scientific reports

OPEN

Mitigating belief projection in explainable artificial intelligence via Bayesian teaching

Scott Cheng-Hsin Yang[1,4]✉, Wai Keen Vong[2,4], Ravi B. Sojitra[3,4], Tomas Folke[1] & Patrick Shafto[1]

State-of-the-art deep-learning systems use decision rules that are challenging for humans to model. Explainable AI (XAI) attempts to improve human understanding but rarely accounts for how people typically reason about unfamiliar agents. We propose explicitly modelling the human explainee via Bayesian teaching, which evaluates explanations by how much they shift explainees' inferences toward a desired goal. We assess Bayesian teaching in a binary image classification task across a variety of contexts. Absent intervention, participants predict that the AI's classifications will match their own, but explanations generated by Bayesian teaching improve their ability to predict the AI's judgements by moving them away from this prior belief. Bayesian teaching further allows each case to be broken down into sub-examples (here saliency maps). These sub-examples complement whole examples by improving error detection for familiar categories, whereas whole examples help predict correct AI judgements of unfamiliar cases.

While Artificial Intelligence (AI) can help address socially-relevant problems[1–3], it is important for humans to be able to scrutinize AI decisions so we may audit, understand, and improve performance; indeed, this is legally mandated in certain contexts[4,5]. The best performing AI algorithms rely on complex decision rules based on features that feel alien to most humans[6]. The abstruseness of these AI models impedes their adaptation in high-leverage contexts, emphasizing the need for successful explanations that facilitate human understanding and prediction of the AI's behavior.

A popular class of methods to explain AI systems is *explanation-by-examples*. Explanation-by-examples takes as input an AI model to be explained and the data that it has been trained on and produces as output a small sub-set of training data that exert high impact on the inference of the explainee. For example, if the aim is to explain whether a deep-learning model would classify a given image as a cat or a dog, explanation-by-examples selects the cat and dog images that are most representative of those categories. The utility of explanation-by-examples is supported by research that confirms humans' ability to induce principles from a few examples[7–10] as well as the extensive use of examples in education[11–13]. The explanation-by-examples approach has many desirable proper-ties: It is fully model-agnostic and applicable to all types of machine learning[14–16]; it is domain- and modality-general[17,18]; and it can be used to generate both global explanation[19–23] and local explanation[24–26]. Although the technology of explanation-by-examples for XAI has been developed for at least two decades[27,28], empirical tests and connections to its ecological roots in the social sciences have been limited.

Explanation-by-examples can be considered a social teaching act, which can be formally captured by *Bayesian teaching*[29]. In Bayesian teaching, there are two parties, a teacher (explainer) who selects examples and an explainee (learner) who draws inferences. The teacher selects examples intended to maximize the explainee's probability of a correct inference based on the teacher's model of the explainee's current beliefs and their inductive biases[30–32]; the explainee uses Bayesian updating to make predictions given these examples[15,33–35]. Existing work on explanation-by-examples has demonstrated explanation effectiveness relative to several baseline conditions[14,20,36,37]; however, there is rarely a principled, *apriori* rationale as to why the proposed improvements should work. By explicating the computations used to model the explainer, the explainee, and the explanation selection process, Bayesian teaching provides testable predictions on the effectiveness of explanatory examples in different contexts.

We use image classification on *the ImageNet 1K dataset*[38] as the testbed. The model to be explained is ResNet-50[39]. Following an ideal-observer approach[40,41], we instantiate Bayesian teaching by selecting examples

[1]Department of Mathematics and Computer Science, Rutgers University, 101 Warren Street, Newark, NJ 07102, USA. [2]Center for Data Science, New York University, 60 5th Ave, New York, NY 10011, USA. [3]Department of Management Science and Engineering, Stanford University, Stanford, USA. [4]These authors contributed equally: Scott Cheng-Hsin Yang, Wai Keen Vong and Ravi B. Sojitra. ✉email: scott.cheng.hsin.yang@gmail.com
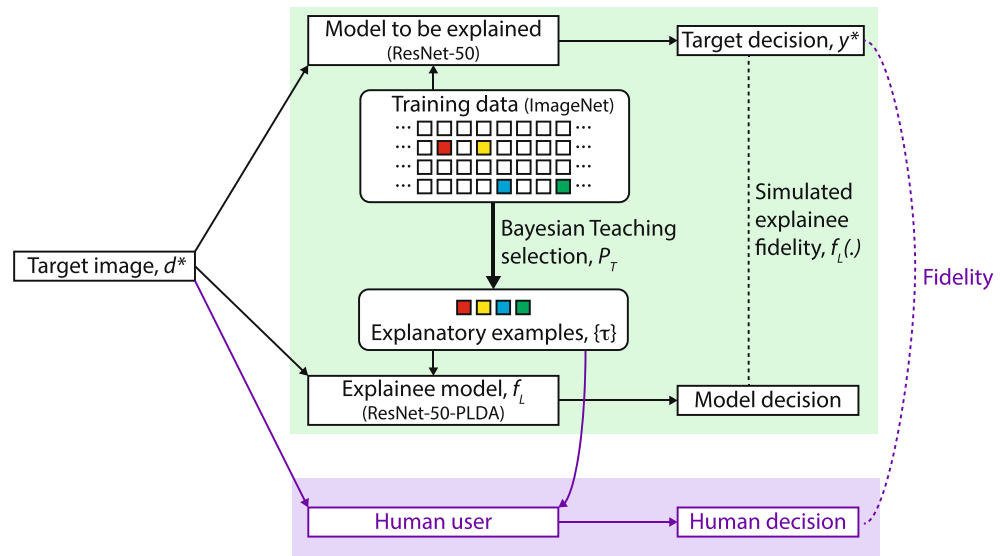
Figure 1. Overview of Bayesian teaching. Green section: Bayesian teaching selects a few examples from the training data as explanatory examples. The explanatory examples are selected such that when the explainee model is trained on those examples, the fidelity between the model decision and the target decision (simulated explainee fidelity) matches a desired value. The same logic is applied to generate the saliency maps, where pixels are treated as examples at a finer granularity. Purple section: The explanatory examples and saliency maps are also shown to human participants. Explanation effectiveness is evaluated by how the examples influence the fidelity between participants' decisions and the target decisions. Symbols correspond to those in Eq. (1); arrows indicate input-output relationship; and dotted line indicate comparison. The models and data used are indicated in parenthesis. This figure was created using Adobe Illustrator CS6 (v. 16.0.0)[43].

with differing degrees of helpfulness as judged by the fidelity between the explainee model and the target model. For the explainee model, we used a ResNet-50-PLDA model, which is a ResNet-50 model where the last softmax layer is replaced by a probabilistic linear discriminate analysis (PLDA) model. This alteration introduces the probabilistic training required by Bayesian teaching while keeping the architecture of ResNet-50, which is known to accurately fit human labels[39]. In the context of image classification, Bayesian teaching can be expressed as

$$P_T\left(\{\tau\}|y^*, d^*\right) \propto f_L\left(y^*|d^*, \{\tau\}\right), \tag{1}$$

where d^* is a target image; y^* is the label predicted by the model to be explained, hence the target decision; $\{\tau\}$ is a set of explanatory examples; f_L is the explainee model; the probability produced by $f_L(\cdot)$ is the simulated explainee fidelity; and $P_T(\cdot)$ determines the probability of selecting a set explanatory examples. See Fig. 1 for an overview and the "Methods" section for further details. Bayesian teaching also allows for selection of examples at different levels of granularity. For the current task, we consider the selection of entire images as well as pixels in an image as explanations. The latter pixel-selection process derived from Bayesian teaching turns out to be mathematically equivalent to a type of feature attribution method called Randomized Input Sampling for Explanations[42]. Thus the two levels of example granularity evaluated in this paper coincide with two popular methods of explanation—explanation-by-examples and saliency maps.

To give a concrete example, consider a trial where a participant tries to predict whether the target AI classifier will classify a certain image as a barn or a flagpole (see Fig. 2). Bayesian teaching operates by selecting the four example images—two of a barn and two of a flagpole—from the training set that are most likely to make the explainee model reach the same judgement as the target model, i.e., high simulated explainee fidelity. The target image and the examples are overlaid with saliency maps where each pixel is weighted by the probability that showing it will guide the explainee model to the same conclusion as the target model.

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: (1) Explanations selected by Bayesian teaching improve the *fidelity* between human prediction of AI classification and actual AI classification; (2) the Bayesian Teacher can correctly infer which explanations humans will prefer; and (3) the Bayesian Teacher can accurately predict both which explanation will improve fidelity and which explanations will decrease it. Additionally, we show how the prior beliefs of human participants can be mitigated by appropriate explanations. Consistent with existing work from psychology[44,45], we find that human participants project their own beliefs onto the AI system. This belief-projection manifests as (4) fidelity being higher when the AI is correct relative to when it is wrong, (5) this impact of AI correctness on fidelity being particularly pronounced for familiar categories, and (6) these effects being mitigated by appropriate explanations. We provide justifications and intuitions for these six points in the

2

following paragraphs. To the best of our knowledge, this is the first paper to empirically explore the implications of human belief-projection for explainable AI.

The core prediction of Bayesian teaching is that explanations which lead the explainee model to correct predictions will help humans to better understand the AI. We test this by evaluating whether participants exposed to helpful examples and saliency maps are better able to predict the AI system's classifications than participants who do not view any explanations. Returning to our example in Fig. 2, this means that a participant who is shown the example images and saliency maps are more likely to correctly predict that the AI classified the image as a flagpole rather than a barn, relative to a participant who is only shown the target image without any explanation. This is a generous test of Bayesian teaching, but a necessary one, because failing this test would make all subsequent results moot. Provided that the explainee model match human users reasonably well, we expect that examples selected to be helpful by the Bayesian Teacher will be preferred over examples that are selected to be unhelpful or at random. A stricter test of the appropriateness of the Bayesian teaching is whether it can predict both explanations that improve the fidelity of human predictions and those that lead to reduced fidelity. Such calibration implies that it is not every explanation improves fidelity, but that explanations need to be curated to reach a desired result. In our experimental setup this would manifest in examples that are judged to be helpful or detrimental by the Bayesian Teacher increasing or reducing the fidelity of the participants' predictions to the AI judgments, respectively.

