

1 May I Ask a Follow-up Question? Understanding the Benefits of
2 Conversations in Neural Network Explainability

3 Tong Zhang

4 Computer Science and Engineering, Nanyang Technological University, Singapore

5 X. Jessie Yang

6 Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan

7 Boyang Li

8 Computer Science and Engineering, Nanyang Technological University, Singapore

9 **Manuscript type:** Original Research

10 **Running head:** Benefits of Conversations in Neural Network Explainability

11 **Corresponding author:** Boyang Li, School of Computer Science and Engineering,
12 Nanyang Technological University, 50 Nanyang Avenue, Singapore, 639798, Email:
13 boyang.li@ntu.edu.sg

14 **Acknowledgments:** This work has been supported by the Nanyang Associate
15 Professorship and the National Research Foundation Fellowship
16 (NRFNRFF13-2021-0006), Singapore. Any opinions, findings, conclusions, or
17 recommendations expressed in this material are those of the authors and do not reflect
18 the views of the funding agencies.

Abstract

Research in explainable AI (XAI) aims to provide insights into the decision-making process of opaque AI models. To date, most XAI methods offer one-off and static explanations, which cannot cater to the diverse backgrounds and understanding levels of users. With this paper, we investigate if free-form conversations can enhance users' comprehension of static explanations in image classification, improve acceptance and trust in the explanation methods, and facilitate human-AI collaboration. We conduct a human-subject experiment with 120 participants. Half serve as the experimental group and engage in a conversation with a human expert regarding the static explanations, while the other half are in the control group and read the materials regarding static explanations independently. We measure the participants' objective and self-reported comprehension, acceptance, and trust of static explanations. Results show that conversations significantly improve participants' comprehension, acceptance (Davis, 1989), trust, and collaboration with static explanations, while reading the explanations independently does not have these effects and even decreases users' acceptance of explanations. Our findings highlight the importance of customized model explanations in the format of free-form conversations and provide insights for the future design of conversational explanations.

Keywords: Explainable AI (XAI), Conversation, Explainability, Interactive Explanation, Human-AI Interaction, XAI for Computer Vision

Introduction

The rapid advancement of Artificial Intelligence (AI) is largely powered by opaque deep neural networks (DNNs), which are difficult to interpret by humans (Bodria et al., 2023). The lack of transparency prevents verification of AI decisions by human domain experts and is especially concerning in areas of high-stake decisions, such as healthcare and law enforcement, where erroneous algorithmic decisions could lead to severe consequences (Cai, Winter, Steiner, Wilcox, & Terry, 2019; Caruana et al., 2015; Zheng et al., 2023) and erosion of public trust (Powles & Hodson, 2017; Quinn, Senadeera, Jacobs, Coghlan, & Le, 2021). To improve the explainability of AI models, numerous eXplainable Artificial Intelligence (XAI) methods have been proposed (for detailed reviews, we refer readers to Bodria et al. (2023); Danilevsky et al. (2020); F. Yang, Du, and Hu (2019)). It has been reported that explainability enhances user understanding (Bansal et al., 2021) and trust (González et al., 2021; R. Luo, Du, & Yang, 2022) in AI models, improves human-AI collaboration in decision-making (Lai & Tan, 2019; Nguyen, Taesiri, & Nguyen, 2022), and helps AI developers identify and rectify model errors (Adebayo, Muelly, Liccardi, & Kim, 2020; Idahl, Lyu, Gadiraju, & Anand, 2021). Despite these successes, a number of recent studies find that the explanations often do not resolve user confusion regarding the neural networks they are purported to explain (Bansal et al., 2021; Lakkaraju, Slack, Chen, Tan, & Singh, 2022; Liao, Gruen, & Miller, 2020; Poursabzi-Sangdeh, Goldstein, Hofman, Vaughan, & Wallach, 2021; Shen, Huang, Wu, & Huang, 2023; Slack, Krishna, Lakkaraju, & Singh, 2023; Wang & Yin, 2021; Y. Zhang, Liao, & Bellamy, 2020). These seemingly conflicting findings warrant further investigation.

