Non-Algorithmic Methods for Explainable Artificial Intelligence

Shane T. Mueller
Michigan Technological University
Gary Klein
Macrocognition, LLC
Robert Hoffman
Institute for Human and Machine Cognition
Tauseef Ibne Mamun
Michigan Technological University
Mohammadreza Jalaeian
Macrocognition, LLC

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. 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.

Cite as:
Mueller, S.T., Klein, G., Hoffman, R.R., Mamun, T., and Jalaeian, M. (2021). "Non-Algorithmic Methods for Explainable Artificial Intelligence." Technical Report, DARPA Explainable AI Program. *Applied Artificial Intelligence Letters* (in press).

Keywords: self-explanation, collaboration, stakeholders, evaluation methodology

Abstract

For Explainable Artificial Intelligence (XAI) there are many ways to support understanding that are not algorithmic. Explanation support can derive from aspects of the interface or architecture. Support can be provided by methods for evaluating the system and determining whether it is effective. Support can be tutorials, help documents, and even conversations among developers or users that help them understand the AI. We refer to these, generally, as Non-Algorithms. Non-Algorithmic explanation is useful for three main reasons. First, substantial cognitive/psychological aspects of explanation must be supported by non-algorithmic approaches. Second, Non-Algorithms can be useful on their own for augmenting existing and fielded AI systems without requiring re-engineering with explanation algorithms. Third, current algorithmic XAI systems may be made more powerful by using Non-Algorithms. The purpose of This Report is an overview of several approaches that have been implemented: The Self-Explanation Scorecard for evaluating machine-generated explanations, the Stakeholder Playbook for tailoring explanations to different stakeholder groups, the Cognitive Tutorial Authoring Guide for guidance in the design of training, the Discovery Platform to enable users to explore the behavior of the AI tool, and the Collaborative XAI (CXAI) tool to allow users to share their understanding and experiences with an AI system.

Outline

1. Introduction

The commonest approach of Explainable AI (XAI) research starts with two algorithms: one that performs some complex behavior (the AI or ML system) and a second one designed to explain it. We refer to this second system as Algorithmic XAI. Yet for any explanation system, there are likely to be many elements that support the explanation and understanding that are not algorithmic. This article describes a number of Non-Algorithmic approaches that have been implemented. Given the XAI emphasis on algorithmic XAI, consideration of non-algorithmic explanation is useful. Non-Algorithms can be valuable on their own for augmenting existing and fielded AI systems without requiring re-engineering with explanation algorithms. Furthermore, current algorithmic XAI systems may be made more powerful by using Non-Algorithms.

In Table 1 we identifies some XAI Non-Algorithm methods that we have developed. One set of methods involves guidance for evaluation, measurement, and validation of XAI systems. Unlike many AI systems which can be assessed on an existing data set (time, accuracy, efficiency), true evaluation and validation of explanations need psychological experimentation with end users (Klein, Hoffman and Mueller, 2019; Mueller, Veinott and Hoffman, 2021). Are the algorithm-generated explanations satisfactory in helping users to develop good mental models and achieve appropriate trust and reliance on the AI? Our framework for measuring explanation effectiveness lays out a suite of measures (including goodness criteria, satisfaction, trust, mental model knowledge, and performance) that can be applied once an AI or XAI system has been developed (Hoffman, Mueller, Klein and Litman, 2018). However, we have also identified a number of formative evaluation methods that can be applied early during conceptualization and development without requiring in situ user evaluation, including measurement according to explanation goodness, stakeholder analysis, and self-explanation. Using these methods, system developers can themselves evaluate their XAI systems.

There are also Non-Algorithms that can be used to provide explanatory material, adumbrate AI-generated explanations, or support the user's sensemaking or self-explanation effort. These methods include a system we call the Discovery Platform, a collaborative approach called CXAI that allows users to explain things and help one another, and a methodology for developing a Cognitive Tutorial, which provides global explanations about a system for novice users.

Table 1. Non-Algorithmic methods for supporting the development of XAI systems.

