AFRL-RI-RS-TR-2022-159

MODEL EXPLANATION BY OPTIMAL SELECTION OF TEACHING EXAMPLES

RUTGERS UNIVERSITY

NOVEMBER 2022

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2022-159 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
PETER ROCCI
Work Unit Manager

/ S /
SCOTT PATRICK
Deputy Chief
Intelligence Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
NOVEMBER 2022	FINAL TECHNICAL REPORT	MAY 2022	JUNE 2022

4. TITLE AND SUBTITLE

Model Explanation by Optimal Selection of Teaching Examples

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-17-2-0146	N/A	62303E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R296

6. AUTHOR(S)

Patrick Shafto, Scott Cheng-Hsin Yang

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Rutgers University
101 Warren St.
Newark NJ 07103

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RIED 525 Brooks Road Rome NY 13441-4505	AFRL/RI	AFRL-RI-RS-TR-2022-159

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our contributions to XAI can be summarized by four lines of work: (1) developing novel XAI techniques, (2) demonstrating successful XAI systems with empirical validation, (3) advancing our understanding of the psychology of explainability, and (4) contributing to the mathematical foundations of XAI. The theme of our approach is to incorporate human inference into all aspects of XAI.

15. SUBJECT TERMS

Explainable AI, machine learning, behavioral science

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	44

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
PETER ROCCI	**N/A**

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 SUMMARY

The field of explainable artificial intelligence (XAI) has evolved and matured over the past decade. In its early days the field focused on developing new XAI techniques. Coverage of machine learning algorithms and diversity of explanation modalities were the main thrust. As XAI techniques increased in number, the field began to question more seriously the effectiveness of the explanation techniques. Consideration about explanation effectiveness brought a core issue of XAI into the spotlight: XAI is a human-centered problem. This realization called for an understanding of how humans use XAI systems, interpret the generated explanations, and develop trust while interacting with XAI systems. While novel techniques are still being created, the current state of XAI urgently needs results and insights into the human aspect of this problem.

Our contributions to XAI can be summarized by four lines of work: (1) developing novel XAI techniques, (2) demonstrating successful XAI systems with empirical validation, (3) advancing our understanding of the psychology of explainability, and (4) contributing to the mathematical foundations of XAI. The theme of our approach is to incorporate human inference into all aspects of XAI.

We propose *Bayesian Teaching* as the overarching framework for generating explanations. Bayesian Teaching is a theory of communication between an explainer and an explainee. The theory splotlights the model of the human explainee, and the framework generates explanations to help the modeled human explainee reach the desired target inference. As an explanation-generation technique, Bayesian Teaching is a post-hoc, model-agnostic method capable of generating both local and global explanations. The explanations are influential examples or sub-examples derived from the training data and thus naturally interpretable to end users. Bayesian Teaching is also a formal language for decomposing, taxonomizing, and unifying XAI methods.

We empirically demonstrate the effectiveness of Bayesian Teaching in four experiments. The empirical evaluation probes the objective effectiveness of explanations by asking participants to predict the generalization behavior of the AI. Importantly, the evaluation also tests whether Bayesian Teaching could *predict the degree of explanation effectiveness* as judged by participants. Our suite of experiments covers a good range of explanation types (local, global, exemplar, saliency maps), tasks (prediction, certification, debugging), and domains (classification of everyday objects, emotions, and medical images). A high point is Experiment 2, where we recruited practicing radiologists to test whether an XAI system can help them understand and trust a pneumothorax image classifier via a sequence of prediction and certification tasks. Statistical analyses of all the experimental results confirm the general effectiveness of explanations and also reveal crucial modifications to explanation effectiveness that can be attributed to participants' prior beliefs.

Based on the explainee-centered theory of Bayesian Teaching and the experimental results, we make big strides towards a psychological theory of explainability. We identify and provide evidence for three core components of such theory: First, people tend to project their own beliefs onto the AI, that is, they believe a priori that the AI would make similar decisions for similar reasons as they do. Second, effective explanations mitigate this belief projection and shift

people's belief to align more with the AI's behavior. Third, the psychological mechanism of this belief mitigation hinges on a comparison between the XAI explanation and participants' self-generated explanations. In particular, the comparison takes the form of generalization between the two explanation sources in a psychological space, which follows Shepard's universal law of generalization. These insights provide, for the first time, a quantitative handle on how XAI explanations alter people's understanding of AI.

Lastly, our contributions to the theoretical understanding of XAI are two-fold: First, we provide the mathematical foundation to a generalized theory of Bayesian Teaching called *cooperative communication*. We prove a series of desiderata for cooperative communication—consistency, stability, optimality—and conduct extensive simulations on communication efficiency and robustness. The mathematical and simulation results establish cooperative communication as an effective and realistic model of explanation. Second, we show how to gain insights into deep neural networks by analyzing deep Gaussian processes. Through derivation of closed-form properties of deep Gaussian processes and simulation studies, we gather a set of analyzable and visualizable mathematical objects—moments, hyperdata, and effective kernels—that summarize the behavior and learned representation in deep learning systems. In conclusion, our contributions are comprehensive, covering the engineering, experimental, psychological, and mathematical aspects of XAI. The theme of our approach is to incorporate human inference into XAI. We believe our work constitutes a sizable advance in making XAI a truly human-centered technology.

2 INTRODUCTION

The goal of explainable artificial intelligence (XAI) is to make black-box AI understandable to humans. In 2016, when DARPA issued a BAA for XAI, the focus was primarily on making machines explainable across a wide range of domains. This thread largely reflected the development of the field since then: a plethora of XAI techniques were developed for all types of machine learning algorithms. However, as the field progressed and applications looming, a crucial thread of pressing issues surfaced: how do we know that the explanations generated are effective at all? This realization has inspired researchers to expose problems with existing XAI methods and empirically validate their usefulness. While these efforts is making XAI more human-centered and user-centric as it should, the core question raised persists. This question of human's interpretation of explanation is arguably the most urgent issue in XAI that requires resolution.

Paralleling the current development of the field of XAI, our contributions over the past five years in the DARPA XAI program can be summarized by four threads of work: (1) the development of XAI techniques, (2) working demonstrations of XAI systems with empirical validation, (3) understanding the psychology of explainability, and (4) studying the theoretical foundations of XAI. We propose *Bayesian Teaching*—a formal theory of communication between an explainer and an explainee—as our overarching framework for explanation generation (see Section 3.1). Bayesian Teaching is a post-hoc explanation generation method that can be applied to explain any machine learning algorithm [1]. The explanations generated take the same modality as the training data and are thus naturally legible to humans as the training data are typically curated by people [1]. In addition to being an explanation-generation method, Bayesian Teaching also offers

a formal language to taxonomize all XAI methods (see [2] for detail). Such a formal language facilitates unification and generalization of XAI methods [2,3] as well as exposes hidden assumptions about the human mental / inference model used in them [2].

We demonstrate the effectiveness of Bayesian Teaching as an explanation-generation method in four experiments (Sections 3.2–3.5). The experiments test both local explanations for individual AI decisions [3,4] and global explanations for the AI's overall behavior [5]. The experiments also cover a wide range of classification domains, including everyday categories [3,4], abstract categories [5], as well as medical imaging [6]. The experiments focus on establishing an *objective* measure of human understanding of AI (by probing human understanding of AI generalization behavior) but also probe subjective preference of the XAI system. For the medical imaging study in Experiment 2, we also emulate a potential use case of using XAI systems for AI certification. We performed careful statistical analysis for all experiments and reported results with the appropriate statistics. Our empirical results confirm that the explanations generated by Bayesian Teaching help people understand AI behavior (see Results and Discussion, especially Sections 4.1–4.4). Importantly, not only can Bayesian Teaching generate explanations, it can also predict whether an explanation is helpful or harmful, a feature derived from its explicit modeling of human explainee inference.

A unique contribution of our work is the development and validation of a psychological theory of explainability. While there are now numerous XAI techniques, we still have no way to assess which methods will do well for a given use-case without performing costly empirical evaluation in every case. Naive empiricism is clearly insufficient to solve this issue because of the range of applications and the speed at which new XAI methods emerge. In other words, a glaring lack in the field of XAI is a theory of explainability, which is necessary to determine to what extent empirical findings and XAI techniques generalize. We make big strides in this direction by formalizing and validating a theory of explainability. The three core components of our theory are: humans tend to project their own beliefs onto the AI; explanations facilitate a mitigation of this a priori belief projection; and the degree of mitigation is governed by a comparison of XAI-generated explanations to humans self-generated explanation [3,7]. The theories generate testable predictions which we validate experimentally (see Sections 4.1, 4.3, and 4.5). The theory also inspired an exploration of developing algorithmic evaluation of explanation interpretability via modeling experts' decision rationale (Section 4.6; [8]).

Lastly, we advance the theoretical foundation of XAI and provide theoretical insights into the behavior of deep learning models. We establish a mathematical foundation for cooperative communication, a generalization of Bayesian Teaching (Section 4.7; [9-13]). This foundation consists of key mathematical proofs of the consistency, stability, and optimality of cooperative communication, all of which are strengthened with comprehensive simulation results. We gain theoretical insights into deep neural networks by analyzing deep Gaussian processes (Section 4.8; [14-17]). The latter exhibits similar expressiveness but is more analyzable compared to the former. We derive closed-form summary statistics of the entire model as well as sections of the model. These closed-form statistics provide insights into the overall behavior of and learned representation in deep learning systems.

3 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Theory

In this section we introduce Bayesian Teaching. As an explanation-generation technique, Bayesian Teaching is versatile: it is model-agnostic and works with any machine learning model; it can generate both local and global explanations; and its explanations are sampled from the training data and naturally illegible to human users [1]. Bayesian Teaching is also a formal language to decompose all XAI methods. The benefits of such formal decomposition include taxonomization and unification of existing XAI methods, unit testing XAI components, and recombination of XAI components to create novel techniques [2]. Furthermore, Bayesian Teaching connects XAI to cognitive science by making the presence of the explainee model explicit. The form of the explainee model—its posterior, prior, and likelihood—provides a systematic roadmap for developing a psychological theory of explainability [7].

3.1.1 Bayesian Teaching.

Bayesian Teaching formalizes explanation as a communication act between the explainer (teacher) and the explainee (learner) by the following equation:

$$P_T(x|\theta,s) = \frac{P_L(\theta|x,s)P(x|s)}{\sum_{x\prime} P_L(\theta|x\prime,s)P(x\prime|s)} \quad (1)$$

The equation describes how a teacher P_T should select an explanation x to best explain a target inference θ, contingent on their model of the learner P_L. Specifically, it says that the probability of choosing an explanation x to explain the target inference θ is proportional to the probability that the explanation x would lead the explainee model P_L to the target inference θ. Thus, Bayesian Teaching *explicitly* decomposes the explanation generation process into four components: the target inference θ; the explanation x; the explained / learner model, which is captured by the posterior $P_L(\theta \mid x,s)$; and the explainer / teacher model, which is captured by the selection posterior $P_T(x \mid \theta,s)$. The symbol s denotes the situation under which the explanation is taking place; in this report, s typically refers to an image that the AI classified.

The target inference θ is an aspect of the model that human users wish to understand. Possible target inferences range from global aspects of the model, such as model parameters, to intermediate components, such as the model's latent variables [5], to local targets, such as the model's prediction on a particular data point [3]. Most of our work focuses on the AI's decisions on a particular data point (local explanation), exception for Section 4.4, which targets the model's latent variables (global explanation).

The explanation x is the object provided to the end user to induce understanding about θ. Common explanation media include instances from the training data, features of the data, and simplified models that accurately describe the target model for some subset of the problem space. For the domain of image classification, which is the focus of this report, we take the training images at different levels of granularity to produce different types of explanations, including the images themselves (explanatory examples) and pixels of an image (saliency maps).

The use of training data as explanation has the benefit of being universally acceptable, since the training data are typically curated by human users for human users. Such explanations can also be interpreted as pedagogical examples—samples selected to best transmit AI knowledge θ to the learner model. As such, these explanatory examples /pixels have justifiable roots in cognitive science, education, and cultural evolution.

The learner/explainee model $P_L(\theta \mid x,s)$ is a computational model that describes how the user makes inferences about θ when given the explanation x. All XAI methods have a learner model, explicit or implicit. Bayesian Teaching makes the learner model **explicit** to allow validation of this crucial component (e.g., see Sections 4.1.3 and Section 4.4.1). The quality of the explanation generated depends on the quality of the learner model used. An inaccurate learner model will lead to unreliable and confusing explanations because the mapping between x and θ would be inaccurate. Conversely, a perfect learner model could lead to optimal explanation. The learner model is also called the **simulated explainee / learner**, because the model is indeed a formal attempt to simulate the explainee's inference process.

The form of the learner model suggests that any model that can be expressed as $P_L(\theta \mid x,s)$ can be input into Bayesian Teaching. In fact, all machine learning algorithms can be expressed in this form, as the form only requires that there is a mapping between the explanation and the target inference describable in probability theory. In Experiments 1–4, we use the AI to be explained as the learner model. The explanations generated from this learner model can be interpreted as summary training signals that helped the AI reach its current state. Other common learner models include simple models that are known to be intuitive to humans (e.g., linear models, small decision trees) and smaller models with the same type of architecture as the target AI (e.g., model distillation). We discuss this crucial component more in Section 3.1.2, showing the form of typical explainee model for explanation generation and the form for modeling actual human inference.

The teacher/explainer model $P_T(x \mid \theta,s)$ specifies the explanation-generation process. As Equation (1) suggests, this selection process is largely determined by the learner model. This is intuitive because a good teacher should consider the learner when selecting an explanation. To find the optimal explanation, one could search for the x that maximizes $P_L(\theta \mid x,s)$, and hence $P_T(x \mid \theta,s)$. This solution is equivalent to finding the x that maximizes the numerator of Equation (1) and thus avoids the computation of the denominator. Other approaches to inference include sampling from the $P_T(x \mid \theta,s)$, which will provide a sense of the relative effectiveness of near-optimal explanations. For most of the works reported here, we sample many x and bin them into different levels of helpfulness according to their $P_T(x \mid \theta,s)$ values.

