Scorecard for Self-Explaining Capabilities of AI Systems

Gary Klein, Ph.D.
Macrocognition, LLC
Robert R. Hoffman, Ph.D.
Institute for Human and Machine Cognition
Shane T. Mueller, Ph.D.
Michigan Technological University

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:
Klein, G., Hoffman, R.R., and Mueller, S.T. (2021). " Scorecard for Self-Explaining Capabilities of AI Systems" Technical Report, DARPA Explainable AI Program.

Abstract

This report describes a Self-Explaining Scorecard for appraising the self-explanatory support capabilities of XAI systems. The Scorecard might be useful in conceptualizing the various ways in which XAI system developers are supporting users, and might also help in comparing and contrasting the various approaches.

Outline

1. Background

A person who is trying to understand how an AI system works is struggling to form and refine a mental model which explains how the AI works, how it fails, and how failures can be overcome (Borders, Klein & Besuijen, 2019; Hoffman, et al., 2019). The XAI literature review (Mueller, et al., 2018) highlighted the important difference between global explanation (*"How does it work?"*) and local explanation (*"Why did it make that particular decision?*).

Progress reports from the Task Area-1 Performer Teams (Hoffman, Klein and Mueller, 2019a) showed that a majority of the systems involved the use of local explanations. These local explanations (e.g., heat maps and other types of local explanations) essentially serve as cues or clues, leaving it to the user to discern how those might explain how the XAI systems work. Local explanations provide information to enable users to form their own explanations and build their own mental models about how the AI systems were working.

This clearly suggested that the concept of "self-explanation" would be important in the XAI activity. Research has shown that self-explanation has a significant and positive impact on learners' understanding (Calin-Jagerman & Ratner, 2005; Chi & Van Lehn, 1991: Chi, et al., 1989, 1994; Chi, Roy & Hausmann, 2008; Lombrozo, 2016; Rittle-Johnson, 2006). The purpose of self-explanation can be to satisfy curiosity, to enable the learner to develop a richer mental model, to make more accurate predictions of the AI's behavior, to better calibrate trust in the AI, and/or to improve performance by using the AI as appropriate.

2. Analysis

As we studied a number of XAI system projects, we were struck by both the commonalities and the contrasts. We tried to codify these common themes and contrasts, and the result was a "Scorecard" (version of October 2019).

This Report presents the current version of this Self-Explaining Scorecard. This version resulted from discussions with XAI system developers in which the Levels were applied to their explanation methods and approaches. The XAI system developers generally found the Levels interesting, allowing them to reflect on how they might enhance their systems' explanation capabilities. Additionally, some subtle distinctions were called out that were missing in the October 2019 iteration of the Scorecard.

We now describe eight Levels for self-explaining support, along with our rationale for scaling the Levels.

The Self-Explanation Scorecard

LEVELS	EXPLANATION FORM
1. Null	No material is provided to support self-explaining.
2. Surface features	Heat maps, bounding boxes, linguistic features, semantic bubbles illustrate some of the analyses done by the AI. Surface features by themselves don't help much in understanding how the AI works, but in conjunction with positive cases and failures they can be useful. The user typically self-explains by accommodating surface feature information with other forms of information described below.
3. Successes	Instances or demonstrations of the AI generating correct predictions or recommendations. These might or might not take the form of text.
4. Mechanism	Global descriptions of how the AI works can refer to mechanisms or architecture. Typically this form of explanation is text, but may include example instances using other forms (e.g., diagrams, salience maps, etc.). This form of explanatory information is typically included in the initial instructions about the XAI system and its uses.
5. AI Reasoning	These are ways to "look under the hood" of the AI to get some idea of how it is making decisions. This can be shown via choice logic, decision rules, goal stacks, parse graphs. These can show the ways the AI weights different pieces of information in order to make a choice. Goal stacks show the goals that are most activated when the AI makes its decisions. Explanations of these types might or might not include text.
6. Failures	Instances or demonstrations of AI mistakes. These are often very informative as they illustrate limitations of the AI and also illustrate how the AI works (and doesn't work). Failures can also be with respect to the explanations, i.e., user feedback to the AI about whether an explanation is correct. Explanations of these types might or might include text, or be in the form of text.
7. Comparisons	Comparisons can be expressed using analogs (highlighting similarities and differences.) or counterfactuals. Comparisons can contrast choice logic or factor weights (Level 5) for different conditions or for successes vs failures. Goal stacks can be used to contrast successes and failures. Explanations of these types might or might include text, or be in the form of text.
8. Diagnosis of Failures	These are even more informative than the failures alone, they are Description of the reasons for failures. In addition, letting the user

	manipulate the AI and to infer diagnoses; capability to manipulate inputs, weights, etc. in order to see the effects on the AI outputs. The use of manipulations allows users to create failure conditions and to make their own inferences about diagnoses. Explanations of these types might or might include text, or be in the form of text.

3. Basis for the Ordination

The ordination of the Levels 1-8 is not simply "More-to-Less" explanation or "Weaker to Better" explanation " or "Sparser-to- Richer" explanation. While the differences among the levels might be thought of as Low-to High "degree of support for self-explanation," there are subtleties and complications that need to be considered with regard to how the levels are ordered in this analysis.

For example, moving up the Levels does not mean that the user needs to engage in less mental effort because the explanations are more complete. Rather, it means that the user has to engage in different sorts of effortful sensemaking. Going from Level 1 to Level 8, there is:

a) Greater consideration of the user's needs,
b) Somewhat greater sophistication to the inferences required of the user, and yet at the same time there is
c) Greater support for the user who is trying to understand how to use the AI as a tool (e.g., how to anticipate confusions).