We postulate that two major factors contribute to the ineffectiveness of AI explanations. First, the explanations do not properly account for average users' knowledge of machine learning, which may be insufficient to establish causal relations between the explanations and the model behaviors (He, Hong, Zheng, & Zhang, 2023; Ma et al., 2023; Poursabzi-Sangdeh et al., 2021; Springer & Whittaker, 2019). Communication theory posits that effective communication requires the senders and receivers to establish common ground (Clark & Brennan, 1991; Clark & Marshall, 1981). However, experts usually

find it hard to accurately estimate what laypeople know (Miller, 2019; Wilkesmann & Wilkesmann, 2011; Wittwer, Nückles, & Renkl, 2008). To make matters worse, underestimating and overestimating the receivers’ knowledge level are equally detrimental to communication (Lakkaraju et al., 2022; Wittwer et al., 2008). As a result, the explanations designed by experts are almost always at a mismatch with the laypersons’ actual knowledge level.

Second, users of XAI have diverse intentions and information needs (Ehsan et al., 2021; He et al., 2023; Liao et al., 2020; Wang & Yin, 2021). For example, Liao and Varshney (2021) identifies five different objectives of users of explanations, including model debugging, assessing the capabilities of AI systems, making informed decisions, seeking recourse or contesting the AI, and auditing for legal or ethical compliance. One static explanation usually cannot satisfy all objectives and purposes. Therefore, researchers have suggested injecting interactivity to model explanations in order to establish common ground, address knowledge gaps, and create customized explanations that adapt to the users (Abdul, Vermeulen, Wang, Lim, & Kankanhalli, 2018; Cheng et al., 2019; Guesmi et al., 2023; Lakkaraju et al., 2022; Rohlfing et al., 2020; Schmid & Wrede, 2022).

Existing work on interactive explanations can be broadly categorized into two types. The first type, interactive machine learning (Amershi, Cakmak, Knox, & Kulesza, 2014; Fails & Olsen Jr, 2003), allows users to provide feedback and suggestions to the machine learning model using model explanations. Their primary goal is to improve machine learning performances, rather than explaining model behaviors to layperson users. In this setting, explanations have been shown to improve user satisfaction (Smith-Renner et al., 2020) and feedback quality (Kulesza, Burnett, Wong, & Stumpf, 2015; Liang, Zou, & Yu, 2020). The second type aims to elucidate model behaviors by allowing users to freely modify input features and observe how outputs change while showing feature attribution explanations (Cheng et al., 2019; Hohman, Head, Caruana, DeLine, & Drucker, 2019; H. Liu, Lai, & Tan, 2021; Tenney et al., 2020). This type of interactivity has been shown to improve user understanding (Cheng et al., 2019) and perceived usefulness (H. Liu et al., 2021) of AI models. However, the effective use of these interactive approaches still

requires a rudimentary understanding of machine learning, such as the generic relation between input and output, or what model properties the interpretations reveal. These interactive explanations cannot answer most types of follow-up questions laypeople may have.

Free-form conversations that accompany static explanations are arguably the most versatile mode of interaction as they allow users to ask arbitrary follow-up questions and receive explanations tailored to their backgrounds and needs (Feldhus, Ravichandran, & Möller, 2022; Lakkaraju et al., 2022; Liao et al., 2020). Through interviews with decision-makers, Lakkaraju et al. (2022) discover that they have a strong preference for explanations in natural language dialogue. They argue that conversational explanations satisfy five requirements of interactive explanations and are ideal for users with limited machine learning knowledge. With the progress in conversational characters (Ni, Young, Pandelea, Xue, & Cambria, 2023; Shuster et al., 2022; T. Zhang et al., 2022), especially knowledge-based question answering (Lan et al., 2021; M. Luo, Fang, Gokhale, Yang, & Baral, 2023; L. Zhang et al., 2023) powered by large language models (Ouyang et al., 2022; Touvron et al., 2023; Zhao et al., 2023), AI systems that can answer questions about their own decisions appear to be within our reach in the near future. However, before investing effort to develop such a chatbot, it would be beneficial to empirically quantify the effects of conversational explanations.