The explicit breakdown into components also offers clarity on how to **validate** XAI approaches. Equation (1) shows that teaching is fully determined by θ, x, the learner model given by $P_L(\theta \mid x,s)$, Ω, and $P(x \mid s)$. Thus, validation of these components implies the validation of the teaching process. By virtue of Bayes' rule, a user task suitable for evaluating the learner model will also be suitable for evaluating the teaching process. A direct way to evaluate the learner model is to evaluate how well the learner model aligns with actual users. This alignment can be measured by the fidelity between the modeled response to a given explanation x (i.e., $P_L(\theta \mid x,s)$) and the user's actual response. All our experiments measure explanation effectiveness by asking human users

to predict the AI's prediction. In this set up, **fidelity** is expressed as the degree of alignment between the AI's prediction (the modeled response) and the participant's prediction of the AI's prediction (user's actual response). Additionally, Bayesian Teaching suggests that an explanation can only be deemed effective if it is observed to shift a user's belief towards a target inference. The **belief-shifting** framework implies that XAI interventions are best tested when there is a misalignment between user beliefs about the target AI system and the ground truth.

3.1.2 Explainee Models.

As mentioned in the previous section, the central piece to choosing effective explanations is an accurate model of the explainee's inference $P_L(\theta \mid x,s)$, which captures how the explainee makes inference about the machine after receiving the explanation. We offer two distinct approaches to create an explainee model. The first approach models the human explainee's inference with machine learning algorithms, whereas the second approach models it with grounded and tested theories in cognitive science. The two approaches have complementary strengths and weaknesses. The first approach is good for generating explanations in essentially any domain where AI is used, but provides little guarantee to human interpretability of the explanation generated. In contrast, the second approach offers an accurate model of how people interpret the explanation received, but currently lacks the data richness for algorithmic explanation generation. Below we describe the formulation of the two approaches.

The first approach projects the machine's inference and learning mechanisms onto people as if they would learn from explanation *x* in the same way that the machine itself does. Mathematically, the explainee model takes on this form:

$$P_L(\theta|x,s) = \sum_w P(\theta|w,s)P(w|x) \tag{2}$$

Here *w* denotes the set of parameters that fully specifies the machine. The $P(w \mid x)$ describes the machine's learning mechanism, that is, how *w* is updated given the explanation *x* as extra training data. The $P(\theta \mid w,s)$ corresponds to the machine's prediction mechanism, that is, the machine's classification of the image *s* under the particular model specification *w*. The sum over *w* is the formal Bayesian treatment to marginalize out uncertainties in the model specification. Almost all existing XAI techniques fall under this category, where the human explainee is modeled by some machine learning algorithm either explicitly (such as with distillation and mimic learning approaches) or implicitly. We also adopt this approach for the generation of explanations in Experiments 1–4.

The second approach formalizes how humans interpret explanations based on the cognitive science literature (see [7] for full detail). The literature suggests 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 follows Shepard's universal law of generalization in a similarity space. A natural formulation of the above is the following:

$$P_L(\theta|x,s) = \frac{p(x|\theta,s)P(\theta|s)}{\sum_{\theta'} P_L(x|\theta',s)P(\theta'|s)} \tag{3a}$$

$$P(x|\theta,s) = \lambda\, exp[-\lambda(1 - sim[x,x^H])] \tag{3b}$$

$$sim[x, x^H] = \frac{<x,x^H>}{|x|_2 \, |x^H|_2} \quad (3c)$$

In Equation (3a), the prior $P(\theta \mid s)$ is the explainee's inference of the AI's classification θ on image s without any explanation. The likelihood $p(x \mid \theta,s)$ is the probability that the explainee themself would provide the observed explanation x as the explanation for assigning class θ to image s. The sum over θ includes the class of interest and the alternative class(es) in contrast. Following Shepard's universal law of generalization, Equation (3b) specifies the likelihood to decay as a function of dissimilarity between the observed explanation x and the self-generated explanation x^H that is measured experimentally (Section 3.6 Experiment 5). Following Sloman, Equation (3c) specifies the similarity measure by how well the prominent features in the AI generated explanation x match the human-generated explanation x^H. To our knowledge, this is the first quantitative theory that describes how a human explainee interprets a received explanation. We validate this theory with Experiment 5 (see Sections 3.6 and 4.5; [7]).

3.2 Experiment 1: ImageNet

In this experiment we use Bayesian Teaching to generate explanations for a ResNet-50 model trained on ImageNet. We evaluate the effectiveness of the explanation by asking participants to predict the AI classification. The prediction task covers both correct and mistaken AI classifications, a wide range of object categories with varied participant familiarity ratings, and a diverse set of explanation conditions. We also test the utility of the explainee model by measuring how well the helpfulness of the explanations predicted by the explainee model matches actual user response. Comparison among the experimental conditions suggests that XAI explanations function by mitigating human users' tendency to project their own belief onto the AI. The experimental results are presented in Section 4.1.

Our technical innovations include generating local explanatory examples for ResNet-50 with varying degree of helpfulness, and unifying explanation-by-examples and saliency maps as the same XAI method with different explanatory granularity. Our experimental innovations include: the design of an objective measure for explanation effectiveness inspired by psychophysics, the measurement of the predictive accuracy of the explainee model, the comparison of different explanation modalities, and a detailed investigation of belief mitigation. Below we give a sketch of the experimental design. See [3] for full details.

The AI to be explained. The machine learning model to be explained is a ResNet-50 model. For this study, we used the pre-trained version of ResNet-50 in Keras with ImageNet weights. For the selection of examples, the Bayesian Teaching framework expects the model to be able to make probabilistic inference on the 2 alternative forced choice (2AFC) task. To achieve this, we replace the fully connected classification layer of the ResNet-50 model with a probabilistic linear discriminant analysis (PLDA) model, which we trained using a transfer-learning-like procedure.

Stimuli. Each experiment consisted of 150 trials. For 50 of the trials, the predictions of the model (or the robot) matched the ground-truth labels of the target images. For the remaining 100, the model predictions did not match the ground-truth labels. We selected the target images and the classification categories based on the model's confusion matrix, with the aim to cover a wide range of model behavior. The experiment ended up including a total of 83 unique categories. We also conducted a small experiment to gather familiarity ratings of the categories from 7 coders.

Please refer to [3] for technical details on the selection of explanatory examples and the generation of saliency maps. Briefly, θ is the ResNet-50's classifications on the target images; x are explanatory examples and/or saliency maps; and the explainee model is the ResNet-50-PLDA model in the form of Equation (2), where w are the PLDA's latent variables.

Experimental design. At the beginning of the experiment, participants were told that a robot has been trained to classify images but sometimes makes mistakes. They were asked to help by guessing how the robot will classify images. On each trial, a target image was displayed along with information about two categories, and the participants were asked to perform the 2AFC task by choosing which of the two categories they think the robot would classify the target image as.

The experimental conditions determined what information was presented during each trial and varied three dimensions: labels, examples, and saliency maps. Figure 1 shows a trial in the experimental condition with all the elements—labels, examples, and saliency maps—and describes how the conditions impact what elements are presented. More precisely, the conditions are characterized by five binary features: informative or generic labels, with or without examples, helpful or random examples (if present), with or without saliency maps, and blur or jet saliency maps (if present). The structured column and row labels of Table 1 show the naming conventions for the different conditions in terms of these features.

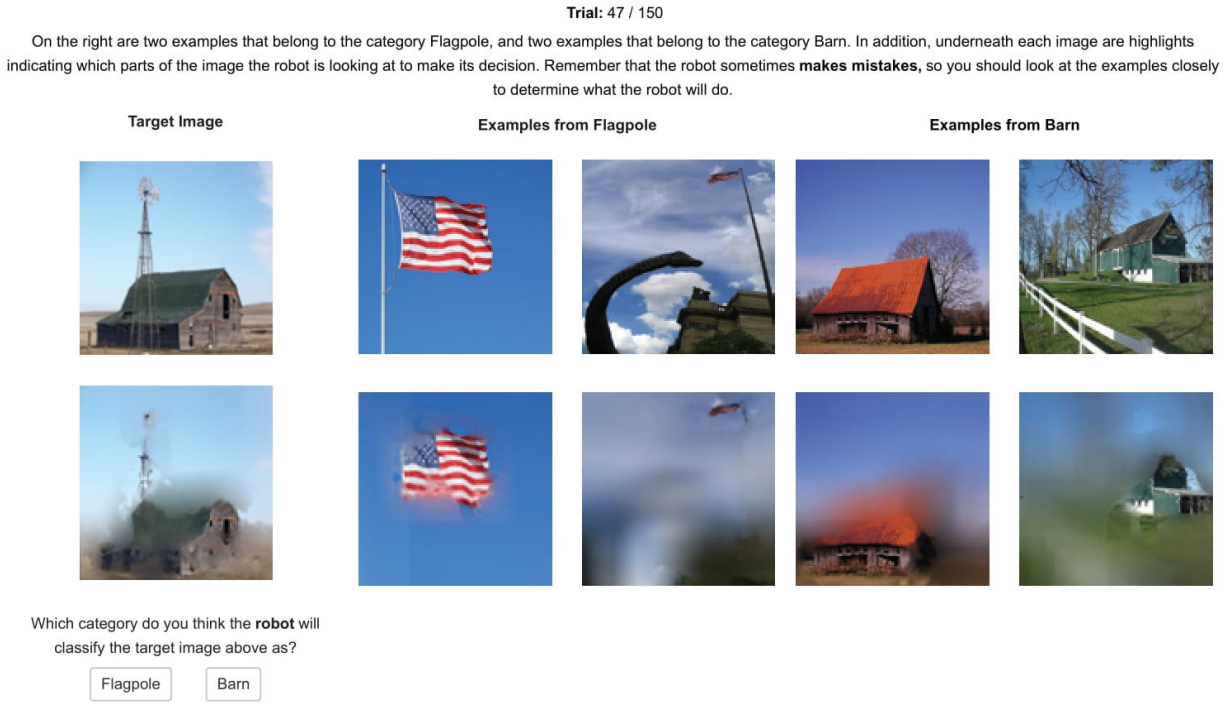

Figure 1. A snapshot of Experiment 1

Table 1. Conditions and the number of participants in Experiment 1

		SPECIFIC LABELS			GENERIC LABELS	
		NO EXAMPLES	EXAMPLES			
			HELPFUL	RANDOM	HELPFUL	RANDOM
NO MAP		N = 76	N = 35	N = 34	N = 38	N = 36
MAP	BLUR	N = 65	N = 33	N = 36	N = 35	N = 34
	JET	N = 71	N = 33	N = 35	N = 35	N = 35

3.3 Experiment 2: Pneumothorax

In this experiment we evaluate Bayesian Teaching in a high-risk domain—classification of pneumothorax from x-ray images. This domain is DoD relevant as pneumothorax is a common battlefield injury, and AI may help speed up the diagnosis and treatment process. As in Experiment 1, we design the experiment to test explanation effectiveness using the prediction task. We also emulate a real-world task, where participating radiologists were asked whether they would certify AI diagnosis and about the rationale for their decision. We then perform quantitative and qualitative analysis on the radiologists' predictions of AI classifications, their own diagnoses, their certification decisions, and the rationale for these decisions. Results are reported in Section 4.2. Please see [6] for full details.

Our technical innovations include expanding the application of Bayesian Teaching to a medical imaging domain and to a new class of model with U-Net architecture. Our experimental innovations include: the recruitment of practicing radiologists as expert participants; the design and implementation of an XAI interface with common radiograph viewing functionality; and a careful within-condition design that allowed us to analyze the relationship among the radiologists' own diagnosis, their prediction of the AI given the explanations, and their certification behavior. Below is a brief sketch of the experiment.

The AI to be explained. The AI to be explained is a deep neural network called AlbuNet used to diagnose pneumothorax in x-ray images. AlbuNet was trained on x-ray images with radiologists' markings of regions of pneumothorax from SIIM-ACR Pneumothorax Segmentation dataset. To provide a diagnosis, AlbuNet first computes the probability that pneumothorax is present for each pixel of the target image. It then takes these pixel-by-pixel probabilities (hereafter referred to as AlbuNet probabilities) and makes a binary classification for the full image by judging whether the number of pixels with AlbuNet probability greater than some threshold A is greater than some threshold B. We developed a probabilistic version of the original thresholding model and applied Bayesian Teaching to the probabilistic thresholding model to generate explanatory examples. The saliency maps are taken directly from the pixel-by-pixel AlbuNet probabilities.

Experimental design. Our experiment consisted of three blocks of trials, following consent forms and general instructions, see Figure 2. The first block evaluated how well the participants could predict the AI diagnoses. The two subsequent blocks evaluated if the explanations developed appropriate trust by asking them to certify the AI for different cases. One of these

blocks involved examples and saliency maps whereas the other just involved saliency maps. Each block consisted of 8 trials, with target images counterbalanced based on the AI's judgment, so that they included two each of true positives, true negatives, false positives, and false negatives. In each block the presentation order of the trials was randomized and differed between participants.

In the first block, each trial began with participants diagnosing the target image on a continuous rating scale with the endpoints labeled as "Certain pneumothorax present" and "Certain pneumothorax absent". They could zoom in on the target image and invert its colors, and they had unlimited time to make their judgment. After making their diagnosis participants were shown four examples (one at a time). The examples were presented in a fixed order, starting with a true positive AI diagnosis followed by a true negative, a false positive and a false negative. For each example participants were informed of the ground truth, the AI classification, and the AI probability judgment of pneumothorax. Participants were also presented with a saliency map of the example image, and were able to zoom and invert colors of the example. Participants decided how long to view each example. Once participants had viewed all examples, they were shown the target image with the AI's saliency map, after which we asked them to predict the AI diagnosis of the target image. Participants made their prediction on a "reminder screen" (see Figure 3) that reminded them how they themselves had diagnosed the target, while also showing them miniatures of all the examples, and the target image, with the option to expand, zoom and invert any of the images. They made their prediction on a continuous rating scale similar to the diagnostic scale described earlier. Once they had made their prediction, they received feedback on the accuracy of their prediction.

