A Psychological Theory of Explainability

Scott Cheng-Hsin Yang [* 1] **Tomas Folke** [* 1] **Patrick Shafto** [1 2]

Abstract

The goal of explainable Artificial Intelligence (XAI) is to generate human-interpretable explanations, but there are no computationally precise theories of how humans interpret AI generated explanations. The lack of theory means that validation of XAI must be done empirically, on a case-by-case basis, which prevents systematic theory-building in XAI. We propose a psychological theory of how humans draw conclusions from saliency maps, the most common form of XAI explanation, which for the first time allows for precise prediction of explainee inference conditioned on explanation. Our theory posits that absent explanation humans expect the AI to make similar decisions to themselves, and that they interpret an explanation by comparison to the explanations they themselves would give. Comparison is formalized via Shepard's universal law of generalization in a similarity space, a classic theory from cognitive science. A pre-registered user study on AI image classifications with saliency map explanations demonstrate that our theory quantitatively matches participants' predictions of the AI.

1. Introduction

Modern AI systems are applied to high-impact domains, such as medicine (Pesapane et al., 2018) and finance (Gogas & Papadimitriou, 2021). These systems, powered by deep neural networks, are notoriously opaque (Adadi & Berrada, 2018), making supervision and safe deployment challenging (Guidotti et al., 2018). The field of explainable AI has produced many techniques for improving the legibility of AI decisions to human users and regulators (Gunning & Aha,

---
[*]Equal contribution   [1]Department of Mathematics and Computer Science, Rutgers University–Newark, New Jersey, USA [2]School of Mathematics, Institute for Advanced Study, New Jersey, USA. Correspondence to: Scott Cheng-Hsin Yang <scott.cheng.hsin.yang@gmail.com>, Tomas Folke <tomas.folke@gmail.com>.

2019; Arrieta et al., 2020). XAI has focused on developing new methods that show high performance on technical metrics related to faithfulness (Hooker et al., 2019; Yeh et al., 2019; Sundararajan et al., 2017) and explanation complexity (Blanc et al., 2021; Ribeiro et al., 2016; Qi et al., 2020), but there is no way to assess which methods will do well for a given use-case (Doshi-Velez & Kim, 2017). In other words, there is no theory of explainability.

The goal of XAI is for humans to understand a target AI system (Gunning & Aha, 2019; Miller, 2019; Doshi-Velez & Kim, 2017). This "understanding" can be formalized as congruence between the AI's input-output mapping, and the human mental model of that mapping. Good explanations shift the human mental model to achieve this congruence. As a consequence, a theory of explainability can be naturally formalized as Bayesian updating. The initial human (mis-) conception of the AI serves as a prior, and the explanations provided modify this prior via a likelihood function that captures human inferential processes. An alternative approach to explainability would be to estimate the relationship between explanation and human inference as a ML problem, for example, by training on saliency maps as input, and human inference of AI classification as output. However, such an approach does not model human inference explicitly, and as a result suffers from all the standard problems with black box models, such as unknown generalization properties, fragility to out-of-distribution observations, and challenges relating to debugging and auditing.

To avoid building black-boxes to explain black-boxes, a formal theory of explainability must be constrained by cognitive science. We propose that humans model AI systems the same way they model any other agent, allowing us to draw on psychological work on belief-formation, generalization, and theory-of-mind. Our theory states that people project their own beliefs onto the AI and update their beliefs based on how they generalize self-generated explanations to XAI explanations in a similarity space (Yang et al., 2021; Shepard, 1987; Sloman & Rips, 1998). We built a cognitive model formalizing these ideas, and compared its predictions to human inference in a user study. We asked users to infer AI classification on images given saliency-map explanations, and found that our image-level model predictions correlated strongly with user responses (Spearman's $\rho = .86$). Variations to the similarity space or the generalization law that

reduced the theory's psychological plausibility also harmed the model's predictive performance. Our theory quantitatively predicts human inferences from explanation, and can thus guide the development and deployment of explainable AI techniques. Furthermore, our results show that well-established constructs from cognitive science offer realistic user mental models for explainable AI, illustrating the value of bridging the two fields.

2. Related work

Calls for human-centered explainability. There has been a proliferation of survey papers of XAI techniques in recent years that have attempted to synthesize existing knowledge and suggest taxonomies of XAI methods (see Linardatos et al. (2021) for a recent example that lists and discusses most previous surveys). Some of the most influential surveys and position papers put human understanding at the core of their definition of explainability (Miller, 2019; Doshi-Velez & Kim, 2017; Adadi & Berrada, 2018), and bemoan the lack of mathematical formalization. In contrast, many attempts to formalize desiderata for XAI have focused on faithfulness, which concerns how accurately the explanation captures the behavior of the AI system to be explained (Hooker et al., 2019; Yeh et al., 2019; Li et al., 2021; Sundararajan et al., 2017), rather than human interpretability. In the absence of a formal theory, interpretability can only be evaluated by user studies (Hase & Bansal, 2020; Jesus et al., 2021; Adebayo et al., 2020), which are expensive and challenging to run. Because of the range of applications and the speed at which new XAI methods emerge, naive empiricism is insufficient to determine which XAI method is most appropriate for a specific problem. Theory is necessary to determine to what extent empirical findings generalize.

Practical issues with explanations. Empirical studies have shown that practitioners can over-interpret explanations to shepherd false beliefs or under-use explanations because of lack of contextualization (Kaur et al., 2020). Furthermore, explanations from different methods often disagree (Krishna et al., 2022), and some explanations might mislead experts,leading to mistakes and miscalibrated trust (Lakkaraju & Bastani, 2020). All these issues emphasize the importance of having a quantitative theory of how humans interpret explanations.

Interpretability and explanation sparsity. When researchers do consider the interpretability of explanations, they often treat it as synonymous with the sparsity of the explanation (Blanc et al., 2021; Ribeiro et al., 2016; Qi et al., 2020). While it is generally true that sparse explanations are typically preferable to more complex ones, it is a mistake to only consider sparsity when attempting to achieve interpretability. First, sparsity comes in many forms, and its impact on interpretability may vary depending on

the users' expertise, explanation presentation, and the task. While some studies have attempted to investigate the impact of various forms of sparsity on interpretability (Narayanan et al., 2018; Poursabzi-Sangdeh et al., 2021), the domain is too large for exhaustive experimentation. Second, there is often a trade-off between the sparsity of an explanation and its expressivity, so penalizing explanation complexity is not always beneficial (Doshi-Velez & Kim, 2017). Most importantly, an exclusive focus on sparsity ignores any inferential biases of the explainee, which is essential to how explanations are understood. An explicit model of explainee inference subsumes sparsity by targeting the quantity we truly care about: what human infers about the AI's input-output mapping based on the explanations. The lack of explicit models of explainee inference results in uncertainty about the effectiveness of XAI methods in the context of new domains.