If human participants project their beliefs onto the AI system, they will expect the AI classifier to be highly accurate because they themselves perform well at image classification[38]. In our experiment this translates to humans who predict AI classifications achieving higher sensitivity (correctly predicting AI's correct classifications) than specificity (correctly predicting AI's mistakes), absent explanation. In the context of our example: since the target image is showing a barn, a participant not given any explanation should typically (incorrectly) predict that the AI will classify the image as a barn rather than a flagpole. However, this effect should not be uniform across trials because some categories are easier to distinguish than others. Since more familiar categories should be easier to distinguish, and since participants expect the model to get the right answer for trials they themselves find easy, belief projection implies that *familiarity should increase fidelity for model hits*. Conversely, *familiarity should decrease fidelity for model errors*. Introducing a different example, a participant who is familiar with dogs will find the discrimination between yorkshire terrier and silky terrier easy, whereas someone less familiar with dogs might struggle with the first-order categorization, and consequently be more willing to consider the AI classifier making a mistake.

If explanations generated by Bayesian teaching operates by mitigating belief-projection, we would expect them to reduce the gap between sensitivity and specificity by increasing the latter (improving error detection). Additionally, the belief-projection implies that examples improve fidelity the most for unfamiliar categories, whereas saliency maps improve fidelity most for familiar categories. The reason why examples are most beneficial for unfamiliar categories is that they could strengthen category distinctions for unfamiliar categories with fuzzier mental representations. In the context of the two breeds of terrier: someone who is unfamiliar with dogs can leverage the examples to better understand what features distinguish the two breed, and compare that to the features of the target image. Saliency maps, on the other hand, might be most diagnostic for familiar cases because they highlight features that were consequential to the AI system, and determining the appropriateness of these features requires familiarity with the categories. In the context of the barn versus flagpole example: most people can reliably distinguish between them, so can notice that the saliency map of the target indicates that the AI classifier pays less attention to the house relative to the whethervane, suggesting a potential misclassification.

Results

Methodological overview. User understanding in the context of classification can be captured by how well the user can predict the model's judgement. Throughout this paper we will refer to this predictive capacity as fidelity, referring to the agreement between an agent's prediction (either a participant or theexplainee model) and the judgement of the classifier. A natural measure of explanation effectiveness is how much the explanations increase such fidelity, relative to a control condition. We designed a two-alternative forced choice (2AFC) task in which participants were asked to predict the model's classification of a target image between two given categories. No trial-by-trial feedback was provided to participants. It is important to note that in this task high fidelity does not imply that participants' judgements match the ground truth of the image, which we refer to as first-order accuracy or simply accuracy. It is possible for a participant to have high accuracy (in that their judgements often match the ground-truth category of the image) but poor fidelity (in that their judgements rarely match the AI's).

We designed a total of 15 conditions that vary along three dimensions: (1) presence of informative labels (two levels: [GENERIC LABELS] or ([SPECIFIC LABELS]), (2) types of examples (three levels: [NO EXAMPLES], [HELPFUL] or [RANDOM]), and types of saliency maps (three levels: [NO MAP], [JET] or [BLUR]). The labels dimension indicate whether the images shown where given informative labels (e.g. "Border terrier" or "Norwich terrier") or generic labels (Category A or Category B). The examples dimension indicate whether examples of the two image categories were shown, and if so, if they were selected to be helpful or were drawn from a uniform distribution of helpfulness as determined by Bayesian teaching. The saliency map dimension indicates if the images were overlaid with saliency maps that highlighted which pixels the AI classifier focused on to make its classification. If saliency maps were included, they were either visualized as a semi-transparent jet color map or as an image filter where unimportant pixels where blurred. We found no significant difference between the [BLUR] and [JET] conditions; thus, for increased clarity we use the [MAP] condition, which contains both variants, in the main text. See Supplementary Discussion D2 for the main analyses in the paper repeated with [BLUR] and [JET] coded separately.

		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

Table 1. Naming convention of conditions and the number of participants in each condition. In the main text, conditions are referred to with brackets and the "&" logical operation. For example, [NO EXAMPLES] & [NO MAP] refers to the condition with 76 participants (see "Experimental conditions" in "Methods" section for more detail).
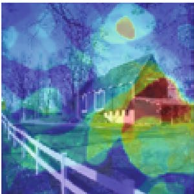
Figure 2. A snapshot of the experiment. See "Methods" section and Table 1 for the naming conventions of the conditions used below. The experimental condition above the black line is [SPECIFIC LABELS] & [HELPFUL] & [BLUR]. Under the black line is the [JET] equivalent of the second row, which is obtained by replacing the blurring maps with the jet color maps. Experimental conditions with generic labels are obtained by replacing specific labels—'Flagpole' and "Barn" in this case—with generic category names—"Category A" and "Category B." Experimental conditions without the saliency maps, i.e., [NO MAP], show only the first row of images. Conditions without examples, i.e., [NO EXAMPLES], show only the first column of image (s). All images and saliency maps shown were 224-by-224 pixels. The prediction of the model to be explained on the target image is "Flagpole" in this case. All photographs are obtained from the open-source *ImageNet 1K* dataset[38]. This figure was created using Adobe Illustrator CS6 (v. 16.0.0)[43].
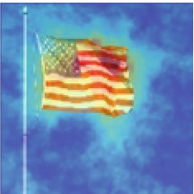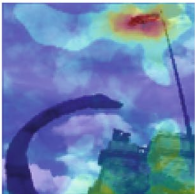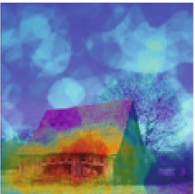
Table 1 shows the sample size of each condition. Figure 2 shows a trial where the categories are represented with informative label, helpful examples, and blur saliency maps.

Each trial has three more distinct features beyond the condition it belongs to: the category accuracy, the simulated explainee fidelity, and a familiarity score. Category accuracy refers to the classification accuracy on the category which the target ResNet-50 model predicts that the target image belongs to (see Supplementary Table T1). Note that in contrast to the *category accuracy* which is an accuracy on the category-level, we use the

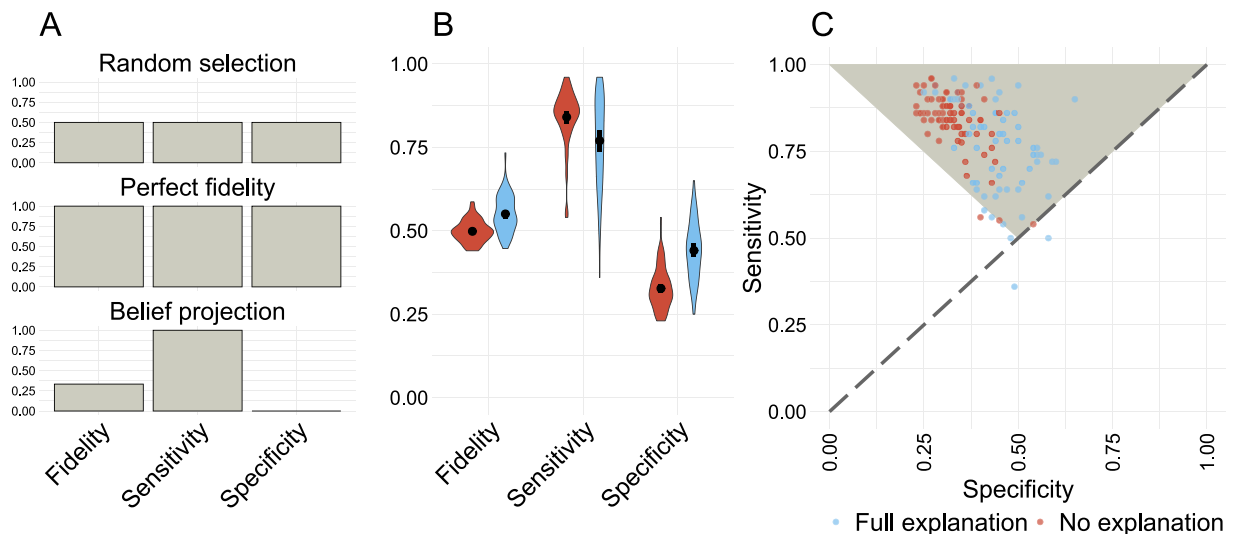**Figure 3.** Bayesian teaching improves fidelity by mitigating belief projection. The effectiveness of examples generated by Bayesian teaching, evaluated by comparing the fidelity of the participants who obtained a full explanation ([SPECIFIC LABELS] & [HELPFUL] & [MAP]; 66 participants; 9899 observations) with a control group ([SPECIFIC LABELS] & [NO EXAMPLES] & [NO MAP]; 76 participants; 11,394 observations). (**A**) Three idealised fidelity profiles, showing the fidelity of: a random agent, a perfect agent, and an agent with perfect access to the ground truth who assumes that the AI classifier always mirror their own predictions (belief projection). (**B**) Human fidelity most closely match the belief projection profile, but the interventions increase specificity (and slightly reduce sensitivity) by making participants better at spotting the AI's errors. The violinplots show the distribution of fidelity within conditions. Black dots show the group mean with error bars signifying 95% bootstrapped confidence intervals. (**C**) Individual participants' sensitivity and specificity. The vertices of the triangle show 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. This figure was created using the ggplot2 package (v. 3.3.2)[46] in R (v. 4.0.3)[47].

term *model correctness* to refer to whether the target model made a correct judgement on a specific trial. The simulated explainee fidelity of a trial (only available in the [EXAMPLES] conditions) is an estimate of the probability that the explainee model's classification would match the target ResNet-50 model's classification, given the categories and examples presented. Finally, in a separate study seven raters indicated their familiarity with each category pairing by stating whether they thought they could correctly match images of the two categories presented to their respective labels. The familiarity score is the mean value across all seven raters. See the "Methods" section for a more technical explanation of these features.

Bayesian teaching improves fidelity. To evaluate whether the XAI interventions improved fidelity we compared participants who obtained a full explanation ([SPECIFIC LABELS] & [HELPFUL] & [MAP]) with a control group that received no explanations ([SPECIFIC LABELS] & [NO EXAMPLES] & [NO MAP]). 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 modelled 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 (see "Methods" section). Absent intervention, participants behave most like the third, belief-projecting, agent (Fig. 3).