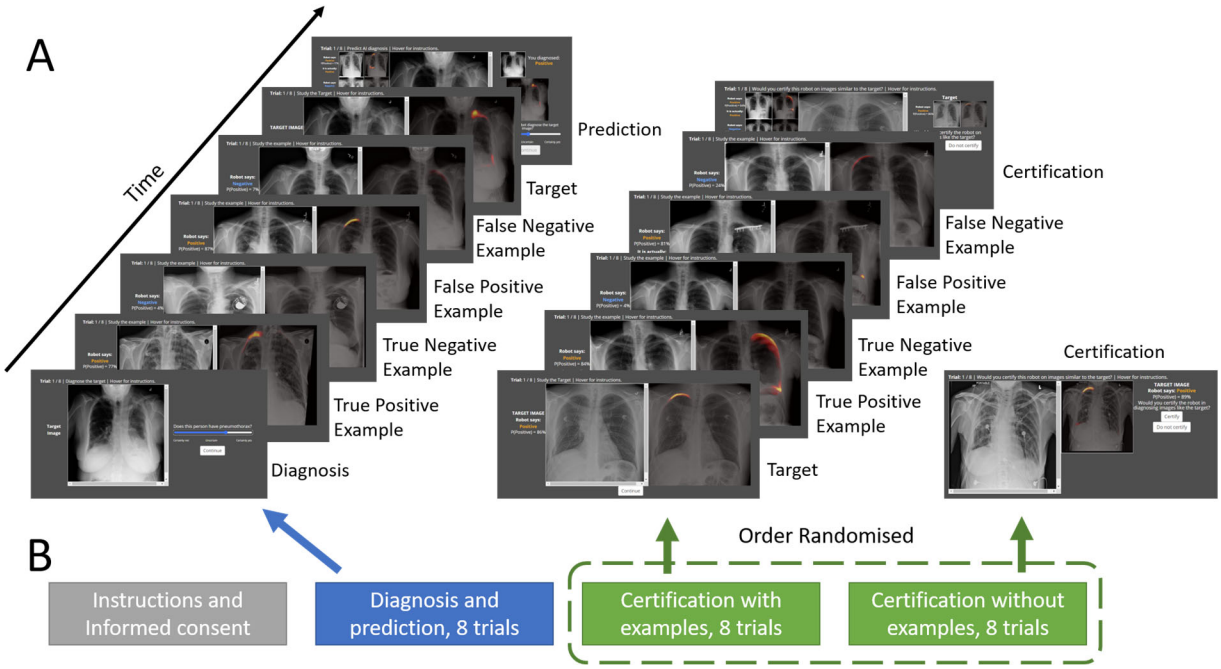

Figure 2. Trial structure and experimental design of Experiment 2

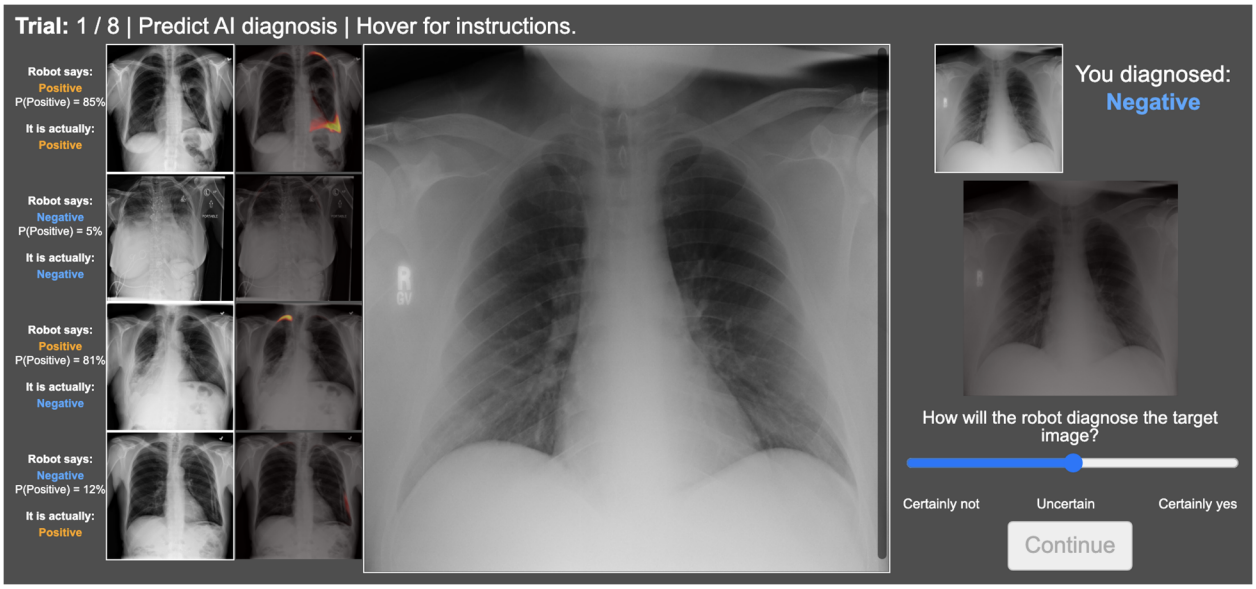

Figure 3. A screenshot of the XAI interface for radiologists in Experiment 2

The order of the two certification blocks were randomized between participants. In one certification block participants were given the same information as in the prediction phase, in the other they only viewed a saliency map of the target image. As opposed to the first block, in the certification blocks they were also shown the AI's judgment on the target image. Participants were asked whether they would certify the AI for images similar to the target (a binary judgment), report whether they agreed with the AI's diagnosis, and finally justify their certification decision. They could select multiple justifications among the following alternatives: (1) The robot got the correct answer, (2) The robot was appropriately confident, (3) The robot looked in the right place, (4) The examples are informative, (5) I am not certain I should certify, (6) Other. If participants selected (4) – (6) they had to elaborate in free text, but regardless of their choices they always had the option to elaborate in free text if they wished.

3.4 Experiment 3: Natural Adversarial ImageNet

In this experiment we extend Experiment 1 to test the effectiveness of explanation on natural adversarial images. Since these images are easy for humans (i.e., natural) and difficult for the AI (i.e., adversarial), they act as a more-sensitive testbed of explanation effectiveness and emulate core components of XAI-aided debugging. The experimental procedure largely follows Experiment 1 (Section 3.2). Modifications include an additional trial type (model errors due to adversarial images in addition to model hits and errors on standard images) and a reduction of experimental conditions (now 4 conditions: no explanation, explanatory examples, saliency maps, and both). A key technical innovation is that the ResNet-50-PLDA model used in Experiment 1 is replaced by a ResNet-50 with a fully probabilistic final layer, the latter of which has higher classification accuracy. Experimental results are presented in Section 4.3. Please see [4] for full details.

3.5 Experiment 4: Facial Expression

In this experiment we extend Bayesian Teaching to generate *global* explanations for *abstract* image categories—emotions in facial expressions. In contrast, the previous three experiments present only local explanations on concrete categories. As in Experiment 1, we evaluate explanation effectiveness as well as the utility of the explainee model. Experimental results are presented in Section 4.4. Please see [5] for full details.

The AI to be explained is a probabilistic linear discriminant analysis (PLDA) model. The θ is the learned latent variable of the PLDA on the whole training data, instead of its decision on an image. The explanation x are examples from the training data, called **teaching set**. The explained model is the PLDA model trained with a small teaching set.

Experimental design. On each trial, participants were presented with a target image and asked to classify it into one of two categories (A or B), where one category matched the category of the target image, and the other was randomly selected from one of the other five emotion categories. The participants were presented with a teaching set of three example images to represent each category. The teaching set presented could be helpful, random, or unhelpful, depending on the experimental condition. The helpfulness of the teaching set is determined by the explainer's selection probability. Figure 4 shows a snapshot of an experimental trial: Participants were shown a target image (left), along with a teaching set of examples from both the target category (angry, bottom right) and the other category (surprise, top right), and asked to predict how the model would respond based on the examples provided.

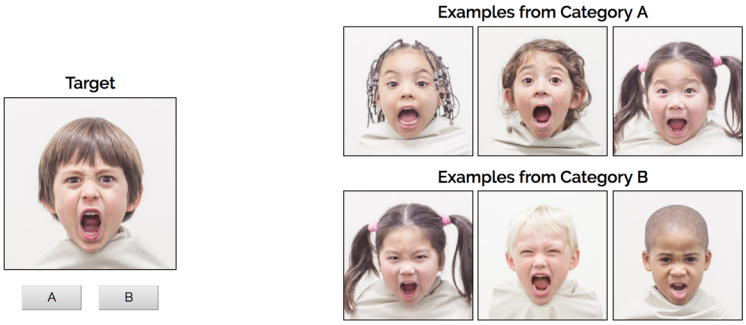

Figure 4. An example trial from Experiment 4

3.6 Experiment 5: Psychological Theory of Explainability

In contrast to the previous four experiments which focus on evaluating explanation effectiveness, in this experiment we investigate how humans interpret XAI-generated explanations by testing the formal psychological theory of explainability described by Equations (3a)–(3c). The experimental design is theory-driven: the prior in Equation (3a) is measured by a control condition without explanation; a key input to the likelihood in Equations (3b) and (3c) is measured by a novel drawing experiment; and the posterior is measured by an intervention condition with explanation. The measured posterior is then compared to the theory's posterior to test the accuracy of the theory. The experimental design is the first of its kind. It offers a novel platform for empirical investigation of the psychology of explainability, a key missing piece in the current state of XAI. The experimental results are presented in Section 4.5. Please see [7] for full details.

Experimental design. There are two types of experiment: classification experiment and drawing experiment. Participants in the 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 judgment but were also exposed to a saliency map explanation x that highlights regions of image that highly influence the AI's decision. In the 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 (the x^H in Equations (3b) and (3c)) that humans would use to interpret the observed saliency maps (the x in Equations (3a)--(3c)).

Figure 5 shows the relationship between the theory and experiments. Human interpretation of an explanation, $P^*(\theta \mid x,s)$, 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, $P(\theta \mid x,s)$ in Equation (3a). 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, $P(\theta \mid s)$. 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, $p(x \mid \theta, s)$. 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.

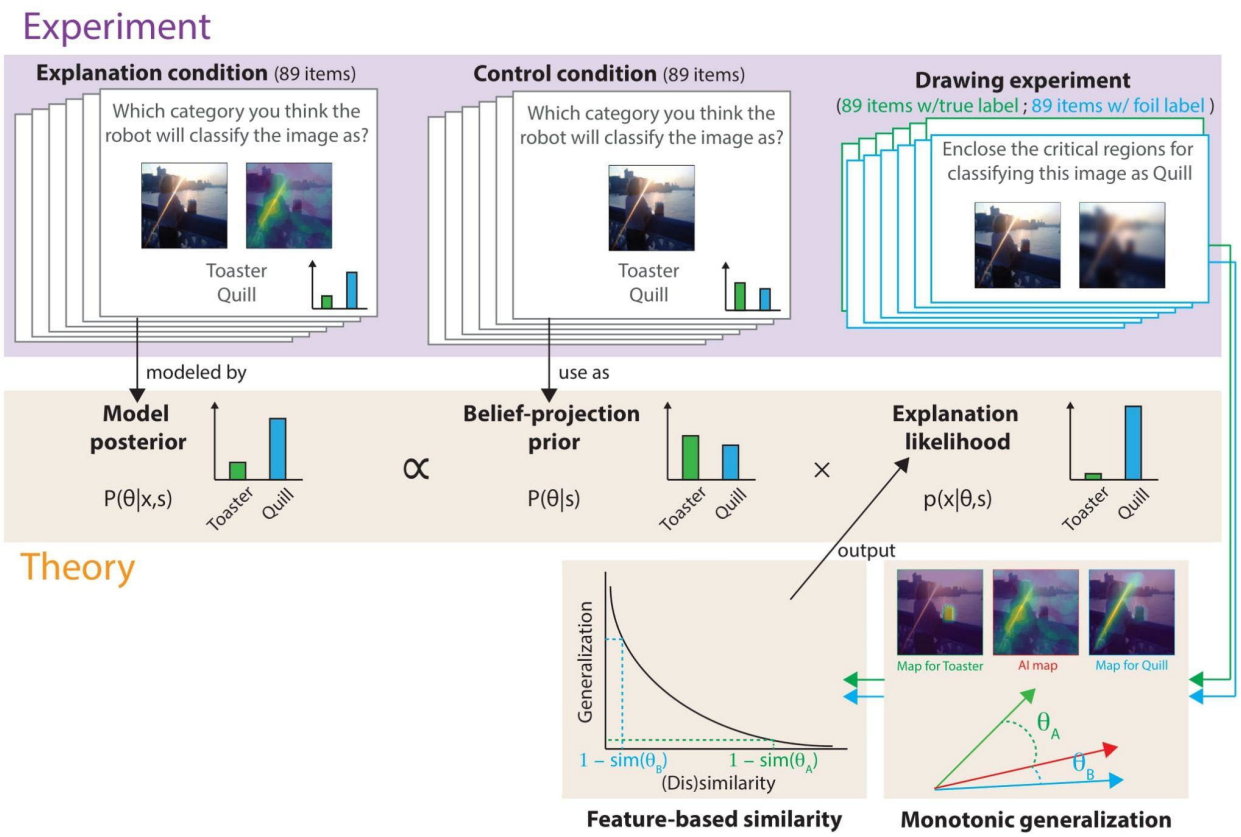

Figure 5. The relationship between theory and experiments in Experiment 5

4 RESULTS AND DISCUSSIONS

In the Introduction we outlined four threads to our contributions: (1) the development of XAI techniques, (2) working demonstrations of XAI systems with empirical validation, (3) insights into the psychology of explainability, and (4) advances in the theoretical foundations of XAI. Section 3.1 provides an overview of our contribution towards the first thread, where we described Bayesian Teaching as a novel, versatile XAI technique to generate explanations. In the remaining sections of Methods we also highlighted several experimental innovations that allowed us to evaluate explanation effectiveness and test the psychology of explainability. The following experimental results (Sections 4.1–4.6) echo the second and third threads. Then, we present theoretical results that advance the mathematical foundation of XAI and reveal analytic insights into the behavior of deep neural networks in Sections 4.7–4.8.