Black-box models of inference from explanations. There has been some attempts to approximate the impact of explanations on human users, by simulating the user with black-box statistical techniques (Pruthi et al., 2022; Chen et al., 2021; Yang et al., 2021). Chen et al. (2021) provide a sophisticated example of this approach. They used supervised learning to teach an explainee model to associate explanations with labelled outcomes of interest, and then tested the explainee model on novel explanations generated by the same method. However, they only qualitatively compared their simulated results to actual user data (collected by other labs). Furthermore, even if they had shown that their explainee model predictions quantitatively matched human explainee inference, their model is itself a black-box, so it could not be used to inform better explanation generation, and the model's generalization properties to new contexts is uncertain. Yang et al. (2021) argued for explainee models informed by human psychology, but still used a neural net to model the explainee. Their empirical results also showed that their explainee model qualitatively predicted human judgements, but their explainee predictions only weakly correlate with the human data. Interestingly, while they argued that humans model the AI the same way as they model other agents, the authors did not implement these ideas in their formal model of the explainee.

3. Theory

3.1. Hypotheses

Here we introduce a quantitative theory of user inference of AI decisions from explanations. Our theory states that people project their own beliefs onto the AI, and update beliefs based on explanation via generalization in similarity space in order to predict AI behavior. Specifically, the belief updating involves comparing the observed explanation to the *projected* explanation a person would give to justify a

given judgement.

Theories of social cognition and neuroscience suggest that we model other agents by simulation (Gallese & Goldman, 1998; Buckner & Carroll, 2007; Schurz et al., 2021), and empirical evidence suggests that if we lack specific information about a person's preferences or beliefs, our initial assumption is that they will be similar to us (Tarantola et al., 2017; Suzuki et al., 2016; Harris et al., 2018). Evidence from human-computer interaction suggests that people treat computers as social agents (Nass & Moon, 2000), even attributing them personality (Nass et al., 1995). For these reasons, we hypothesize that human users will not model the AI as a completely unknown entity, but will rather project their own beliefs onto the AI system (Hypothesis 1). Successful explanations should inform this belief-projection so as to improve the fidelity between user beliefs and AI behavior when belief-projection is misleading compared to when it is not (Hypothesis 2) (Yang et al., 2021). Throughout the paper, fidelity refers to the probability that the human's prediction of AI classification matches the AI's actual classification. The effect of an explanation on human beliefs depends on how well projected explanations match observed explanations, which is hypothesized to be quantitatively predicted by generalization in feature similarity space (Hypothesis 3).

A successful explanation updates the explainee's belief of AI behavior towards the actual AI behavior. As such, inference from explanation is naturally modelled by Bayes' rule. The Bayesian formulation puts a natural minimum criterion that any successful theory of explanation needs to exceed: a successful cognitive model should capture humans' beliefs post-exposure to explanation better than a model that only relies on their prior beliefs. We hypothesize that our theory with a psychologically informed likelihood will match human inference from explanation better than a prior-only model that is based on human beliefs about AI classifications when no explanations are presented (Hypothesis 4).

The quantitative degree of belief update depends on generalization in similarity space (Goldstone & Son, 2012). The comparison between projected and observed explanations is assessed in similarity space. There has been extensive work on the mathematical form of psychological similarity judgements (Goldstone & Son, 2012). We use a symmetric, feature-based similarity measure proposed by Sloman (Sloman & Rips, 1998). To evaluate whether psychological plausibility of the theory's similarity space impacts its predictive performance, we contrast the Sloman similarity to L1 norm — a raw distance metric commonly used in computer science. We hypothesize that a model that compares projected and observed explanations in a psychologically natural similarity space will match human beliefs better than L1 distance would (Hypothesis 5).

Shepard showed that generalization between stimuli in simi-

larity space follows a monotonically decreasing, approximately exponential, function (Shepard, 1987). To test whether this law of generalization holds for inference from explanation, we contrast a model with an exponential likelihood to one with a beta-distribution where the parameters are constrained to violate monotonicity. We hypothesize that the monotonically decaying likelihood will capture human beliefs better than the non-monotonic alternative would (Hypothesis 6).

3.2. Formalism

Let c be a class of interest, and $\mathbf{x}$ be an input image with $W \times H$ pixels. The black-box AI model to be explained, $f(\mathbf{x}, c)$, takes a pair of $\mathbf{x}$ and c as inputs, and outputs the probability that x is an instance of c. A saliency-map XAI method, $g(\mathbf{x}, c, f)$, takes the triplet $\mathbf{x}$, c and f as inputs, and outputs an explanation $\mathbf{e} \in [0, 1]^{W \times H}$ that is the same size as $\mathbf{x}$. Each pixel of the saliency map explanation $\mathbf{e}$ represents the importance of that pixel in $\mathbf{x}$ to f's classification of $\mathbf{x}$ as c. The main quantity we set out to model is the human explainee's inference about the AI's output given the $\mathbf{x}$, c, $\mathbf{e}$, and the alternative class(es) that are involved, which we denote as $P(c \mid \mathbf{e}, \mathbf{x})$. Cast in a Bayesian framework, the explainee's inference can be expressed as a posterior:

$$P(c \mid \mathbf{e}, \mathbf{x}) = \frac{p(\mathbf{e} \mid c, \mathbf{x})P(c \mid \mathbf{x})}{\sum_{c'} p(\mathbf{e} \mid c', \mathbf{x})P(c' \mid \mathbf{x})}. \qquad (1)$$

The prior $P(c \mid \mathbf{x})$ is the explainee's inference of the AI's classification without any explanation. The likelihood $p(\mathbf{e} \mid c, \mathbf{x})$ is the probability that the explainee themself would provide the observed saliency map $\mathbf{e}$ as the explanation for assigning class c to image $\mathbf{x}$. The sum over c' includes the class of interest and the alternative class(es) in contrast.

Figure 1 depicts how the theory is constructed and validated. First, we instantiate the theory in a two-forced alternative choice (2AFC) task, limiting the alternative class to a single foil class. Using such as task, we measured the prior in a control condition where participants inferred the AI's classification without any explanation. Quantitatively, the prior $P(c \mid \mathbf{x})$ is set to be the average response across participants in the control condition for image $\mathbf{x}$. The likelihood is constructed from a set of drawing experiments that measured the explanations that the participants themselves would generate. These participant-generated explanations, $\mathbf{e}^P$, are then compared to the observed XAI-generated explanation $\mathbf{e}$ in a similarity space to produce the likelihood $p(\mathbf{e} \mid c, \mathbf{x})$. Finally, the prior and likelihoods are combined using Bayes' rule to output the posterior $P(c \mid \mathbf{e}, \mathbf{x})$. On an image-by-image base, the posterior is then compared against its experimental measurements, $P^*(c \mid \mathbf{e}, \mathbf{x})$, in an explanation condition, where participants' inferred the AI's classification with the XAI-generated explanation $\mathbf{e}$.

Likelihood construction. The intuition behind the construction of the likelihood is that humans interpret the observed explanation by comparing it to the explanations that they themselves would generate. If the observed explanation is similar to the explanation that they would generate for a particular class, the explanation will push the explainees' inference to favor that class. In order to quantify this intuition, we formalized this comparison with the law of generalization in a psychologically plausible similarity space.

