I-CEE: Tailoring Explanations of Image Classification Models to User Expertise

Yao Rong[1], **Peizhu Qian**[2], **Vaibhav Unhelkar**[2], **Enkelejda Kasneci**[1]
[1]Technical University of Munich
[2]Rice University
{yao.rong, enkelejda.kasneci}@tum.de, {pqian, vaibhav.unhelkar}@rice.edu

Abstract

Effectively explaining decisions of black-box machine learning models is critical to responsible deployment of AI systems that rely on them. Recognizing their importance, the field of explainable AI (XAI) provides several techniques to generate these explanations. Yet, there is relatively little emphasis on the user (the explainee) in this growing body of work and most XAI techniques generate "one-size-fits-all" explanations. To bridge this gap and achieve a step closer towards human-centered XAI, we present I-CEE, a framework that provides **I**mage **C**lassification **E**xplanations tailored to User **E**xpertise. Informed by existing work, I-CEE explains the decisions of image classification models by providing the user with an informative subset of training data (i.e., example images), corresponding local explanations, and model decisions. However, unlike prior work, I-CEE models the *informativeness* of the example images to depend on user expertise, resulting in different examples for different users. We posit that by tailoring the example set to user expertise, I-CEE can better facilitate users' understanding and simulatability of the model. To evaluate our approach, we conduct detailed experiments in both simulation and with human participants ($N = 100$) on multiple datasets. Experiments with simulated users show that I-CEE improves users' ability to accurately predict the model's decisions (simulatability) compared to baselines, providing promising preliminary results. Experiments with human participants demonstrate that our method significantly improves user simulatability accuracy, highlighting the importance of human-centered XAI.

Introduction

As AI systems receive increasingly important roles in our life, human users are challenged to comprehend the decisions made by these systems. To ensure user safety and proper use of AI systems, experts across disciplines have recognized the need for AI transparency (Yang et al. 2017; Ehsan et al. 2021; Russell 2021). Solutions for AI transparency – e.g., techniques for explainable AI (XAI) – are essential as most AI models can be viewed as a "black box," whose decision-making process cannot be easily interpreted or understood by human users. Among the different settings of XAI, our work focuses on explaining image classification tasks (Barredo Arrieta et al. 2020). Existing XAI techniques for image classification widely use attribution explanations, such as GradCAM (Selvaraju et al. 2017), SHAP (Lundberg and Lee 2017) or LIME (Ribeiro, Singh, and Guestrin 2016). While these techniques inform our work, they all miss one key element: human factors, potentially due to the complexity of modeling human users.

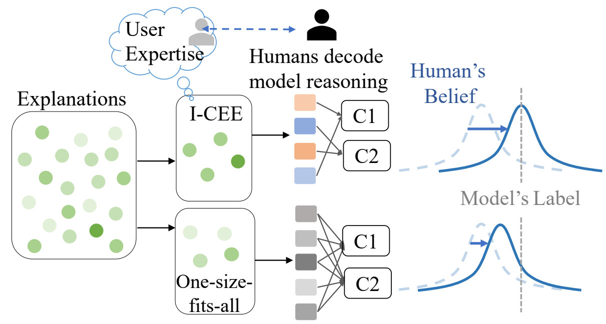


Figure 1: I-CEE tailors the explanation process to each user by considering their expertise. By selecting the most informative explanations based on user expertise, I-CEE can better enhance user simulatability of ML model's decisions.

We advocate that human modeling is critical to XAI research because explainability is inherently centered around humans (Liao and Varshney 2021). A few works focusing on explaining reinforcement learning policies use cognitive science theories to model the human user and generate explanations based on the human model (Baker and Saxe 2011; Huang et al. 2019; Lage et al. 2019b; Qian and Unhelkar 2022). Closer to our focus, the works of Yang, Folke, and Shafto (2022) and Yang et al. (2021) utilize a Bayesian Teaching framework to model human perception and then generate human-centered explanations. One limitation of these works is that all human users are treated the same by the modeling method, presuming that an identical set of explanations will work for *all* users. In contrast, we attempt to generate tailored explanations for each user by modeling their *task-specific expertise*. Our approach to modeling user expertise is informed by human annotator models used in active and imitation learning (Welinder et al. 2010; Beliaev et al. 2022). Similar to these works, our user model aims to capture both the decisions and reasoning process (expertise in concepts used for image classification) of the human user in the context of a given classification task.

To bridge the research gap that personalization is missing in the explanation process, we propose the framework **I**mage **C**lassification **E**xplanations tailored to User **E**xpertise (I-CEE). Informed by existing XAI methods for image classification, our framework utilizes the *explanation-by-examples* paradigm and provides attribution explanations (local explanations) for a subset of training data. However, in I-CEE, the approach of selecting the example explanations differs and is user-specific. For a given image classification task, I-CEE first discovers a set of m task-relevant concepts. It then models the user's task-specific expertise as a m-dimensional vector, where each entry lies between $[0, 1]$ and represents their expertise in the corresponding concept. Based on this user model, I-CEE finally selects the set of local explanations that can best fill user's knowledge gaps.