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 ($\beta = 0.21(0.03)$, $z = 6.99$, $p < 0.0001$). 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%]; $\beta = 0.49(0.05)$, $z = 9.20$, $p < 0.0001$). 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%]); $\beta = -0.43(0.12)$, $z = -3.68$, $p = 0.0002$), but not enough to offset the specificity gains. 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.

Participants prefer examples that are helpful according to Bayesian teaching. Having established that examples generated by Bayesian teaching improved participants' ability to predict AI judgements, we want to 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. Participants showed a small but reliable preference for helpful relative to random examples and a substantial preference for helpful versus misleading examples. Consistent with our hypothesis that helpful examples are most beneficial for unfamiliar categories, our results show that the preference for helpful examples was particularly pronounced when the image categories were unfamiliar (see Supplementary Discussion D1 for all the details).

Bayesian teaching can predict which explanations improve and reduce fidelity. Bayesian teaching makes explicit the existence of an explainee and suggests that a sound explainee model should have the capacity to track the inference of actual explainees. In our experiment the calibration between the explainee model and the participants is captured by the relationship between category accuracy and participant accuracy. We estimate participant accuracy (their first-order belief about the ground truth) by using their fidelity in the control trials (their second-order belief about the AI classifier with no exposure to explanation). The assumption that their attempt to predict the AI classifier may serve as a proxy of their first-order accuracy is justified given the tendency to belief-project observed in previous sections. We found that participant fidelity (interpreted as accuracy for the control trials) was positively correlated with category accuracy for trials where the model was correct ($\beta = 1.74(0.20)$, $z = 8.67$, $p < 0.0001$), indicating good calibration between the model and participants in this situation (see Supplementary Fig. F1). We also found a negative interaction between category accuracy and model correctness ($\beta = -2.57(0.23)$, $z = -11.03$, $p < 0.0001$). This suggests the poor calibration in the special case in which the model's overall accuracy on the predicted category is high but it misclassifies the particular trial. In sum, these results imply that category accuracy is a good proxy of human ground truth judgements at the aggregate level, which in turn suggests that our explainee model is appropriate for our participants.

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 [EXAMPLES] trials only, 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 ($\chi 2(1, 4) = 71.68$, $p < 0.0001$). 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 ($\chi 2(3, 7) = 7371.28$, $p < 0.0001$). This indicates that how well the category accuracy and/or the modelled helpfulness of the examples shown predicted fidelity differed for trials with correct or incorrect AI judgements.

To explore how model correctness interacted with category accuracy and simulated explainee fidelity, we explored the parameters of the third regression model. Participants are typically better at predicting the AI classifier when it is correct relative to when it is wrong ($\beta = 0.53(0.06)$, $z = 9.15$, $p < 0.0001$). This aligns with our previous results, which suggest that participants have a sense of the ground truth for most trials, and assume that the AI classifier would make the same judgement that they would make. Category accuracy is positively associated with participant fidelity when the AI is wrong ($\beta = 0.59(0.05)$, $z = 12.30$, $p < 0.0001$), and even more strongly associated with fidelity when the AI classifier is correct ($\beta = 0.93(0.09)$, $z = 10.68$, $p < 0.0001$; see Fig. 4). Because there was a significant positive relationship between ResNet accuracy and participant fidelity for both the control trials and the example trials, it seems plausible that the calibration between model and participant observed in the control condition survives the introduction of explanatory examples, at least partially. Finally, while statistically controlling for category accuracy, simulated explainee fidelity did not predict fidelity on trials when the AI classifier was wrong ($\beta = -0.01(0.03)$, $z = -0.16$, $p = 0.89$) but did so for trials when the AI classifier was correct ($\beta = 0.77(0.05)$, $z = 14.19$, $p < 0.0001$). Because the simulated explainee fidelity determined which examples were shown, the fact that this variable could accurately predict human fidelity above and beyond ResNet accuracy implies that the Bayesian Teacher can successfully predict which explanations improve *or impair* the fidelity of participant judgements.

Bayesian teaching improves fidelity through belief-mitigation. The previous results indicate that examples deemed helpful by the Bayesian Teacher improve participant predictions of the AI classifier's judgements. Additionally, participants prefer examples that are helpful according to the Bayesian Teacher, and this preference is particularly pronounced for unfamiliar categories. Next, we will explore *how* explanatory examples improve fidelity, and evaluate the relative importance of the different explanation features employed. The preceding results imply that people belief-project by default: that is, they use their own beliefs as priors for the AI classifier's beliefs. The interventions 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.
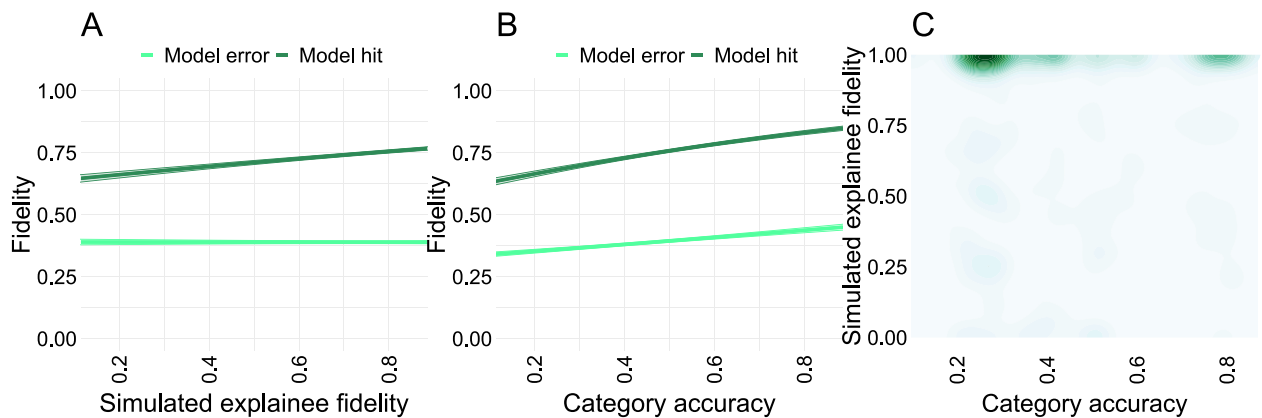
Figure 4. The helpfulness of the presented examples as determined by the Bayesian teacher predicts human fidelity across trials with examples (419 participants; 62,820 observations). (**A**) The simulated explainee fidelity—the helpfulness of the explanatory examples expected by the Bayesian teacher—correlates significantly with participant fidelity for correct trials but not for incorrect trials. This suggests 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. (**B**) Category accuracy is positively associated with participant fidelity, both for trials when the AI classifier is correct and when it is wrong. A similar trend is observed in the control condition (see Supplementary Fig. F1). This suggests that humans and ResNet-50 find the same categories difficult to discriminate, implying that the ResNet architecture can serve as an appropriate model of human participants in this task. The difference in fidelity between when the AI classifier is correct and when the AI classifier is wrong suggests that it is harder to teach incorrect judgements, at least in this context. (**C**). Two-dimensional kernel density with 25 density bins showing the distribution of trials in terms of category accuracy and simulated explainee fidelity. In this study the two are independent. Note that the higher density near perfect simulated explainee fidelity was due to all the helpful examples being selected based on this variable, so they constitute a majority of our example trials. This figure was created using the ggplot2 package (v. 3.3.2)[46] in R (v. 4.0.3)[47].

To further evaluate whether explanations improve fidelity by mitigating belief-projection, we compared how the interventions impacted fidelity and first-order accuracy in the complete data set. Specifically we contrasted [SPECIFIC LABELS] vs [GENERIC LABELS], [MAP] vs [NO MAP], and [EXAMPLES] vs [NO EXAMPLES], while controlling for category accuracy and familiarity score. We ran separate analyses for when the AI classifier was correct and when the AI classifier was wrong, corresponding to the distinction between sensitivity and specificity in previous sections. We will treat the ground truth as a proxy of participant first-order beliefs, a defensible assumption given the reported human accuracy on ImageNet in previous works[38]. Based on this assumption, interventions increasing fidelity while also increasing mismatches to the ground-truth, would shift participant predictions of the AI classifier away from their first-order judgements. The [SPECIFIC LABELS] are associated with higher fidelity than the [GENERIC LABELS] regardless of whether the AI classifier is correct ($\beta = 0.24(0.08)$, $z = 3.06$, $p = 0.002$) or not ($\beta = 0.07(0.03)$, $z = 2.13$, $p = 0.03$). Because these effects are small and orthogonal to belief projection, they will not be discussed further.

The presence of the saliency maps in the [MAP] condition improves fidelity when the AI classifier is wrong ($\beta = 0.43(0.03)$, $z = 14.24$, $p < 0.0001$), but reduces fidelity (to a lesser extent) when the AI classifier is correct ($\beta = -0.56(0.07)$, $z = -7.98$, $p < 0.0001$; see Fig. 5). In both cases, saliency maps reduced the first order-accuracy of the participants (model hit: $\beta = -0.56(0.07)$, $z = -7.98$, $p < 0.0001$; model error: $\beta = -0.43(0.03)$, $z = -14.24$, $p < 0.0001$), meaning that they were less likely to report that the AI classifier's judgements matches 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).