PURPOSE	METHOD	IMPLEMENTED EXAMPLE
The design of explanations	Mapping explanations to requirements	Stakeholder Playbook guidance to tailoring algorithm-generated explanations to the needs of different stakeholders
		Self-explanation Scorecard to map XAI-generated explanations on to user's sensemaking requirements
	Measurement	Evaluation of Explanation Goodness, Satisfaction, Trust, Mental Model adequacy, Curiosity
Support for the explaining process	Explanation as exploration	The Discovery Platform to explore edge cases and counterfactuals
	Explanation as collaboration	CXAI (collaborative XAI) to allow users to share surprises and discoveries and to pose questions

	Global explanation	Cognitive Tutorial leverages expert knowledge to provide global explanations and practice exercises
Rigorous experimental evaluation	Methodological guidance	Handbook of Experimental Design for rigorous assessment of XAI systems with human users
	Data analysis	Methods for determining practical significance of the results of evaluation experiments

In conceiving of XAI systems, a simple initial model might be that the XAI would generate an explanation, which would be presented to the user, and then performance, trust, and reliance would improve. In contrast, XAI Non-Algorithms are motivated by psychological research on explanation and sensemaking and work on intelligent tutoring systems (see Clancey and Hoffman, 2021; Klein, Hoffman and Mueller, 2019; Mueller, et al., 2018). Explanatory systems need to empower users by giving them information, but also support interaction and exploration that allow them to form and refine explanations that they need for their particular goals. This represents a different design concept from one that focuses just on creating an explanation algorithm.

We briefly describe the high-level motivations and implementations of several of the novel explanation Non-Algorithms listed in Table 1.

2. The Stakeholder Playbook

The initial focus of XAI has been on explaining AI systems to end-users. The Stakeholder Playbook was created in recognition of the possibility that various "stakeholders" would also need explanations, but also that different stakeholders would need different kinds of explanations depending on their roles and responsibilities. The purpose of the Stakeholder Playbook is to enable system developers to appreciate the different ways in which stakeholders might need or want to "look inside" of the AI/XAI system. For example, some stakeholders, like end-users, might need to understand the boundary conditions of the system (its strengths and limitations). Program managers might need to understand an AI/XI system, not for their own understanding but to enable them to succinctly explain the system to other people. Leaders of system development teams need to be able to develop appropriate optimism, informed by appropriate skepticism.

While interest in the stakeholder-dependence of explanations has burgeoned in the last few years, there have been only tentative attempts to investigate the matter empirically. By hearing first-hand from the different stakeholders about what they need in terms of explanations, developers will be better able to help stakeholders develop good mental models of a system. We conducted cognitive interviews with 18 experienced professionals concerning their interactions with AI systems. The group included program managers, developers, end-uses, legal advocates, and others. Participants were asked just a few questions, including: "What do you feel you need to know about an AI system in order to properly exercise your responsibilities?" and "Can you briefly describe any experiences you have had with AI systems where more knowledge would have helped?"

The interviews resulted in a number of surprises. One of the first surprises we encountered involved the demographics. All the interviewees wore more than one "hat." A given interviewee might make a comment pertinent from the perspective of the system developer, but then make another comment that pertained to the explanation requirements of an end-user. Thus, it is better to refer to roles than to stakeholder types or groups. That said, we clustered responses according

to the following "hats": jurisprudence specialists, system developers, system development team leaders, procurement or contracting officers, trainers, system evaluators, and policy makers. The answers to the interview questions resulted in a great many discoveries. Here are just three examples:

- *Not everyone actually needs or wants an explanation*. Only three of the Participants spontaneously said that they want explanations of how the AI works. Far more frequent were assertions about the explanation needs of stakeholders' *other* than themselves.
- *Stakeholders are more likely to need to know about the data than about the AI system that processes the data*. Understanding the data the AI system uses would be more helpful than poking under the hood to examine the innards of the system. They wanted to know what data were used to train the AI/ML. They want to know about any system biases.
- *Sensemaking by exploration is of greater interest that prepared explanations.* A number of Participants commented about how they preferred to manipulate ("poke around") and explore the AI system behavior under different scenarios, to "get a feel for it." Stakeholders want to be provided with more examples of the AI encountering different situations. End-users said that they would benefit from local explanations that are exploratory rather than discursive: The visualization of tradeoffs (e.g., in a scheduling algorithm) would support appropriate reliance and the capacity to anticipate conditions under which anomalous events might occur and the recommendation may be misguided.