In the thread of working demonstrations, we show that Bayesian Teaching improves people's prediction of AI behavior across several image classification domains (Sections 4.1.1, 4.2.1, 4.3, and 4.4.2), including medical imaging. The results show interesting interactions between explanation modalities, AI accuracy, and category familiarity (Sections 4.1.4 and 4.3). Furthermore, we show that Bayesian Teaching can predict the helpfulness of Explanations (Sections 4.1.3 and 4.4.1). Explanations that are deemed objectively more helpful by Bayesian Teaching are also preferred by participants (Section 4.1.2).

In the thread of understanding the psychology of explainability, we first show that people project their own beliefs onto the AI, that is, they believe a priori that the AI would make similar decisions as they do (Sections 4.1.1, 4.1.4, and 4.5). They also tend to certify the AI if the AI's decisions match their own decisions (Sections 4.2.2). We then show that explanations can mitigate this belief projection and shift people's belief to align with the AI's behavior to some extent (Sections 4.1.1, 4.1.4, 4.2.1, and 4.3). The psychological mechanism of this belief mitigation hinges on a comparison between the XAI explanation and participants' self-generated explanations (Section 4.5). Our formal model of the comparison predicts the direction and magnitude of the mitigation effect. These results provide a quantitative handle on *what* the explanations are changing (the beliefs projected onto the AI) and *how* (or *how much*) XAI explanations improve understanding (via a psychological comparison with self-generated explanations as reference points).

The success of the psychological theory of explainability described above hinges on an accurate description of the human explainee's own explanations. We obtained these descriptions by empirical measurement. To make the theory useful at scale, one would benefit from having a computational description of the human's explanation generating process. Section 4.6 is an attempt in this direction by leveraging *textbook features* in the medical domain as an approximate model of expert reasoning.

In the thread of advancing theoretical foundation, we first present a generalized theory of Bayesian Teaching called cooperative communication (Section 4.7). We summarize mathematical results on the optimality, consistency, and stability / robustness of cooperative communication as well as simulation results that broaden the coverage of the theory. These results establish a rigorous foundation for cooperative communication, supporting the idea that explanation via Bayesian Teaching is an effective way to transmit information. In Section 4.8 we study the behavior of deep neural networks by analyzing an intimately related family of deep learning models called deep Gaussian processes. Through derivation of closed-form properties and simulation studies, we gather a set of analyzable and visualizable mathematical objects—moments, hyperdata, and effective kernels—that summarize the representation learned in deep learning systems.

Below, we present the experimental results in the order introduced in the Methods section for ease of reference. These results are then followed by the theoretical ones.

4.1 Results from Experiment 1: ImageNet

Bayesian Teaching contributes to the literature on XAI by formalizing the role of the explainee. Explicitly considering the explainee highlights how XAI methods can be validated, and how explanations informed by the explainee model can mitigate human prior beliefs about the AI system. We showcase three criteria to validate explainable AI from the Bayesian Teaching perspective: Explanations selected by Bayesian Teaching improve the fidelity between human prediction of AI classification and actual AI classification (Section 4.1.1); the Bayesian Teacher can correctly infer which explanations humans will prefer (Section 4.1.2); and the Bayesian Teacher can accurately predict both which explanations will improve fidelity and which explanations will decrease it (Section 4.1.3).

Additionally, we show how the prior beliefs of human participants can be mitigated by appropriate explanations. Consistent with existing work from psychology, we find that human participants project their own beliefs onto the AI system. This belief-projection manifests as: fidelity is higher when the AI is correct relative to when it is wrong; this impact of AI correctness on fidelity is particularly pronounced for familiar categories; and appropriate explanations can mitigate these effects (Section 4.1.4). In particular, examples improve fidelity the most for unfamiliar categories by confirming model hits, whereas saliency maps improve fidelity most for familiar categories by exposing model errors. Please refer to [3] for full details.

4.1.1 Bayesian Teaching Improves Fidelity.

To evaluate whether the XAI interventions improved fidelity, we compared participants who obtained a full explanation (examples and saliency maps) with a control group that received no explanations. When interpreting these results in relation to belief projection it is instructive to consider three idealized scenarios. An agent who picked categories at random would have 50% fidelity, sensitivity (correctly predicting AI classifications when the AI classifier is correct), and specificity (correctly predicting the AI's mistakes). An agent who modeled the AI classifier perfectly would have 100% fidelity, sensitivity, and specificity. Finally, an agent with perfect first-order accuracy who projected their own beliefs onto the AI classifier would have 100% sensitivity, 0% specificity, and 33% overall fidelity because the experiment contains twice as many AI errors as AI correct classifications. Absent intervention, participants behave most like the third, belief-projecting, agent (Figure 6A-B).

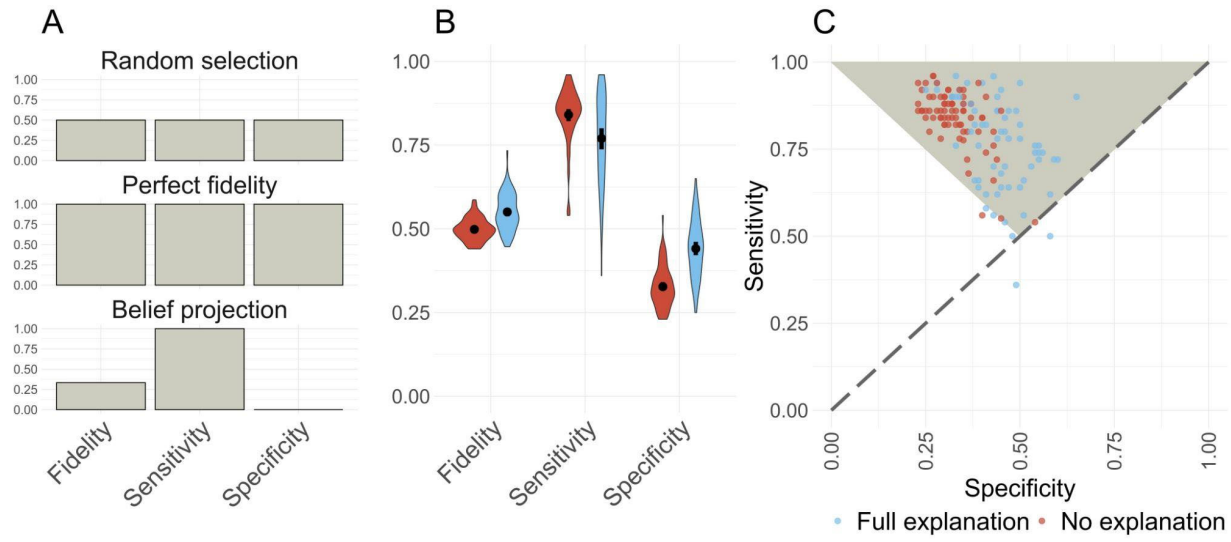

Figure 6. Bayesian Teaching improves fidelity by mitigating belief projection

Figure 6B shows that the explanation interventions increase overall fidelity by increasing specificity (participants are better able to spot the AI's mistakes), at the cost of some sensitivity. Participants in the control condition have a mean fidelity of 49.83% [95% CI = 48.83% -

50.84%], significantly lower than the 55.04% [95% CI = 52.58% - 57.48%] fidelity of the experimental group ($\beta = 0.21(0.03)$, $z = 6.99$, $p < .0001$). This is primarily driven by higher specificity in the experimental group (43.98% [95% CI = 39.68% - 48.37%] relative to the control group's 32.54% [95% CI = 30.96% - 34.13%]; $\beta = 0.49(0.05)$, $z = 9.20$, $p < .0001$). The greater vigilance of the experimental group came with a minor cost to sensitivity for the experimental group (78.90% [95% CI = 71.59% - 84.80%]) and for the control group (85.26% [95% CI = 83.12% - 87.22%]); $\beta = -0.43(0.12)$, $z = -3.68$, $p = .0002$), but not enough to offset the specificity gains. Figure 6C shows individual participants' sensitivity and specificity. The vertices of the triangle correspond to the fidelity of a belief-projecting agent with perfect access to the ground truth (upper left), an agent with a perfect model of the AI classifier (upper right), and an agent choosing at random (lower middle). The control group is clustered at high sensitivity and low specificity towards the upper left, whereas the experimental group is shifted to the right. However, the experimental group also shows greater variance, signifying inter-individual differences in the intervention effectiveness. Collectively, these results imply that participants attempt to predict the AI by projecting their own beliefs, and that the explanations improve fidelity by mitigating this belief projection.

4.1.2 Participants Prefer Examples That Are Helpful According to Bayesian Teaching.

Next, we evaluate whether participants preferred helpful to random and misleading examples. To test this, we ran a second study where participants chose between helpful examples versus random examples or versus misleading examples, where helpfulness was determined by Bayesian Teaching. We also evaluate whether helpful examples are most beneficial for unfamiliar categories, as hypothesized.

Figure 7A shows that participants showed a small but reliable preference for helpful relative to random examples (53.05% [95% CI = 51.08% - 55.01%], $z=3.03$, $p = .002$) and a substantial preference for helpful versus to unhelpful examples (64.14% [95% CI = 61.68% - 66.59%], $z=10.95$, $p < .0001$). These two conditions were reliably different ($\chi 2 = 36.94$, $p < .0001$), implying that the Bayesian Teacher is not only capable of selecting helpful examples, but can also select examples that are actively confusing. Evaluating whether these preferences are particularly pronounced for unfamiliar examples, we found that participants were more likely to prefer helpful examples when the choice categories were unfamiliar to them ($\beta = -0.57(0.08)$, $z = -7.02$, $p < .0001$; Figure 7B), irrespective of whether helpful examples were contrasted with random or unhelpful examples.

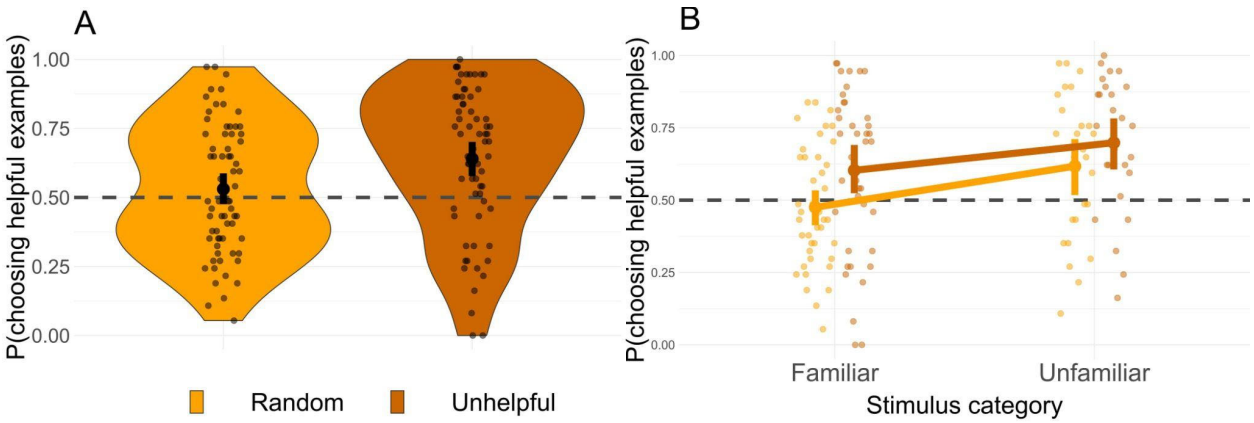

Figure 7. Participant preference for helpful examples

4.1.3 Bayesian Teaching Can Predict Which Explanations Improve and Reduce Fidelity.

Bayesian Teaching should be able to modify participant fidelity by selecting explanations of varying helpfulness. To test this in practice, we ran three nested hierarchical logistic regression models of increasing complexity. Each regression model predicted participant fidelity (whether the participant correctly predicted the AI classifier on a given trial) from the trials with only explanatory examples (not saliency maps), as these are the only trials impacted by the simulated explainee fidelity, which measures the degree to which the examples would lead the explainee model to the targeted inference. The first regression model served as a null-model, not using simulated explainee fidelity as a predictor, only including category accuracy and a dummy variable encoding AI correctness (whether the AI prediction for that trial matched the ground truth or not). The second regression model added simulated explainee fidelity as a predictor, capturing the hypothesis that the helpfulness of the examples as determined by Bayesian Teaching covaries with participant fidelity. The third regression model added two two-way interactions between model correctness (model hit and error) and category accuracy, and model correctness and simulated explainee fidelity, capturing the hypothesis that helpful examples had differential impact on error detection relative to hit confirmation. We found that the second regression model fitted the fidelity data better than the first regression model ($\chi 2(1, 4) = 71.68$, $p < .0001$). This means that the Bayesian Teacher's perception of the helpfulness of the presented examples predict participant fidelity above and beyond category accuracy. The third regression model outperformed the second regression model ($\chi 2(3, 7) = 7371.28$, $p < .0001$). This indicates that how well the category accuracy and/or the modeled helpfulness of the examples shown predicted fidelity differed for trials with correct or incorrect AI judgements. Figure 8 shows that the Bayesian Teaching framework can predict explanations that are informative or misleading for trials that are correctly classified by the model, but not for trials that are incorrectly classified. This trial-accuracy modulation is likely due to the AI being a more accurate model of the human explainee for the trials where AI is correct.

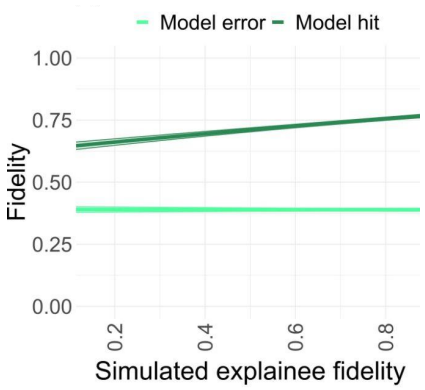

Figure 8. Bayesian Teacher predicts human fidelity

4.1.4 Bayesian Teaching Improves Fidelity Through Belief-Mitigation.

Next, we will explore *how* explanations improve fidelity and evaluate the relative importance of the different explanation features employed. The results from Figure 6 imply that people belief-project by default: that is, they use their own beliefs as priors for the AI classifier's beliefs. Explanations shift these priors, allowing the participants to distinguish their first-order beliefs about the correct classification from their second-order beliefs about the decisions of the AI classifier.