Comparing all [EXAMPLES] trials to all [NO EXAMPLES] trials, the presence of examples do not significantly improve fidelity when the AI classifier is correct ($\beta = -0.13(0.08)$, $z = -1.61$, $p = 0.11$) or when the AI classifier is wrong ($\beta = 0.02(0.03)$, $z = 0.69$, $p = 0.49$). However, in the conditions where examples were present, helpful examples improve fidelity for trials when the AI classifier was correct ($\beta = 0.77(0.08)$, $z = 10.11$, $p < 0.0001$), but not for trials when the AI classifier was wrong ($\beta = 0.06(0.04)$, $z = 1.77$, $p = 0.08$). The positive influence of the helpful but not the random examples illustrates that it is not the mere presence of examples that improves fidelity, but that examples have to be carefully selected to be beneficial. Note also 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 (see Fig. 5). In other words, the saliency maps and the examples serve separate and complementary functions in explaining AI judgements to the participants.
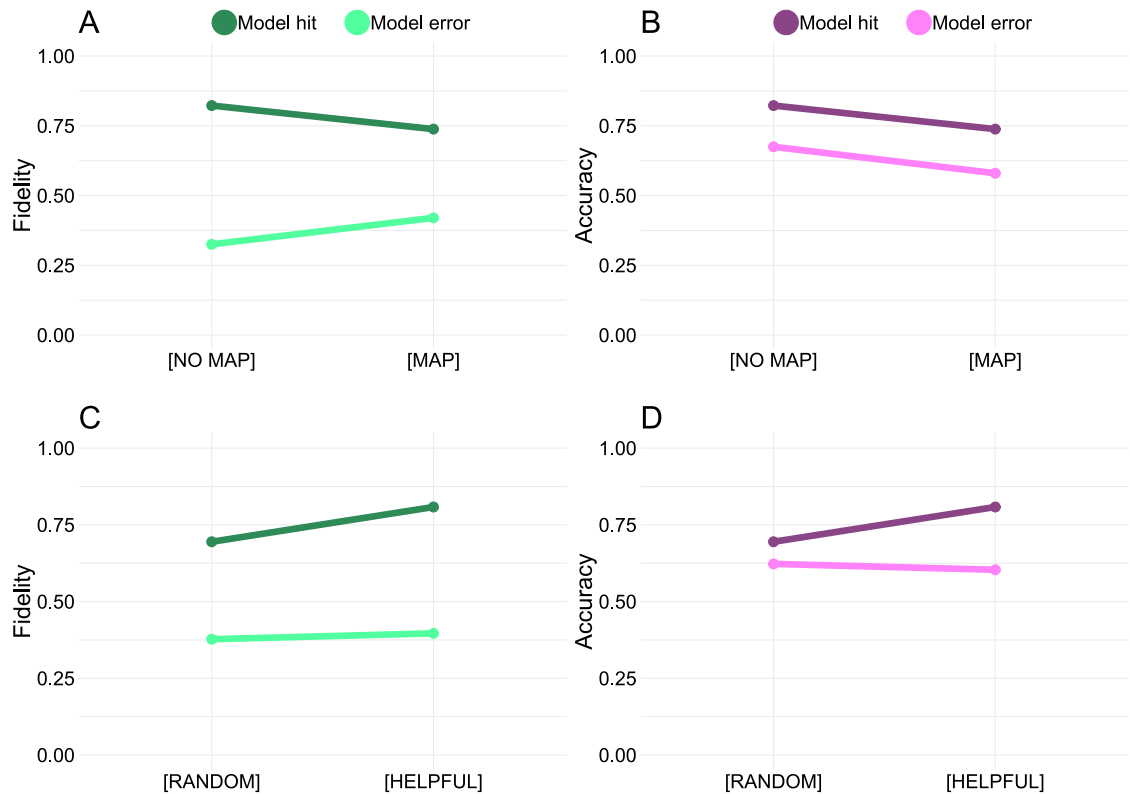
Figure 5. The fidelity between the participant predictions and the AI classifications is higher when the AI is correct than when the AI is wrong. (**A,B**) are based on the entire data set, comparing all [MAP] conditions to all [NO MAP] conditions (631 participants; 94,582 observations). (**C,D**) exclude the [NO EXAMPLES] trials and contrast all [HELPFUL] trials with all [RANDOM] trials (419 participants; 62,820 observations). (**A**) The saliency maps improve fidelity for trials when the AI classifier is wrong but reduce fidelity when the AI classifier is correct. (**B**) The saliency maps make people less likely to predict that the AI classification of the target image matches the ground truth. Together, (**A,B**) imply that the saliency maps help people to consider that the AI classifier might make mistakes. (**C**) In trials with examples, helpful examples tend to help people accurately model the AI classifier in cases when the AI classifier is correct, but have a limited impact when the AI classifier is wrong. (**D**) Consequently, helpful examples make participants more likely to pick the ground truth option when the AI classifier is correct, but do not really impact the probability of selecting the ground truth option when the AI classifier is wrong. Collectively, these results suggest that helpful examples and saliency maps improve human understanding of the AI classifier in distinct and complementary ways: saliency maps improve error detection, whereas helpful examples enable participants to accurately determine when the AI classifier is correct. Errorbars represent 95% bootstrapped confidence intervals. All point estimates have confidence intervals, though some are too narrow to see clearly. This figure was created using the ggplot2 package (v. 3.3.2)[46] in R (v. 4.0.3)[47].

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 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 ($\beta = 1.10(0.04)$, $z = 29.28$, $p < 0.0001$), but negatively associated with fidelity for AI errors ($\beta = -0.92(0.02)$, $z = -42.82$, $p < 0.0001$; Fig. 6).

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 Fig. 6): the presence of saliency maps reduces the positive impact of familiarity on fidelity when the AI classifier is correct ($\beta = -0.51(0.08)$, $z = -6.31$, $p < 0.0001$). Conversely, saliency maps reduce the negative impact of familiarity on fidelity when the AI is wrong ($\beta = 0.70(0.05)$, $z = 15.22$, $p < 0.0001$; Fig. 6). 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.
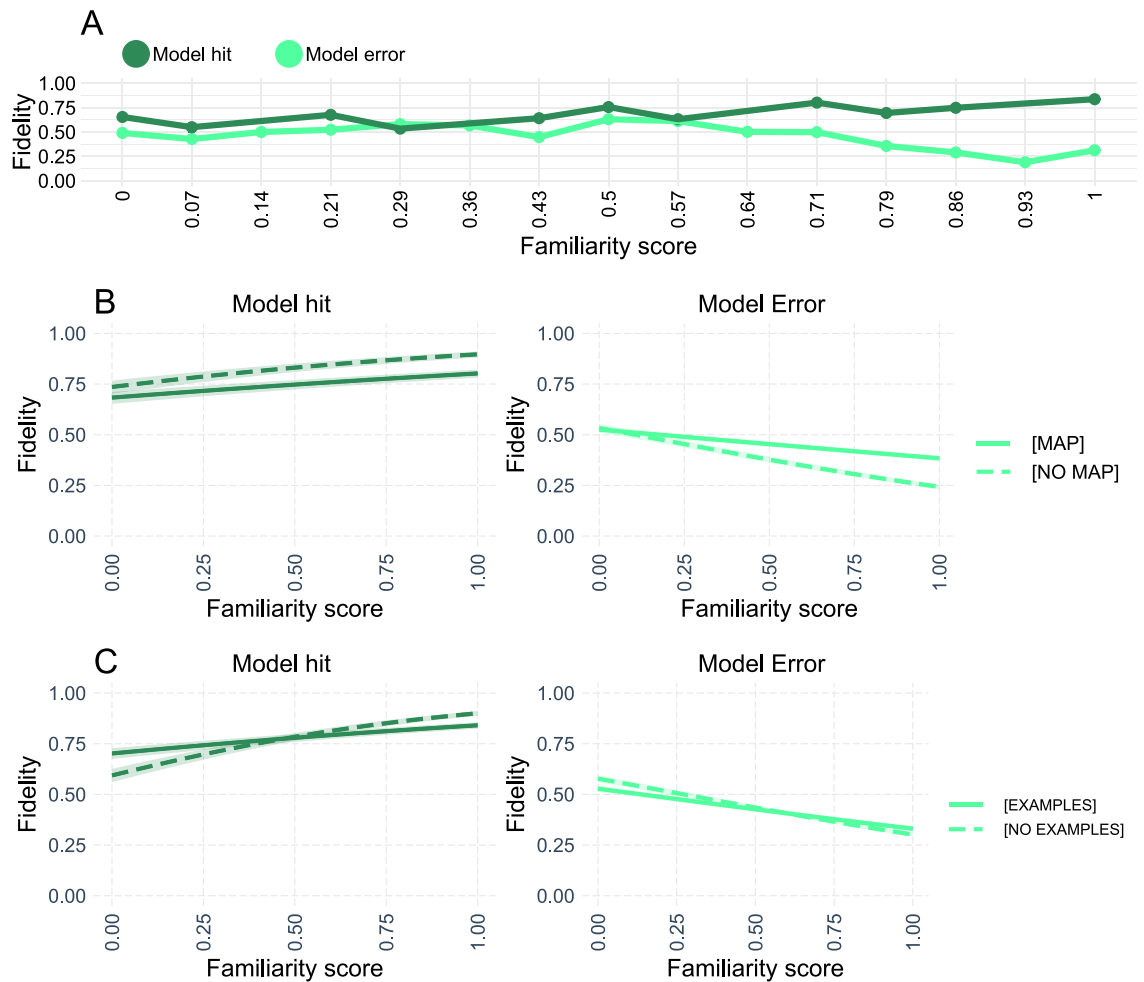
Figure 6. Familiarity score predicts fidelity based on the full data set (631 participants; 94,582 observations). (**A**) The fidelity between the participant predictions and the AI classifier's judgements increase with category familiarity when the AI is correct, but decrease with familiarity when it is wrong. This provides further evidence that participants project their own beliefs unto the AI classifier, as they are more likely to predict that the AI makes the correct choice on trials they themselves find easy. (**B**) Saliency maps decrease the impact of familiarity on participant judgements. For model hits this leads to decreased fidelity, whereas for model errors it leads to improved fidelity. This pattern provides further evidence that the saliency maps work by shifting participants away from using their first-order judgments to model the AI's classifications. (**C**) Examples also decrease the impact of familiarity on participant judgements. For model hits this improves fidelity for unfamiliar items but decreases fidelity for familiar items, with the opposite pattern for model errors. These results suggest that examples are most beneficial for unfamiliar items when the AI classifier is correct. Errorbars represent 95% bootstrapped confidence intervals. All point estimates have confidence intervals, though some are too narrow to see clearly. Shaded areas represent analytic 95% confidence intervals. This figure was created using the ggplot2 package (v. 3.3.2)[46] in R (v. 4.0.3)[47].

Though the presence of examples 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 ($\beta = -1.01(0.08)$, $z = -12.71$, $p < 0.0001$) and when the AI classifier is wrong ($\beta = 0.33(0.05)$, $z = 7.35$, $p < 0.0001$). However, in contrast to the saliency maps, examples seem to be most helpful for unfamiliar trials when the AI classifier is correct, see Fig. 6C. This effect may imply that the examples help participants develop a working representation of the unfamiliar categories, which they are otherwise lacking.

