EXPLAINING EXPLANATION FOR "EXPLAINABLE AI"

Robert R. Hoffman
Institute for Human and Machine Cognition
Pensacola, FL

Gary Klein
Macrocognition, LLC
Washington, DC

Shane T. Mueller
Michigan Technological University
Houghton, MI

ABSTRACT

What makes for an explanation of "black box" AI systems such as Deep Nets? We reviewed the pertinent literatures on explanation and derived key ideas. This set the stage for our empirical inquiries, which include conceptual cognitive modeling, the analysis of a corpus of cases of "naturalistic explanation" of computational systems, computational cognitive modeling, and the development of measures for performance evaluation. The purpose of our work is to contribute to the program of research on "Explainable AI." In this report we focus on our initial synthetic modeling activities and the development of measures for the evaluation of explainability in human-machine work systems.

INTRODUCTION

The importance of explanation in AI has been emphasized in the popular press, with considerable discussion of the explainability of Deep Nets and Machine Learning systems (e.g., Kuang, 2017). For such "black box" systems, there is a need to explain how they work so that users and decision makers can develop appropriate trust and reliance. As an example, referencing Figure 1, a Deep Net that we created was trained to recognize types of tools.

Figure 1. Some examples of Deep Net classification.

Outlining the axe and overlaying bird silhouettes on it resulted in a confident misclassification. While a fuzzy hammer is correctly classified, an embossed rendering is classified as a saw. Deep Nets can classify with high hit rates for images that fall within the variation of their training sets, but are nonetheless easily spoofed using instances that humans find easy to classify. Furthermore, Deep Nets have to provide some classification for an input. Thus, a Volkswagen might be classified as a tulip by a Deep Net trained to recognize types of flowers. So, if Deep Nets do not actually possess human-semantic concepts (e.g., that axes have things that humans call "blades"), what do the Deep Nets actually "see"? And more directly, how can users be enabled to develop appropriate trust and reliance on these AI systems?

Articles in the popular press highlight the successes of Deep Nets (e.g., the discovery of planetary systems in Hubble Telescope data; Temming 2018), and promise diverse applications "... the recognition of faces, handwriting, speech... navigation and control of autonomous vehicles... it seems that neural networks are being used everywhere" (Lucky, 2018, p. 24).

And yet "models are more complex and less interpretable than ever... Justifying [their] decisions will only become more crucial" (Biran and Cotton, 2017, p. 4). Indeed, a proposed regulation before the European Union (Goodman and Flaxman, 2016) asserts that users have the "right to an explanation." What form must an explanation for Deep Nets take?

This is a challenge in the DARPA "Explainable AI" (XAI) Program: To develop AI systems that can engage users in a process in which the mechanisms and "decisions" of the AI are explained. Our tasks on the Program are to:

(1). Integrate philosophical studies and psychological research in order to identify consensus points, key concepts and key variables of explanatory reasoning,

(2). Develop and validate measures of explanation goodness, explanation satisfaction, mental models and human-XAI performance,

(3) Develop and evaluate a computational model of how people understand computational devices, and

then evaluate the model using the validated measures,
(4). Generate a corpus of cases in which people try to explain the workings of complex systems, especially, computational systems work, and
(5) From the case analysis create a "naturalistic decision making" model of explanation that can guide the development of XAI systems by computer scientists.

In this presentation we report progress on our synthesis of ideas and concepts of explanation, the development of models of the explanation process, and the development of metrics.

LITERATURE SYNTHESIS

A thorough analysis of the subject of explanation would have to cover literatures spanning the history of Western philosophy: Disciplines including philosophy and psychology of science, cognitive psychology, psycholinguistics, and expert systems. The archive we created includes over 700 papers.

Psychology

The challenge of XAI entrains concepts of representation, modeling, language understanding, and learning. Concepts that are entrained include abductive inference, causal reasoning, mental models, and self-explanation. Potentially measurable features of explanation include: various forms of explanation (e.g., contrastive explanation, counterfactual reasoning, mechanistic explanation, etc.); various utilities or uses of explanation (e.g., diagnosis, prediction), the limitations or foibles of explanatory reasoning (e.g., people will believe explanations to be good even when they contain flaws or gaps in reasoning) (Lombrozo & Carey, 2006).

Many researchers present a list of the features that are believed to characterize “good” explanations (e.g., Brezillon and Pomerol, 1997). These include context- or goal-relevance, reference to cause-effect covariation and temporal contiguity, and plausibility. There are also some contradictions in the literature: Some assert that good explanations are simple; others assert that good explanations are complete. Clearly, good explanations fall in the sweet spot between detail and comprehensibility.

