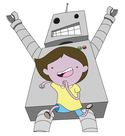
RUTGERS UNIVERSITY | NEWARK

shaftolab.com



A Psychological Theory of Explainability

Scott Cheng-Hsin Yang*[1], Tomas Folke*[1] & Patrick Shafto[1,2]
[1]Department of Mathematics & Computer Science, Rutgers University–Newark
[2]School of Mathematics, Institute for Advanced Study, New Jersey
* Equal contribution

ICML | 2022

Thirty-ninth International Conference on Machine Learning

Introduction

The goal of eXplainable Artificial Intelligence (XAI) is to make AI decision **understandable to humans**.

Current state of XAI:

- √ Techniques to generate explanations (many)
- √ Analysis of the techniques (increasingly more)
- √ Empirical validations of the techniques (increasingly more)
- X How humans interpret the explanations given (non-existent)

We offer a theory of how humans interpret XAI explanations.

The core idea: Humans **project their beliefs onto the AI**; thus, they interpret the explanation provided by **comparing it to the explanations that they themselves would give**.

Formulation

We formulate and test this theory in the context of using saliency maps to explain image classifiers. We measure humans' interpretation behavior by asking participants to predict the AI classification given the explanation (**Explanation condition**). The goal of the theory is to predict the human responses: P(c|**e**,**x**), the human prediction of AI classification *c* of image **x** given explanation **e**.

The model prediction can be expressed via **Bayes' rule**:

$$P(c \mid \mathbf{e}, \mathbf{x}) \propto P(c \mid \mathbf{x})\; p(\mathbf{e} \mid c, \mathbf{x})$$

The likelihood is the probability that the explainee themself would provide the observed explanation **e** as the explanation for assigning class *c* to image **x**. It follows **Shepard's universal law of (monotonic) generalization** [1]:

$$p(\mathbf{e} \mid c, \mathbf{x}) = \lambda \exp[-\lambda\ (1 - sim[\mathbf{e}(c, \mathbf{x})\ ,\ \mathbf{e}'(c, \mathbf{x})])]$$

Comparison occurs in a psychologically plausible (feature) space [2]:

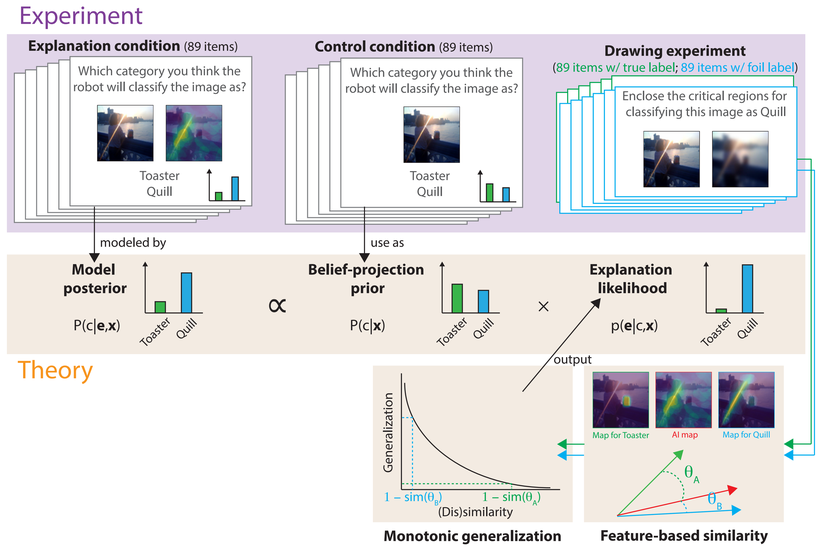
$$sim[\mathbf{e}(c, \mathbf{x})\ ,\ \mathbf{e}'(c, \mathbf{x})] = \frac{\langle \mathbf{e}, \mathbf{e}' \rangle}{\|\mathbf{e}\|_2 \|\mathbf{e}'\|_2}$$

The prior P(c|**x**) is measured through the **Control condition**. The **e** refers to the AI saliency map; the **e'** refers to the self-generated map, measured through the **Drawing experiment**.