Discussion

Bayesian teaching provides a novel way to think about XAI by explicitly modeling the explainee and their prior beliefs. It suggests that explanations can be evaluated in terms of how well they shift explainees' beliefs away from their prior towards a target. We have presented evidence that a Bayesian Teacher can successfully predict which explanations will improve the fidelity between human predictions and target classifications as well as be preferred by human users. Crucially, our results show that the Bayesian Teacher is well-calibrated to human users: it both knows which explanations will improve predictions about the AI and which explanations are problematic. This

calibration provides strong evidence that the selection process of the Bayesian Teacher has a causal effect on explainee understanding. Not all examples are created equal, so they need to be appropriately curated.

Multiple strands of evidence from our results suggest that in the absence of explanations people project their own first-order beliefs onto the AI classifier. Specifically, we find that participants in the control condition show higher sensitivity than specificity, and that this discrepancy becomes more extreme the more familiar participants are with the trial categories. The finding that participants predict the AI system by projecting their own beliefs onto the AI links research on explainable AI to the rich psychological literature on social prediction. In many social prediction tasks (in contrast to mechanistic prediction tasks) people use their own preferences, judgements, and beliefs as priors for other agents[44,45,48–50]. Our results imply that such belief-projection can be mitigated by Bayesian teaching. The most compelling evidence that explanations mitigate belief projection is that the impact of familiarity on fidelity is reduced by explanations: explanations make participants more likely to catch AI mistakes on trials they themselves found easy.

Bayesian teaching also gives a coherent framework for comparing and contrasting explanatory methods that hereto have been considered independent: explanation-by-examples and feature attribution. We apply Bayesian teaching to study explanation-by-examples, a popular method for XAI that previously has lacked a sound theoretical footing. Explanation-by-examples has many strengths: it is model-agnostic, domain-general, and easy to use with other XAI methods. Viewed through a Bayesian teaching lens, this method can be generalized to include feature attribution, another popular post-hoc method, by splitting each example into its component features (i.e. pixels in this study) and considering each pixel individually. When applied to images, such feature attribution on the pixel level generates saliency maps, which is arguably the most popular method for XAI in the image domain. The connection between feature attribution and pixel selection by Bayesian teaching opens up the possibility to reinterpret all feature attribution methods (e.g.[51,52]) as a form of teaching. By treating images and saliency maps as explanatory examples at different levels of granularity, we discover that the two explanations show complementary effects. Namely, example images are effective explanations for confirming the model's correct classification of unfamiliar categories, and saliency maps are effective explanations for exposing the model's incorrect classification of familiar categories.

The lack of a coherent theory is currently stifling XAI as methods are developed around technical innovations without any *apriori* hypothesis as to whether they are appropriate for the specific use case[53]. Bayesian teaching both exposes this blind spot and offers a solution: effective explanation is a communication act which depends on a knowledgeable teacher, a good model of the explainee, and an awareness of the context in which inference takes place. Consequently, the framework encourages systematic evaluation of XAI interventions on these dimensions, and provides a way to systematically diagnose how interventions could be improved. In our study we show how such an evaluation applies to explanation-by-examples. We modeled the explainee by a ResNet-50 architecture, focused on two contextual variables (familiarity and model correctness), and surfaced how explanations generated by Bayesian teaching can mitigate mistaken prior beliefs. These results highlight the promise of the Bayesian teaching approach, since the function of explanation is to shape the explainee's inductive reasoning[54]. Furthermore, Bayesian teaching exemplifies how XAI can be improved by considering links to other fields such as education and cognitive science. A balanced synergy between the social sciences and the more technical literature of AI is much needed, as XAI is simultaneously a machine-learning problem and a human-centered endeavor.

Methods

The objective of this study was to explore the effects of explanations, in the form of examples and saliency maps, on users' understanding of high-performing machine learning models (referred to as AI throughout the paper) in the domain of image classification. We probe users' understanding by a two-alternative-forced-choice (2AFC) task in which users are asked to predict the model's classification of a target image into one of two categories. Experimental conditions vary in terms of the information presented on the screen during each classification. The information presented differs along three dimensions: types of labels, types of examples, and types of saliency maps. All the examples and saliency maps are generated by the Bayesian teaching framework. The fidelity of the participant is captured by sensitivity, specificity, and accuracy.

The model to be explained.　The machine learning model to be explained is a ResNet-50 model[39]. For this study, we used the pre-trained version of ResNet-50 in Keras with ImageNet weights. For the selection of saliency maps, the Bayesian teaching framework expects the model to be able to make probabilistic inference on the image classification task presented in the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). The ResNet-50 model has this capability, and we can use the ResNet-50 model without any modification. However, for the selection of examples, the Bayesian teaching framework expects the model to be able to make probabilistic inference on the 2AFC task, and the ResNet-50 model is deterministic. We replace the fully-connected classification layer of the ResNet-50 model with a probabilistic linear discriminant analysis (PLDA) model[55]. This new PLDA layer is trained using a transfer-learning-like procedure. Training images were first passed through the ResNet-50 model and transformed into feature vectors. Then, the PLDA layer was fit to these feature vectors and the corresponding class labels following the algorithm presented in[55]. Using the training dataset *ImageNet 1K* from the ILSVRC2012[38], this ResNet-50-PLDA model has a top-1 accuracy of 52.86% and a top-5 accuracy of 76.29%. For the actual experiment, we focused on a subset of 100 categories that include the most difficult, easiest, and most confusable categories (see the next subsection for details). Unless otherwise stated, all the model predictions used to design the experiment is based on the ResNet-50-PLDA model trained on the training data in only these 100 categories.

Stimuli selection. 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. First we calculated the ResNet-50-PLDA model's confusion matrix on *ImageNet 1K*, which contains 1000 categories. Then, we randomly selected 25 categories from each of the following four subsets: the 100 categories on which the model was most accurate, the 100 categories that were most confusable with these most accurate categories, the 100 categories on which the model was least accurate, and the 100 categories that were most confusable with these least accurate categories. This resulted in 100 categories. We recorded the model's predicted labels of all the training images in these 100 categories and marked all images for which the model predictions were also among these 100 categories.

From this subset where both the image and the top model prediction belonged to our 100 categories, we randomly sampled 50 images where the model prediction matched the ground truth labels and 100 images for which the model predictions did not match the ground-truth labels. For the 50 trials with correctly classified target images, the two classification options participants could choose from were the correct model-predicted category and one of the two most confusable categories (out of our 100 selected categories). Which one of the two most confusable categories was presented were selected randomly for each trial. For the 100 incorrectly classified trials the two classification options were simply the ground-truth category and the incorrect model prediction. This procedure resulted in a total of 83 unique categories used in the experiment (Supplementary Table T1). This number is smaller than 100 because not all confusable categories are unique and not all categories were kept during the random sampling. Figure 7 depicts the trial generating process. The pairs of categories used in the experiments are listed in Supplementary Table T2.

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 2 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. Below, we provide more details on the conditions.

Specific or generic labels. Conditions with informative or generic labels are referred to as [SPECIFIC LABELS] and [GENERIC LABELS], respectively. In the [SPECIFIC LABELS] conditions, the English labels of the two categories (e.g., "Flagpole" and "Barn" in Fig. 2) are given. In the [GENERIC LABELS] conditions, the two categories are named "Category A" and "Category B."

With or without examples. Conditions with and without examples are referred to as [EXAMPLES] and [NO EXAMPLES], respectively. In the [EXAMPLES] conditions, two examples are sampled from each of the two categories to represent the category. Thus, five images—one target image and four example images—are on display in each trial in these conditions. In the [NO EXAMPLES] conditions, only the target image is shown.

Helpful or random examples. Conditions with the helpful examples and random examples are referred to as [HELPFUL] and [RANDOM], respectively. The selection of the examples are based on the *simulated explainee fidelity*, which is the numerator of the Bayesian teaching probability, $f_L(\cdot)$. The simulated explainee fidelity characterizes the probability that the four examples will lead a "explainee model" to classify the target image as the ResNet-50-PLDA model would. The Bayesian teaching probability and its numerator $f_L(\cdot)$ are rigorously defined in Eqs. (2) and (3), respectively, in the subsection below called "Selection of examples with Bayesian teaching" section. In the [HELPFUL] conditions, the four examples are chosen such that $f_L(\cdot) > 0.8$. In the [RANDOM] conditions, the four examples are chosen such that the $f_L(\cdot)$ values across the 150 trials are uniformly distributed over the five bins that evenly partition the [0,1] interval.

With or without saliency maps. Conditions with and without saliency maps are referred to as [MAP] and [NO MAP], respectively. A saliency map is an image mask that shows the contribution of each pixel to the model's classification decision. Details on the generation of the saliency maps are provided in the subsection below called "Selection of saliency maps with Bayesian teaching" section. In the [MAP] conditions, a saliency map is shown for every image displayed. In the [NO MAP] conditions, no saliency map is shown.