For each category we were able to distill explanation requirements. For example, Trainers require access to a rich corpus of cases, but especially "edge cases" that allow the end-user to learn how to handle them but also to anticipate when the AI system is entering a brittle zone. For each category we were also able to distill some "cautions." For example, end-users often require access to the system development team to answer their questions (a Requirement) but for end-users, explanation is never a 'one-off'—continuing explanation is required as the input data, the work system context, or the operational environment change (a Caution). The Stakeholder Playbook itself is a three-page document that can be provided upon request. The full technical report is also available and provides details of the method and results, including ample quotations from the participants.

3. The Self-Explanation Scorecard

The Self-Explanation Scorecard is useful for formative evaluation of an explanation concept. An XAI developer can evaluate their envisioned system according to the following criteria:

1. Features. Does the XAI highlight importance of features that the AI uses in a decision?
2. Successes. Does the XAI show examples of successful operation?
3. Mechanisms. Does the XAI describe mechanisms, rules, or architecture?
4. AI Reasoning. Does the XAI provide functional description of algorithms/processes?
5. Failures. Does the XAI show examples of failure?
6. Comparisons. Does the XAI allow user to compare conditions to draw causal inferences?
7. Diagnosis of failures. Does the XAI provide analysis of why failures occur?

These criteria are roughly ordinal in that they go from lowest to highest in terms of the "depth of analysis": surface to deep information that people find useful when generating and

refining their own understanding of a complex system. Items listed earlier in the Scorecard tend to be simpler, while later items are cognitively more complex and offer deeper insights into the system.

Using the Scorecard, developers can assess their own explanation designs, encouraging them to think about their intended explanations early in the design process in terms of their explanatory depth, and decide whether an interface or algorithm change might support deeper levels of self-explanation.

The Scorecard can also be used to examine existing systems. To illustrate this, we coded several published systems according to the Scorecard criteria, with results that are shown in Table 2. Two of the coauthors independently assessed the systems on each dimension of the Scorecard.

Table 2. Codings of the explanations provided by some published XAI systems . The plus signs mean that the explanations fell at the indicated Level, the minus signs mean that the explanations did not. The ± signs indicate disagreement.

XAI SYSTEM	SCORECARD LEVEL						
	1	**2**	**3**	**4**	**5**	**6**	**7**
Bird Classifier	+	+	-	-	-	±	-
ANN-CBR Twin	-	-	+	±	+	-	-
Partial Dependence	+	+	-	-	+	+	-
Baobab View	+	+	+	+	+	-	+
Deconvnet	+	+	+	-	-	-	±
GA^2M	+	+	+	-	-	+	-

Of the 42 codings, 14 fell at Levels 1-3 and only six fell at Levels 4-7. Three of the 42 codings were disagreements (marked with a ±), representing a Cohen's $\varkappa$ of 0.81. These were all situations in which both raters thought the explanation was present in the paper, but one rater felt this explanation was not actually generated by the XAI system itself, but rather by the authors in support of scientific communication. This finding was actually more pervasive than just these cases of disagreement. For many of the explanation types coded as "not present" in the system, the paper provided a self-explanation type, often by comparing different versions of models or examining ground truth, which that might not always be available at all, let alone accessible by the user. the important lesson here is that some algorithms can support numerous "depth of explanation" levels of explanation, but these are often not presented to the system users. Rather, they are reserved for presentation only to the researchers' peers. This accounts for much of the difficulty we encountered determining whether a particular Level of self-explanation potential characterized some of the systems. Multiple kinds of self-explanation can be supported by an XAI system, but these are not always exposed to the user, and the Scorecard might be useful in identifying alternative ways to present information that might help users.

4. The Discovery Platform

This Non-Algorithmic Method was developed to support the exploration of the behavior of an AI or XAI system, and thereby satisfy some of the requirements of user self-explanation. These include:

- Commonalities and patterns. Patterns allow user to understand typical cases.
- Exceptions. Understanding outliers, anomalies, and exceptions help user isolate and anticipate problematic cases.
- Failures. It should be easy to identify errors and mistakes
- Contrasts. Contrasting cases allow for easy comparisons, enabling counterfactual and causal reasoning.
- Confusions. Identifying high-confusion classes helps anticipate weakness areas of the system.
- Representations, instances, and examples. Thumbnails and examples should be visible and browseable.

For an AI system that has a clear training and test corpus—typical for most image classifiers–there will be high-level patterns of performance that cannot be revealed by the usual approach of providing a local justification (e.g., in the form of feature highlighting or a heatmap). The Discovery Platform concept was inspired by conversations we had with XAI system developers who had often browsed hundreds or thousands of image cases of their own data set and had discovered systematic problems (and strengths) with their system. As a consequence they developed special-purpose browsers to help them debug and diagnose their system, an instantiation of the notion of "explanatory debugging."