Following Shepard's universal law of generalization (Shepard, 1987), we expect the likelihood to decay as a function of dissimilarity between the observed saliency map $\mathbf{e}$ and the projected saliency map $\mathbf{e}^P$ from the drawing experiment. For simplicity, we use the celebrated exponential form:

$$p(\mathbf{e} \mid c, \mathbf{x}) = \lambda \exp\left[-\lambda\left(1 - sim\left[\mathbf{e}(c, \mathbf{x}), \mathbf{e}^P(c, \mathbf{x})\right]\right)\right], \tag{2}$$

where $\lambda > 0$ is the only free parameter in our cognitive model, used to calibrate the rate at which generalization decays with dissimilarity. The saliency maps $\mathbf{e}$ and $\mathbf{e}^P$ are expressed as a function of c and $\mathbf{x}$ to clarify where the class and image contribute to the equation. Following Sloman (Sloman & Rips, 1998), we adopt a simple parameter-free form of similarity (cosine similarity) given by

$$sim\left[\mathbf{e}(c, \mathbf{x}), \mathbf{e}^P(c, \mathbf{x})\right] = \frac{\sum_i \mathbf{e}_i(c, \mathbf{x})\, \mathbf{e}_i^P(c, \mathbf{x})}{\sqrt{\sum_i \mathbf{e}_i(c, \mathbf{x})^2}\sqrt{\sum_i \mathbf{e}_i^P(c, \mathbf{x})^2}}, \tag{3}$$

where the i index the pixel of the saliency maps. See Appendix A.2 for details on the similarity calculations.

Ablation models. In order to evaluate whether the likelihood in the full cognitive model successfully captures belief-updating in response to specific observed saliency maps, we create a series of ablation models in which the likelihood deviates from psychological plausibility. The first ablation model is a **prior-only model** that excludes the likelihood term, such that the posterior is equal to the prior. The second model is an **L1-norm model** that replaces Sloman similarity in Equation 3 with the L1-norm distance $\sum_i |\mathbf{e}_i(c, \mathbf{x}) - \mathbf{e}_i^P(c, \mathbf{x})|$. The L1-norm model evaluates the impact of using a pixel-based comparison in L1 norm, which is believed to be less natural for humans than a feature-based comparison captured by the Sloman similarity. Lastly, the third model is a **Beta-distribution model** in which the exponential distribution of Equation 2 is replaced with a Beta distribution where its two shape parameters are set to equal each other. Shepard showed that generalization follows a monotonic decreasing function (Shepard, 1987); thus, the Beta-distribution model tests whether this law of generalization holds when linking participants' similarity judgements to their specific responses. The Beta distribution takes the form:
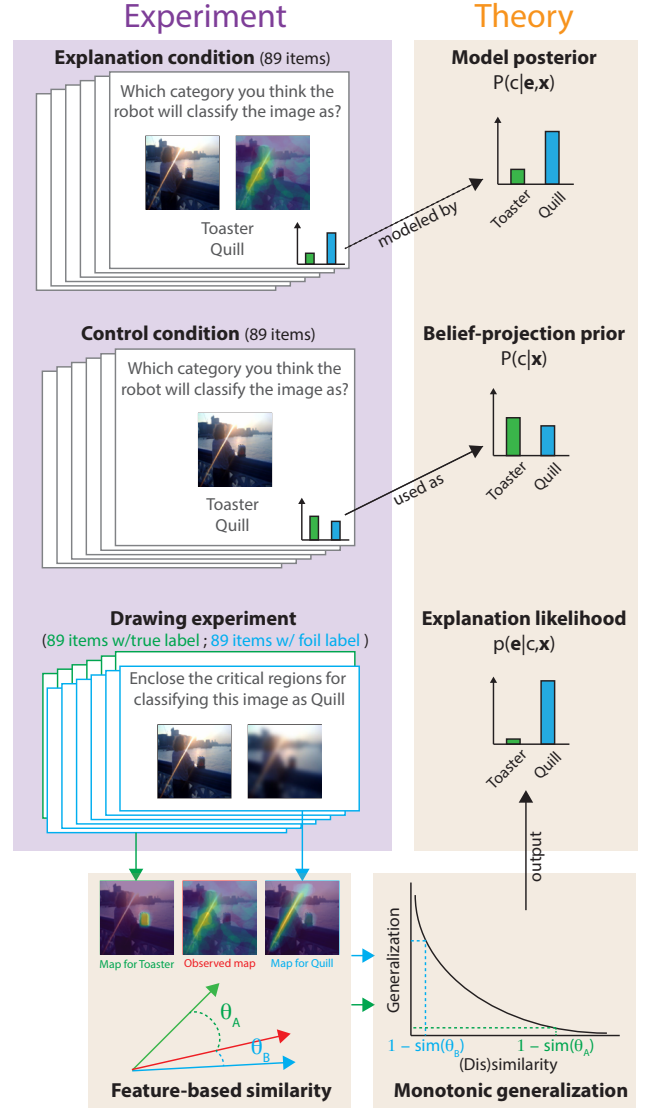

Figure 1. The relationship between the theory and experiments. Human interpretation of an explanation, $P^*(c \mid \mathbf{e}, \mathbf{x})$, is measured by the participants' responses when viewing a saliency map explanation (the classification experiment's explanation condition) and modeled by the theory's posterior, $P(c \mid \mathbf{e}, \mathbf{x})$. The participants' responses to the same stimuli absent explanation (the classification experiment's control condition) are taken to be the belief-projection prior of the cognitive model, $P(c \mid \mathbf{x})$. Different participants enclosed important regions of the same images, contingent on a given class (the drawing experiment). The regions of interest recorded in the drawing experiment are used to compute the explanation likelihood, $p(\mathbf{e} \mid c, \mathbf{x})$. The computation involves calculating how well the XAI-generated saliency map generalizes to the average participant-generated saliency maps in the feature-based similarity space. The figure illustrates a trial in which the explanation helped participants to shift their belief from favoring that the AI classified the image as Toaster to a strong and correct belief that the AI classified the image as Quill.

$$p(\mathbf{e} \mid c, \mathbf{x}) = \frac{\Gamma(2\lambda)}{\Gamma(\lambda)\Gamma(\lambda)} sim\big[\mathbf{e}(c,\mathbf{x})\,,\,\mathbf{e}^P(c,\mathbf{x})\big]^{\lambda-1} \qquad (4)$$
$$\times \big(1 - sim\big[\mathbf{e}(c,\mathbf{x})\,,\,\mathbf{e}^P(c,\mathbf{x})\big]\big)^{\lambda-1},$$

where $\Gamma(\cdot)$ is the gamma function, and $\lambda > 0$ is a free shape parameter. By restricting the two parameters of the Beta distribution to be equal ($\alpha = \beta = \lambda$), we force the likelihood to violate monotonicity and thus contrast against the consistently decaying exponential form. In Section 4.4 we detail how the hypotheses relate to the quantities introduced in this section.