In the current study, we conduct Wizard-of-Oz experiments to investigate how conversations assist users in understanding static explanations of image classification models, improving acceptance and trust in XAI methods, and selecting the best AI models based on explanations. Specifically, a total of 120 participants join our experiments. We first present them with static explanations for an image classification task and measure their objective understanding and subjective perceptions of static explanations. After that, half of the participants, who are assigned to the experimental group, seek to clarify any doubts with an online textual conversation with an AI system, played by human XAI experts. The other half of the participants, assigned to the control group, read materials about the static explanations independently. After the conversation or reading session,

participants complete the same pre-session measurements. From the results, we estimate the effects of conversational explanations.

The experimental measurements include both an objective component and a subject component of the users’ understanding and perception. In the objective evaluation, from three candidate neural networks, the users need to choose one network that would be the most accurate on test data so far unobserved, using information from the static explanations. This task, known as model selection, is one of the most fundamental tasks for machine learning practitioners (Anderson & Burnham, 2004). By design, the three candidate networks make exactly the same predictions on the same inputs but have different rationales for the predictions, as revealed by the static explanations. Hence, the only way for the users to make the right choice is to correctly understand the explanations. The subjective evaluation contains 13 questions requiring users to self-report three aspects of their perceptions of the static explanations: comprehension, acceptance, and trust.

Results show that free-form conversations with XAI experts in the Wizard-of-Oz setting significantly improve comprehension, acceptance, trust, and collaboration with static explanations. Our study underscores the effects of free-form conversations on neural network explainability in practice and provides insights into the future development of conversational explanations. To the best of our knowledge, this is the first study of how free-form conversations may facilitate neural network explainability in practice.

Related Work

In this section, we review three bodies of research that motivate our study. First, we explore the existing work of static Explainable Artificial Intelligence (XAI). Second, we discuss interactive explanations, especially the limitations of existing methods and the need for conversations to enhance explainability. Lastly, we examine different types of human-AI collaboration and the design of the subjective evaluation during collaboration.

Static Explanation

Explainable Artificial Intelligence (XAI) refers to those models that can explain either the learning process or the outcome of AI predictions to human users (F. Yang et al.,

2019). Static XAI involves models that provide a fixed, one-time explanation, without the capability for further user interaction or exploration. They are usually categorized into two groups: self-explanatory models and post-hoc methods. Post-hoc methods can be categorized into feature attribution methods and example-based methods. Self-explanatory models are inherently transparent, offering clarity in their decision-making processes and facilitating explainability (Bodria et al., 2023; Danilevsky et al., 2020). Examples of such models include linear regression, logistic regression, decision trees, Naive Bayes, attention mechanism (Bahdanau, Cho, & Bengio, 2014), decision sets (Lakkaraju, Bach, & Leskovec, 2016), rule-based models (Rudziński, 2016; H. Yang, Rudin, & Seltzer, 2017), among others. However, the requirements of self-explanatory models place constraints on model design, which may cause them to underperform in complex tasks. Conversely, the majority of recent XAI methods are post-hoc XAI methods, which can be used for an already developed model that is usually not inherently transparent (Adadi & Berrada, 2018; Bodria et al., 2023; Y. Chen, Li, Yu, Wu, & Miao, 2021; Ribeiro, Singh, & Guestrin, 2016; Selvaraju et al., 2017; Verma, Dickerson, & Hines, 2020). These methods often do not attempt to explain how the model works internally, but instead, employ separate techniques to extract explanatory information. Post-hoc XAI methods can be viewed as reverse engineering processes that employ other independent explanatory models or techniques to extract explanatory information without altering, elucidating, or even understanding the inner workings of the original black-box model. There are two main groups of methods to generate post-hoc XAI explanations, i.e., feature attribution methods and example-based methods.

Feature Attribution Methods. Feature attribution methods (Alvarez-Melis & Jaakkola, 2017; Cortez & Embrechts, 2013; Hu, Chen, Nair, & Sudjianto, 2018; Ignatiev, Narodytska, & Marques-Silva, 2019; N. Liu, Huang, Li, & Hu, 2018; Lundberg & Lee, 2017; Ribeiro et al., 2016; Selvaraju et al., 2017; Shih, Choi, & Darwiche, 2018; Simonyan, Vedaldi, & Zisserman, 2013; Sundararajan, Taly, & Yan, 2017) explain model predictions by investigating the importance of different input features to final predictions. There are two main types of feature attribution methods, gradient-based methods (Cortez &