The presence of the saliency maps improves fidelity when the AI classifier is wrong ($\beta = 0.43(0.03)$, $z = 14.24$, $p < .0001$), but reduces fidelity (to a lesser extent) when the AI classifier is correct ($\beta = -0.56(0.07)$, $z = -7.98$, $p < .0001$; see Figure 9A). In both cases, saliency maps reduced the first order-accuracy of the participants (model hit: $\beta = -0.56(0.07)$, $z = -7.98$, $p < .0001$; model error: $\beta = -0.43(0.03)$, $z = -14.24$, $p < .0001$), meaning that they were less likely to report that the AI classifier's judgements match the ground truth of the image. This implies that the saliency maps encourage participants to consider that the AI classifier might be mistaken. One potential explanation for this observation is that the saliency maps show when the AI classifier attends to non-sensible features (i.e., parts that are not representative of either of the categories) as well as ambiguous features (e.g., thin metal strips that are present in both the "Electric Fan" and "Buckle" category).

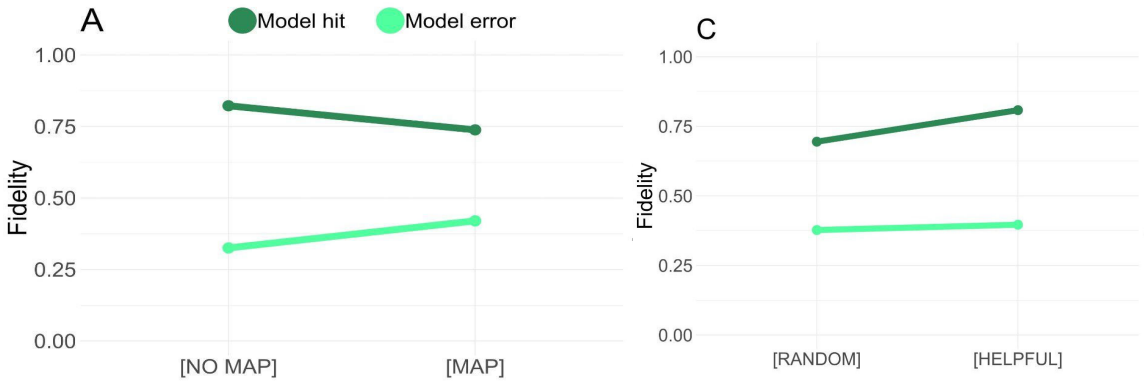

Figure 9. The fidelity based on AI correctness

In the conditions where examples were present, helpful examples improve fidelity for trials when the AI classifier was correct ($\beta = 0.77(0.08)$, $z = 10.11$, $p < .0001$), but not for trials when the AI classifier was wrong ($\beta = 0.06(0.04)$, $z = 1.77$, $p = .08$). Note that the effect of helpful examples is the opposite to what we found for the saliency maps: Whereas saliency maps help participants to identify trials when the AI classifier has made a mistake by exposing inappropriate sub-image-level features, the examples help reinforce participant's prior beliefs for trials in which the AI classifier is correct (Figure 9). In other words, the saliency maps and the examples serve separate and complementary functions in explaining AI judgements to the participants.

The familiarity scores capture the ease of the discrimination task in that they are higher for trials involving categories that humans are familiar with. These scores provide further clues as to whether participants project their own beliefs onto the AI: If humans use their first-order classifications to model the AI, participants should assume that the AI classifier gets the correct answer for trials that they themselves find easy. This is indeed what we find: familiarity is positively associated with fidelity when the and the AI classifier is correct ($\beta = 1.10(0.04)$, $z = 29.28$, $p < .0001$), but negatively associated with fidelity for AI errors ($\beta = -0.92(0.02)$, $z = -42.82$, $p < .0001$; Figure 10A).

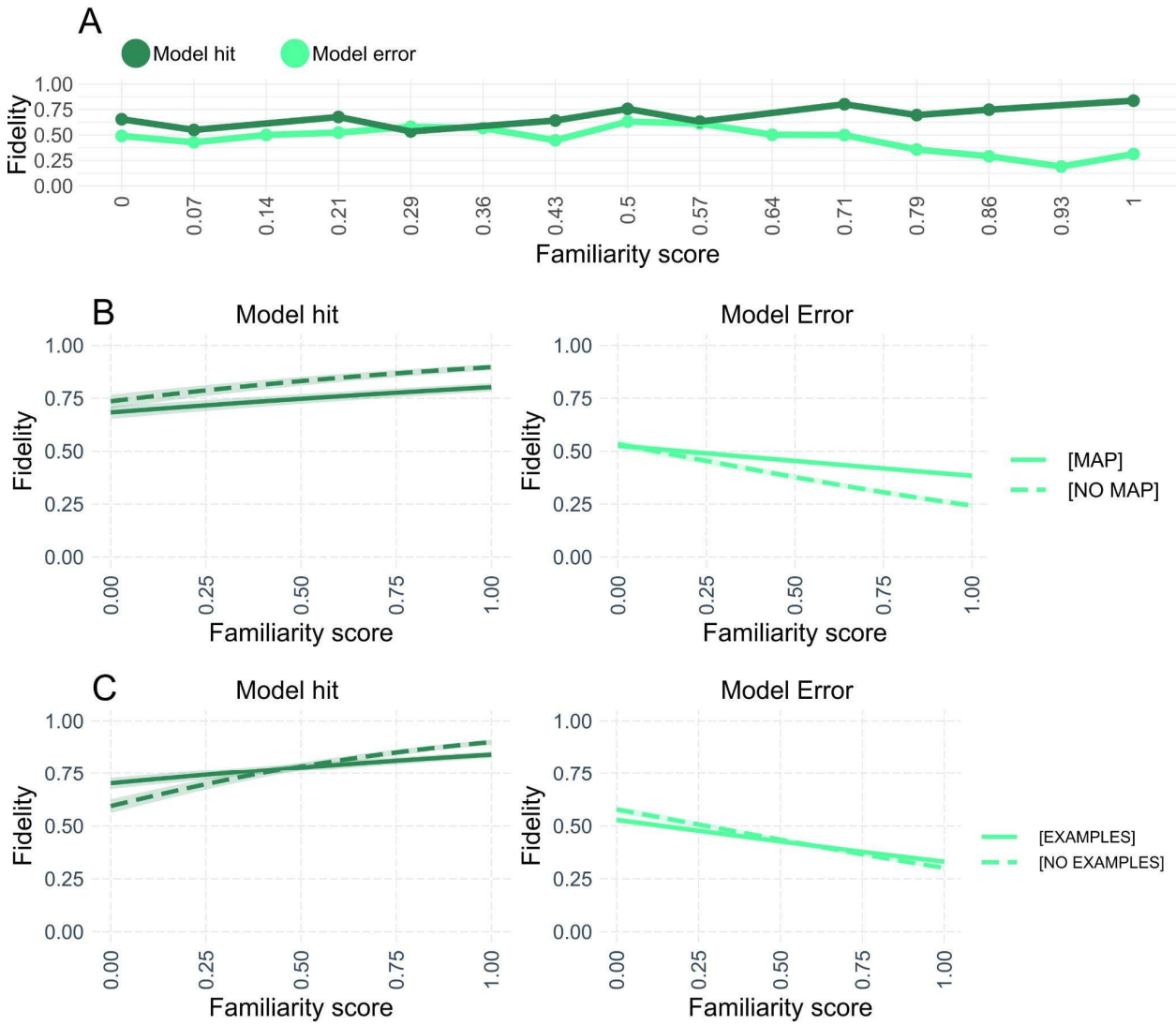

Figure 10. Familiarity score, model correctness, and explanation modalities affect fidelity

Previously, we showed that saliency maps improved fidelity on trials when the AI classifier was wrong. This could be explained by saliency maps helping participants distinguish between their first-order judgements of the ground truth and their second-order beliefs about the model classification. This explanation can be evaluated by testing whether the impact of the familiarity scores on fidelity are attenuated by the saliency maps. In other words, if participants are more likely to predict that the AI classifier is correct on trials that they themselves find easy, and the saliency maps work by helping people realize that the AI classifier use decision-processes that differ from their own, the saliency maps should make participants more willing to consider that the AI classifier might be wrong for trials they themselves find easy. This is what we find (see Figure 10B): the presence of saliency maps reduces the positive impact of familiarity on fidelity when the AI classifier is correct ($\beta = -0.51(0.08)$, $z = -6.31$, $p < .0001$). Conversely, saliency maps reduce the negative impact of familiarity on fidelity when the AI is wrong ($\beta = 0.70(0.05)$, $z = 15.22$, $p < .0001$). Collectively, these results suggest that the presence of saliency maps helps participants model the AI as an agent with distinct beliefs that may conflict with their own.

Although the presence of examples (including random and unhelpful ones) did not generally impact fidelity, it is possible that they impacted judgements specifically for unfamiliar categories. Like the saliency maps, examples typically reduced the impact of familiarity on fidelity, both when the AI classifier is correct ($\beta = -1.01(0.08)$, $z = -12.71$, $p < .0001$) and when the AI classifier is wrong ($\beta = 0.33(0.05)$, $z = 7.35$, $p < .0001$). However, in contrast to the saliency maps, examples seem to be most helpful for unfamiliar trials when the AI classifier is correct, see Figure 10C. This effect may imply that the examples help participants develop a working representation of the unfamiliar categories, which they are otherwise lacking.

4.2 Results from Experiment 2: Pneumothorax

Limited expert time is a key bottleneck in medical imaging. Due to advances in image classification, AI can now serve as decision-support for medical experts, with the potential for great gains in radiologist productivity and, by extension, public health. However, these gains are contingent on building and maintaining experts' trust in the AI agents. Explainable AI may build such trust by helping medical experts to understand the AI decision processes behind diagnostic judgements.

We applied Bayesian Teaching to a deep neural net used to diagnose pneumothorax in x-ray images. The explanations were integrated into an interface that carries basic functionalities for viewing x-ray images. We designed an experiment that aimed to test **(1)** participants' understanding of the AI (captured by how well they could predict the AI's decisions) and **(2)** the development of appropriate trust (captured by when they chose to certify the of AI's decisions). **Radiologists were recruited** to evaluate whether medical professionals benefited from the explanations generated by Bayesian Teaching. Our results confirmed the utility of explanation both for understanding the AI and for developing appropriate trust in the AI system. Please refer to [6] for full details.

4.2.1 Bayesian Teaching Helps Radiologists Predict AI Behavior.

To assess the first-order diagnostic accuracy of our radiologists, we compared two nested Bayesian linear models: one predicting diagnoses from only participant-wise random intercepts (capturing participant-specific response biases) and one model that added a fixed effect for the ground truth of the target image (capturing the discriminant ability of the radiologists). The second model fitted the data better than the first model, as indicated by a higher leave-one-out expected log point-wise predictive density (ELPD-LOO) than the first model ($\Delta = 6.5$, $se = 2.4$), indicating that radiologists could successfully diagnose pneumothorax. The posterior mean of this effect was 21.19 (95% credible interval = 7.69-33.83) on a 100-point scale, suggesting that radiologist judgements (their diagnoses using the continuous rating) differed on average about 20 points between trials when pneumothorax was present and trials when pneumothorax was absent. For a descriptive overview of these results see Figure 11 A.

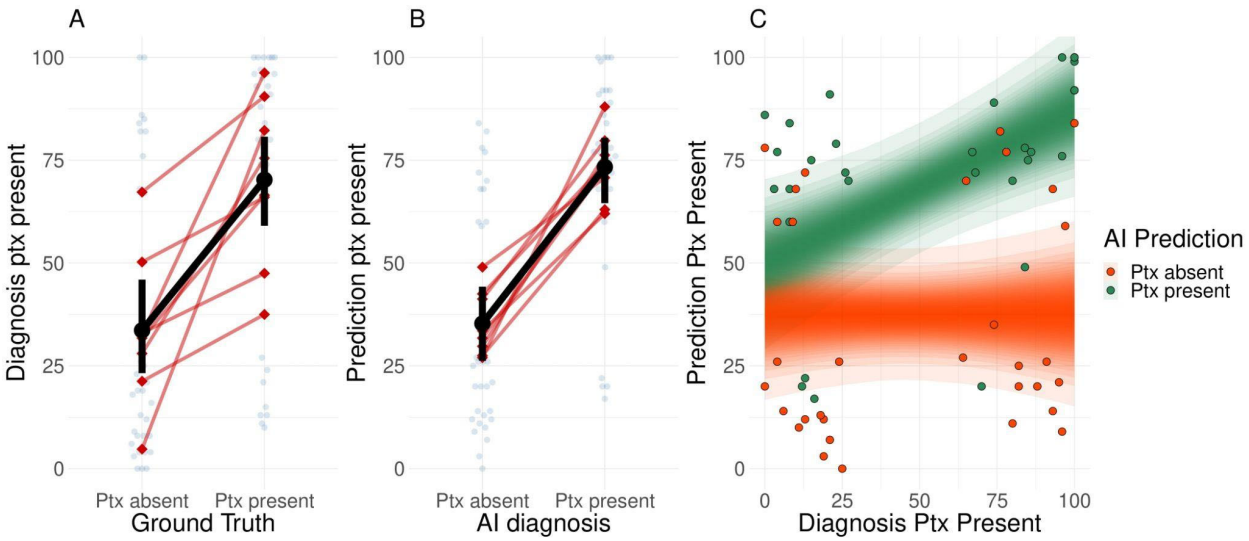

Figure 11. Diagnosing and predicting pneumothorax

The radiologists correctly predicted the AI's judgment on 6 out of 8 trials on average (range = 5–7), see Figure 11 B. Previous work on non-experts suggests that absent intervention humans expect the AI's judgment to mirror their own (see Section 4.1). To account for this we fitted three nested Bayesian linear models predicting radiologist predictions of the AI diagnoses. The null-model contained participant-wise intercepts and a fixed effect of the radiologist's diagnosis for that trial. The second model added a fixed effect for the AI's classification (coded as 0 for "pneumothorax absent" and 1 for "pneumothorax present"). The third model added an interaction term between the AI's classification and the radiologist diagnosis. Radiologists could effectively predict the AI even when accounting for their first-order diagnosis as illustrated by the second model fitting the data better than the first model (ELPD-LOO $\Delta = 11.8$, se = 3.5). The third model fitted the data marginally better than the second model (ELPD-LOO $\Delta = 0.5$, se = 2.3), implying that the first order diagnostic judgements might impact predictions of the AI differently when the AI is correct relative to when it is wrong, but we have too few observations to reach a strong conclusion.