As depicted in Figure 1, by selecting the set of local explanations that can best increase the user's task-specific expertise, I-CEE aims to accelerate user's understanding of the decision-making process of the machine learning model. In contrast, most existing work in XAI either selects random or one-size-fits-all local explanations, thereby foregoing the opportunity to accelerate model understanding by providing tailored explanations. The contributions of this work can be summarized as follows:

- We identify the opportunity for tailored explanations for explaining decisions made by image classification models and develop a novel framework named I-CEE that realize this opportunity. This work represents an advancement towards human-centered explanations.
- To evaluate I-CEE, we test the simulatability of explanations generated by our framework on four datasets. Results demonstrate that our framework achieves better simulatability (i.e., users' ability to predict the model's decisions) relative to state-of-the-art XAI baselines[1].
- We evaluate our framework through detailed human-subject studies ($N = 100$). Experimental results indicate that our framework can more effectively help users understand the ML model's decision-making than the state-of-the-art technique Bayesian Teaching (Yang et al. 2021), and is subjectively more preferred by the participants, highlighting the advantages of our framework.

Related Work

Human-centered Explainable AI. Recent surveys indicate a growing activity in XAI research (Doshi-Velez and Kim 2017; Liao and Varshney 2021; Rong et al. 2023). The field recognizes the central role of humans in their explanations, leading to increasing adoption of human-centered evaluations of explanation techniques (Lage et al. 2019a). Besides evaluations, a few techniques have also considered human factors in generating explanations (Lage and Doshi-Velez 2020; Lage et al. 2019b; Huang et al. 2019; Qian and Unhelkar 2022; Yang, Folke, and Shafto 2022). Among these, the most related framework is that of Bayesian Teaching, which focuses on image classification and selects explanations by modeling the users as a Bayesian agent (Yang et al. 2021). However, this work does not model differences between users' reasoning or prior expertise. In contrast, we consider personalized user models to better fit the specific explanation needs of different users. Our design is informed by research in pedagogy and active machine learning.

Pedagogical Theories on Learning from Errors. XAI has been viewed as a teaching process, where the XAI technique serves the role of the teacher and the user that of the student (Qian and Unhelkar 2022). To teach learners effectively, pedagogical research confirms that a teacher needs to assess a learner's prior knowledge and design instructions accordingly (Owens and Tanner 2017; Ambrose et al. 2010). A common indicator of incorrect knowledge is errors, caused by an incorrect association or understanding. To correct the errors, feedback on the correct answers along with explanations have been found to be crucial and most helpful (Metcalfe 2017). These findings in learning sciences have laid the groundwork for our XAI framework, motivating our example selection approach; in particular, I-CEE emphasizes explaining the images on which it estimates the user will make errors. Additionally, as the confidence in an error increases, learning from the error also increases (Butterfield and Metcalfe 2001; Metcalfe and Finn 2011). This is an effect known as the hypercorrection effect. To reflect the hypercorrection effect in our framework, we choose images where the user has low confidence in the correct label (i.e., high confidence in the incorrect label), and argue that using these examples will result in better learning outcomes.

Active Learning. In the context of machine learning (ML), techniques for active learning aim to achieve high model accuracy while minimizing the required labeling effort (Settles 2009; Ren et al. 2021). Active learning is valuable in domains where a limited amount of training data is labeled, and it has been used beyond classification tasks such as in sequence labeling (Settles and Craven 2008) or image semantic segmentation (Sinha, Ebrahimi, and Darrell 2019). While active learning pertains to training machines, we observe that insights from the field are highly relevant for XAI (which seeks to train humans about an AI model). By making this novel connection, we leverage a central component of active learning techniques – *query strategies* – to inform the development and evaluation of I-CEE.

Problem Statement

Consider an ML classifier, denoted as f or the *target model*, trained on dataset $\mathcal{D}$ of image-label pairs $(\mathbf{x}, y)$. The classifier $f : \mathbb{R}^d \rightarrow \{1 : K\}$ maps an input image $\mathbf{x} \in \mathbb{R}^d$ to a label $y \in \{1 : K\}$, i.e., $f(\mathbf{x}) = y$, where K is the number of classes. For a subset of images, the predicted label y may not match the true label y^*. To explain such target models, different feature attribution methods have been proposed that generate local explanations (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). These local explanation assigns each input pixel an importance value, denoted as $\mathbf{e} \in \mathbb{R}^d$, which is usually visualized as a saliency map. In

[1]Code is available at https://github.com/yaorong0921/I-CEE.

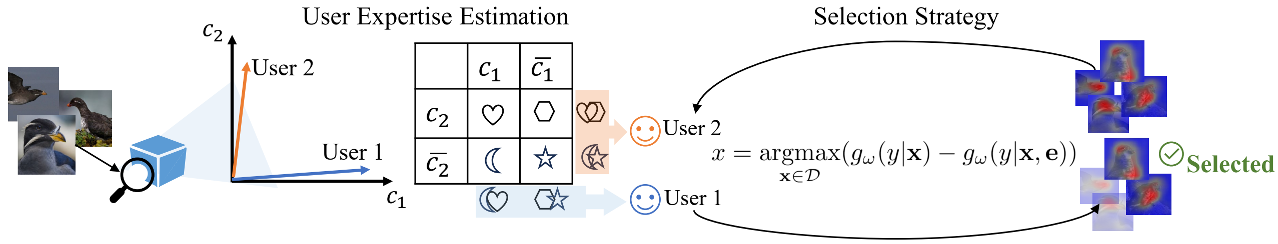


Figure 2: Overview of I-CEE. Left: The target model is first projected into a concept space, which is then used to estimate user expertise. Two users are illustrated. User 1 uses the concept c_1 in the reasoning process and can differentiate only two classes (highlighted in blue). Likewise, User 2 is able to distinguish two classes based on c_2 (in orange). Right: Based on user models, explanations with images $(\mathbf{x}, \mathbf{e})$ in the training set that maximize Hypercorrection Effect are selected and delivered to the users.

the *explanation-by-example* paradigm, the user is shown a set of images sampled from the training data, its local explanation, and its prediction, i.e., $(\mathbf{x}, \mathbf{e}, y)$. As the user has limited time to understand the model, it is important to select the set of most informative example images.