Embrechts, 2013; Lundberg & Lee, 2017; Selvaraju et al., 2017; Simonyan et al., 2013; Sundararajan et al., 2017) and surrogate methods (Alvarez-Melis & Jaakkola, 2017; Hu et al., 2018; Ignatiev et al., 2019; N. Liu et al., 2018; Ribeiro et al., 2016; Shih et al., 2018). Gradient-based methods use gradients/derivatives to evaluate the contribution of a model input on the model output. An example method is Grad-CAM (Selvaraju et al., 2017). It superimposes a heatmap on the regions of important input features by weighting the activations of the final convolutional layer by their corresponding gradients and averaging the resulting weights spatially. Besides directly calculating the importance score of input features, several methods propose to use a simple and understandable surrogate model, e.g., a linear model, to locally approximate the complex deep neural model. Surrogate models can explain the predictions from the complex deep neural model due to their inherent interpretable nature. LIME and its variants are typical methods for generating local surrogate models. LIME (Ribeiro et al., 2016) builds a linear model locally around the data point to be interpreted and generates an instance-level explanation for the output.

Example-based Methods. Example-based methods (Y. Chen et al., 2021; Jeyakumar, Noor, Cheng, Garcia, & Srivastava, 2020; Mothilal, Sharma, & Tan, 2020; Poyiadzi, Sokol, Santos-Rodriguez, De Bie, & Flach, 2020; Tran, Ghazimatin, & Saha Roy, 2021; Verma et al., 2020) refer to those that explain predictions of black-box models by identifying and presenting a selection of similar or representative instances. Those examples can be selected or generated from different perspectives, such as training data points that are the most influential to the parameters of a prediction model or the predictions themselves (C. Chen et al., 2019; Y. Chen et al., 2021; Yoon, Arik, & Pfister, 2020), counterfactual examples that are similar to the input query but with different predictions (Karimi, Barthe, Balle, & Valera, 2020; Mothilal et al., 2020; Poyiadzi et al., 2020; Sharma, Henderson, & Ghosh, 2019; Tran et al., 2021; Verma et al., 2020; Wachter, Mittelstadt, & Russell, 2017), or prototypes that contain semantically similar parts to input instances (Bien & Tibshirani, 2011; Croce, Rossini, & Basili, 2019; Doshi-Velez, Wallace, & Adams, 2015; Jeyakumar et al., 2020; B. Kim, Khanna, & Koyejo, 2016; Mikolov,

Sutskever, Chen, Corrado, & Dean, 2013).

In this work, we mainly focus on feature attribution methods as they directly highlight the importance of input features, making the decision-making process of models more intuitive (S. S. Y. Kim, Watkins, Russakovsky, Fong, & Monroy-Hernández, 2023) than example-based methods for laypeople. Specifically, we select Grad-CAM from gradient-based methods and LIME from surrogate methods to conduct conversational explanations with participants.

Interactive Explanation

Several studies emphasize the need for interactivity in XAI methods (Abdul et al., 2018; Lakkaraju et al., 2022; Rohlfing et al., 2020; Schmid & Wrede, 2022). For instance, Lakkaraju et al. (2022) find that decision-makers strongly prefer interactive explanations. Similarly, a literature analysis by Abdul et al. (2018) suggests that interactions can help users progressively explore and gather insights from static explanations. Rohlfing et al. (2020) reason that explanations should be co-constructed in an interaction between the explainer and the explainee, adapting to individual differences since the human understanding process is dynamic. From an interdisciplinary perspective, Schmid and Wrede (2022) underscore the necessity of user-XAI interactions to adapt to diverse information requirements.