To more fully explore the relationship between radiologist predictions, radiologist diagnoses, and the classification of the AI, we studied the posterior coefficients of the third model (Figure 11 C). Radiologists' predictions tend to be more positive when the AI did classify pneumothorax than when it did not, even when they themselves found pneumothorax very unlikely (posterior mean = 12.47, 95% credible interval = -2.38–27.22). When the AI did not classify pneumothorax as present there was no relationship between radiologist diagnoses and their prediction of the AI (posterior mean = 0.00, 95% credible interval = -0.22–0.22). But for the trials when the AI classified pneumothorax as present there was a positive relationship between radiologist diagnoses and their prediction of the AI (posterior mean = 0.37, 95% credible interval = 0.10–0.64). The difference in intercepts indicate that the explanations worked: radiologist predictions were typically more positive on the rating scale for positive AI classifications than negative AI classifications, when accounting for their own diagnostic judgment. The difference in slopes indicate that the participants' own diagnoses serve as priors for the prediction of the AI for target images where the AI is correct, but not when it is wrong.

4.2.2 Radiologists Certify the AI When the AI Matches Their Own Judgment.

To evaluate appropriate trust, we aim to address three key questions related to radiologist certifications: 1) Are radiologists more likely to certify the AI for images where it makes a correct diagnosis than where it makes mistakes? 2) Are they more likely to certify correct trials for the block with explanatory examples relative to the block without explanatory examples? 3) What justifications do participants provide for their certification judgements and what do these tell us about their decision processes? We address the first two questions with Bayesian regression models, to maintain analytic coherence. Because the third question is more qualitative and open-ended, we only explore it descriptively.

To test whether radiologists are more likely to certify images that the AI classifies correctly, and whether the examples impact these judgements, we fit and evaluate three Bayesian hierarchical logistic regression models. The null model predicted certification judgements (certified coded as 1, not certified coded as 0) from random intercepts at the participant level; the second model added AI correctness as a fixed-effect predictor; the third model added fixed effects for the explanation block and an interaction term between explanation block and AI correctness.

The main-effect model accounted for the data better than the null-model (ELPD-LOO Δ = 7.0, se = 4.0) or the interaction model (ELPD-LOO Δ = 1.8, se = 1.1). These results imply that participants are more likely to certify trials when the AI classifies the target correctly (Mean OR = 5.12, 95% Credible interval = 2.21–10.72), but that there is no reliable difference in certifications between the block with examples and the block without examples, see Figure 12A. Also note that certification probabilities tend to be below chance when the AI is incorrect but above chance when the AI is correct. Because the two blocks are not reliably different, we will collapse them in our subsequent discussion on certification justifications.

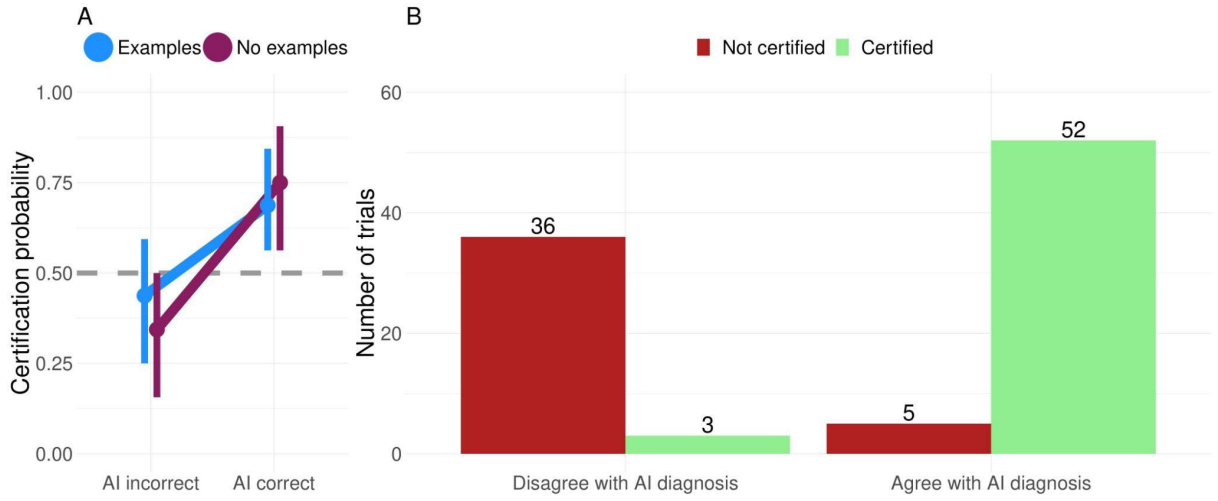

Figure 12. Certifying the AI.

Our participants certified 46 cases where the AI was correct versus 25 cases when the AI was wrong. They chose not to certify 39 cases when the AI was wrong, relative to 18 cases when the AI was correct. In 16 out of 25 cases when participants certified the AI despite it being wrong, they justified their certification in terms of agreement with the AI. This suggests that these certification judgements are grounded in participant errors, which in turn implies that the ground truth is not a reliable proxy of participant belief in this task. Because we are interested in how radiologists justify certification in relation to their own diagnostic judgements, we focus on 6 of the 8 participants that were explicitly asked if they agreed with the AI before they justified their certification decisions.

Certification decisions appear to be primarily driven by agreement with the AI, see Figure 12 B. For the three cases where participants certified the AI despite disagreeing with it, they left open-ended responses clarifying their thinking. All of these responses suggest that they believed the AI actually got the overall classification correct, but had either been too confident or not confident enough regarding pneumothorax elsewhere in the lung, based on the saliency map. In the five cases where participants chose not to certify, the most common justifications were either that the AI was looking in the wrong place or open-ended responses. Here the responses again allude to getting judgements right for parts of the lung but making mistakes elsewhere, or that the AI has been performing poorly for a certain type of cases (e.g., lungs with prior surgical intervention or other pathology); therefore, despite the AI getting the particular case right, they would not want to certify it for similar images. Collectively these results illustrate that the explanations enable radiologists to engage in complex reasoning about the AI judgements and capacity.

4.3 Results from Experiment 3: Natural Adversarial ImageNet

In Section 4.1 we showed that humans tend to assume that the AI's decision process mirrors their own and that explanations can mitigate this prior assumption. Here we provide further evidence of the mitigation effect using natural adversarial images (see [4] for full details.). Adversarial images are images that cause the AI to be confidently wrong, despite being easily classified by humans. Because humans assume that AI classifiers share their perceptions and beliefs by default, adversarial images are harder for humans to identify as they themselves are not fooled by such cases. Here we test whether explanations help people to predict misclassifications of natural adversarial images, which is an essential prerequisite for effective human oversight of AI systems.

We aim to determine how well humans can predict AI classifications as a function of whether the target image is adversarial, and what explanation features they have access to. To test this we first compared the performance of three nested logistic hierarchical regressions. The simplest model represents the null hypothesis that predictive accuracy differed between participants, target categories, and trial types, but that the explanations did not impact predictive performance. This *null model* was formalized such that human predictive accuracy at the trial level was based on two random intercepts based on participant and target category, respectively, and a fixed effect of trial type (standard correct, standard incorrect, and adversarial incorrect; treating adversarial incorrect as the reference condition). The second model represents the hypothesis that the explanations impacted the predictive performance of the participants, but that explanation

effectiveness was constant across trial types. This *explanation model* expanded on the null model by adding main effects for whether participants were exposed to saliency map explanations and example explanations. The final model represented the hypothesis that the impact of the two explanation features (examples and saliency maps) were not additive, and that they varied between trial types. This *interaction model* built on the explanation model by adding interaction terms for the two explanation features and the trial types. The explanation model captured prediction accuracy better than the null model according to a likelihood ratio test ($\chi 2(2) = 45.37$, $p < .0001$), and the interaction model outperformed the explanation model ($\chi 2(7) = 209.62$, $p < .0001$). These results are consistent with the hypothesis that explanations did impact performance differently for different trial types. To explore these effects more thoroughly we studied the coefficients of the interaction model, see Figure 13.

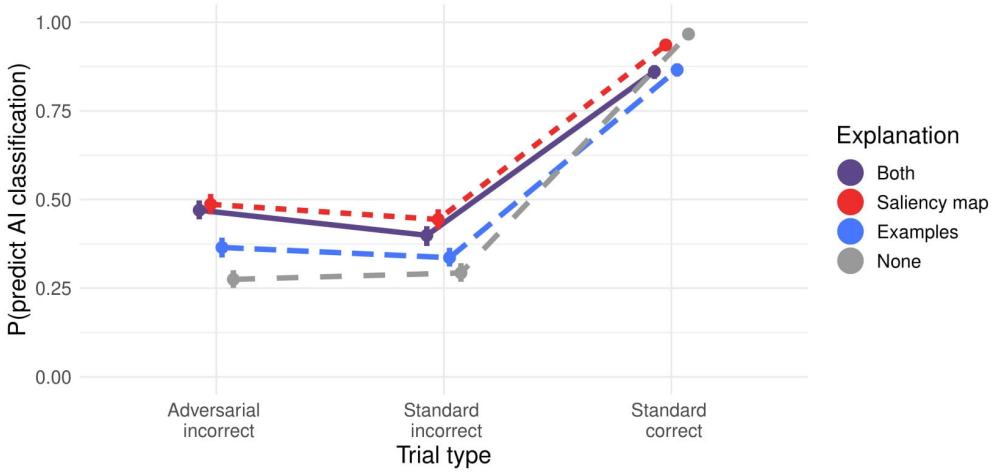

Figure 13. Effect of exemplar and saliency map explanations on predictive performance

Absent intervention, human predictive accuracy is similar for standard incorrect trials and adversarial incorrect trials, but much higher for standard correct images. This may imply that absent explanations, humans tend to assume that the AI makes correct classifications, in line with previous results showing that the default human assumption is that AI classifications will match their own. While both saliency maps and examples significantly improve predictive performance on adversarial images, this improvement is about four times larger for the saliency maps. Additionally, the effect of the two explanation features is not additive. Comparing the relative benefit of explanations on standard incorrect trials versus adversarial trials, we note that the improvement from saliency maps is significantly smaller for standard trials relative to adversarial trials. The improvement from examples is also smaller for standard incorrect trials, but not significantly so. Finally, for the standard correct trials all interventions are associated with a decrease in performance. The decrease in performance for the standard correct trials is likely due to the familiarity modulation shown in Figure 10 in Section 4.1.4, where explanations decrease fidelity (equivalent to human predictive accuracy) for categories that people are very familiar with.

4.4 Results from Experiment 4: Facial Expression

In this section we explore using Bayesian Teaching to provide a global explanation for AI classification of abstract image categories—facial expressions. We explain a prototype-based machine learning model (Probabilistic Linear Discriminant Analysis), formalize the explanation as finding pedagogical examples for this ML model, and run a classification experiment to test the effectiveness on explanatory examples chosen. The effectiveness of explanation is measured by how well humans can predict the machine learning model's predictions. Our results indicate that explanatory examples selected by Bayesian Teaching are helpful for explaining what the ML model learned about abstract image categories. Please see [5] for full details.

4.4.1 Participant Judgments Correlate with the Behavior of the Simulated Learner.

Similar to Section 4.1.2, we first ask: did participant's judgments actually match the behavior of the simulated learner? If so, then the Bayesian Teaching approach holds promise in generating teaching sets (aka explanatory examples) that influence human responses. To verify this, for each trial we examined the probability that the simulated learner would choose the correct category (correct is w.r.t. the target/ML model's prediction of the target image) given the two sets of examples selected by Bayesian Teaching, and compared this to how well human behavior matched the target/ML model. The result is illustrated in Figure 14 (left). The results indicate that the simulated learner matched how humans responded in the task ($r(262) = 0.49$, $p < .001$).

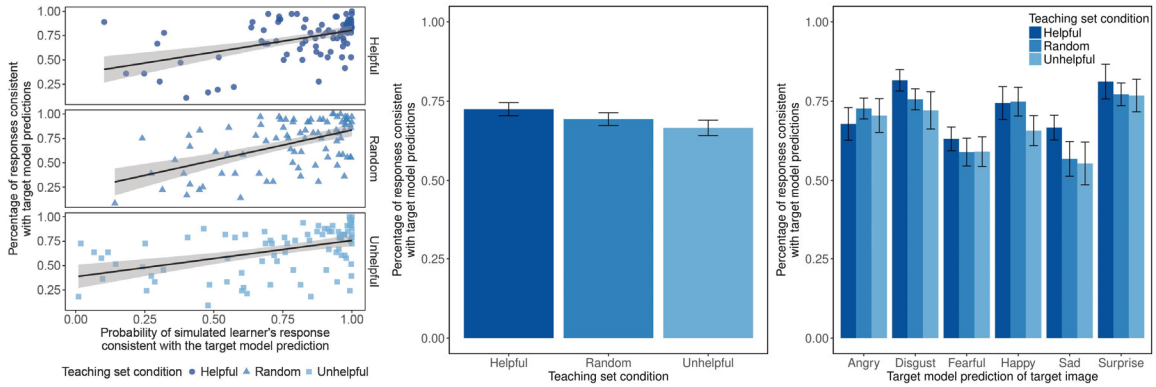

Figure 14. Bayesian Teaching for facial expression classification

4.4.2. Explanatory Examples Help Participants Predict Model Behavior.

Next, we ask: did the various *teaching set* (aka explanatory examples) conditions lead to differences in how well participants' responses matched the model predictions? Mean performance across the three teaching set conditions is shown in Figure 14 in the middle. Performance was highest for participants in the Helpful condition (mean = 72.5%, SD = 2.1%), followed by the Random condition (mean = 69.3%, SD = 2.0%) and finally the Unhelpful condition (mean = 66.6%, SD = 2.4%). We conducted a planned contrast across the different teaching set conditions (with Helpful = 1, Random = 0 and Unhelpful = -1) and found a

significant effect of teaching set condition on accuracy to the target model predictions ($F(1, 102) = 6.29$, $p = .013$).