4. Experiments

We tested our theory and formalized cognitive model in the domain of image classification. Participants saw an image $\mathbf{x}$ and were asked to report which class c they believed the AI would assign to $\mathbf{x}$ in a two-alternative-forced-choice task (see Figure 1). Each choice was between the ground-truth label of the image, and a foil label that was either the AI's classification (when the AI made a mistake) or the second-most likely category according to the AI (when the AI made a correct classification). Participants in this **classification experiment** were randomly assigned to one of two conditions: in the control condition they inferred the AI classification without seeing an explanation; in the explanation condition they made the same judgement but were also exposed to a saliency map explanation $\mathbf{e}$ that highlights regions of $\mathbf{x}$ that highly influence the AI's decision (Yang et al., 2021; Petsiuk et al., 2018). To estimate the projected saliency maps contingent on a specific classification $\mathbf{e}^P(c, \mathbf{x})$, we ran a **drawing experiment**. Participants were asked to enclose which regions of the image they thought were important for a given classification. This drawing experiment involved two between-subject conditions: one for the true labels and one for the foil labels. The average of these human-generated regions were taken to be the projected saliency maps that humans would use to interpret the observed saliency maps.

The classification experiment and the drawing experiment shared the same 89 images, with one image shown per trial. The AI, a ResNet-50 model (He et al., 2016), correctly classified 30 images and misclassified the remaining 59.

4.1. Classification Experiment

Task structure. Participants completed a 2AFC where they had to predict how the AI classified an image. The experiment consisted of two between-participant conditions: a *control* condition without explanation and an *explanation* condition with saliency map explanation.

Image selection. Participants viewed 89 images. These images belonged to three distinct types all sharing the same classes. The types were correctly classified images drawn from ImageNet (Russakovsky et al., 2015), misclassified images drawn from ImageNet, and misclassified images drawn from the Natural Adversarial ImageNet dataset (Hendrycks et al., 2021). The latter contains a 200-class subset of the original ImageNet 1000 classes. From these 200 classes, we selected 30 classes that span the spectrum of model accuracy based on the predictions of ResNet-50 on the normal ImageNet's validation set. We selected one image from each class, for each of our three image types. All images were randomly sampled from the chosen classes and datasets. In our pilot work, we found no clear distinctions between AI mistakes based on the adversarial images and the AI mistake based on the standard ImageNet images, so in this study we treat all mistake trials as a single type.

To form the 2AFC trials, we selected two possible labels for each image. One of these labels was always the ground truth label of the image, whereas the other was a foil. For correctly classified images, the foil was the class that ResNet-50 found most confusable according to the confusion matrix constructed on ResNet-50's predictions on normal ImageNet's validation set. For misclassified images, the foil was the class that ResNet misclassified the image as.

The above procedure produces 90 trials that specify the image, the ground-truth class, and the foil class. The original experiment that inspired this classification experiment involved both example images and saliency maps as explanations. One of the trials was excluded because no good explanatory examples were found, leaving a total of 89 trials. Since the symmetry of trial types is not central to the hypotheses we test in this study, we adopt that setup as is.

Saliency map generation. We used the method of Bayesian Teaching to generate saliency map explanation (Yang et al., 2021). The saliency value of each pixel of the explanation is related to the probability that the corresponding pixel in the image helps the AI to arrive at the targeted prediction. See Appendix A.1 for technical details.

Participants. The study protocol was approved by Rutgers University IRB. All research was performed in accordance with the approved study protocol. An IRB-approved consent page was displayed before the experiment. Informed consent was obtained from all participants prior to the start of the experiment. Our preregistered aim was to test participants until we had at least 40 participants per condition who passed our inclusion criteria. For the control condition we tested 50 participants and excluded 9, leaving us with a final sample of 41 participants. For the explanation condition, we tested 59 participants and excluded 13, leaving us with a final sample of 46.

Exclusion criteria. In line with our preregistration, we ex-

cluded participants based on total response time. We had different thresholds for different conditions since the explanation condition involved integrating more information than the control condition. For the control condition we excluded any participant that had a total response time less than three minutes, corresponding to approximately 2 seconds per item. For the explanation condition we excluded participants that took less than five minutes to complete the experiment, corresponding to an average response time of 3.4 seconds per item.

4.2. Drawing Experiment

Task structure. Participants were shown the 89 images from the classification experiment. They generated a mask for each image from that experiment, by drawing on a blurred version of the image to indicate what regions of the image was important for a given classification. The experiment consisted of two between-participant conditions depending on which class participants were asked to explain (ground-truth or foil). In the ground-truth condition, these instructions read: "enclose CATEGORY_NAME"; in the foil condition these instructions read: "enclose the critical regions you believe the robot attended to when determining the image contains CATEGORY_NAME".

Participants. The study protocol has been approved under the same IRB used for the classification experiment. Our aim was to include data from at least 40 participants per condition post exclusions. Because we had high attrition rates during our pilot, we tested 160 participants participants in the ground-truth condition and 80 participants in the foil condition. We excluded 95 from the ground-truth condition and 8 from the foil condition, leading to ultimate sample sizes of 65 and 72, respectively.

Exclusion criterion. In line with our preregistration, we excluded participants based on how much their masks deviated from the aggregate image-level masks across participants. The intuition behind this (validated in our pilot) is that people who pay attention to the image tend to report that the same regions as important, whereas people who do not pay attention deviate from this consensus. We quantified how much participants deviate from the consensus on an image by the following steps: First, for each participant compute the L2 norm between the aggregate and the participant's specific mask. Second, compute the participant-level mean of these L2 norms. Third, compute a Z-score for each participant using the participant-wise mean and assuming a half-normal distribution. Any participant with a Z-score greater than 1.5 is excluded for that image. The L2 norm, the participant-wise means, and the Z-scores are then recomputed on the post-exclusion data, and this process is repeated until all participants have a Z-score less than or equal to 1.5. We arrived at this threshold from exploratory

analysis of our pilot data.

Preregistration. The classification experiment, the drawing experiment, the mathematical models, and our hypothesis tests were all preregistered prior to collecting the data for the results presented in the paper (link to pre-registration to be provided upon acceptance or reviewer request; the preregistration document and our analysis code has been anonymized and uploaded in the supplementary folder). We only deviated from our preregistration in one respect: we used a more generous inclusion criterion for the drawing experiment. Using our preregistered threshold reduced our sample size, but did not otherwise impact our results, see Appendix A.4 for details.

4.3. Model evaluation

Fitting model parameter. To fully specify the models introduced in Section 3.2, we need to assign values to λ in Equations 2 and 4. We select λ values for these models by minimizing the sum of the squared error. The error term was computed on the image-level between the model posterior probabilities and the empirical responses from the explanation condition in the classification experiment.

Leave-one-out cross-validation. For the final three hypotheses, we expected that a cognitive model with psychologically informed components — prior belief-projection, feature-based similarity, and monotonic generalization — would match human responses in the explanation condition better than alternatives would. To compare the predictive performance of the full model to the alternatives, we used leave-one-out cross-validation (LOO-CV) to control for model complexity. For cognitive models with a λ parameter, we fitted the model on 88 of the 89 trials, and used the obtained λ to compute the squared error between the fitted model and the remaining data point. For each left-out item, we computed the squared error between the model's posterior and the empirical response from the explanation condition in the classification experiment (LOO-CV MSE). We then used paired t-tests to evaluate if the LOO-CV MSE was significantly lower for the full model than the ablated alternatives.