Within the explanation-by-example paradigm, we consider the problem of selecting the set of most informative example images (and corresponding explanations). Formally, our problem assumes three inputs: the target model f, a data set $\mathcal{D}$ ($|\mathcal{D}| = N$), and a feature attribution method to generate local explanations. Given these inputs, we seek to generate a subset $S \subset \mathcal{D}$ of training data composed of $M \ll N$ images that best facilitate *simulatability*, i.e., help users predict the decisions of the ML model. As the problem objective hinges on a human-centered metric, its successful resolution warrants a human-centered approach.

I-CEE: Image Classification Explanations Tailored to User Expertise

We now present our approach to solve this problem: I-CEE, which is composed of two phases (Figure 2). First, our framework models the user by estimating their task-specific expertise (lines 3-4, Algorithm 1). Second, by simulating the user using this model and a query strategy, I-CEE selects informative example images and explanations (lines 5-8).

User Expertise Estimation

The process of a user predicting an ML model's labeling decisions can be viewed as one of image annotation, where the annotators might possess distinct areas of strengths or *expertise* affecting their giving labels (Welinder et al. 2010). For instance, some users find textual patterns to be more recognizable than shapes while others find shapes to be more intuitive. During the annotation process, humans frequently use "concept-based thinking" in reasoning and decision making: identifying similarities among various examples and organizing them systematically based on their resemblances (Yeh et al. 2020; Armstrong, Gleitman, and Gleitman 1983; Tenenbaum 1999). Recognizing these aspects of human reasoning and informed by annotator models proposed in active learning, we model a user by estimating their expertise in applying different task-relevant concepts. We first discover the underlying concepts in the feature space of the target model. Using the discovered concepts, we model a user with a vector representing their ability to utilize each concept when annotating images.

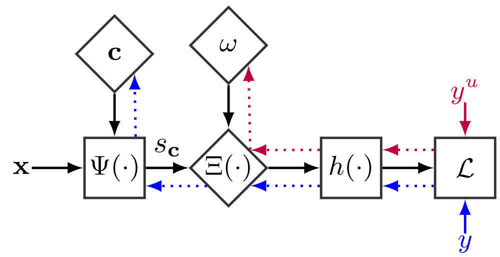


Figure 3: User Modeling: Square nodes are deterministic, while diamond nodes are trainable. Loss back-propagated for concept discovery (Eq. 3) is marked in blue, while that for expertise estimation (Eq. 4) is in red.

Figure 3 provides an overview of the user model. To arrive at the model, I-CEE begins with applying the concept discovery algorithm on the target model (Yeh et al. 2020) that aims to recover m concept $[\mathbf{c}_1, \cdots \mathbf{c}_m]$, such that

$$f(\mathbf{x}) = h(\Psi(\mathbf{x})) = h(\Xi_\theta(s_\mathbf{c}(\mathbf{x}))) \tag{1}$$

where $\Psi(\mathbf{x}) \equiv [\psi(\mathbf{x}^1), \ldots, \psi(\mathbf{x}^T)]$ are T activation vectors, $h(\cdot)$ represents the mapping from the intermediate output of activation vectors to image labels,[2] $s_\mathbf{c}(\cdot)$ is the concept score

$$s_\mathbf{c}(\mathbf{x}) = \langle \psi(\mathbf{x}^i), \mathbf{c}_j \rangle |_{j=1}^{m} |_{i=1}^{T} \in \mathbb{R}^{m \cdot T} \tag{2}$$

that estimates the alignment between each concept and activation vector pair, and $\Xi_\theta : \mathbb{R}^{T\cdot m} \to \mathbb{R}^{T\cdot n}$ is a trainable mapping that converts concept scores back into the activation space. Both the concept vectors and concept scores are unit normalized. For concept discovery (i.e., computing $\mathbf{c}, \theta$), the following cross-entropy loss is minimized:

$$\mathcal{L}_{(\mathbf{c},\theta)} = -\sum_{i=1}^{N} y_i \log(h(\Xi_\theta(s_\mathbf{c}(\mathbf{x}_i)))), \tag{3}$$

[2] Ψ and h can also be viewed as the intermediate and final layers of the image classification neural network, respectively. As h and Ψ are not trained as part of the user model, we do not explicitly denote their parameters (such as weights and biases) in our notation.

Algorithm 1: I-CEE

Input: Target model $f(\cdot)$, data $\mathcal{D}$, user annotation y^u.
Output: A set of example images and explanations $\mathcal{S}$.
Discover concepts by solving Eq. 3.
Estimate user expertise by solving Eq. 4.
for $\mathbf{x} \in \mathcal{D}$ **do**
    Calculate Hypercorrection Effect for $\mathbf{x}$ using Eq. 5.
end for
Return top-K image samples.

where y is the prediction from the target model $f(\cdot)$.