Blur or jet saliency maps. Conditions with the blur saliency maps and jet saliency maps are referred to as [BLUR] and [JET], respectively. The two types of map differ only in the rendering of the mask but not in the generation of the mask. The jet saliency map renders the importance of each pixel by colors following the jet color map. In order of decreasing importance, the jet color map goes from red to green to blue. The jet color map, overlaid on an image with some level of transparency, is one of the most commonly used renderings of saliency maps.
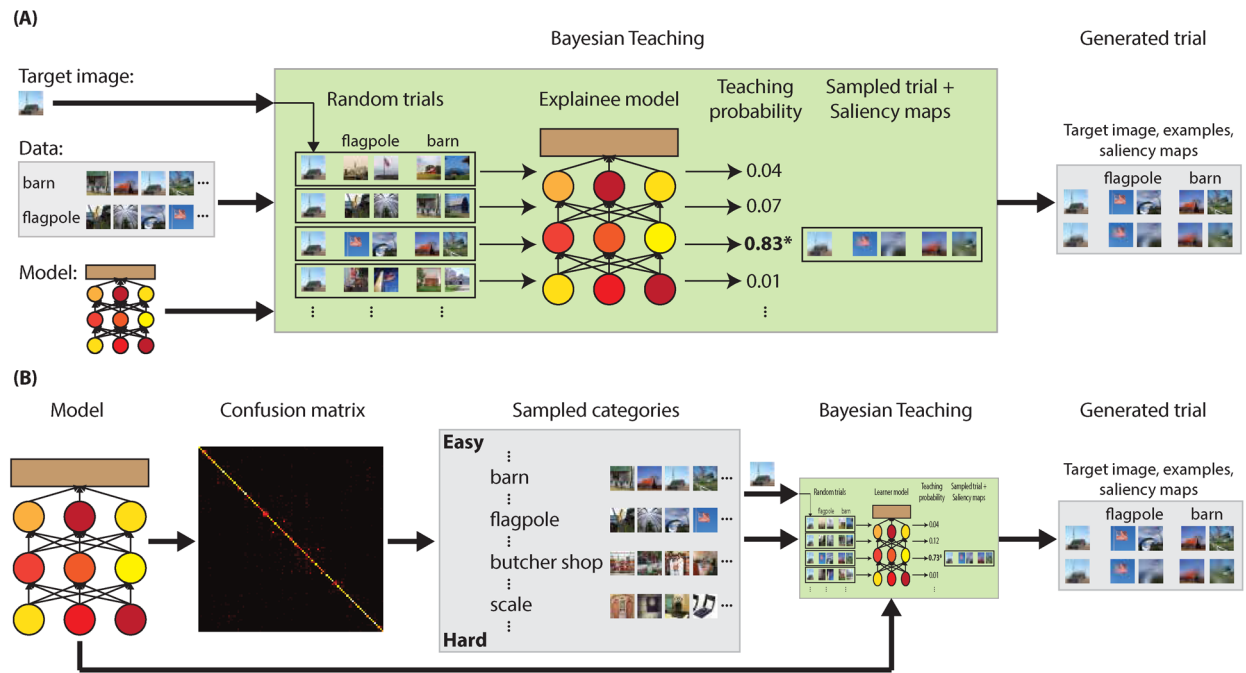
Figure 7. Flowchart of trial generation. (**A**) Selection of examples and saliency maps with Bayesian teaching. The inputs to Bayesian teaching are: the model to be explained, data sets from two categories, and a target image that belongs to one of the two categories. The green box depicts the inner working of Bayesian teaching. Random image pairs are selected from each of the input categories. Along with the target image, two sets of image pairs, one set from each category, are selected at random to form a trial. The explainee model, which is set to have the same architecture as the input model, takes in a large number of random trials to produce the simulated explainee fidelity (unnormalized teaching probabilities according to Eq. (3)). Here, a trial with high fidelity (probability) is selected, exemplifying the trial generation process in the [HELPFUL] condition. Saliency maps are generated for the target image and the four selected examples using Eq. (6). The final output is a set of ten images: a target image, two examples selected from each of the two input categories, and the saliency maps of the above five images. (**B**) Trial generation steps peripheral to Bayesian teaching. Our model to be explained is a ResNet-50 trained on *ImageNet 1K*. A confusion matrix on the 1000 ImageNet categories was computed using the model. Using the confusion matrix, we sampled 25 categories where the model has high accuracy (the "Easy" categories), 25 categories where the model has low accuracy (the "Hard" categories), and the categories that are most confusable with the above 50 categories. To generate a trial, we select at random two categories from the 100 candidates mentioned above as well as a target image that belongs to one of the two selected categories. The model, the target image, and the data associated with the two categories are fed into Bayesian teaching to produce a trial. See Methods for the full details. This figure was created using Adobe Illustrator CS6 (v. 16.0.0)[43].

Two disadvantages of jet saliency maps are that the colors of the map can interfere with the colors of the image and that the unimportant regions remain visible to the user and can attract involuntary visual attention. For these reasons, we created the [BLUR] conditions in which the saliency maps are rendered by blurring the image. Furthermore, blurring is a more naturalistic visual effect than any color map masking because our visual system constantly experiences a large difference in visual acuity between our fovea and peripheral vision. The implementation details of both renderings are provided in the subsection below on saliency map selection.

Naming convention. As shown in Table 1, not all combinations of the five binary features are allowed. Conditions with generic labels and no examples are not tested because that would make the 2AFC task a game of pure guessing. Furthermore, conditions without examples cannot be paired with helpful or random examples, and conditions without saliency maps cannot be paired with blur or jet maps. This leaves a total of 15 experimental conditions.

The naming convention for the conditions is based on filter queries using the database structure presented in Table 1. To give a few examples: [HELPFUL] refers to the aggregate of the six conditions in columns 2 and 4; [MAP] refers to the aggregate of the 10 conditions in rows 2 and 3; [HELPFUL] & [BLUR] refers to the aggregate of the two conditions in row 2 column 2 and row 2 column 4; and [HELPFUL] & [BLUR] & [SPECIFIC LABELS] refers to the one condition in row 2 column 2.

Participants. The study protocol was approved by Rutgers University IRB. All research was performed in accordance with the approved study protocol. An IRB-approved consent page was displayed before the experi-

ment. Informed consent was obtained from all participants. The experiment began after the participants gave consent.

656 participants (404 male, 249 female, 3 other) were recruited from Amazon Mechanical Turk and paid $2.50 for completing the experiment, which took roughly 15 minutes to complete. The mean age of participants was 34.8 years (SD = 10.1), ranging from 18 to 72 years. The participants were randomly assigned to each condition, with the aim to obtain 36–40 participants per condition. 25 participants were excluded from analysis for completing the experiment too quickly (less than one second per trial), resulting in a final sample of 631 participants completing 150 trials each. The [NO EXAMPLES] conditions received twice the sample size of the other conditions, so that they would match the sample size of the [EXAMPLES] conditions, which had two distinct versions ([HELPFUL] and [RANDOM]). Table 1 shows the number participants in each of the 15 conditions.

The number of participants in every condition is shown in Table 1. All participants in the [HELPFUL] conditions experienced the same set of 150 trials, i.e., the same 150 combinations of target image, category pairs, and example images, but in randomized order. All participants in the [RANDOM] conditions experienced another set of 150 trials, also in randomized order. All the category pairs used are listed in Supplementary Table T2. Participants in the [NO EXAMPLES] condition experienced one of these two sets of trials, selected at random. Note that because there are no examples but only English labels in the [NO EXAMPLES] conditions, the two sets of trials are functionally equivalent.

Selection of examples with Bayesian teaching. The goal of Bayesian teaching is to select small subsets of the training data such that the inference made by a explainee model using this small subset will be similar to the inference made by a target model using the entire training data. For this study, the target model is the ResNet-50-PLDA model trained on the 100 selected categories as described earlier. The inference task is to classify the target image among the 100 categories. The inference task of the explainee model is the 2AFC image classification task presented in each trial. For the explainee model, we search for an ideal-observer model[40,41] that would capture the participant's inference in the 2AFC task. A good candidate is the ResNet-50-PLDA because it is trained on human-labeled data and achieves high accuracy on predicting humans' labelling behavior. This means that the target model and explainee model share the same parameters (the ResNet-50 weights and PLDA parameters mentioned after Eq. (4)), and the use of Bayesian teaching is focused on explaining the image classification inference based on roughly 100 K training examples, i.e., all the training data in the 100 selected categories, with only four training examples, i.e., those selected to be displayed on each trial of the experiments in the [EXAMPLES] conditions.

We introduce some notation to define the Bayesian teaching probability formally. The two categories that define the 2AFC task in each trial consist of the predicted category of the ResNet-50-PLDA model and an alternative category, which we denote by y^* and y, respectively. The two examples sampled from the model-predicted category are denoted by τ^{y^*}, and the two sampled from the alternative category are denoted by τ^y. Let the explainee model be denoted by f_L and the target image be denoted by d^*. The Bayesian teaching probability, P_T, is defined as the probability that the selected examples, τ^{y^*} and τ^y, will lead the explainee model to classify the target image as the target model would. Mathematically, this probability can be expressed using Bayes' rule as:

$$P_T\left(\tau^{y^*}, \tau^y \mid y^*, d^*\right) = \frac{f_L\left(y^* \mid \tau^{y^*}, \tau^y, d^*\right)}{\sum_{\left(\tau^{y^*}, \tau^y\right)' \in \Omega} f_L\left(y^* \mid \left(\tau^{y^*}, \tau^y\right)', d^*\right)}. \tag{2}$$

The sum in the denominator is over all possible candidate sets of the four examples. The set of all candidate sets is denoted by Ω. Equation (2) assumes a uniform prior over Ω so that the prior terms in the numerator and denominator cancel out. Technically, Ω is the Cartesian product of all possible pairings of images in the category y^* with all possible pairings of images in the category y, which is on the order of 10^{11} for the dataset in use. Our goal here is to select τ^{y^*} and τ^y such that $f_L(\cdot)$, the numerator of Eq. (2), would provide good coverage of the full range of [0,1]. This would ensure the existence of valid examples for both the [RANDOM] and [HELPFUL] conditions. We found that the full range can usually be covered by forming a Cartesian product of 1000 random pairings from each category (10^6 combinations). In general, given a target value of $f_L(\cdot)$, one could use genetic algorithm[56] or other types of discrete optimization method to select the examples. To sample in proportion to P_T, one could use Markov Chain Monte Carlo and variational inference techniques[14,57,58]. These optimization and advanced inference methods would also be more efficient in the case that more than a few examples for each category is desired.

Using Bayes' rule again, we express the explainee model's inference of the target image's label given the target image and examples, the numerator in Eq. (2), as

$$f_L\left(y^* \mid \tau^{y^*}, \tau^y, d^*\right) = \frac{f\left(d^* \mid \tau^{y^*}\right)}{f\left(d^* \mid \tau^{y^*}\right) + f(d^* \mid \tau^y)}, \tag{3}$$

where $f(d^* \mid \tau^k)$ is the probability that the target image, d^*, belongs to the category from which the two example images, τ^k, are sampled. Under the PLDA model, one can write this probability in closed form as a normal distribution[21]:

$$f\left(d^* \mid \tau^k\right) = \mathcal{N}\left(u^* \,\middle|\, \frac{\Psi}{2\Psi + I}\left(u_1^k + u_2^k\right), \frac{\Psi}{2\Psi + I} + I\right). \tag{4}$$

Here, u is an image transformed in two steps. First, the image is passed through ResNet-50 and transformed into a feature vector; then, this feature vector undergoes an affine transformation with shift vector **m** and rotation and scaling matrix A to become u. Thus, in Eq. (4), u^* is a transformed target image, and (u_1^k, u_2^k) are a pair of transformed examples sampled from category k. The quantities **m** and A in the second transformation and the Ψ in Eq. (4) are parameters of the PLDA model obtained by training on the images in the 100 selected categories. The precise definitions of these parameters and the training procedure are presented in Fig. 2 in Ioffe's PLDA paper[55].