4.4. Statistical analysis strategy

Hypothesis 1. Our first hypothesis was that participants' prior belief of the AI classification, $P(c \mid \mathbf{x})$, would not be uniform, but that the participants would expect the AI to make similar classifications to themselves. Because humans tend to find this type of image classification easy, the result of this belief projection would be that participants in the control condition of the classification experiment would pick the ground truth label more often than would be expected by chance (Yang et al., 2021). We tested this hypothesis with chi-squared test on the proportions of participant responses

that matched the ground truth.

Hypothesis 2. Our second hypothesis was that successful explanations would update the belief-projection prior by highlighting the discrepancy between what features participants expected the AI to attend to and the features it actually found important. Because this discrepancy is generally larger when the AI makes a mistake, we hypothesized that explanations would help participants identify AI mistakes. However, sometimes the AI will attend to strange features even when it is correct, meaning that explanations would be less helpful (or possibly harmful) for identifying correct AI classifications.

To evaluate this hypothesis, we computed the average image-level fidelity between the participants' predictions of the AI and the AI's actual classifications for each of the experimental conditions, resulting in 178 observations. In terms of the quantities introduced in Section 3.2, the fidelity in the control and explanations conditions are $P(c = c_{AI} \mid \mathbf{x})$ and $P^*(c = c_{AI} \mid \mathbf{e}, \mathbf{x})$, respectively, where c_{AI} is the AI's classification of $\mathbf{x}$. We then performed statistical analysis on the image-level fidelity using linear regression:

$$Fidelity \sim Normal(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_1 A_i + \beta_2 E_i + \beta_3 (A_i \times E_i). \quad (5)$$

Here, A_i is a binary independent variable coded as 1 if the AI correctly classified the image of that trial, and 0 otherwise. E_i is another binary independent variable coded as 1 if the response belonged to the explanation condition, and 0 if it belonged to the control condition. i corresponds to a single image in one of the conditions and ranges from 1 to 178, which is the total number of images for both conditions.

Hypothesis 3. Our third hypothesis was that our cognitive model would capture the change in fidelity from the control to the explanation condition, depending on AI correctness as outlined for hypothesis 2, despite the model being blind to whether the AI made a mistake on a given trial.

To evaluate whether our cognitive model could qualitatively recover the empirical patterns, we ran a variation of the regression analysis for Hypothesis 2, where we replaced the fidelity of participant choices with the fidelity of our cognitive model's predictions for the explanation trials; that is, we analyzed the model-generated $P(c = c_{AI} \mid \mathbf{e}, \mathbf{x})$ instead of the measured $P^*(c = c_{AI} \mid \mathbf{e}, \mathbf{x})$.

Hypothesis 4. Our fourth hypothesis was that the likelihood in our cognitive model captures belief-updating from specific explanations, meaning that the full model should reliably outperform the prior-only model. Performance is captured by the match between model predictions and human responses in the explanation condition. That is, we expect the MSE between $P^*(c \mid \mathbf{e}, \mathbf{x})$ and $P(c \mid \mathbf{e}, \mathbf{x})$ to be smaller than that between $P^*(c \mid \mathbf{e}, \mathbf{x})$ and $P(c \mid \mathbf{x})$.

Hypothesis 5. Our fifth hypothesis was that our full model, which used symmetric Sloman similarity, would match human responses better than the L1-norm model would. That is, we expect the MSE between $P^*(c \mid \mathbf{e}, \mathbf{x})$ and $P(c \mid \mathbf{e}, \mathbf{x})$ to be smaller than that between $P^*(c \mid \mathbf{e}, \mathbf{x})$ and $P_{\mathrm{L1}}(c \mid \mathbf{e}, \mathbf{x})$, where $P_{\mathrm{L1}}(c \mid \mathbf{e}, \mathbf{x})$ denotes the posterior obtained from the L1-norm model. In this context, L1-norm captures dissimilarity as the sum of the unsigned pixel-wise differences, which serves as a good foil since pixels are too granular to be psychologically meaningful features.

Hypothesis 6. Finally, our sixth hypothesis was that translating the similarity judgement between the observed explanation and the projected explanation to a response about AI classification is determined by the probability that the response to the projected explanation would generalize to the observed explanation. Shepard compellingly demonstrated that generalization decays monotonically, so we hypothesized that a generalization distribution that obeys this law would capture the human responses better than a distribution that does not. To formally test this, we compared our full model — that used an exponential distribution — to the Beta-distribution model that used a likelihood that was constrained to violate monotonicity. We expect the MSE between $P^*(c \mid \mathbf{e}, \mathbf{x})$ and $P(c \mid \mathbf{e}, \mathbf{x})$ to be smaller than that between $P^*(c \mid \mathbf{e}, \mathbf{x})$ and $P_{\mathrm{Beta}}(c \mid \mathbf{e}, \mathbf{x})$, where $P_{\mathrm{Beta}}(c \mid \mathbf{e}, \mathbf{x})$ denotes the posterior from the Beta-disribution model.

Not preregistered analysis. Though not a preregistered hypothesis, we wanted to test how well the model posteriors $P(c \mid \mathbf{e}, \mathbf{x})$ matched the empirical data $P^*(c \mid \mathbf{e}, \mathbf{x})$ from the explanation condition on an image-by-image level. We tested this with Spearman correlation between the fidelity of the empirical data and the fidelity of the model posteriors.

5. Results

All of our preregistered hypotheses were supported (see Appendix A.3 for details). First, absent explanation participants responded that the AI would correctly classify the image in 73% of the trials ($\chi^2 = 802.28$, $p < .0001$), which is consistent with belief-projection because it implies that participants expected the AI to get most trials right in a task that they themselves find easy. Second, explanations improved the fidelity between participant responses and AI classifications when the AI makes a mistake ($\beta = 0.14$, SE $= 0.03$, $t = 4.69$, $p < .0001$; see Figure 2A), and the impact of explanations on fidelity was reduced when the AI is correct ($\beta = -0.17$, SE $= 0.05$, $t = -3.40$, $p < .001$). Third, our model predictions qualitatively match the empirical data, as our cognitive model also predicts that explanations will increase fidelity on mistake trials ($\beta = 0.12$, SE $= 0.03$, $t = 3.90$, $p < .001$), and that the impact of explanations should be less pronounced when the AI is correct ($\beta = -0.14$, SE
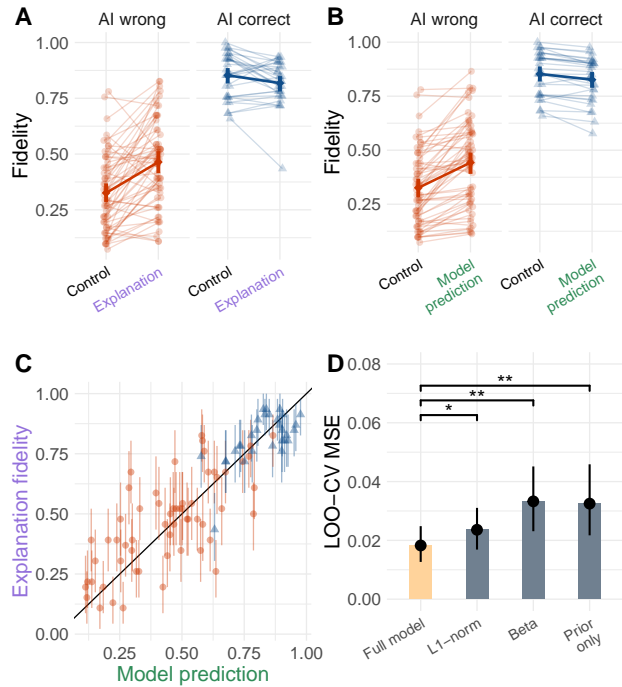
Figure 2. **(A)** Experimental results from the classification experiment. Explanations increase the fidelity between participant responses and AI classifications on trials when the AI is wrong, but slightly decrease fidelity when the AI is correct. Semi-transparent points show image-level data; solid points show condition-level means. **(B)** Model predictions of participant responses. Our cognitive model recovers the qualitative patterns of the empirical data. **(C)** Scatter plot between the empirical image-level fidelity and our model predictions. **(D)** Leave-One-Out Cross-Validation MSE for our model compared to ablated alternatives. All error bars show bootstrapped 95% confidence intervals.