[1] Shepard, R. N. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

[2] Sloman, S. A. Feature-based induction. *Cognitive psychology*, 25(2), 231-280, 1993.

Schematic of the approach



Hypotheses

H1. Participants will project their own beliefs onto the AI, resulting in low fidelity between human beliefs and AI behavior for trials when the AI is wrong (Chi-square test; p < .0001).

H2. Good explanations increase fidelity, especially when the original fidelity is low, i.e., here when AI is wrong (GLM: p < .001).

H3. Model prediction recovers H2 (GLM: p = .006).

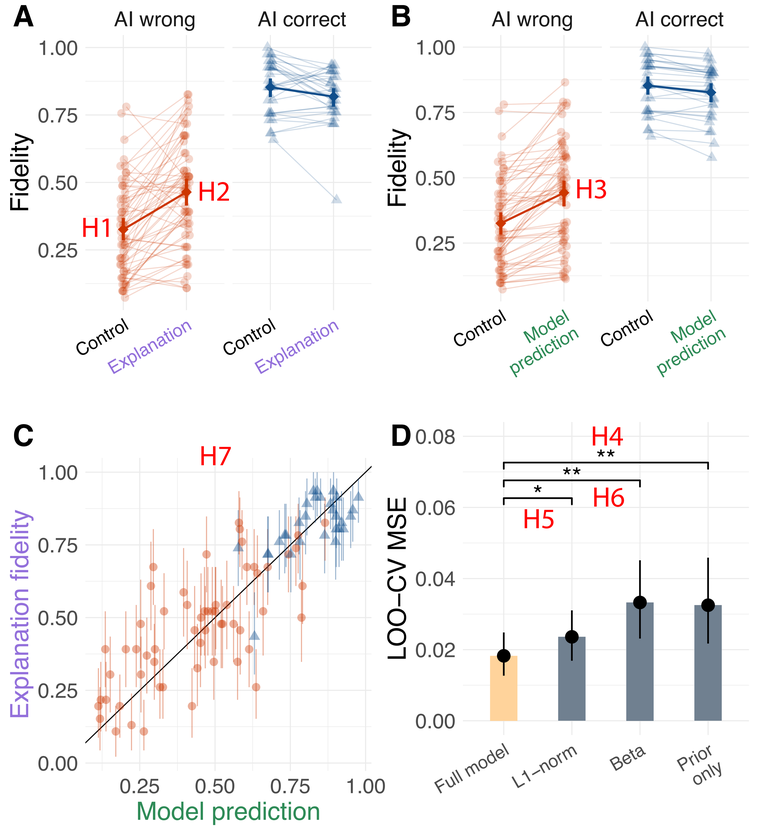
H4. The likelihood captures belief-updating from specific explanations, meaning that the full model is better than a prior-only model at predicting human behavior (LOO-CV MSE: p = .006).

H5. Comparison between explanations is done in a psychological space, implying that less-natural space (L1-norm) will be worse (LOO-CV MSE: p = .02).

H6. Generalization follows Shepard's universal law and decays monotonically with increasing psychological distance, implying that distributions that violate this decay (Beta(λ,λ)) will be worse (LOO-CV MSE: p = .003).

H7. The theory predicts human response well across a wide range of stimuli, classes, and explanations (Spearman ρ = .86; p < .0001).

Results



Conclusion

We have provided a simple, psychologically grounded, quantitative theory of human interpretation of explanations for AI systems. We note three broad implications: **(1)** Such a theory can improve the accuracy of explanation by integrating theories from cognitive science to better model human behavior. **(2)** A general theory of human inference from explanations reduces the need for validation experiments by virtue of being reusable across XAI methods. **(3)** A psychological theory of explanation improves understanding of explainability by better integrating the human and machine components of this problem.