A number of conceptual psychological models of the explanation process have been presented in the research literature. The first step in the model of Krull and Anderson (1994), the noticing of an event, is reminiscent of the first step in C.S. Peirce's model of abduction (1891), that is, the observation of something that is interesting or surprising. Subsequent steps are Intuitive Explanation, Problem Formulation and Problem Resolution. The model is not specific about what is involved in these steps, but is explicit about the role of motivation and effort.

Johnson and Johnson (1993) studied an explanation process in which experts explained to novices the processes of statistical data analysis. Transcripts of explainer-learner dialogs were analyzed. A key finding was that the explainer would present additional declarative or procedural knowledge at those points in the task tree where sub-goals had been achieved. The Johnson and Johnson model is expressed as a chain of events in which the explainer provides analogies, instructions, and justifications.

Artificial Intelligence

AI has has a history of work on explanation. (A review of the literature, with a bibliography, is available from the authors.) Starting with the first generation of expert systems, it has generally been held that explanations must present easy-to-understand coherent stories in order to ensure good use of the AI or good performance of the human-machine work system (Biran & Cotton, 2017; Clancey, 1986).

Attempts to explain Deep Nets have often taken contrastive approaches. These include occlusion (e.g., Zeiler & Fergus, 2014), which shows how classifications differ as regions are removed from an image, and counter-examples (e.g., Shafto, Goodman, & Griffiths, 2014). A limitation of these approaches is that they conflate explanation and justification. So, for example, one team of computer scientists might “explain” how their Deep Net works by showing a matrix of node weights at the multiple layers within a network. This works as a justification of the architecture to computer scientists but does not work for explaining the Deep Net to a human user who is not a computer scientist. Furthermore, the focus of the contrastive approaches is "local" explanation, that is, explaining why the AI made a particular determination for a particular case. An example would be to show the user a heat map that highlights the eyes and beak of a bird, accompanied by a brief statement that the beak and eye features make this bird a sparrow. This is different from "global” explanation, which is aimed at explaining how an AI system works in general (e.g., Doshi-Velez and Kim, 2017). Finally, explainability is often conflated with interpretability, which is a formal/logical notion in computer science. The fact that a computer system is interpretable does not mean that it is human understandable; the formal interpretation has explanatory value only to computer scientists.

From these literatures, we have identified some key concepts that serve as guidelines to consider in the development of XAI systems.

KEY CONCEPTS

(1). Explaining is a Continuous Process. Humans are motivated to "understand the goals, intent, contextual awareness, task limitations, [and] analytical underpinnings of the system in an attempt to verify its trustworthiness" (Lyons, et al., 2017). One of the consensus points coming from the philosophy of science is that explanations have a heuristic function: They guide further inquiry. The delivery of an explanation is not always an end point. Indeed, it must be thought of as a continuous process since the XAI system that provides explanations must enable the user to develop appropriate trust and reliance in the AI system with continued experience. The user must be able to actively explore the states and choices of the AI, especially when the system is operating close to its boundary conditions, including when it makes errors (see Amerishi, et al., 2015). How can XAI work in concert with the AI to empower learning-during-use?

(2). Explaining is a Co-adaptive Process. Many conceptual models, such as that of Johnson and Johnson (1993) assume that the explanation process is a one-way street: The explainer presents information and instruction to the explainee. In addition, conceptual models typically assume that an explanation can be "satisfying," implying that it is a process with clear-cut beginning and end points (the delivery of instructional material that the user simply assimilates). An alternative view is that explanation is a collaboration or co-adaptive process involving, in the case of XAI, the learner/user and the system. "Explanations improve cooperation, cooperation permits the production of relevant explanations" (Brezillon and Pomerol, 1997, p. 7; Moore & Swartout, 1991). This is the concept of "participatory explanation," similar to the notion of "recipient design" in the conversation analysis literature, i.e., that messages must be composed so as to be sensitive to what the recipient of the message is understanding (Sacks & Schegloff, 1974). An assumption in some of the first generation of AI-explanation and intelligent tutoring systems was that it is only the human who has to learn, or change, as a result of explanations offered by the machine.

(3). Explanation Triggers. Not everything needs to be explained, and explanations are quite often triggered by violations of expectation. Explanations among people serve the purpose of clarifying unexpected behavior, and so a good explainable system may need to understand what are the appropriate triggers of explanation.

(4). Self-explanation. Psychological research has demonstrated that self-explanation improves learning and understanding. This finding holds for both self-motivated explanation and self-explanation that is prompted by the instructor (Chi, Leeuw, Chiu, & LaVancher, 1994).

(5). Explanation as Exploration. An important mode of explanation is helping the user understand the boundaries of the intelligent system (Mueller, et al. 2011). System developers are often reluctant to tell the user what the system cannot do—until they misuse it. Famously, Tesla's autopilot system is touted as a self-driving car, except when accidents occur and the user is blamed for operating it in circumstances in which it was not intended to be used. Clarifying boundary conditions can help produce appropriate trust, so that the user knows when to rely on the system, and when to take over.