We implemented a prototype version of the Discovery Platform using the R web interface "Shiny". The prototype system uses a simple image support vector machine (SVM) classifier on the MNIST hand-written digit data set, and we sampled 10,000 test cases to produce a browseable data set on which the system achieved 50% accuracy (the SVM in actuality achieved an 89% accuracy rate). The Discovery Platform provides an interface allowing the user to select cases based on input class and classifier label, and then to sample 5-15 cases based on different simple criteria related to the system's judged probability across outcome cases. One of the panels of the Discovery Platform is shown in Figure 1.

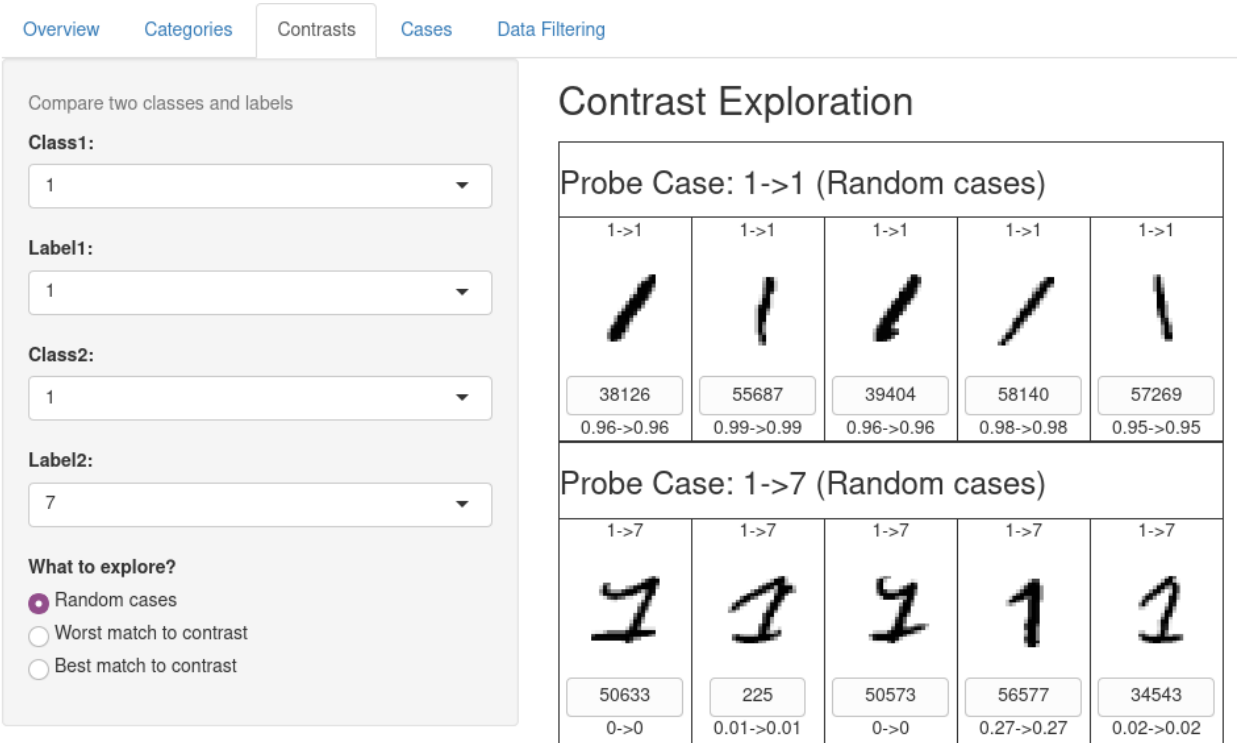

Figure 1. Screenshot of the "contrast explorer" pane of the Discovery Platform.

5. Collaborative XAI (CXAI)

Wildly successful social Q&A systems such as StackExchange suggest a non-algorithm for the explanation of complex software systems. They provide a searchable and browseable database of questions that other users can answer, and those answers can be upvoted to provide social credit to users, and benefit others who have similar questions. We thought of that a collaborative platform for explaining AI might be successful as:

- A way of supporting explanation for a team using AI tools that do not have algorithmic XAI solutions,
- A way for augmenting an XAI system to help users understand edge cases and surprises, and
- A way for user evaluations to serve as feedback for the further refinement of the AI/XAI system.