After completing concept discovery (which is a one-time process), the expertise estimation for each user takes place within the concept space. We freeze all model parameters ($\Psi(\cdot)$, $s_{\mathbf{c}}(\cdot)$, $\Xi_\theta(\cdot$ and $h(\cdot)$) trained using Eq. 3 to learn an expertise vector $\omega \in \mathbb{R}^m$ for each user. The variations among users are manifested through different values of ω, as their diverse domain knowledge influences the way they utilize concepts to arrive at predictions. Concretely, we ask users to annotate images and use ω to simulate their predictions. The expertise vector ω for a user is learned by minimizing the following cross-entropy loss:

$$\mathcal{L}_\omega = -\sum_{i=1}^{N} y_i^u \log(h(\Xi_\theta(\omega \cdot s_{\mathbf{c}}(\mathbf{x}_i))), \tag{4}$$

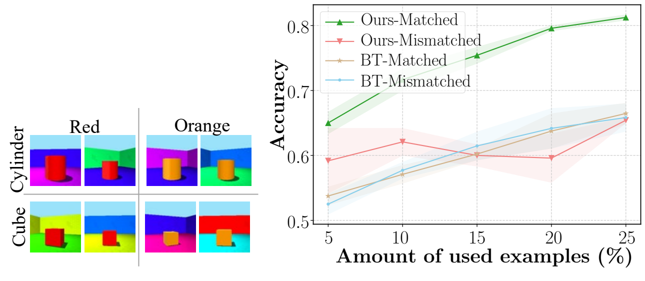
where y^u denotes annotated labels collected from the user. Once ω is learned, we obtain a user model denoted as $g_\omega(\cdot) = h(\Xi_\theta(\omega \cdot s_{\mathbf{c}}(\cdot))$. If $\omega_1 \approx \omega_2$, it implies that these two users (Users 1 and 2) have very similar "reasoning process" as the utilization of concepts is very similar. Likewise, if $\omega \approx \mathbf{1}_m$, this user employs a very similar reasoning mechanism as the target model f.

Selection Strategy

Our goal is to select a set of informative examples that can most improve the user's simulatability. To estimate the informativeness of the examples, we employ the concept of the hypercorrection effect in educational psychology: learning from this error example is more effective (Butterfield and Metcalfe 2001; Metcalfe and Finn 2011). As the human needs to learn how the model makes the decision, the model's prediction is viewed as the "correct" answer whereas the human's disagreed initial belief is the "error". Feedback on the correct answer along with explanations has been found to be crucial and most helpful in learning new knowledge (Metcalfe 2017). To reflect the hypercorrection effect in I-CEE, we choose images where the user has lower confidence in the model's predicted label after knowing the model's reasoning and argue that using these examples will lead to higher learning outcomes. Concretely, I-CEE aims to identify a set of examples $\mathcal{S} \subseteq \mathcal{D}$ which consists of samples with the top maximal Hypercorrection Effect:

$$x = \underset{\mathbf{x} \in \mathcal{D}}{\operatorname{argmax}}(g_\omega(y|\mathbf{x}) - g_\omega(y|\mathbf{x}, \mathbf{e})), \tag{5}$$

where $g_\omega(\cdot)$ represents the user model, $\mathcal{D}$ denotes the training dataset, and $\mathbf{e}$ and y are the local explanation and machine prediction corresponding to the image $\mathbf{x}$.



(a) Synthetic dataset (b) (Mis-)Matched expertise

Figure 4: (a): Overview of four classes in the synthetic dataset. (b): User simulatability accuracy when trained with examples that match/mismatch with the user expertise.

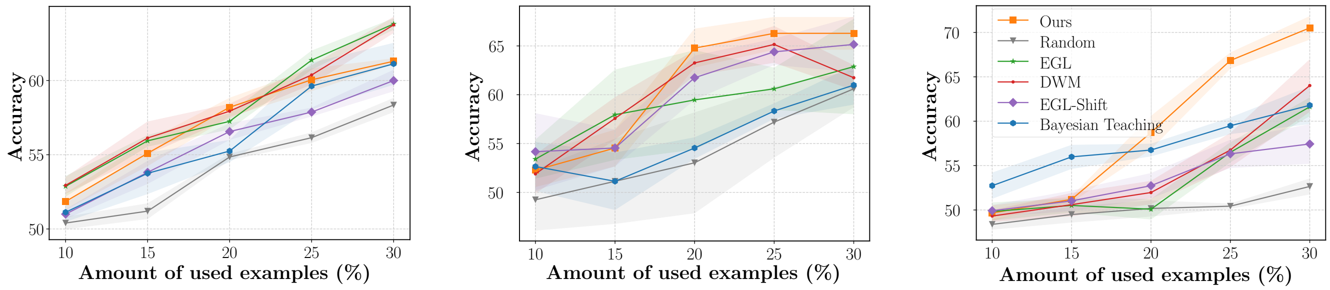
Experiments with Simulated Users

Before conducting a user study, we first evaluate our approach through extensive experiments with simulated users on one synthetic and three realistic image classification tasks. To facilitate reproducibility, Appendix includes more details about the experimental setup.