To integrate interactivity and explainability, two primary methodologies emerge. One group of methods focuses on using explanations to help users provide feedback about improving machine learning models. In these methods, the interactivity lies in the cycle of model explanation, user feedback, and model improvement. Explanations aim to help users better understand model decisions and provide valuable feedback. As a result, machine learning models can be incrementally trained with additional loss terms from explanatory feedback (Kulesza et al., 2015; Lertvittayakumjorn, Specia, & Toni, 2020; Liang et al., 2020; Ross, Hughes, & Doshi-Velez, 2017; Schramowski et al., 2020; Smith-Renner et al., 2020) or with added data points (Alkan et al., 2022; Biswas & Parikh, 2013; Teso, Bontempelli, Giunchiglia, & Passerini, 2021; Teso & Kersting, 2019). However,

these methods are aimed at machine learning practitioners who can well understand and utilize explanations. Another group focuses on enhancing user understanding of explanations by allowing them to modify the model input and observe changes in the corresponding output. Such interactivity has been shown to improve user comprehension and the perceived utility of AI models (Cheng et al., 2019; H. Liu et al., 2021). For instance, Tenney et al. (2020) and Hohman et al. (2019) propose different user interfaces that allow for debugging and understanding machine learning models by examining input-output relationships. However, a rudimentary understanding of machine learning is still required for effective utilization of these interfaces, such as the generic relation between input and output, or what model properties the interpretations reveal.

HCI researchers have recently proposed that XAI methods should align with the ways humans naturally explain mechanisms. Specifically, Lombrozo (2006) argues that an explanation is a byproduct of a conversational interaction process between an explainer and an explainee. Miller (2019) argues that explanations should contain a communication process, where the explainer interactively provides the information required for the explainee to understand the causes of the event through conversations. Building on this perspective of human explanations, recent works envision "explainability as dialogue" to provide explanations suitable for a wide range of layperson users (Feldhus et al., 2022; Lakkaraju et al., 2022; Liao et al., 2020). While there is much theoretical analysis about the significance of conversations in explainability, practical investigations into their impact on users remain limited. In this context, two previous works have investigated the practical effect of conversations for explainability (Shen et al., 2023; Slack et al., 2023). Shen et al. (2023) apply conversational explanations to scientific writing tasks, observing improvements in productivity and sentence quality. Slack et al. (2023) design dialogue systems to help users better understand machine learning models on diabetes prediction, rearrest prediction, and loan default prediction tasks. Despite these advances, the conversations in these studies are generated based on templates and cope with limited predefined user intentions. In this study, we explore the role of free-form conversations in enhancing users' comprehension of static explanations, and how they affect users' ac-

ceptance, trust, and collaboration with these explanations.

Human-AI Collaboration

Human-AI collaboration is an emerging research area, which explores how humans and AI systems can work together to achieve shared goals (Herse, Vitale, & Williams, 2023; S. S. Y. Kim et al., 2023; Xu, Dainoff, Ge, & Gao, 2023). Prior studies within this domain have investigated collaborations between humans and various AI systems, from robots (Bhat, Lyons, Shi, & Yang, 2024; Carissoli, Negri, Bassi, Storm, & Fave, 2023; Gero et al., 2020; Häuslschmid, von Bülow, Pfleging, & Butz, 2017; L. Liu, Guo, Zou, & Duffy, 2022) to virtual agents (Ashktorab et al., 2020; Cai et al., 2019; D’Avella, Camacho-Gonzalez, & Tripicchio, 2022; Numata et al., 2020). The tasks involved span a broad scope, including text (Bansal et al., 2021) and image (S. S. Y. Kim et al., 2023) classifications, medical diagnosis (Cai et al., 2019), deception detection (Lai & Tan, 2019) and cooperative games (Ashktorab et al., 2020; Feng & Boyd-Graber, 2019; Gero et al., 2020). An area of particular interest within these collaborations is the role of explanations in influencing human-AI decision-making (Bansal et al., 2021; Lai & Tan, 2019; Nguyen, Kim, & Nguyen, 2021; Nguyen et al., 2022).

Our study aligns with existing work on human-AI collaboration (Bansal et al., 2021; Feng & Boyd-Graber, 2019; Lai & Tan, 2019; Nguyen et al., 2021, 2022). In our work, participants need to collaborate with explanations to choose the most accurate neural networks among others. Instead of exploring the role of explanations in collaboration, we mainly examine the potential of conversations in aiding users to effectively use explainability techniques and understand their outputs.

Method

Our study aims to investigate the impact of conversations on the explainability of AI models by observing participants’ comprehension, acceptance, trust of the static explanations, and ability to use the explanations to select the most accurate neural networks before and after the conversation. Our study has received approval from the Institutional Review Board at Nanyang Technological University (#IRB-2023-254).