This third functionality is particularly interesting as it regards the explanation process as a human-machine collaboration, a two-way street (see Clancey and Hoffman, 2021). We have developed a prototype system which we call Collaborative XAI (CXAI). The primary interface panel is shown in Figure 2.

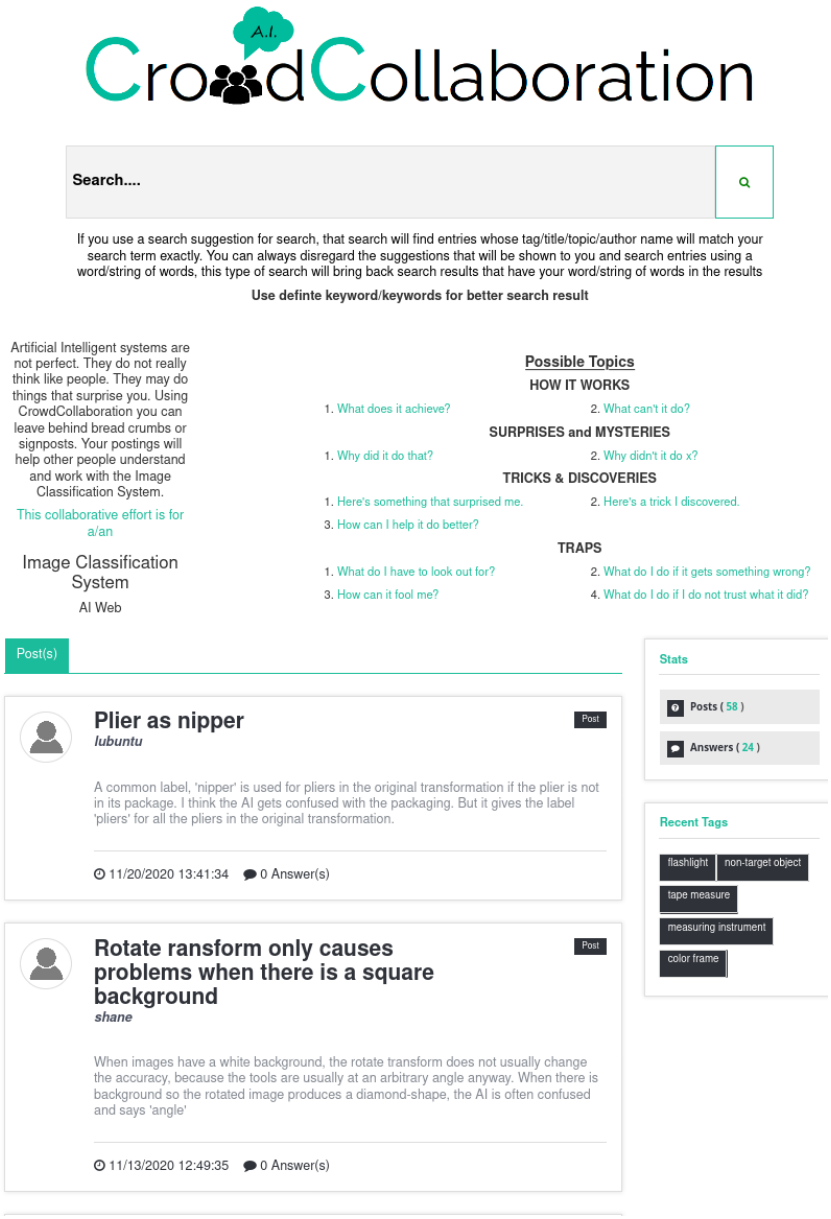

Figure 2. Screenshot of the interface panel of the CXAI system.

Via this interface, the user can pose a query based on one of the "triggers" for explanation that have been discussed in the psychology literature on explanation and our own analysis of a large corpus of "real world" explanations of complex systems (Klein, Hoffman and Mueller, 2019; see also Lim and Dey, 2009). When users experience a surprise, they ask such questions as "Why did it do that?" or "Why didn't it do something else?"

6. Cognitive Tutorials for AI and XAI

Researchers in the field of XAI have been discussing the distinction between global and local explanation from the outset (although the distinction can be traced to earlier work on causal reasoning). This distinction covers the focus of the explanation–with global describing how the system architecture works in general (perhaps covering algorithms, training materials, specific sets of rules, and patterns of behavior), and local describing how a particular case was handled.