We further explored how well participants' responses matched the model's predictions for each emotion separately, as shown on the right in Figure 14. A two-way ANOVA revealed significant main effects for teaching set condition ($F(2, 612) = 8.87$, $p < .001$) and emotion category ($F(5, 612) = 32.48$, $p < .001$). There was a marginal, but not significant interaction between the two variables ($F(10, 612) = 1.84$, $p = .051$), suggesting that the effect of teaching set condition was consistent across emotion categories.

4.5 Results from Experiment 5: Psychological Theory of Explainability

Modern AI systems, powered by deep neural networks, are notoriously opaque, making supervision and safe deployment challenging. The field of explainable AI has produced many techniques for improving the legibility of AI decisions to human users and regulators. XAI has focused on developing new methods that show high performance on technical metrics related to faithfulness and explanation complexity, but there is no way to assess which methods will do well for a given use-case. In other words, there is no theory of explainability.

The goal of XAI is for humans to understand a target AI system. 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. We propose that humans’ inferential processes about AI systems occur in the same way that people model any other agent. Thus, we propose a theory that draws 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. 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 $\rho = .86$).

We tested six pre-registered hypotheses regarding our theory (please see [7] for full details): **(1)** 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. **(2)** Successful explanations should inform the above 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. **(3)** The effect of explanation on human beliefs, i.e., the shift in belief post explanation, is quantitatively predicted by the proposed theory. **(4)** A 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. **(5)** A theory that compares projected and observed explanations in a psychologically natural similarity space will match human beliefs better than that in a less natural similarity space (e.g., L1 distance). **(6)** A monotonically

decaying likelihood will capture human beliefs better than the non-monotonic alternative would, following Shepard's universal law of generalization.

All of our hypotheses were supported. **(1)** 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. **(2)** 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 15 A), and the impact of explanations on fidelity was reduced when the AI is correct ($\beta = -0.17$, SE = 0.05, $t = -3.40$, $p < .001$). **(3)** 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 = 0.05, $t = -2.77$, $p = .006$; see Figure 15 B). 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 15 C).

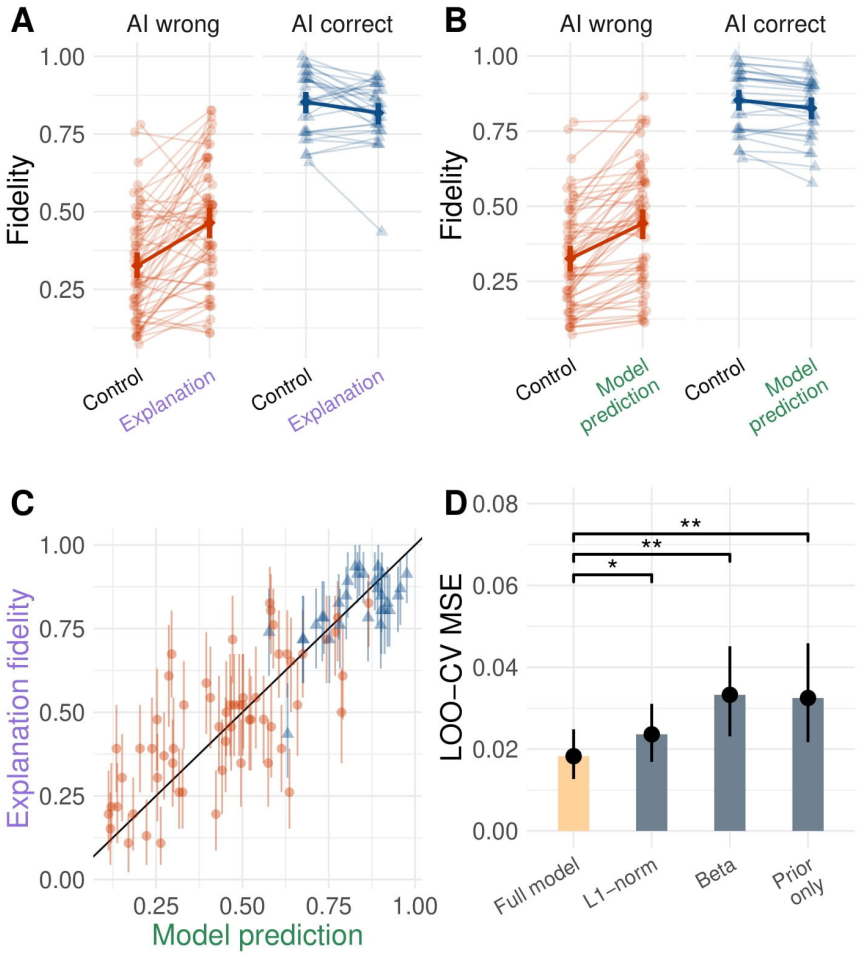

Figure 15. Validation of our psychological theory of explainability

Moreover, model comparisons demonstrate the importance of the psychologically informed components of our cognitive model as predicted in hypotheses 4-6. **(4)** 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 15 D), implying that our likelihood function captured explanation-specific belief-updating. **(5)** 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. **(6)** 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.

Comparing Figure 15 A 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. 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 15 A). In particular, some explanations were poor and resulted in a reduction in fidelity. The full model could capture the change in fidelity across all these variations, including explanations of poor quality.

4.6 Results from Computational Study of Melanoma Classification

Because user studies are expensive and challenging to run, especially when targeting expert users—such as radiologists to evaluate medical imaging AI—automatic metrics of explainability are attractive since they allow for some analysis and pre-screening of XAI methods. Most existing automatic metrics of explainability focus on *faithfulness*, which measures whether the explanations accurately highlight features that are important to the AI decision. In contrast, there exists no automatic metrics for *interpretability* that evaluate how understandable the explanation is to the human user.

We propose to evaluate the human interpretability of saliency methods along two dimensions, one based on low-level visual coherence, and the other based on whether the regions highlighted by the explanation match user expectations, captured by *textbook features*). To assess **perceptual interpretability** we rely on research on human perception of randomness. Research shows that perceived randomness is largely correlated with the probability of alternation, i.e., the strength and frequency of how neighboring pixels differ in appearance. Our **semantic interpretability** metric rests on the assumption that people find saliency maps more interpretable the more they overlap with the features humans find important for the target class. For certain high-stakes classification problems the features that humans pay attention to are well-documented. We refer to these features as textbook features. Textbook features are common in the medical domain, where clinical textbooks list a number of features that are diagnostic of a particular pathology. For example, dermatologists screen for melanoma by using the ABCDE-rule, which determines the likelihood of melanoma based on five features.

Rules such as the ABCDE-rule can easily be captured by simple classifiers, in contrast to the majority of human decision-making that is more complex. We term classifiers that are based on well-established features and standardized decision-rules textbook feature models (TFMs). Crucially, TFMs exist for important decision-problems where outcomes are reliably linked to a limited set of coherent features. In other words, TFMs are simple models that capture some considerations of expert human decision-makers. Therefore, TFMs can be used as proxies for human decision-makers when developing and evaluating XAI explanation methods.

We demonstrate the usefulness of visual and semantic interpretability and a well-established faithfulness metric by exploring the relationships among them in the context of melanoma classification. **(1)** We first use TFMs to validate whether the textbook features are encoded properly: a TFM model that is accurate above chance level captures human expert knowledge to some extent. **(2)** We show the relationships among visual interpretability, semantic interpretability, and an existing faithfulness metric (iAUC). The relationships suggest that there is no single best saliency method across all lesion images. **(3)** We show how these metrics can be used to adaptively combine different saliency methods to maximize user-understanding across the data distribution. Please refer to [8] for full details.

4.6.1 Performance of Textbook Feature Models Is Similar to Junior Dermatologists.

If the textbook features are reliably diagnostic of melanoma, the TFM will classify melanoma above chance. As expected, the TFM is more accurate in assigning melanoma to the test images than the null model (Accuracy TFM = 62%, Accuracy null = 50%, Log-Likelihood Ratio = 19.97). For comparison, dermatologists' accuracy can range from 56% to 80% depending on experience. The TFM's accuracy is similar to the accuracy of a dermatologist with 3–5 years of experience, consistent with the idea that the TFM, while incomplete, is a sensible proxy for dermatologists' decision making.

4.6.2 Relationships Among Interpretability Metrics and Faithfulness.

Visual interpretability vs semantic interpretability. We do not see a strong correlation between the two interpretability metrics across methods (pooled Pearson correlation = .06), suggesting that the two metrics are mostly independent and it is informative to measure both when evaluating the interpretability of a method. We do observe two clusters along the visual interpretability dimension: RISE and Integrated Gradients have much higher visual incoherence relative to the other four saliency methods. The mean visual incoherence score of Integrated Gradients and RISE is 4.5 and 6.6 times greater than the mean of the other cluster, respectively. On the dimension of semantic interpretability, the textbook feature overlap score is overall negative, indicating that most saliency methods generate maps with more saliency concentrated outside the textbook features than within the features relative to a uniform map. This result suggests that the classifier attends to regions that do not fully align with the regions highlighted by the textbook features. The Occlusion method for the benign cases is the one method that has mean positive textbook feature overlap.

Visual interpretability vs Faithfulness. There is a positive relationship between faithfulness and visual incoherence for benign cases, and a negative relationship for melanoma cases (Figure 16 B). A negative relationship between visual incoherence and iAUC is desirable. Such a negative correlation implies that methods are perceptually interpretable when they are also faithful, and less so when they are less faithful. We suspect that the positive correlation for the benign class is caused by the tendency to classify a random set of pixels as benign (because the benign class is more prevalent, it corresponds to the model's default assumption). Thus, when incoherence is high (meaning that the distribution of salience pixels is more random), the iAUC tends to be high for the benign class. Conversely, the negative relationship for the melanoma class is likely caused by that class's strong reliance on a specific set of pixels. For the melanoma class, low incoherence and high iAUC together suggest the saliency map is both structured and targeting the right pixels.

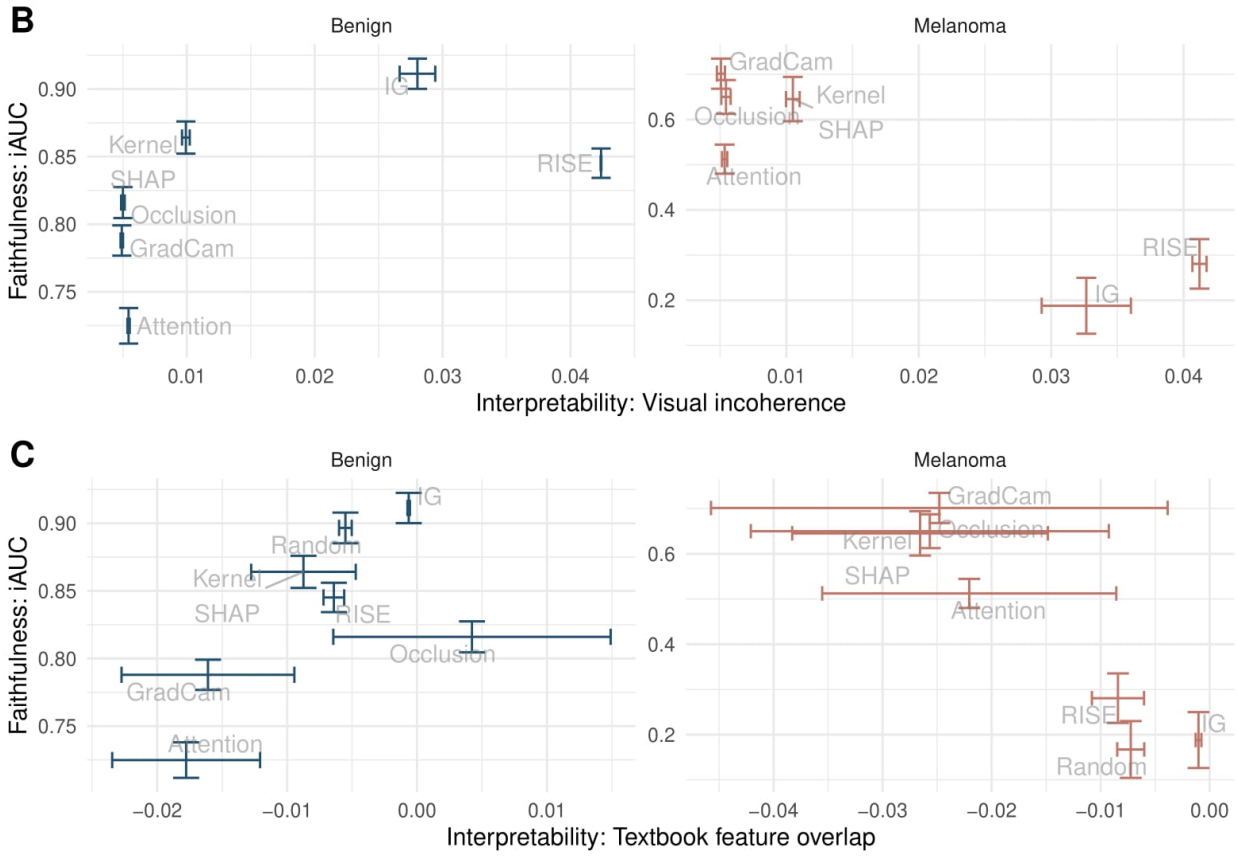

Figure 16. Faithfulness vs. perceptual and semantic interpretability

Semantic interpretability vs Faithfulness. Textbook feature overlap and iAUC are positively correlated for the benign class, and negatively correlated for the melanoma class (Figure 16 C). Because high textbook feature overlap is desirable, we would like to see positive relationships between these two metrics, as it would imply that explanations that are semantically meaningful to experts are also faithful to the DNN. We see this positive relationship for the benign class but observe a negative relationship for the melanoma class. This negative relationship for the

melanoma class, together with the fact that most saliency methods have higher iAUC scores than randomly generated masks but lower textbook feature overlap than uniform masks, indicates a trade-off between faithfulness and semantic interpretability. The trade-off implies that the DNN is more likely to attend to pixels outside the feature masks than inside the feature masks for the melanoma class, suggesting potential discrepancies between the DNN's and experts' decision process.