Figure 3. Examples of XAI saliency maps and participant drawn maps. Title of each column show the truth and foil class presented in the classification experiment for that image. The * indicates the AI classification; thus, first (last) two columns correspond to AI being correct (incorrect).

= 0.05, $t = -2.77$, $p = .006$; see Figure 2B). To obtain a general estimate of the model effectiveness in predicting human judgments, we ran a (not preregistered) Spearman correlation between fidelity based on the empirical data and fidelity based on the model predictions, which was statistically significant (Spearman's $\rho = .86$, $p < .0001$; see Figure 2C).

Moreover, model comparisons demonstrate the importance of the psychologically informed components of our cognitive model as predicted in hypotheses 4-6. Fourth, the full model predicts aggregate participant responses significantly better than a prior-only model (LOO-CV MSE difference[95% CI] = 0.014[0.004-0.024], $t(88) = 2.84$, $p = .006$; see Figure 2D), implying that our likelihood function captured explanation-specific belief-updating. Fifth, the full model outperforms the L1-norm model (LOO-CV MSE difference[95% CI] = 0.005[0.001-0.010], $t(88) = 2.33$, $p = .02$), implying that the psychological plausibility of the

similarity space improves predictive accuracy. Sixth, the full model outperformed the Beta-distribution model (LOO-CV MSE difference[95% CI] = 0.015[0.005-0.025], $t(88) = 2.96$, $p = .003$), consistent with the monotonic decay in generalization behavior being present in interpreting XAI explanations. Table 1 summarizes the hypotheses and statistical results.

Comparing Figure 2A and B, we note that the cognitive model captured the effect of the explanations over the whole range of explanation quality tested. Firstly, the XAI saliency maps varied in visual quality, with some being more diffuse than others (Figure 3). The variation in quality is also apparent from the wide spectrum of fidelity and in the change in fidelity from the control to the explanation conditions being both positive and negative (mean=8%, SD=17%; Figure 2A). In particular, some explanations were poor and resulted in a reduction in fidelity (33/89 images). The full model could capture the change in fidelity across all these variations, including explanations of poor quality.

6. Conclusion

Systematic, scientific efforts enable the goal of safe, effective, and ethical AI by providing generalizable theories that inform development of effective explanations across domains. We have provided a simple, psychologically

Table 1. Summary of hypotheses and results. See Section 4.4 for details on the hypotheses.

Hypothesis	Test	P-value
1) Participants expect the AI to mostly be correct	Chi-square	$< .0001$
2) Participants find explanations to be more helpful when the AI is wrong ($\beta_3 < 0$)	GLM	$< .001$
3) Full cognitive model recovers Hypothesis 2	GLM	.006
4) Full cognitive model matches participants better than the prior-only model	LOO-CV MSE	.006
5) Full cognitive model matches participants better than the L1-norm model	LOO-CV MSE	.02
6) Full cognitive model matches participants better than the Beta-distribution model	LOO-CV MSE	.003
Modelled fidelity matches participants' measured fidelity	Spearman correlation	$< .0001$ ($\rho = .86$)

grounded, quantitative model of human interpretation of explanations for AI systems. Our theory proposes that humans reason about AI systems in the same way they reason about human agents: by projecting beliefs. Human predictions of AI behavior contingent on explanations are based on generalization functions in similarity space, consistent with those identified in cognitive science. Our theory is, therefore, general and can be applied and tested across XAI methods and application domains.

7. Broader impacts

Most previous attempts at formalizing explainability has focused on general qualities inherent to explanations, but ignored the agents who actively interprets the explanations. For example, XAI methods prefer less complex explanations to not incur a heavy cognitive load, but do not account for the explainee's inferential biases. Here, we have argued for a more specific formalization of explainability that considers and exploits the distinct nature of human belief-updating. To the best of our knowledge, our theory is the first empirically validated, mathematical formulation of how humans infer AI judgements from provided explanations.

The work presented here can be extended in several directions for wider applicability and scalability. First, the task we used to evaluate user understanding focused on predicting AI image classification. This is foundational to — but a few steps removed from — real-world applications of XAI such as auditing and debugging (Kaur et al., 2020). Further cognitive and behavioral modelling of inference from explanation in the context of such applications can help bridge this gap. Second, our work focuses on saliency maps, whereas many real-world applications of attribution methods attempt to explain tabular and text data. Our framework can be extended to abstract, non-visual features encoded in tabular and text data by drawing on the large psychological literature on similarity judgements in different domains. Empirical validation of our theory in these contexts requires further research. Third, our framework currently relies on measurements of human generated explanations to predict the interpretability of AI explanations, which limits scalability. If approximations of human generated explanations could be automatically generated, the utility of our theory for solving practical XAI problems would increase.

Lastly, we note three broad implications of a psychological theory of explanation: (1) Such a theory improves the accuracy of explanation by integrating theories from cognitive science to better model human behavior. Having a formal framework that captures how humans interpret explanations allows researchers to develop ML methods that leverage the mathematical template to better reason about and accommodate human partners in human-machine teams. (2) A general theory of human inference from explanations reduces the need for validation experiments by virtue of being reusable across XAI methods. To give an example, our theory can be used to seed experiments validating other saliency methods by indicating explanations that are likely to be successful. (3) A psychological theory of explanation improves understanding of explainability by better integrating the human and machine components of this problem. ML researchers can leverage the constraints in the explanation space highlighted by our theory to design machines that could avoid explanations that are unlikely to be informative to human users.

Software and Data

All experiments, mathematical models, analysis code, and hypothesis tests were preregistered (https://osf.io/4n67p). User study data that support the findings of this study are available upon request.

Acknowledgements

References

Adadi, A. and Berrada, M. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

Adebayo, J., Muelly, M., Liccardi, I., and Kim, B. Debugging tests for model explanations. *Advances in Neural Information Processing Systems*, 33:700–712, 2020.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

Blanc, G., Lange, J., and Tan, L.-Y. Provably efficient, succinct, and precise explanations. *Advances in Neural Information Processing Systems*, 34, 2021.