To summarize this subsection, Eq. (2) defines the Bayesian teaching probability, and Eq. (3) defines its numerator (simulated explainee fidelity), $f_L(\cdot)$, used to select examples in the [EXAMPLES] conditions. A high $f_L(\cdot)$ means that the selected examples will lead the model of the explainee to classify the target image as the category predicted by the model to be explained with high probability. Conversely, a low $f_L(\cdot)$ means that the selected examples will lead the explainee model to classify the target image as the other category in the 2AFC with high probability. Note that $f_L(\cdot)$ is trial specific, as this probability is a function of the target image, d^*, the model predicted label of the target image, y^*, and the four examples, (τ^{y^*}, τ^y), which precisely define a trial.

Selection of saliency maps with Bayesian teaching.   A saliency map is an image mask that shows how important each pixel of the image is to the model's inference. In the [MAP] conditions, we generate a saliency map for every image displayed. To generate a saliency map, one needs to specify a model, an inference task, and a definition of importance. We used ResNet-50 as the model and the classification of an image into the 1000 categories in *ImageNet 1K* as the inference task. Using the Bayesian teaching framework, we define importance to be the probability that a mask, m, will lead the model to predict the image, d, to be in category, y, when the mask is applied to the image. This is expressed by Bayes' rule as

$$Q_T(m \mid y, d) = \frac{g_L(y \mid d, m)p(m)}{\int_{\Omega_M} g_L(y \mid d, m)p(m)}. \tag{5}$$

Here, $g_L(y \mid d, m)$ is probability that the ResNet-50 model will predict the d masked by m to be y; $p(m)$ is the prior probability of m; and $\Omega_M = [0, 1]^{W \times H}$ is the space of all possible masks on an image with $W \times H$ pixels. The prior probability distribution $p(m)$ on m is a sigmoid-function squashed Gaussian process prior.

Instead of sampling the saliency maps directly from Eq. (5), we find the expected saliency map for each image by Monte Carlo integration:

$$\begin{aligned} E[M \mid y, d] &= \int_{\Omega_M} m \, Q_T(m \mid y, d) \\ &\approx \frac{\sum_{i=1}^N m_i \, g_L(y \mid d, m_i)}{\sum_{i=1}^N g_L(y \mid d, m_i)}, \end{aligned} \tag{6}$$

where m_i are samples from the prior distribution $p(m)$, and $N = 1000$ is the number of Monte Carlo samples used. To see why an expected map is desirable, imagine the following case. Suppose that an image contains 7 goldfish and its category is "goldfish." In this case, a mask that reveals any one of the goldfish will have a high Q_T value. However, it is more desirable that the mask would reveal all the goldfish in the image. The expectation provides this by averaging the masks appropriately weighted by their Q_T values.

Now, we describe the step-by-step procedures for generating the saliency map for an image, d. First, d is resized to be 224-by-224 pixels, which is the size displayed in the experiments (Fig. 2). A set of 1000 2D functions are sampled from a 2D Gaussian process (GP) with an overall variance of 100, a constant mean of -100, and a radial-basis-function kernel with length scale 22.4 pixels in both dimensions. The sampled functions are evaluated on a 224-by-224 grid, and the function values are mostly in the range of $[-500, 300]$. A sigmoid function, $1/(1 + \exp(-x))$, is applied to the sampled functions to transform each of the function values, x, to be within the range $[0, 1]$. This results in 1000 masks. The mean of the GP controls how many effective zeros there are in the mask, and the variance of the GP determines how fast neighboring pixel values in the mask change from zero to one. The 1000 masks are the m_i's in Eq. (6). We produce 1000 masked images by element-wise multiplying the image d with each of the masks. The term $g_L(y \mid d, m_i)$ is the ResNet-50's predictive probability that the ith masked image is in category y. Having obtained these predictive probabilities from ResNet-50, we average the 1000 masks according to Eq. (6) to produce the saliency map of image d. If d is a target image, the y used to generate the saliency map is the ResNet-50-PLDA model's prediction. If d is an example, the y is the category from which the example is sampled.

In the [JET] conditions, the saliency maps are rendered in the Matplotlib package with the "jet" colormap and an alpha value of 0.4 and overlaid on the images (see Fig. 2; images at the bottom row). In the [BLUR] conditions, a saliency map is rendered by blurring the image for which it is generated (Fig. 2; images in the middle row). To generate the blur, each pixel value of a saliency map, z, is assigned a blurring window width, $w(z) = \text{ceil}(30/(1 + \exp(20z - 10)))$. The jth pixel value of the rendered saliency map is the average pixel value of a patch of the original image, where the patch is w-by-w in size and centered on the jth pixel of the original image. If the jth pixel is close to an edge of the image, the patch becomes rectangular, and the average is over whichever pixel values that are inside the w-by-w sized window.

To conclude this subsection, we make a few final remarks. First, a PLDA layer is unnecessary in the generation of saliency maps because the ResNet-50 model is capable of generating the probabilities $g_L(y \mid d, m)$ in Eq. (5). In contrast, the ResNet-50 model cannot be used directly to generate the probabilities $f_L(y^* \mid \tau^{y^*}, \tau^y, d^*)$ in Eq. (2). Second, while the 2AFC task may be suitable for generating a saliency map for the target image, it cannot be used to generate saliency maps for the examples. This is the main reason that here we used the inference

task of the image classification task that the ResNet-50 model is trained on. Lastly, Eq. (6) is the same as Eq. (5) in the RISE approach introduced by Petsiuk, Das, and Saenko[42], which presents a state-of-the art method for generating saliency maps. Our implementation and their implementation differ only in the way the individual masks are sampled. In our implementation, we sampled functions from a GP prior and turned them into masks by applying a sigmoid function. In[42], random binary matrices are first sampled and subsequently up-sampled to the desired mask size through bilinear interpolation. The expectation is done in the same way.

Familiarity coding. In addition to the splits by conditions presented in Table 1, analysis also rely on scores of human familiarity with the image categories. The familiarity of each of the [HELPFUL] and [RANDOM] trials was manually coded by 7 raters. Each rater was asked to code the trial as "familiar" if they thought they could correctly match the category labels to the images presented in that trial, and "unfamiliar" otherwise. A familiarity score for each pairing of categories was then constructed by assigning each raters judgements as 1 for familiar and 0 for unfamiliar, and computing the mean across raters. The 300 trials across the [HELPFUL] and [RANDOM] conditions resulted in 167 unique category pairings (counting the ordering of target versus other category), and their familiarity scores are presented in Supplementary Table T2.

Statistical analysis. Whenever we report testing how well participants predict the model classifications (fidelity), or how often their judgements correspond to the image ground truth (accuracy) we used hierarchical logistic regressions with random intercepts per participants and fixed effects for the remaining terms. For analyses of sensitivity and specificity analyses, we still used a logistic regression framework but only included trials corresponding to true positives and false negatives, or true negatives and false positives, respectively. Sensitivity captures how well participants predict trials when the AI is correct, and specificity capture how well participants predict trials when the model is wrong.

To illustrate, *Bayesian teaching improves fidelity* used the following equations on the full set of trials, and on a subset of the trials to capture sensitivity and specificity respectively:

$$\Pr\big(\text{ParticipantChoice}_i = \text{AIChoice}_i\big) = \text{logit}^{-1}\big(\alpha_{j[i]} + \beta_1 \text{ExplanationCondition}_i + \epsilon_i\big), \text{ for } i = 1,\ldots,I$$

$$\text{logit}^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}$$

$$\alpha_j \sim N(U_j, \sigma_\alpha^2), \text{ for } j = 1,\ldots,J.$$

where the agreement between a participant's choice and the AI classifier's choice is a binary variable coded as 1 when participant correctly predict the AI classification and 0 otherwise, i is the observation index, j is the participant index. ExplanationCondition is a binary dummy variable coded as 1 if participants experienced heatmaps and helpful examples and 0 if they did not experience any explanations.

For the *Participants prefer helpful examples* section we compared three hierarchical logistic models: (A) An intercept only model that treated intercepts as nested within participants (B) an intercept only model that treated intercepts as nested within participants and conditions, (C) Model two, with an added fixed effect for the familiarity score. We then compared the negative log-likelihoods of these models to determine which best accounted for the observed data.

We evaluated whether *Bayesian teaching can lead participants to both correct and incorrect inference* by predicting fidelity in the conditions containing examples by fitting three nested models:

$$\Pr\big(\text{ParticipantChoice}_i = \text{AIChoice}_i\big) = \text{logit}^{-1}\big(\alpha_{j[i]} + \beta_1 \text{CategoryAccuracy}_i + \varepsilon_i\big), \text{ for } i = 1,\ldots,I$$

$$\textit{Pr}(\text{participant choice}_i = \text{AI choice}_i) = \text{logit}^{-1}(\ldots + \beta_2 \text{SimExplaineeFidelity}_i + \varepsilon_i), \text{ for } i = 1,\ldots,I$$

$$\Pr\big(\text{participant choice}_i = \text{AI choice}_i\big) = \text{logit}^{-1}(\ldots + \beta_3 \text{ModelCorrectness}_i$$
$$+ \beta_4 \text{ModelCorrectness}_i \text{CategoryAccuracy}_i$$
$$+ \beta_5 \text{ModelCorrectness}_i \text{SimExplaineeFidelity}_i + \varepsilon_i), \textit{ for } i = 1,\ldots,I.$$

where SimExplaineeFidelity is the expected probability that the participant pick the same response as the target model, conditional on seeing the examples, CategoryAccuracy, is the average classification accuracy of the target ResNet-50 model for the target category, and ModelCorrectness is a dummy variable coding if ResNet made a correct classification on this particular trial. We then compared the negative log likelihoods of these three models, and reported the coefficients of the best-fitting model (the interaction model).