In XAI work generally, local explanations have been used synonymously with justifications. Formal justifications take sense to computer scientists, and explain to them why the system developers "did it that way." But this in not at all the same as explaining things to users, who generally are not computer scientists. Local justification in a psychological sense is useful for trying to understand issues of fairness, justice, and to identify remedies: A local explanation of why a loan was denied will hopefully let a developer understand if the denial stemmed from something improper, and will help an applicant determine what they can do to improve their chances of approval. However, it is perhaps at odds with developing long-term general trust and understanding in the system, because local explanations are myopic and analytical.

One method we have explored for producing global explanations is what we call a Cognitive Tutorial, which advocates using experiential training in the form of a user guide about the cognitive operations of the AI . The Cognitive Tutorial recognizes that users will come to the AI with misconceptions about how it works–often assuming it works in the same way a human would. However, these systems often succeed and fail in unexpected ways. The goal of the cognitive tutorial is to use experiential training to help the user understand the competence boundaries of the system–along dimensions that include modeling/representation, algorithms, data, and output/visualization.

We have developed a Cognitive Tutorial Authoring Guide, which can be provided upon request. It details the specific steps and procedures for identifying learning objectives and implementing the tutorial modules: How to Use It, how to Use It Improperly, Common Misconceptions, and Novel Problem Exercises.

7. Conclusion

The combination of Algorithmic and Non-algorithmic approaches to XAI is likely to be more successful that any purely Algorithmic approach. This derives from the psychology of explanation: explanation as an exploratory and collaborative process for ensuring that AI technology is part of an understandable, learnable, usable, and useful human-machine work system.

References

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730. New York: Association for Computing Machinery.

Clancey, W.J., and Hoffman, R.R. (2018). "Methods and Standards for Research on Explainable Artificial Intelligence Research: Lessons from Intelligent Tutoring Systems Research." Technical Report from Task Area-2, Explainable AI Program, DARPA, Alexandria, VA.

Ford, C., Kenny, E.M., and Keane, M.T. (2020). Play MNIST for me! User studies on the effects of post-hoc, example-based explanations and error rates on debugging a deep learning, black-box classifier. [*arXiv preprint arXiv:2009.0634p]*

Hendricks, L.A, Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating visual explanations. In *Proceedings of the European Conference on Computer Vision*. [arXiv:1603.08507v1]

Hoffman, R.R., Mueller, S..T, Klein, G, and Litman J. (2081). "Metrics for explainable AI: Challenges and Prospects." [arXiv:1812.04608 2018.]

Klein G., Hoffman, R.R., and Mueller, S.T. (2019)."The Plausibility Cycle: A Model of Self-Explaining How AI Systems Work." Technical Report from Task Area-2, Explainable AI Program, DARPA, Alexandria, VA.

Klein, G., Hoffman, R.R., & Mueller, S. (2019). Naturalistic psychological models of explanatory reasoning: how people explain things to others and themselves. Presentation at the *International Conference on Naturalistic Decision Making*. Dayton, OH: Shadowbox LLC.

Krause J., Perer, A, and Bertini, E. (2016). Using visual analytics to interpret predictive machine learning models. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning***.** [arXiv:1606.05685]

Lim, B.Y., and Dey, A.K. (2009). Assessing demand for intelligibility in context-aware applications. In P*roceedings of the 11th International Conference on Ubiquitous Computing* (pp. 195-204). New York: Association for Computing Machinery.

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2018). "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI." Technical Report from Task Area-2, Explainable AI Program, DARPA, Alexandria, VA [https://arXiv.org/abs/1902.01876]

Mueller, S.T., Veinott, E.S., Hoffman, R.R., Klein, G., Alam, L., Mamun, T., and Clancey, W.J. (2020). Principles of explanation in human-AI systems. In *Proceedings of the Workshop on Explainable Agency in Artificial Intelligence* (AAAI-2020) [arXiv:2102.04972].

Van Den Elzen S. and Van Wijk J. J. (2011). BaobabView: Interactive construction and analysis of decision trees. In: *Proceedings of the IEEE Conference On Visual Analytics Science And Technology* (pp. 151-160). New York: IEEE.

Zeiler, M.D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of European Conference on Computer Vision* (pp. 818-833). [arXiv:1311.2901]