(6). Contrast Cases. When forming explanations of intelligent systems, it can be as important to tell what is not being done as to tell what is being done. Contrastive reasoning has been identified as central to all explanation (e.g., Miller, Howe, & Sonenberg, 2017) and it can be an effective way to help the user understand why an expectation was violated. For example, an explainable GPS system might explain why a turn was made by describing why a (normally shorter) route was not taken.

MEASURES

One purpose of our cognitive modeling is to highlight the key concepts that must be mated with measures and metrics. The creation of AI systems that can explain themselves will require a number of types of measures.

Explanations generated by the AI can be evaluated in terms of the *goodness criteria*, of what makes an explanation good, according to the research literature. From a roster of those criteria we developed an "Explanation Satisfaction Scale," which has been evaluated using the Content Validity Ratio method (Lawshe, 1975), and following that a test of discriminant validity which resulted in a very high Cohen's alpha ~.80. The final scale consists of seven Likert items that reference understandability, satisfyingness, detail, accuracy, completeness, usability, usefulness, and trustworthiness. This scale may be used by AI researchers in the XAI Program to evaluate the explanations that their systems produce but might be used in other applications as well.

Effective use of intelligent systems depends on user *mental models* (Kass & Finn, 1988). These have to be elicited and evaluated. In the XAI Program they can be elicited using some forms of structured interview in which users express their understanding of the AI system, with the protocols compared for their propositional concordance with explanations provided by experts. Based on the literature, we have

developed a guidebook that details a variety of methods for eliciting mental models.

Finally, the evaluation of XAI systems will measure the change in *performance* attributable to the explaining process, via controlled experimentation. Performance can be evaluated in a number of ways. Good explanations should enable the user to:

- Efficiently and effectively use the AI in their work, for the purposes that the AI is intended to serve.
- Correctly predict what the AI system will do for given cases. This can include cases that the AI gets right and also cases it gets wrong (e.g., failures, anomalies).
- Explain how the AI works to other people.
- Correctly assess whether a system determination is correct, and thereby have appropriate trust.
- Judge when and how to rely on the AI even while knowing the boundary conditions of the competence of the AI, and thereby having appropriate reliance.

Experiments will have to evaluate the learning that occurs during training as well as during performance. These experiments will have to take into account the difference between global and local explanations. These key variables are modeled in Figure 1, which appears following the References.

DEVELOPING A NATURALISTIC MODEL

Another aspect of our effort in XAI is to develop a "naturalistic" model of explanation based on the analysis of a corpus of cases in which people create explanations of complex situations or systems. The trigger for local explanations is typically a violated expectancy. "Why did it do that?" signifies a surprise, and calls for an account to revise the violated expectancy. And this process requires the explainer to diagnose what user expectations need revision — where is the learner's mental model flawed or incomplete. Second, many AI systems start with a complete account and then try to whittle this account down into something manageable, but if the trigger for a local explanation is a violated expectancy then the process of explaining is aimed at the flawed expectancy, and no whittling down is needed. Third, what is the stopping point for explaining something? AI systems do not have a clear stopping point whereas our initial review of naturalistic cases suggests that the stopping point is a perspective shift in which the user moves from "Why did it do that?" to "Now I see that in this situation I would have done the same." The current state of art for AI systems does not take perspective shifts into account.

ACKNOWLEDGEMENT

This material is based on research sponsored DARPA under agreement number FA8650-17-2-7711 The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. 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 DARPA, AFRL or the U.S. Government.

REFERENCES

Amershi, S., Chickering, M., Drucker, S.M., Lee, B., Simard, P., & Suh, J. (2015). Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 337–346). New York: Association for Computing Machinery.

Biran, O., & Cotton, C. (2017). Explanation and Justification in Machine Learning: A Survey. *IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*.

Brézillon, P., & Pomerol, J.-C. (1997). Joint cognitive systems, cooperative systems and decision support systems: A cooperation in context. In *Proceedings of the European Conference on Cognitive Science, Manchester* (pp. 129–139).

Chi, M.T., Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self‐explanations improves understanding. Cognitive Science, 18(3), 439–477.

Clancey, W.J. (1986a). From GUIDON to NEOMYCIN and HERACLES in twenty short lessons. *AI Magazine*, *7*(3), 40.

Doshi-Velez, F., & Kim, B. (2017). A Roadmap for a Rigorous Science of Interpretability. ArXiv Preprint ArXiv:1702.08608. Retrieved from https://arxiv.org/abs/1702.08608

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

Johnson, H., & Johnson, P. (1993). Explanation Facilities and Interactive Systems. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (pp. 159–166). New York: Association for Computing Machinery.

Kass, R., & Finin, T. (1988). The Need for User Models in Generating Expert System Explanation. *International Journal of Expert Systems*, *1*(4), 345–375.