In the *Bayesian teaching improves fidelity through belief-mitigation* section we fitted four logistic hierarchical regression models to the full data. These models shared the following form:

$$\Pr(Y_i = 1) = \text{logit}^{-1}\big(\alpha_{j[i]} + \beta_1 \text{FamiliarityScore}_i + \beta_2 \text{CategoryAccuracy}_i \tag{7}$$
$$+ \beta_3 \text{Examples}_i + \beta_4 \text{MAP}_i + \beta_5 \text{Labels}_i + \epsilon_i\big), \text{ for } i = 1,\ldots,I.$$

where FamiliarityScore is a proportion of raters who rated the trial categories as familiar, Examples, MAP and Labels where dummy variables that captured whether examples were shown, whether heatmaps were shown and whether category labels were informative or not, respectively.

These four models were distinguished based on whether the AI was correct or not and whether Y corresponded to whether the participant judgement matched the ground truth or matched the AI's judgement. We fitted similar models to the [EXAMPLES] trials only, with the only difference being that the Examples term, that

previously had captured whether examples were present was replaced with a dummy variable that captured whether the examples presented were helpful or not. Finally, we fitted two more models predicting fidelity from the full data. These are similar to Eq. (6), but added two additional interaction terms:

$$\Pr(Y_i = 1) = \text{logit}^1\left(\ldots + \beta_6\text{MAP}_i\text{FamiliarityScore}_i + \beta_7\text{Examples}_i\text{syFamiliarityScore}_i + \epsilon_i\right), \textit{for } i = 1, \ldots, I.$$

Coefficient tables for these models can be found in Supplementary Tables T3. All hierarchical logistic regression models were fitted using the lme4 package (1.1-23)[59] in R version 4.0.3[47], Figures were created in ggplot 2 version 3.3.2[46].

Data availability
Raw experimental data and analysis code will be deposited at https://github.com/CoDaS-Lab/XAI-BT-SR upon publication.

References

1. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D. *et al*. Accountability of ai under the law: The role of explanation. Preprint at http://arXiv.org/1711.01134 (2017).
2. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T. *et al*. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. Preprint at http://arXiv.org/1711.05225 (2017).
3. Esteva, A. *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542** (7639), 115 (2017).
4. European Commission. *2018 Reform of EU Data Protection Rules* (European Commission, 2018).
5. Coyle, D. & Weller, A. Explaining machine learning reveals policy challenges. *Science* **368** (6498), 1433–1434 (2020).
6. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
7. Mill, J. S. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation* (Longmans, Green, and Company, 1889).
8. Bloom, P. *How Children Learn the Meanings of Words* (MIT Press, 2002).
9. Fei, X. & Tenenbaum, J. B. Word learning as Bayesian inference. *Psychol. Rev.* **114** (2), 245 (2007).
10. Lake, B. M. & Piantadosi, S. T. People infer recursive visual concepts from just a few examples. *Comput. Brain Behav.* **3** (1), 54–65 (2020).
11. Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P. & Glaser, R. Self-explanations: How students study and use examples in learning to solve problems. *Cogn. Sci.* **13** (2), 145–182 (1989).
12. Aleven, V. A. M. *Teaching Case-Based Argumentation Through a Model and Examples* (Citeseer, 1997).
13. Bills, L., Dreyfus, T., Mason, J., Tsamir, P., Watson, A. & Zaslavsky, O. Exemplification in mathematics education. In *Proc. 30th Conference of the International Group for the Psychology of Mathematics Education*, Vol. 1, 126–154 (ERIC 2006).
14. Chen, J., Song, L., Wainwright, M. & Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning* 882–891 (2018).
15. Eaves, B. S., Schweinhart, A. M. & Shafto, P. Tractable bayesian teaching. In *Big Data in Cognitive Science* 74–99 (Psychology Press, 2016).
16. Ho, M. K., Littman, M., MacGlashan, J., Cushman, F. & Austerweil, J. L. Showing versus doing: Teaching by demonstration. In *Advances in Neural Information Processing Systems* 3027–3035 (2016).
17. Hendricks, L. A., Hu, R., Darrell, T. & Akata, Z. Generating counterfactual explanations with natural language. Preprint at http://arXiv.org/1806.09809 (2018).
18. Kanehira, A. & Harada, T. Learning to explain with complemental examples. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 8603–8611 (2019).
19. Kim, B., Rudin, C. & Shah, J. A. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems* 1952–1960 (2014).
20. Kim, B., Khanna, R. & Koyejo, O. O. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems* 2280–2288 (2016).
21. Vong, W. K., Sojitra, R. B., Reyes, A., Yang, S. C.-H. & Shafto, P. Bayesian teaching of image categories. In *Proc. 40th Annual Conference of the Cognitive Science Society* (2018).
22. Wang, T., Zhu, J.-Y., Torralba, A. & Efros, A. A. Dataset distillation. Preprint at http://arXiv.org/1811.10959 (2018).
23. Koh, P. W. & Liang, P. Understanding black-box predictions via influence functions. In *Proc. 34th International Conference on Machine Learning-Volume 70* 1885–1894. www.JMLR.org (2017).
24. Papernot, N. & McDaniel, P. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. Preprint at http://arXiv.org/1803.04765 (2018).
25. Yeh, C.-K., Kim, J.,Yen, I. E.-H. & Ravikumar, P. K. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems* 9291–9301 (2018).
26. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D. & Lee, S. Counterfactual visual explanations. Preprint at http://arXiv.org/1904.07451 (2019).
27. Caruana, R., Kangarloo, H., Dionisio, J. D., Sinha, U. & Johnson, D. Case-based explanation of non-case-based learning methods. In *Proc. AMIA Symposium* 212 (American Medical Informatics Association, 1999).
28. Keane, M. T. & Kenny, E. M. How case-based reasoning explains neural networks: A theoretical analysis of xai using post-hoc explanation-by-example from a survey of ann-cbr twin-systems. In *International Conference on Case-Based Reasoning* 155–171 (Springer, 2019).
29. Yang, S. C.-H. & Shafto, P. Explainable artificial intelligence via bayesian teaching. In *NIPS 2017 Workshop on Teaching Machines, Robots, and Humans* (2017).
30. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **267**, 1 (2018).
31. Shulman, L. S. Those who understand: Knowledge growth in teaching. *Educ. Res.* **15** (2), 4–14 (1986).
32. Chick, H. L. Teaching and learning by example. *Math. Essent. Res. Essent. Pract.* **1**, 3–21 (2007).
33. Shafto, P., Goodman, N. D. & Griffiths, T. L. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cogn. Psychol.* **71**, 55–89 (2014).
34. Eaves, B. S. Jr., Feldman, N. H., Griffiths, T. L. & Shafto, P. Infant-directed speech is consistent with teaching. *Psychol. Rev.* **123** (6), 758 (2016).

35. Yang, S. C.-H., Yu, Y., Givchi, A., Wang, P., Vong, W. K. & Shafto, P. Optimal cooperative inference. In *International Conference on Artificial Intelligence and Statistics* 376–385 (2018).
36. Aodha, O. M., Su, S., Chen, Y., Perona, P. & Yue, Y. Teaching categories to human learners with visual explanations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 3820–3828 (2018).
37. Chen, Y., Aodha, O. M., Su, S., Perona, P. & Yue, Y. Near-optimal machine teaching via explanatory teaching sets. In *International Conference on Artificial Intelligence and Statistics* 1970–1978 (2018).
38. Russakovsky, O. *et al.* ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115** (3), 211–252 (2015).
39. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (2016).
40. Geisler, W. S. Ideal observer analysis. *Vis. Neurosci.* **10** (7), 12 (2003).
41. Geisler, W. S. Contributions of ideal observer theory to vision research. *Vis. Res.* **51** (7), 771–781 (2011).
42. Petsiuk, V., Das, A. & Saenko, K. *RISE: Randomized Input Sampling for Explanation of Black-box Models* (2018).
43. Adobe Inc. *Adobe Illustrator CS6 2012 (v. 16.0.0)*. https://adobe.com/products/illustrator. Accessed 18 December 2019.
44. Gordon, R. M. Folk psychology as simulation. *Mind Lang.* **1** (2), 158–171 (1986).
45. Koster-Hale, J. & Saxe, R. Theory of mind: A neural prediction problem. *Neuron* **79** (5), 836–848 (2013).
46. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
47. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna (2020).
48. Tarantola, T., Kumaran, D., Dayan, P. & De Martino, B. Prior preferences beneficially influence social and non-social learning. *Nat. Commun.* **8** (1), 1–14 (2017).
49. Suzuki, S., Jensen, E. L. S., Bossaerts, P. & O'Doherty, J. P. Behavioral contagion during learning about another agent's risk-preferences acts on the neural representation of decision-risk. *Proc. Natl. Acad. Sci.* **113** (14), 3755–3760 (2016).
50. Bio, B. J., Webb, T. W. & Graziano, M. S. A. Projecting one's own spatial bias onto others during a theory-of-mind task. *Proc. Natl. Acad. Sci.* **115** (7), E1684–E1689 (2018).
51. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (ACM, 2016).
52. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 4765–4774 (2017).
53. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at http://arXiv.org/1702.08608 (2017).
54. Lombrozo, T. The structure and function of explanations. *Trends Cogn. Sci.* **10** (10), 464–470 (2006).
55. Ioffe, Se. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision* 531–542 (Springer, 2006).
56. Back, T. *Evolutionary Algorithms in Theory and Practice: Evolution Strategies, Evolutionary Programming, Genetic Algorithms* (Oxford University Press, 1996).
57. Haario, H., Laine, M., Mira, A. & Saksman, E. Dram: Efficient adaptive mcmc. *Stat. Comput.* **16** (4), 339–354 (2006).
58. Maclaurin, D. & Adams, R. P. Firefly monte carlo: Exact mcmc with subsets of data. In *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015).
59. Bates, D., Sarkar, D., Bates, M. D. & Matrix, L. The lme4 package. *R Pack. Version* **2** (1), 74 (2007).

Acknowledgements

Author contributions

S.C.-H.Y., W.K.V., R.B.S. and P.S. conceived and designed the experiments. W.K.V. and S.C.-H.Y. conducted the experiments. R.B.S. and S.C.-H.Y. prepared materials for the experiments. T.F., W.K.V., and S.C.-H.Y. analyzed the data. S.C.-H.Y., T.F., W.K.V., R.B.S and P.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-89267-4.

Correspondence and requests for materials should be addressed to S.C.-H.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.