments of probable cause and explanatory relevance. Think Reasoning *2*:273–308
34. Hinds PM, Patterson M, Pfeffer J (2001) *Bothered by abstraction: the effect of expertise on knowledge transfer and subsequent novice performance*. J Appl Psychol *86*(6):1232–1243
35. Hoffman RR (2017) A taxonomy of emergent trusting in the human-machine relationship. In: Smith P, Hoffman RR (eds) Cognitive systems engineering: the future for a changing world. Taylor and Francis, Boca Raton, FL, pp 137–164
36. Hoffman RR, Deal SV, Potter S, Roth EM (2010) May/June). The Practitioner's Cycles, part 2: Solving Envisioned World Problems. *IEEE Intelligent Systems*, pp. 6–11
37. Hoffman RR, Jalaeian M, Tate C, Klein G, Mueller ST (in review). Metrics for Explainable AI: The Explanation Scorecard. A method in AI measurement science. [https://www.ihmc.us/wp-content/uploads/2021/11/The-Self-Explanation-Scorecard-2021.pdf]
38. Hoffman RR, Klein G, Jalaeian M, Tate C, Mueller ST (2023) Explainable AI: Roles, stakeholders, desirements and challenges. In Press, *Frontiers in Computer Science*. downloaded 28 march 2023 at [https://www.ihmc.us/rgoups/hoffman]
39. Hoffman RR, Lee JD, Woods DD, Shadbolt N, Miller J, Bradshaw JM (2009), November/December The dynamics of trust in cyberdomains. *IEEE Intelligent Systems*, pp. 5–11
40. Hoffman RR, Mueller ST, Klein G, Litman J (2023) Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Front Comput Sci. [downloaded 29 March 2023 at https://www.frontiersin.org/articles/10.3389/fcomp.2023.1096257/full]
41. Hutson M (2018) Hackers easily fool artificial intelligences. Science 361:215
42. Jesus S, Belem C, Balayan V, Bento J, Saliero P, Bizarro P, Gama J (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. [downloaded 30 January 2023 at arXiv:2101.08758v2]
43. Johnson M, Vera AH (2021) No Ai is an island. *The AI Magazine*, pp. 17–28
44. Kalyanam K, Stefik M, de Kleer J (2020) March). "Partnering with Autonomous Systems to reduce unintended behaviors," presentation to the Air Force Science Board
45. Kass R, Finin T (1988) The need for user models in generating expert system explanations. Int J Expert Syst *1*(4):345–375
46. Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Vaughan W (2020), April J. 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* (pp. 1–14)
47. Kenny E, Ford C, Quinn M, Keane M (2021) Explaining black-box classifiers using post-hoc explanations by example: the effect of explanations and error-rates in XAI user studies. *Artificial Intelligence, 294*, (C):103459
48. Kenny EM, Keene MT (2020) On generating plausible counterfactual and semi-factual explanations for deep learning. [arxXiv: 2009,06399v1]
49. Kim J, Canny J (2017) Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of International*

Conference on Computer Vision (pp. 2942–2950). New York: Springer
50. Klein G, Hoffman RR, Clancey WJ, Mueller ST, Jentsch F (2023) Minimum Necessary Rigor in empirically evaluating human-AI work systems. *The AI Magazine*, in press
51. Klein G, Hoffman RR, Mueller ST (2019) "The Plausibility Cycle: A Model of Self-explaining How AI Systems Work." Report on Award No. FA8650-17-2-7711, DARPA XAI Program. DTIC accession number AD1073994. [https://psyarxiv.com/rpw6e/]
52. Klein G, Hoffman RR, Mueller ST, Newsome E (2021) Modeling the process by which people try to explain complex things to other people. J Cogn Eng Decis Mak *15*:213–232
53. Koh OW, Liang P (2017) Understanding black-box predictions via influence functions. [arXiv:1703.04730]
54. Lage I, Chen E, He J, Narayanan M, Kim B, Gershman S, Doshi-Velez F (2019) An evaluation of the human-interpretability of explanation. [downloaded 29 January 2023 at arXiv:1902.00006]
55. Lakkaraju H, Bastani O (2020) "How do I fool you?" Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* New York: Association for computing Machinery. downloaded 29 March 2023 at [https://www.aies-conference.com/2020/wp-content/papers/182.pdf]
56. Lim BY, Dey AK (2010) Toolkit to support intelligibility in context-aware applications. In P*roceedings of the 12th International Conference on Ubiquitous Computing* (pp. 13–22). New York: Association for Computing Machinery
57. Lipton ZC (2016) The mythos of model interpretability. Queue *16*:31–57
58. Litman JA, Jimerson TL (2004) The measurement of curiosity as a feeling-of-deprivation. J Pers Assess *82*:147–157. [https://doi.org/10.1207/s15327752jpa8202_3]
59. Mai T, Khanna R, Dodge J, Irvine J, Lam K-H, Lin Z, Kiddle N, Newman E, Raja S, Matthews C, Perdriau C, Burnett M, Fern A (2020) Keeping It "Organized and Logical": After-Action Review for AI (AAR/AI). *Proceedings of the ACM International Conference on Intelligent User Interfaces* (pp. 465–476). New York: Association for Computing Machinery. [http://www.ftp.cs.orst.edu/pub/burnett/iui20-AARAI.pdf]