Buckner, R. L. and Carroll, D. C. Self-projection and the brain. *Trends in cognitive sciences*, 11(2):49–57, 2007.

Chen, V., Plumb, G., Topin, N., and Talwalkar, A. Simulated user studies for explanation evaluation. In *Workshop at NeurIPS2021: eXplainable AI approaches for debugging and diagnosis.*, 2021.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Gallese, V. and Goldman, A. Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12):493–501, 1998.

Gogas, P. and Papadimitriou, T. Machine learning in economics and finance. *Computational Economics*, 57(1): 1–4, 2021.

Goldstone, R. L. and Son, J. Y. *Similarity.* Oxford University Press, 2012.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Gunning, D. and Aha, D. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40(2):44–58, 2019.

Harris, A., Clithero, J. A., and Hutcherson, C. A. Accounting for taste: A multi-attribute neurocomputational model explains the neural dynamics of choices for self and others. *Journal of Neuroscience*, 38(37):7952–7968, 2018.

Hase, P. and Bansal, M. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

Hooker, S., Erhan, D., Kindermans, P.-J., and Kim, B. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., and Gama, J. How can i choose an explainer? an application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 805–815, 2021.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–14, 2020.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Lakkaraju, H. and Bastani, O. " how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020.

Li, X.-H., Shi, Y., Li, H., Bai, W., Cao, C. C., and Chen, L. An experimental study of quantitative evaluations on saliency methods. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3200–3208, 2021.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2021.

Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

Nass, C. and Moon, Y. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1): 81–103, 2000.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. C. Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, 43 (2):223–239, 1995.

Pesapane, F., Codari, M., and Sardanelli, F. Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine. *European radiology experimental*, 2(1):1–10, 2018.

Petsiuk, V., Das, A., and Saenko, K. RISE: Randomized Input Sampling for Explanation of Black-box Models. *29th British Machine Vision Conference*, 2018.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–52, 2021.

Pruthi, D., Bansal, R., Dhingra, B., Soares, L. B., Collins, M., Lipton, Z. C., Neubig, G., and Cohen, W. W. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022.

Qi, Z., Khorram, S., and Fuxin, L. Visualizing deep networks by optimizing with integrated gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11890–11898, 2020.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Schurz, M., Radua, J., Tholen, M. G., Maliske, L., Margulies, D. S., Mars, R. B., Sallet, J., and Kanske, P. Toward a hierarchical model of social cognition: A neuroimaging meta-analysis and integrative review of empathy and theory of mind. *Psychological Bulletin*, 147(3): 293, 2021.

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

Sloman, S. A. and Rips, L. J. Similarity as an explanatory construct. *Cognition*, 65(2-3):87–101, 1998.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Suzuki, S., Jensen, E. L., Bossaerts, P., and O'Doherty, J. P. Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proceedings of the National Academy of Sciences*, 113(14):3755–3760, 2016.

Tarantola, T., Kumaran, D., Dayan, P., and De Martino, B. Prior preferences beneficially influence social and non-social learning. *Nature communications*, 8(1):1–14, 2017.

Yang, S. C.-H., Vong, W. K., Sojitra, R. B., Folke, T., and Shafto, P. Mitigating belief projection in explainable artificial intelligence via bayesian teaching. *Scientific reports*, 11(1):1–17, 2021.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32:10967–10978, 2019.

A. Appendix

A.1. Saliency map generation through Bayesian Teaching

Following previous work (Yang et al., 2021), we generated saliency maps by using Bayesian Teaching to select pixels of an image that help a learner model to arrive at the targeted prediction. Let $q_{teacher}(\mathbf{m} \mid c, \mathbf{x})$ be the probability that a mask $\mathbf{m}$ will lead the learner model to predict the image $\mathbf{x}$ to be in class c when the mask is applied to the image. This is expressed by Bayes' rule as

$$q_{teacher}(\mathbf{m} \mid c, \mathbf{x}) = \frac{Q_{learner}(c \mid \mathbf{x}, \mathbf{m})p(\mathbf{m})}{\int_{\Omega_{\mathbf{M}}} Q_{learner}(c \mid \mathbf{x}, \mathbf{m}')p(\mathbf{m}')}.$$

Here, $Q_{learner}(c \mid \mathbf{x}, \mathbf{m})$ is the probability that the ResNet-50 model with pre-trained ImageNet weights will predict the $\mathbf{x}$ masked by $\mathbf{m}$ to be c; $p(\mathbf{m})$ is the prior probability of $\mathbf{m}$; and $\Omega_{\mathbf{M}} = [0,1]^{W \times H}$ is the space of all possible masks on an image with $W \times H$ pixels. We used a sigmoid-function squashed Gaussian process prior for $p(\mathbf{m})$.

Instead of sampling the saliency maps directly from the above equation, we find the expected saliency map for each image by Monte Carlo integration:

$$\begin{aligned}
\mathrm{E}[\mathbf{M} \mid \mathbf{x}, c] &= \int_{\Omega_{\mathbf{M}}} \mathbf{m}\, q_{teacher}(\mathbf{m} \mid c, \mathbf{x}) \\
&\approx \frac{\sum_{i=1}^{N} \mathbf{m}_i\, Q_{learner}(c \mid \mathbf{x}, \mathbf{m}_i)}{\sum_{i=1}^{N} Q_{learner}(c \mid \mathbf{x}, \mathbf{m}_i)},
\end{aligned} \tag{6}$$

where $\mathbf{m}_i$ are samples from the prior distribution $p(\mathbf{m})$, and $N = 1000$ is the number of Monte Carlo samples used. The expected mask is used as the saliency map explanation $\mathbf{e}$.

To generate the saliency map $\mathbf{e}$ for an image $\mathbf{x}$, we first resized $\mathbf{x}$ to be 224-by-224 pixels. A set of 1000 2D functions were sampled from a 2D Gaussian process (GP) with an overall variance of 100, a constant mean of -100, and a radial-basis-function kernel with length scale 22.4 pixels in both dimensions. The sampled functions were evaluated on a 224-by-224 grid, and the function values were mostly in the range of $[-500, 300]$. A sigmoid function, $1/(1 + \exp(-a))$, was applied to the sampled functions to transform each of the function values a to be within the range $[0, 1]$. This resulted in 1000 masks. The mean of the GP controlled how many effective zeros there were in the mask, and the variance of the GP determined how fast neighboring pixel values in the mask changed from zero to one. The 1000 masks were the $\mathbf{m}_i$'s in Equation 6. We produced 1000 masked images by element-wise multiplying the image $\mathbf{x}$ with each of the masks. The term $Q_{learner}(c \mid \mathbf{x}, \mathbf{m}_i)$ was the ResNet-50's predictive probability that the ith masked image was in class c. Having obtained these predictive probabilities, we averaged the 1000 masks according to Equation 6 to produce the saliency map $\mathbf{e}$ of image $\mathbf{x}$.

