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
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 predictions or recommendations.
4. 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.
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.
6. 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.
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.
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.