Synthetic Dataset. We construct a synthetic dataset[3] to validate the design of our proposed method in simulation. This dataset contains four classes and each class is described with two concepts, color and shape, illustrated in Figure 4a. For instance, if a user uses colors to distinguish between different classes (i.e., they have more expertise in using "colors" then "shapes"), then to this user, the red cylinders and red cubes belong to the same class, which differs from the orange ones. Likewise, for a user who has high expertise in using shapes, the cylinders and the cubes are distinguishable for this user regardless of their colors. The other visual features such as angles or background colors are randomly sampled as they are not essential in this decision-making process. For each class, we generate 300 images (80% for training and 20% for testing). We use a ResNet-18 (He et al. 2016) as our classification model and use GradCAM (Selvaraju et al. 2017) for generating explanations. Given their annotation behavior, a simulated behavior is modeled using Eqs. 3-4, i.e., identical to the modeling approach of I-CEE.

Realistic Datasets. We also benchmark I-CEE on three real-world datasets: CIFAR-100 (Krizhevsky, Hinton et al. 2009), CUB-200-2011 (Wah et al. 2011) and German Traffic Sign Recognition Benchmark (GTSRB) (Stallkamp et al. 2012). We construct a simulated user from pre-defined annotations on each dataset who behaves differently from the target model. In particular, for each dataset, our simulated user can distinguish only two classes out of four similar classes. All methods are evaluated based on this user. For instance, on CUB-200-2011, the simulated user labels both Crested and Least Auklet as the same class (Crested Auklet), and Parakeet and Rhinoceros Auklet as the same class (Parakeet Auklet). We use the original training-test splits on these datasets and, similar to the procedure in the synthetic

[3]This dataset is based on 3d-shapes (Kim and Mnih 2018).



(a) CIFAR-100 (b) CUB-200-2011 (c) GTSRB

Figure 5: Comparison with baseline algorithms with simulated users on three datasets. The ratio of used examples p (in percentage) is plotted on the x-axis and simulatability accuracy is on the y-axis. (Results averaged over 5 runs.)

dataset, we use ResNet-50 (He et al. 2016) for classification training and GradCAM for computing explanations.

Baseline Methods

We evaluate I-CEE against a recent human-centered XAI approach: Bayesian Teaching (BT) (Yang et al. 2021). BT simulates a user's behavior (i.e., their prediction of an image class) by deploying a ResNet-50-PLDA (probabilistic linear discriminate analysis (Ioffe 2006)) model. By assuming users perform Bayesian reasoning, it selects example images and explanations to better align user's beliefs to the target model. I-CEE and BT differ in their approaches to both user modeling and example selection.

To evaluate the example selection alone, we also benchmark against query strategies derived from active learning (AL). Unlike traditional AL, in our application of AL query strategies to XAI, the simulated user is the learner and the target model is the annotator. We use Expected Gradient Length (EGL) (Settles, Craven, and Ray 2007), Density-Weighted Method (DWM) (Settles, Craven, and Friedland 2008) as well as a random sampling strategy as baselines. EGL, in the context of this paper, selects samples (x, e) that result in the greatest change to the current model if the annotated label is known. The "change" imparted to the model from the queried samples is measured by the gradient of the objective function with respect to the model parameters. However, the instances chosen by EGL might be outliers that cause significant gradient changes. To alleviate this issue, Settles, Craven, and Friedland (2008) proposes to integrate a density-weighting technique with the query strategy such as EGL. Specifically, each sample is weighted with its average similarity to all other instances in the input dataset. In this work, we extend EGL with the belief shift in the calculated EGL when considering e in the input (denoted as EGL-Shift). Specifically, we compute the difference between EGL of (x, e) and x. With EGL-Shift, we aim to alleviate the influence of an image itself on the training gradient but emphasize the impact of explanations.

Evaluation Metric

To evaluate our method, we use simulatability, which is commonly used as a proxy for testing a user's understanding of the model's decision-making process (Hase and Bansal 2020; Arora et al. 2022; Hase et al. 2020). Simulatability is measured as "to what extent can a user successfully predict a model's prediction." This metric can be used in both simulation experiments and human user studies.

We follow the experimental settings proposed in (Yeh et al. 2018; Koh and Liang 2017) to study the influence of selected examples. Specifically, each method provides an ordered set of example images $\mathcal{S}$, where the ranking is decided by the *informativeness* defined in the respective method. We denote the ratio between number of example images $|\mathcal{S}|$ and the size of training data $\mathcal{D}$ as $p = {|\mathcal{S}|}/{|\mathcal{D}|}$. The simulated user is retrained using these example images $\mathbf{x}$ and their corresponding labels $y = f(\mathbf{x})$, where recall that f is the target model. Given the retrained user model g'_ω, we compute the user's accuracy of predicting the model's predictions on the test set, i.e., the simulatability of the user:

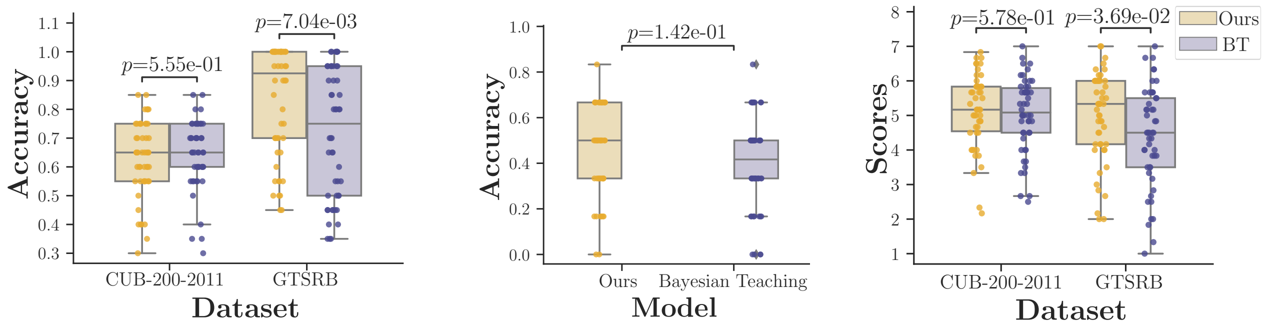
$$\text{Acc} = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{1}(y_i = g'_\omega(\mathbf{x_i})), \tag{6}$$

where N_t is the number of samples in the test set.

Experimental Results

Ablation Study. To validate our model design of $g(\cdot)$, we study (1) whether ω can faithfully reflect the user expertise and (2) the advantages of tailored explanations according to the user expertise. We simulate two users on the synthetic dataset: User 1 only uses color in classification while User 2 only uses shape. We deduce annotations for each user based on attributes for each class (Figure 4a).

After estimating each user, we investigate their expertise vector: ω_1 and ω_2 ($\omega_i \in \mathbb{R}^8$). Each entry in ω_i represents the expertise of the user in one specific concept. The top four largest entries in ω_1 and ω_2 are complementary, corresponding to the fact that each user has the opposite expertise (i.e., each user uses different concepts in the decision-making). To validate the efficacy of the user model via expertise, we run an experiment where we train User 1 using a set of examples specifically chosen based on the User 1 model ("Matched"), against a set of examples chosen for User 2 ("Mismatched"). As demonstrated in Figure 4b, we observe that the simulated user achieves high simulatability accuracy



(a) User Accuracy (all predictions) (b) User Accuracy (inaccurate pred., CUB) (c) User evaluation score.

Figure 6: Results of experiments with human users ($N = 100$) comparing I-CEE with the baseline Bayesian Teaching (BT). (a) Simulatability accuracy on all predictions, (b) Simulatability accuracy on images where the target model made inaccurate predictions in the CUB-200-2011 dataset, (c) User's subjective perception of model explanations.

when they receive examples selected according to their expertise ("Ours Matched"). However, if selecting examples that do not maximize the Hypercorrection Effect tailored to the particular user ("Ours Mismatched"), the simulatability accuracy is low, indicating that such examples fail to provide substantial insights into the target model. Additionally, we compare our user simulation model to that of Bayesian Teaching. We observe little differences between the matched and mismatched settings using the BT framework, suggesting that BT might not be able to accurately simulate the different behaviors of various users. Consequently, it cannot provide examples that effectively improve user simulatability (less performance improvement compared to ours).

Comparison. We compare I-CEE with baselines on three real-world datasets in Figure 5. Evaluation in user prediction accuracy is conducted at $p = [10, 15, 20, 25, 30]\%$. On CIFAR-100, our method always outperforms BT and EGL-Shift but is inferior to EGL and DWM. A potential reason for this result is that the explanation of CIFAR-100 is vague due to the low resolution of images. In this case, Hypercorrection Effect cannot be well captured since explanations are noisy. On CUB-200-2011 and GTSRB, our method outperforms other baselines at most of the percentages. For instance, on CUB our method achieves the best performance after 20%. Note that 20% of the train data consists of 24 images. This is a reasonable number of samples that can be efficiently studied by human users, which we will show in the next section. On GTSRB, we observe an evident performance gap between our method and the competitive baseline BT. A possible explanation for this can be attributed to the architecture of the user model: our model simulates the user via learning ω in the concept space without weakening the capability of the final classifier. On the contrary, BT relies on a PLDA layer to classify images, which can result in sub-optimal performance when the latent features of images are highly similar, such as in traffic signs. This is not desirable because humans are good at distilling critical concepts and filtering out similar but irrelevant visual features. With more precise user modeling, our method demonstrates the capability of offering informative learning samples in most of the cases within the simulation experiments.

Experiments with Human Users

We conduct a human user study using the CUB-200-2011 and GTSRB datasets following the same settings as in the simulation experiments. We choose these two datasets as they are more challenging and the images are in higher resolution. We use Bayesian Teaching (Yang et al. 2021) as a baseline since it is the most state-of-art and closest to our focus. Users are first asked to study two classes (among which there are actually four classes) and write down the features used to distinguish between these classes. This step is to let the user think as the pre-defined simulated user, to whom we have tailored model explanations. Then, 20 model explanations selected by our method (experimental group) or Bayesian Teaching (control group) for users are shown, and we ask them to write down the features they use to determine the model prediction. During the evaluation section, participants first receive a test with 15 questions to predict the model's label (images used here are sampled from the test set and include all four classes evenly). We refer to this section as "objective understanding". Then, participants rate their perceived understanding on seven questions on the 7-Likert scale, which we refer to as "subjective understanding". In the user study, we aim to study the following research questions:

- **R1**: Our framework selects informative samples that can increase human understanding of the model.
- **R2**: Human understanding of the model is affected by task domains.