Krull, D. S., & Anderson, C. A. (1997). The process of explanation. *Current Directions in Psychological Science*, *6*(1), 1–5.

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

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, *28*, 563–575.

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

Lucky, R.W. (2018, January). The mind of neural networks. IEEE Spectrum, p. 24.

Lyons, J.B., Clark, M.A., Wagner, A.R., & Schuelke, M.J. (2017). Certifiable trust in autonomous systems: Making the intractable tangible. *AI Magazine*, *38*(3), 37–49.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum. In Proceedings of the International joint Conference on Artificial Intelligence (IJCAI-17) Workshop on Explainable Artificial Intelligence (XAI).

Moore, J. D., & Swartout, W. R. (1991). A reactive approach to explanation: taking the user's feedback into account. In C. Paris, W.R. Swartyout, & W.C. Mann (Eds.), *Natural language generation in artificial intelligence and computational linguistics* (pp. 3–48). New York: Springer.

Mueller, S.T. & Klein, G. (March-April 2011). Improving users' mental models of intelligent software tools. *IEEE: Intelligent Systems, 26*(2), 77—83.

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis* (pp. 2–4). Washington, DC: Office of the Assistant Director of Central Intelligence for Analysis and Production.

Sacks, H. & Schegloff, E. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50,* 696-735

Shafto, P., & Goodman, N. (2008). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 1632–1637). Austin, TX: Cognitive Science Society Austin.

Temming, M. (2018, 20 January). AI has found an 8-planet system like ours in Keppler data. *Science News*, p. 12.

Zeiler, M. D., Krishnan, D., Taylor, G. W., & Fergus, R. (2010). Deconvolutional networks. In Proceedings of the *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2528–2535). New York: IEEE.

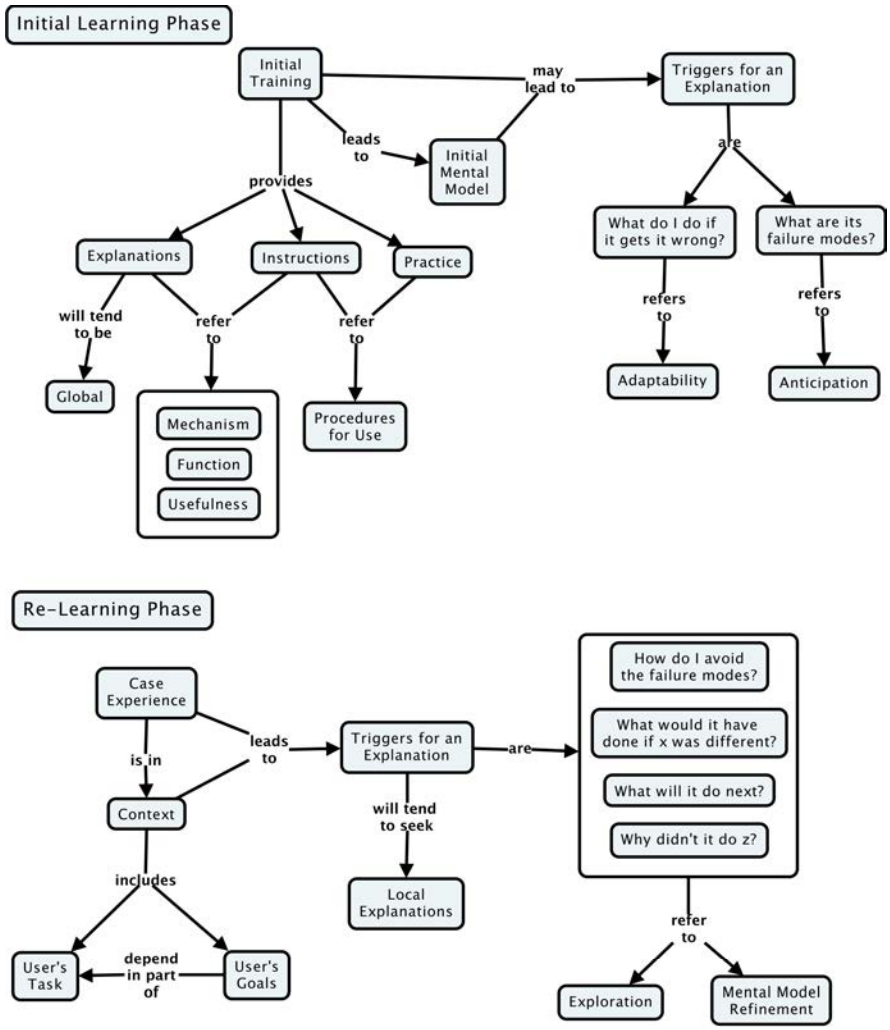

Figure 1. Explanation spans the training and performance contexts, but in doing so requires different kinds of explanation.