4. Blurring the Distinctions

One of our findings from the analysis of instances of explanatory reasoning (Klein, et al., 2019) was that global explanations often include examples, and local explanations often include some hints about how the AI works. Thus, the classic distinction between global and local explanation becomes blurry, however useful it may be in the abstract. Global-versus-Local is not an absolute distinction. Indeed, the value of local explanations is increased if global explanation is also provided to users.

As entries in Table 1 point out, explanations can involve combinations across the Levels, e.g., heat maps along with positive and negative cases. A decision rule (Level 5) can describe when and why an AI system fails (Level 6). In the application of the scorecard, the attribution for a given XAI system defaults to the highest level of any of the components of a combination. For example, a heat map in conjunction with positive cases portraying choice logic would default to Level 5 (AI Reasoning). If an XAI system then included negative cases then their system would default to Level 6 (Failures). It would not constitute level 8 (Diagnosis of failures) but it might stimulate the user to make diagnostic conjectures.

References

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182.

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

Chi, M.T.H., Roy, M., & Hausmann, R.G.M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science 32*, 301–341.

Chi, M.T., & VanLehn, K.A. (1991). The content of physics self-explanations. *The Journal of the Learning Sciences*, *1*(1), 69–105.

Hoffman, R.R., Klein, G., & Mueller, S.T. (2019). "TA-2 Suggestions for Experimental Design Based on the Gate-2 Reports of the TA-1 Performer Teams." Report on Award No. FA8650-17-2-7711, DARPA XAI Program.

Hoffman, R.R., Mueller, S.T., Klein, G., & Litman, J. (2019). "Metrics for Explainable AI: Challenges and Prospects." Technical Report on Award No. FA8650-17-2-7711, DARPA XAI Program.

Klein, G., Hoffman, R., Mueller, S.T. (2019, April). " Naturalistic Psychological Model of Explanatory Reasoning: How People Explain Things to Others and to Themselves." Technical Report on Award No. FA8650-17-2-7711, DARPA XAI Program.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759. https://doi.org/10.1016/j.tics.2016.08.001

Mueller, S.T., Hoffman, R.R., Clancey, W, 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 on Award No. FA8650-17-2-7711, DARPA XAI Program.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1–15.

Williams, M.D., Hollan, J.D., and Stevens, A.L (1983). Human reasoning about a simple physical system. In D. Gentner and A.L. Stevens (Eds.) *Mental models* (pp. 131-154). New York: Psychology Press.

Appendix A

The Explanation Scorecard

The purpose of the Scorecard is to help XAI developers consider more powerful types of information to better support users in understanding how the AI works, and thereby engender appropriate trust and reliance.

Users usually do not go from a given explanation to an immediate and satisfactory understanding. Rather, they think about the explanation they have been given, and the cues that are available to them. This is a process of sensemaking or self-explanation. This process is critical for users to take initiative in learning how to work with AI. The user's purpose can be to satisfy curiosity, to develop a richer mental model, to anticipate the limitations and boundary conditions of the AI, to make more accurate predictions of the AI's behavior, to better calibrate trust in the AI, or to improve performance by using the AI as appropriate.

The Scorecard presents a number of Levels of explanation. At the lower levels are explanations in the terms of the cues or features of individual instances. At the higher levels are explanations that answer more general questions about how the AI works. Going from the lower to higher levels can be thought of as enabling insights about the strengths and weaknesses of the AI system. There is greater consideration of user needs, somewhat greater sophistication to the inferences required of the user, and at the same time there is greater support for the user who is trying to understand how to use the AI as a tool (e.g., how to anticipate confusions).

A detailed Technical Report on the development of the Scorecard is available upon request [rhoffman@ihmc.us]

LEVELS	EXPLANATION FORM
9. Null	No material is provided to support self-explaining.
10. Surface features	Heat maps, bounding boxes, linguistic features, semantic bubbles illustrate some of the analyses done by the AI. Surface features by themselves don't help much in understanding how the AI works, but in conjunction with positive cases and failures they can be useful. The user typically self-explains by accommodating surface feature information with other forms of information described below.
11. Successes	Instances or demonstrations of the AI generating predictions or recommendations.
12. Mechanism	Global descriptions of how the AI works can refer to mechanisms or architecture. Typically is text, but may include example instances. This form of explanatory information is typically included in the initial instructions about the XAI system and its uses.
13. AI Reasoning	These are ways to "look under the hood" of the AI to get some idea of how it is making decisions. This can be shown via choice logic, decision rules, goal stacks, parse graphs. These can show the ways the AI weights different pieces of information in order to make a choice. Goal stacks show the goals that are most activated when the AI makes its decisions.
14. Failures	Instances of AI mistakes breakdowns. These are often very informative as they illustrate limitations of the AI and also illustrate how the AI works (and doesn't work). Failures can also be with respect to the explanations, i.e., user feedback to the AI about whether an explanation is correct.
15. Comparisons	Comparisons can be expressed using analogs (highlighting similarities and differences.) or counterfactuals. Comparisons can contrast choice logic or factor weights (Level 5) for different conditions or for successes vs failures. Goal stacks can be used to contrast successes and failures.
16. Diagnosis of Failures	These are even more informative than the failures alone, they are Description of the reasons for failures. In addition, letting the user manipulate the AI and to infer diagnoses; capability to manipulate inputs, weights, etc. in order to see the effects on the AI outputs. The use of manipulations allows users to create failure conditions and to make their own inferences about diagnoses.