Participants. We recruited 100 participants (average age is 28.8 ± 8.6, 49 females, 50 males, and 1 undefined) using a research platform Prolific, and randomly assigned them to one of the two conditions (50 participants/condition). 51 participants have prior experience with AI from using Alexa, Siri, ChatGPT, or from ML-related courses. All participants passed the attention check during the user study. The study protocol has been approved by the Technical University of Munich IRB. At the beginning of the experiment session,

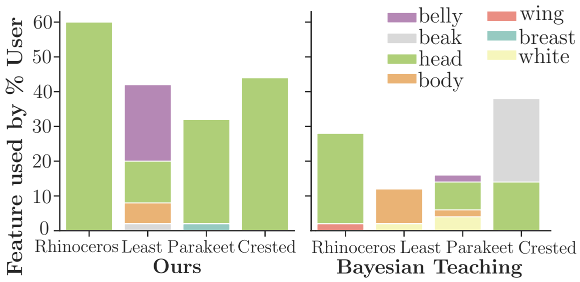


Figure 7: Illustration of features used by human users for distinguishing each class on CUB-200-2011.

we collected informed consent through Prolific. Each participant was compensated with a payment of £4.50 for participation in the user study (within 30 minutes).

Results

Analysis on R1. The results of the simulatability accuracy in each condition on each dataset are shown in Figure 6a. On GTSRB, we observe a statistically significant improvement in using our framework on user simulatability accuracy by 11.5% ($p = 0.007$). On the CUB dataset, we see that users from two conditions achieve similar user prediction accuracy and no significant effect is observed. However, if we inspect the test samples where the target model makes inaccurate predictions (wrong classification) (6 out of 15 images in the test are wrongly predicted), our method demonstrates superior performance compared to BT. Users from the experimental condition achieve an accuracy of 46.3%, whereas users from the control condition achieve 40.3%, as plotted in Figure 6b. These results indicate that users exhibit improved capability in simulating inaccurate predictions from the target model using our method, which is a more challenging task. Additional evidence of the enhancement achieved through our model can be found in Figure 7. We count the words of the features that users think the model uses to distinguish four different classes. When using our framework, the users tend to agree on the same feature (body part of the bird) for each class. For instance, about 68% of the users use "Head" to distinguish Rhinoceros, and about 20% of the users think highly of "Belly" for Least Auklet. Nevertheless, it is more difficult for users in Bayesian Teaching to come to an agreement, for example, for Least Auklet, only around 10% of the participants use "Body" as a feature while other users give diverse descriptions. These results highlight the advantage of the method in improving user understanding of the given target model.

As shown in Figure 6c, the improvement in subjective understanding (rating scores) is not significant on CUB (average rating score is 5.14 in our method and 5.02 in BT). However, we observe that on GTSRB our method surpasses BT significantly with $p = 0.037$. The reason for significant improvement in GTSRB is that our method selects explanations bringing knowledge for distinguishing four classes. But BT chooses examples that reflect important features only for two classes, which hinders users from understanding how the model makes predictions for the other classes.

Analysis on R2. The quantitative result shows that the task domain (dataset) affects the user's objective understanding. However, different tasks influence less subjective understanding, e.g., no significant difference between two datasets when using our method as illustrated in Figure 6c. At the end of the user study, we asked participants for feedback on comparing the perceived helpfulness of model explanations in two datasets. While most of the users in both conditions find the explanations useful, seven users in the experimental condition and fourteen users in the control condition find the explanations on bird species are more helpful than the explanations on road signs. One reason causing this uncertainty in the road sign images is that the salient area is always a circle that covers the road sign, which seems to "be the only one characteristic" for different classes.

Conclusion

We present a human-centered XAI framework, I-CEE, that provides explanations of image classification ML models that are tailored to user expertise. Our framework first discovers task-relevant concepts, uses these concepts to arrive at expertise-based user models, and then selects examples and explanations that help the users to learn the missing concepts so they can accurately predict the machine's image classification decisions. We evaluate our approach through simulation experiments on four datasets, and report on a detailed human-subject study ($N = 100$). In these experiments, we observe that I-CEE outperforms prior art, shows the promise of human-centered XAI, and motivates future research direction for the design of XAI systems.

Limitations and Future Work. Future investigation of our framework can consider the following avenues. First, more complex models of expertise estimation should be studied. In this work, we simulate user expertise by employing the concept-based reasoning approach for image classification proposed in (Yeh et al. 2020). An alternative approach involves utilizing Large Language Models to simulate multiple humans in textual format (Argyle et al. 2023; Aher, Arriaga, and Kalai 2023). Second, the current framework does not consider the sample complexity associated with user expertise estimation. Future work should investigate methods that estimate user expertise with a small number of real-user annotations. Third, we encourage replication of our work to be tested with different datasets, as the power of explanations is dependent on the task domain. Future work should evaluate on datasets that include a more diverse pool of examples, as suggested by some of the participants.

Implications for XAI Systems. This study highlights the importance of personalized XAI, within the explanation-by-example paradigm for image classification. Future work should investigate the potential of personalized XAI in other contexts. We argue that user modeling is essential to provide explanations that target user-specific misunderstanding or confusion. Future XAI systems should leverage and address individual users' preferences and confusion. This involves the development of human-in-the-loop systems, allowing users to actively participate in the process of generating explanations.