60. Miller T (2017) Explanation in Artificial Intelligence: Insights from the social sciences. [arXiv:1706.07269 [Cs]
61. Mohseni S, Zarel N, Raganm DE (2020) A multidisciplinary survey and framework for design and evaluation of explainable AI Systems. [arXiv:1811.11839v5]
62. Mueller ST, Hoffman R, Clancey WJ, Emrey A, Klein G (2019) "Explanation in Human-AI Systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for Explainable AI." Technical Report, Explainable AI Program, Defense Advanced Projects Agency, Washington, DC. [arXiv:1902.01876 [pdf]
63. Mueller ST, Nelson B (2018) A computational model of sensemaking in a hurricane prediction task. *Proceedings of ICCM 2018, the 16th International Conference on Cognitive Modeling* (pp 84–89). [https://acs.ist.psu.edu/iccm2018/ICCM%202018%20Proceedings.pdf]
64. Mueller ST, Veinott ES, Hoffman RR, Klein G, Alam L, Mamun T, Clancey WJ (2020) Principles of explanation in human-AI systems. In *Proceedings of the AAAI Workshop on Explainable Agency in Artificial Intelligence* (AAAI-2020) [arXiv:2102.04972]
65. Nourani M, Honeycutt D, Block J, Roy C, Rahman T, Ragan E, Gogate V (2020) Investigating the importance of first Impressions and Explainable AI with interactive video analysis. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (ACM CHI 2020), pp. 1–8. https://doi.org/10.1145/3334480.3382967]
66. Pollack ME, Hirschberg J, Weber B (1982) User participation in the reasoning processes of expert systems. In *Proceedings of AAAI-82* (pp. 358–361). Menlo Park, CA: Association for the Advancement of Artificial Intelligence
67. Rosenfeld A (2021) Better metrics for evaluating explainable Artificial Intelligence. In U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), *Proceedings of the 21th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)* downloaded 28 March 2023 at [https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p45.pdf]
68. Russell C (2019) Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 20–28). New York: Association for computing Machinery. [https://doi.org/10.1145/3287560.3287569]
69. Samek W, Wiegand T, Müller K-R (2017) Explaining artificial intelligence: understanding, visualizing and interpreting deep learning models. International Telecommunications Union Journal: ICT Discoveries, Special Issue No. 1. [arXiv:1708.08296v1]
70. Schank R (1996) Information is surprises. [www.edge.org/conversation/roger_schank-chapter-9-information-is-surprises]
71. Schön DA (1987) *Educating the reflective practitioner. Jossey-Bass*, San Francisco
72. Selvaraju RR, Lee S, Shen Y, Jin H (2019) Taking a HINT: Leveraging explanations to make vision and language models more grounded. *Proceedings of the International Conference on Computer Vision* (pp. pp. 2591–2600). New York: IEEE
73. Sokol K, Flach P (2020) Explainability fact sheets: A framework for systematic assessment of explainable approaches. [https://doi.org/10.1145/3351095.3372870]
74. van Someren MW, Barnard YF, Sandberg JAC (1994) *The think aloud method. Academic Press*, London
75. Somers S, Mitsopoulos K, Thomson R, Lebiere C (2018) Cognitive-level salience for explainable artificial intelligence. *Proceedings of the 17th International Conference on Cognitive Modeling* (ICCM2018) (pp. 235–240), Madison, WI
76. Stefik M, Youngblood M, Pirolli P, Lebiere C, Thomson R, Price R, Nelson LD, Krivacic R, Le J, Mitsopoulos K, Somers S, Schooler J (2021) Explaining autonomous drones: an XAI journey. *Applied AI Letters, 2*(4)
77. Swartout WR (1981) Producing explanations and justifications of expert consulting programs. Technical Report, Massachusetts Institute of Technology. [http://dl.acm.org/citation.cfm?id=889859]
78. Thomson R, Schoenherr JR (2020) Knowledge-to-Information Translation Training (KITT): An Adaptive Approach to Explainable Artificial Intelligence. In R A Sottilare and J Schwarz (Eds.) *International Conference on Human-Computer Interaction: Track on Adaptive Instructional Systems* LNCS 12214 (pp. 187–204). Cham, Switzerland: Springer
79. Wang P, Givchi A, Shafto P (2020) Manifold learning from a teacher's demonstrations. [arXiv:1910.04615]
80. Wang D, Yang Q, Abdul A, Lim BY (2019) Designing theory-driver user-centric explainable AI. In *Proceedings of CHI 2019* (Paper 601). New York: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300831
81. White A, Garcez d'A (2021) Counterfactual instances explain little. [arXiv:2109.09809v1]
82. Wick MR, Thompson WB (1992) Reconstructive expert system explanation. Artif Intell *54*(1–2):33–70
83. Yeh C-K et al (2019) On the (in)fidelity and sensitivity of explanations. [arXiv:1901.09392v4]