Interim conclusion. Collectively, these results imply that there is no single best saliency method that is both human interpretable and faithful to the AI for this use case, as illustrated by: (i) the lack of consistent negative correlations between visual incoherence and textbook feature overlap, (ii) visual incoherence and faithfulness, and (iii) the lack of a positive correlation between textbook feature overlap and faithfulness.

4.6.3. Adaptive Saliency Map Selection.

Despite the absence of a consistently high-performing saliency method we can leverage our metrics to adaptively select the saliency map that is best suited to a given image from among the saliency methods. We can achieve this by creating a ranking of the saliency methods within each image, ranging from 1 (worst) to 6 (best), since we evaluated six methods in this study. We do this for each of the three metrics and then sum the ranks over them (ranges from 3 to 18). We then select the method with the highest summed rank for each image. For the adaptive approach to be valuable, the selected saliency maps should consistently have high summed ranks (implying high performance across metrics) and be distributed across methods (suggesting that different methods perform well in different subregions of the test data distribution). Figure 17 shows that each saliency method is the best method for at least one image, confirming that different saliency methods work well for different cases. This adaptive method outperforms the static method of choosing the Occlusion, the best method on average (paired-sample Wilcoxon: V = 9537; p = .001).

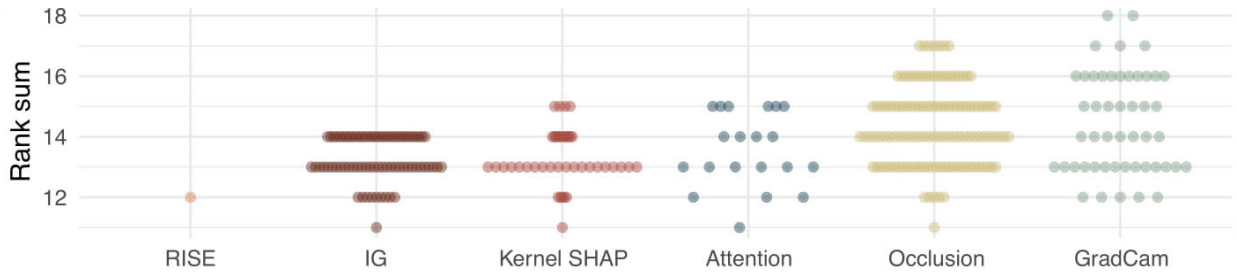

Figure 17. The Distribution of rank sum of the highest ranked saliency methods

4.7 Explanation as Cooperative Communication

In Section 3.1 we positioned Bayesian Teaching as a language to understand all XAI methods. Through Bayesian Teaching we have also framed the problem of XAI as a problem of communication between an explainer and an explainee who is modeled explicitly. In this section we provide an overview of our work on cooperative communication, a generalization of Bayesian Teaching. In Bayesian Teaching the explainer considers the explainee's inference, but the explainee does not reason about the explainer's selection. In contrast, cooperative communication admits the infinite recursive reasoning between the explainer and the explainee. As such, cooperative communication offers mathematical bounds and guarantees on communication / explanation effectiveness.

The scenario of communication entertained here is that the explainer selects the data and the explainee infers from the data the intended hypothesis. The term "cooperative" means that the two agents share the same goal of reaching the target inference, also referred to as the hypothesis intended to be transmitted. Under this setup, we prove the conditions that allowed perfect one-shot transmission of information from the explainer to the explainee [9]. The conditions are: (1) perfectly shared common ground between the explainer and the explainee and (2) a particular form of the data and hypothesis space where the consistency matrix formed by data-hypothesis pairs is triangular.

We then show that the above conditions can be relaxed [10]. That is, cooperative communication is provably stable under perturbations of common ground, and the data-hypothesis consistency matrix can be a rectangular matrix of any shape. The former relaxation ensures robustness of cooperative communication when the agents' beliefs differ, making cooperative communication a viable model in practice. The latter relaxation extends the theory to any discrete model of data and hypothesis. In the same work, we also derive efficient computation of bounds on communication effectiveness directly from the shared data-hypothesis space without the need to simulate the recursive reasoning. Lastly, we find a geometric interpretation to cooperative communication that reveals connections to established areas of importance sampling and optimal transport.

The connection to optimal transport, in particular, to entropy-regularized optimal transport, allows unification of previous models of communication, pedagogical reasoning, rational speech act theory, naive utility calculus, and Bayesian Teaching [11]. More importantly, the connection to optimal transport opens the way to more thoroughly analyze the robustness of cooperative communication [11]. First, we show that entropy-regularized optimal transport, and thereby cooperative communication, is statistically and information theoretically optimal. Then, we show that because optimal transport plans are infinitely differentiable, agents engaging in cooperative communication can reconstruct a corrective cooperative plan using linear approximation once they realized the deviation from the previously assumed common ground. This result not only highlights the stability of communication but also reveals a way to update common ground on-the-go, which is a crucial feature of communication.

Another feature of communication is that communication is rarely a one-shot dump of information but a sequential delivery of data. In [12] we provide the theoretical foundation for

sequential cooperative communication. Specifically, we prove the consistency of sequential cooperative communication—that is, the explainee is guaranteed to reach the intended hypothesis under this communication scheme—and its rate of convergence. We also show that sequential cooperative communication allows the explainee to reach the intended hypothesis with fewer samples than randomly sampled data in most situations. Lastly, we conduct a thorough simulation study on how convergence is influenced by various types and sizes of perturbations. These results present a rigorous foundation for sequential cooperative communication.

In conclusion, we aim to establish the theoretical foundation of cooperative communication, of which explanation is a special case. We prove a series of desiderata for cooperative communication, including its optimality, consistency, and stability. The theory sets up the theoretical foundation for sequential explanation with imperfectly aligned common ground, pointing to new avenues of XAI research. As an interesting side, cooperative communication is not only a multi-agent model interaction but can also be modified to be a single-agent model for information foraging [13].

4.8 Understanding Deep Neural Networks through Deep Gaussian Processes

The success of deep learning models is generally perceived as stemming from greater expressivity which results in powerful generalization. However, deep learning models are viewed as black boxes because their complexity arises from the enormous number of parameters and possible choices for different structures and activation units. Understanding these models remains an open and challenging problem. It has been demonstrated that the characteristics of single-layer neural networks (NNs) can be understood from its effective kernel showing non-stationary and non-local correlation. It is thus appealing to create more interpretable methods through the correspondence between deep learning and kernel-based methods which have the advantage of an explicit mathematical formalization.

In this line of work we investigate a class of kernel-based methods called deep Gaussian processes (DGP). It is well known that a GP is equivalent to a NN with one hidden layer of infinite units. Being the infinite-width limit of a neural network, GP has the expressive power of a NN in learning complex representations, but is also more amenable to mathematical analysis. Thus, we analyze the statistical properties of DGP—layers of GP stacked on top of each other—to gain insights into the behavior of DNN.

Our foundational contribution is the derivation of analytic forms for the second and fourth moments of a variety of DGP [14]. Our derivation shows that by stacking GPs (and by implication, NNs), the fourth moment of a DGP exhibits heavy-tailedness, indicating that DGP / DNN are suitable to capture rare, extreme signals. An analysis of the second moment of DGP shows that the depth of the DGP / DNN gives rise to the ability to capture signals at multiple length scales with long range correlations. An analysis of the derivatives of the moments reveals parameter regimes where the behavior of DGP / DNN becomes chaotic and pathological. Furthermore, we condense a DGP into a simple GP by taking the second moment of the DGP as the kernel of the condensed GP. Such distillation allows us to instantiate the condensed GP computationally to visualize its inductive bias and data-fitting ability. To aid such computational

studies, we also develop a novel training method for certain classes of GPs that is orders of magnitude faster than traditional methods on large data sets [15].

In [16] we make further connections between DNN and DGP by considering a more tractable version of DGP called conditional DGP. The difference between the DGP and the conditional DGP is that the intermediate layers in the latter are conditioned on the hyperdata, which are learned representation of the data embedded in those layers. The hyperdata reveals a spectrum of DGP behavior: In one extreme, when the hyperdata are dense, they make the intermediate GPs approximately deterministic functions, and the conditional DGP recovers a popular kernel-method called deep kernel learning, indicating that the behavior of the conditional DGP can be well-summarized by a single kernel. On the other hand, when the hyperdata are diffuse, the intermediate GPs are representations of random functions passing through the hyperdata, and the conditional DGP can be viewed as an ensemble of deep kernels. In short, these learned hyperdata provide an analyzable and visualizable summary of the distributional properties of the representation generated in deep learning models.

Another way that conditional DGP can provide insights into deep learning representation is that it allows the learning of effective kernels over different sections of the DGP as a summary of the intermediate representation. We study such kernel learning in condition DGP with multi-fidelity data—data consisting of multiple sources and different precision [17]. By deriving *closed-form* kernels for the intermediate GPs layers, we show that low fidelity data lead to non-stationary kernels that generate high-frequency signals. The effective kernels learned for different sections of the conditional DGP show little interference and thus provide a compositional understanding of the representational learned from multi-fidelity data.

In conclusion, we study deep Gaussian processes (DGP)—a close relative of deep neural networks (DNN) that is more amenable to analysis—to inform the behavior of DNN. In our analysis of DGPs, we derive closed analytical forms for the second and fourth moments of the DGP. These forms provide the mathematical foundations for well-known behavior in DNNs, such as their ability to capture non-local and non-stationary data. We also introduce a more tractable version of DGP called conditional DGP. The conditional DGP allows us to summarize behavior of intermediate layers through hyperdata and effective kernels. The hyperdata and effective kernels are analyzable and visualizable mathematical objects that provide rigorous understanding of the representation learned in deep learning systems.

5 CONCLUSIONS

In this report we gave an overview of how our work contributed to the development of novel XAI techniques, demonstrated successful XAI systems with empirical validation, advanced our understanding of the psychology of explainability, and solidified the mathematical foundations of XAI. The theme of our approach is to incorporate human inference into XAI. The human-centeredness of our approach is evident in the explainee-centered framework of Bayesian Teaching, the comprehensive suite of human experiments reported, and the establishment of the theory of cooperative communication. Our effort is driven by a key persisting question in the field of XAI: "How do we know the explanations generated are effective at all?" We believe that the work reported here, especially our contribution to the psychological theory of explainability, paves the way for us to answer this key question.

6 REFERENCES

[1] Yang, S. C.-H., Shafto, P., "Explainable Artificial Intelligence via Bayesian Teaching." *NIPS 2017 workshop on Teaching Machines, Robots, and Humans*, Long Beach, 2017.

[2] Yang, S. C.-H., Folke, T., Shafto, P., "Abstraction, Validation, and Generalization for Explainable Artificial Intelligence," *Applied AI Letters* **2**:e37, 2021.

[3] Yang, S. C.-H., Vong, W. K., Sojitra, R. B., Tomas Folke, Shafto, P., "Mitigating belief projection in explainable artificial intelligence via Bayesian Teaching," *Scientific Reports* **11**:9863, 2021.

[4] Folke, T., Yang, S. C.-H., Li, Z., Sojitra, R. B., Shafto, P., "Explainable AI for Natural Adversarial Images," *ICLR-2021 Workshop on Responsible AI*, 2021.

[5] Vong, W. K., Sojitra, R. B., Reyes, A., Yang, S. C.-H., Shafto, P., "Bayesian Teaching of image categories." *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, Madison, 2018.

[6] Folke, T., Yang, S. C.-H., Anderson, S. P., Shafto, P., "Explainable AI for medical imaging: Explaining pneumothorax diagnoses with Bayesian Teaching," *Proc. SPIE 11746, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*, 117462J, 2021.

[7] Yang, S. C.-H., Folke, T., Shafto, P., "A Psychological Theory of Explainability," *Proceedings of the 39th International Conference on Machine Learning*, PMLR **162**:25007-25021, 2022.

[8] Bokadia, H., Yang, S. C.-H., Li, Z., Folke, T., Shafto, P., "Evaluating perceptual and semantic interpretability of saliency methods: A case study of melanoma," Accepted by *Applied AI Letters*, 2022.

[9] Yang, S. C.-H., Yu, Y., Givchi, A., Wang, P., Vong, W. K., Shafto, P., "Optimal Cooperative Inference," *Proceedings of the 21st international conference on Artificial Intelligence and Statistics*, PMLR **84**:376-385, 2018.

[10] Wang, P., Paranamana, P., Shafto, P., "Generalizing the theory of cooperative inference," *Proceedings of the 22nd international conference on Artificial Intelligence and Statistics*, PMLR **89**:1841-1850, 2019.

[11] Wang, P., Wang, J., Paranamana, P., Shafto, P., "A mathematical theory of cooperative communication," *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020.

[12] Wang, J., Wang, P., Shafto, P., "Sequential cooperative Bayesian inference," *Proceedings of the 37th International Conference on Machine Learning*, PMLR **119**:10039-10049, 2020.

[13] Yang, S.C-H., Vong, W.K., Yu, Y., Shafto, P., "A unifying computational framework for teaching and active learning," *Topics in Cognitive Science* **11**:316-317, 2019.

[14] Lu, C.-K., Yang, S. C.-H., Hao, X., Shafto, P., "Interpretable deep Gaussian Processes with moments," *Proceedings of the 23rd international conference on Artificial Intelligence and Statistics*, PMLR **108**:613-623, 2020.

[15] Lu, C.-K., Yang, S. C.-H., Shafto, P., "Standing Wave Decomposition Gaussian Process." *Physical Review E* **98**:032303, 2018.

[16] Lu, C.-K., Shafto, P., "Conditional Deep Gaussian Processes: empirical Bayes hyperdata learning," *Entropy* **23**:1387, 2021.

[17] Lu, C.-K., Shafto, P., "Conditional Deep Gaussian Processes: multi-fidelity kernel learning," *Entropy* **23**:1545, 2021.