Ethical Statement

In this work, we attempt to put human users at the center of XAI design, with the aim of creating AI systems that can be interpreted by non-expert end users. To safeguard user privacy and user rights, we have received approval from University IRB. We believe that only when AI becomes more accessible, acceptable, and usable, can we realize its full potential to empower the world around us.

References

Aher, G. V.; Arriaga, R. I.; and Kalai, A. T. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, 337–371. PMLR.

Ambrose, S. A.; DiPietro, M.; Bridges, M. W.; Norman, M. K.; and Lovett, M. C. 2010. *How Does Students' Prior Knowledge Affect Their Learning*, chapter 1, 10–39. John Wiley & Sons.

Argyle, L. P.; Busby, E. C.; Fulda, N.; Gubler, J. R.; Rytting, C.; and Wingate, D. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351.

Armstrong, S. L.; Gleitman, L. R.; and Gleitman, H. 1983. What some concepts might not be. *Cognition*, 13(3): 263–308.

Arora, S.; Pruthi, D.; Sadeh, N.; Cohen, W. W.; Lipton, Z. C.; and Neubig, G. 2022. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 5277–5285.

Baker, C.; and Saxe, R. 2011. Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution. *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*.

Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Beliaev, M.; Shih, A.; Ermon, S.; Sadigh, D.; and Pedarsani, R. 2022. Imitation learning by estimating expertise of demonstrators. In *International Conference on Machine Learning*, 1732–1748. PMLR.

Butterfield, B.; and Metcalfe, J. 2001. Errors Committed with High Confidence Are Hypercorrected. *Journal of experimental psychology. Learning, memory, and cognition*, 27: 1491–4.

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

Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–19.

Hase, P.; and Bansal, M. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831*.

Hase, P.; Zhang, S.; Xie, H.; and Bansal, M. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? *arXiv preprint arXiv:2010.04119*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, S.; Held, D.; Abbeel, P.; and Dragan, A. 2019. Enabling Robots to Communicate their Objectives. *Autonomous Robots*, 43.

Ioffe, S. 2006. Probabilistic linear discriminant analysis. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV 9*, 531–542. Springer.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *International Conference on Machine Learning*, 2649–2658. PMLR.

Koh, P. W.; and Liang, P. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, 1885–1894. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lage, I.; Chen, E.; He, J.; Narayanan, M.; Kim, B.; Gershman, S. J.; and Doshi-Velez, F. 2019a. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, 59–67.

Lage, I.; and Doshi-Velez, F. 2020. Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898*.

Lage, I.; Lifschitz, D.; Doshi velez, F.; and Amir, O. 2019b. Exploring Computational User Models for Agent Policy Summarization. *IJCAI : proceedings of the conference*, 28: 1401–1407.

Liao, Q. V.; and Varshney, K. R. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Metcalfe, J. 2017. Learning from Errors. *Annual Review of Psychology*, 68(1): 465–489.

Metcalfe, J.; and Finn, B. 2011. People's Hypercorrection of High-Confidence Errors: Did They Know It All Along? *Journal of experimental psychology. Learning, memory, and cognition*, 37: 437–48.

Owens, M.; and Tanner, K. 2017. Teaching as Brain Changing: Exploring Connections between Neuroscience and Innovative Teaching. *Cell Biology Education*, 16: fe2.

Qian, P.; and Unhelkar, V. 2022. Evaluating the Role of Interactivity on Improving Transparency in Autonomous

Agents. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1083–1091.

Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; and Wang, X. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9): 1–40.

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

Rong, Y.; Leemann, T.; Nguyen, T.-T.; Fiedler, L.; Qian, P.; Unhelkar, V.; Seidel, T.; Kasneci, G.; and Kasneci, E. 2023. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Russell, S. 2021. Human-compatible artificial intelligence. *Human-like machine intelligence*, 3–23.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Settles, B. 2009. Active learning literature survey.

Settles, B.; and Craven, M. 2008. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, 1070–1079.

Settles, B.; Craven, M.; and Friedland, L. 2008. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:.

Settles, B.; Craven, M.; and Ray, S. 2007. Multiple-instance active learning. *Advances in neural information processing systems*, 20.

Sinha, S.; Ebrahimi, S.; and Darrell, T. 2019. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5972–5981.

Stallkamp, J.; Schlipsing, M.; Salmen, J.; and Igel, C. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32: 323–332.

Tenenbaum, J. B. 1999. *A Bayesian framework for concept learning*. Ph.D. thesis, Massachusetts Institute of Technology.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. 2010. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, 23.

Yang, S. C.-H.; Folke, N. E. T.; and Shafto, P. 2022. A psychological theory of explainability. In *International Conference on Machine Learning*, 25007–25021. PMLR.

Yang, S. C.-H.; Vong, W. K.; Sojitra, R. B.; Folke, T.; and Shafto, P. 2021. Mitigating belief projection in explainable artificial intelligence via Bayesian teaching. *Scientific reports*, 11(1): 9863.

Yang, X. J.; Unhelkar, V. V.; Li, K.; and Shah, J. A. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 408–416.

Yeh, C.-K.; Kim, B.; Arik, S.; Li, C.-L.; Pfister, T.; and Ravikumar, P. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33: 20554–20565.

Yeh, C.-K.; Kim, J.; Yen, I. E.-H.; and Ravikumar, P. K. 2018. Representer point selection for explaining deep neural networks. *Advances in neural information processing systems*, 31.