A.2. Similarity calculations

The likelihood function of our cognitive model takes as input the similarity between the observed AI saliency map and the image-level aggregates of the participant-generated maps for the ground truth and foil classes. Therefore, before we explain the similarity calculations, we discuss the mask aggregation and scaling. For the aggregation, we compute the pixel-level mean across all participants who passed the exclusion thresholds outlined above. For the Sloman similarity (Sloman & Rips, 1998), we then min-max scale both the observed AI saliency maps and the aggregated participant-generated masks, so that the highest pixel value in each map/mask is 1, and the lowest pixel value is 0. For the L1 norm, we scale the observed maps and participant-generated masks so that all pixel-values for each map/mask sums to one. We compute separate similarity measures by comparing each of the observed AI saliency maps with the participant-generated masks for the ground-truth label and the foil label.

The Sloman similarity captures the overlap between features (regions important for the classification) in the observed AI saliency map and those in the participant-generated masks (Sloman & Rips, 1998). Specifically, the Sloman similarity takes this form:

$$sim\big[\mathbf{e}(c, \mathbf{x}),\, \mathbf{e}^P(c, \mathbf{x})\big] = \frac{\sum_i \mathbf{e}_i(c, \mathbf{x})\mathbf{e}_i^P(c, \mathbf{x})}{\sqrt{\sum_i \mathbf{e}_i(c, \mathbf{x})^2}\sqrt{\sum_i \mathbf{e}_i^P(c, \mathbf{x})^2}},$$

where $\mathbf{e}$ is the observed AI saliency map, and $\mathbf{e}^P$ is the aggregated human-generated mask, which is referred to as the projected map in the main text. The i indexes the pixels in the map/mask. The numerator captures feature overlap as the intersection between highly salient regions. The denominator normalizes this area of intersection to be between [0,1].

We compute the L1 norm dissimilarity by summing the absolute pixel-wise difference between the observed AI saliency map and the participant-generated mask:

$$\sum_i |\mathbf{e}_i(c, \mathbf{x}) - \mathbf{e}_i^P(c, \mathbf{x})|,$$

where c indexes the ground-truth or foil class.

A.3. Detailed statistical results

A.3.1. HYPOTHESIS 1

In line with our preregistered hypothesis, participants believed that the AI would correctly classify the image in 73% of the trials, which is significantly different from chance according to a chi-squared test (χ^2 = 802.28, $p < .0001$).

A.3.2. HYPOTHESIS 2

In our preregistration we hypothesized that β_2 would be significantly positive, and that β_3 would be significantly negative. Both of these hypotheses were supported, meaning that we found that explanations improved the fidelity between participant responses and AI classifications when the AI makes a mistake (β_2 = 0.14, SE = 0.03, t = 4.69, $p < .0001$), and that the impact of explanations on fidelity was reduced when the AI is correct (β_3 = -0.17, SE = 0.05, t = -3.40, $p < .001$). See Table 2.

Label	Parameter	Coefficient (SE)
Intercept	β_0	0.33***(0.02)
AICorrect	β_1	0.53***(0.04)
ExplanationCondition	β_2	0.14***(0.03)
AICorrect $\times$ ExplanationCondition	β_3	-0.17***(0.05)
R^2		0.65
Adj. R^2		0.64
I		178

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 2. Image-level empirical explanation results. AICorrect and ExplanationCondition correspond to A_i and E_i in Equation 5, respectively.

A.3.3. HYPOTHESIS 3

Our preregistered hypothesis was that these results should match the empirical results in the explanation condition. Specifically, β_2 should be significantly positive, and β_3 should be significantly negative. As hypothesized, the model posteriors recover the effect of explanations on fidelity during mistake trials (β_2 = 0.12, SE = 0.03, t = 3.90, $p < .001$), as well as the moderation effect between explanations and AI correctness (β_3 = -0.14, SE = 0.05, t = -2.77, $p = .006$). See Table 3.

A.3.4. HYPOTHESIS 4

In line with our hypothesis, the full model significantly outperforms a prior-only model (LOO-CV MSE difference [95% CI] = 0.014[0.004-0.024], $t(88)$ = 2.84, $p = .006$).

A.3.5. HYPOTHESIS 5

In line with our hypothesis, our full model outperforms the L1-norm model (LOO-CV MSE difference [95% CI] = 0.005[0.001-0.010], $t(88)$ = 2.33, $p = .02$).

Label	Parameter	Coefficient (SE)
Intercept	β_0	$0.33^{***}(0.02)$
AICorrect	β_1	$0.53^{***}(0.04)$
ExplanationCondition	β_2	$0.12^{***}(0.03)$
AICorrect $\times$ ExplanationCondition	β_3	$-0.14^{**}(0.05)$
R^2		0.65
Adj. R^2		0.65
I		178

$^{***}p < 0.001; {}^{**}p < 0.01; {}^{*}p < 0.05$

Table 3. Image-level modelled explanation results. AICorrect and ExplanationCondition correspond to A_i and E_i in Equation 5, respectively.

A.3.6. HYPOTHESIS 6

In line with our hypothesis, our full model outperformed the Beta-distribution model (LOO-CV MSE difference [95% CI] = 0.015[0.005-0.025], $t(88)$ = 2.96, p = .003), suggesting that the monotonic decay in generalization behavior is also observed in interpreting XAI explanations.

A.3.7. NOT PREREGISTERED ANALYSIS

The Spearman correlation between fidelity based on the empirical data and fidelity based on the model posteriors was statistically significant (Spearman's ρ = .86, p < .0001).

A.4. Statistical analysis: Pre-registered exclusions

Upon completing data collection we learned that our exclusion criterion ($Z > 1.2$) for the drawing experiment data had been too strict, leading us to exclude 82% of participants for the ground-truth condition and 65% of participants for the foil condition. We adapted a more generous threshold ($Z > 1.5$) that allowed us to meet our sample size targets and report the results from these larger samples in the main text.

Below are the results of the analyses on only the data that met the preregistered inclusion threshold. Because the inclusion criteria only differed for the drawing experiment, the results for Hypothesis 1 and 2 are exactly the same as the previous section.

A.4.1. HYPOTHESIS 3

See Table 4.

Label	Parameter	Coefficient (SE)
Intercept	β_0	$0.33^{***}(0.02)$
AI Correct	β_1	$0.53^{***}(0.04)$
Explanation Condition	β_2	$0.11^{***}(0.03)$
AI Correct $\times$ Explanation Condition	β_3	$-0.13^{*}(0.05)$
R^2		0.65
Adj. R^2		0.65
I		178

$^{***}p < 0.001; {}^{**}p < 0.01; {}^{*}p < 0.05$

Table 4. Image-level modelled explanation results based on preregistered exclusion thresholds. AICorrect and ExplanationCondition correspond to A_i and E_i in Equation 5, respectively.

A.4.2. HYPOTHESIS 4

MSE difference [95% CI] = 0.014[0.005-0.024], $t(88)$ = 2.98, p = .003.

A.4.3. HYPOTHESIS 5

MSE difference [95% CI] = 0.007[0.001-0.014], $t(88) = 2.38$, $p = .02$.

A.4.4. HYPOTHESIS 6

MSE difference [95% CI] = 0.014[0.004-0.024], $t(88) = 2.85$, $p = .005$.

A.4.5. NOT PREREGISTERED ANALYSIS

Spearman correlation at the trial level between empirical fidelity for the explanation condition and posterior fidelity of the cognitive model (Spearman's $\rho = .86$, $p < .0001$).