TABLE 1: *ACADEMIC DISCIPLINES OF OUR PARTICIPANTS AND THE NUMBER OF PARTICIPANTS IN EACH GROUP. THERE ARE 120 PARTICIPANTS FROM 4 DIFFERENT DISCIPLINE GROUPS.*

Academic Discipline	Number of Participants
Business	23
Engineering	16
Humanities	55
Science	26

Participants

A total of 120 participants joined our study. All were 21 years old or older, fluent in English, and had not been involved in research about XAI previously. We recruited our participants in two ways: by posting advertisements on an online forum and by emailing students and staff across various departments and schools. They are from a wide range of disciplines to promote diversity. For ease of reporting, we categorize their disciplines into four groups:

- Business, including Business and Accountancy.
- Engineering, including Civil and Environmental Engineering, Computer Science, Electrical and Electronics Engineering, Maritime Studies, and Food Science.
- Humanities, including Psychology, Economics, Communication Studies, Linguistics and Multilingual Studies, and Sociology.
- Science, including Biology, Chemistry, Chemical Engineering and Biotechnology, Sport Science & Management, Mathematics, Medicine, and Physics.

Table 1 shows statistics of the academic disciplines that the participants enrolled in.

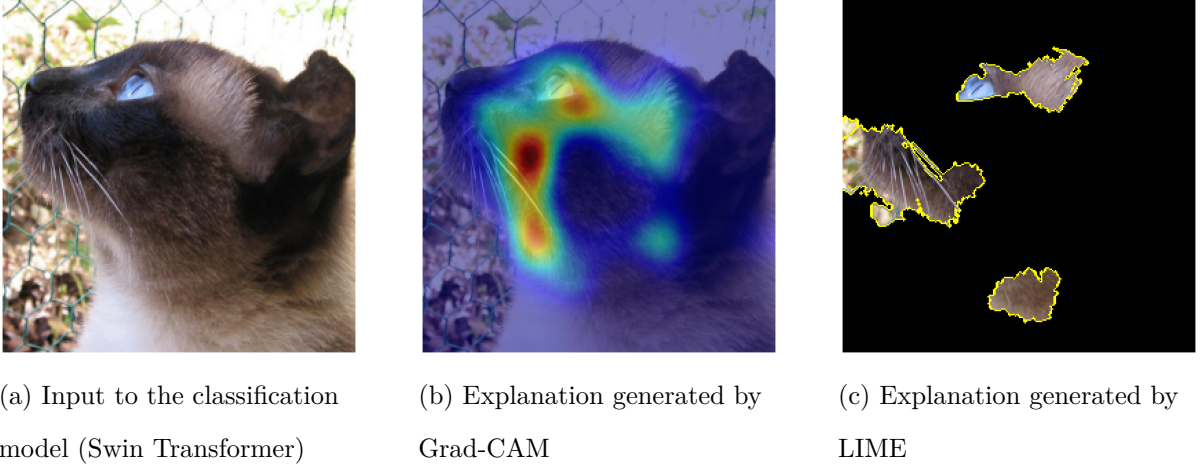


Figure 1. Example explanations generated by Grad-CAM and LIME. (a) is the input to the classification model (Swin Transformer), (b) is the explanation generated by Grad-CAM, and (c) is the explanation generated by LIME. The predicted class of the model is "Siamese cat".

Experimental Task

In our study, we focus on the image classification task on the ImageNet dataset (Deng et al., 2009). Image classification task is a cornerstone in the field of computer vision (CV) that has been the subject of various human-AI collaborative studies (Jeyakumar et al., 2020; Taesiri, Nguyen, & Nguyen, 2022). We train three classification models with different top-1 classification accuracies: Swin Transformer (Z. Liu et al., 2021) (84.1%), VGG-16 (Simonyan & Zisserman, 2015) (71.6%), and AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) (56.5%). To generate explanations for model predictions, we select two explanation techniques from two main categories of feature attribution explanation methods: LIME (Ribeiro et al., 2016) (a surrogate method) and Grad-CAM (Selvaraju et al., 2017) (a gradient-based method). We focus on feature attribution explanations as we believe the relationship between input features and model predictions is more intuitive to understand than example-based methods for laypeople (S. S. Y. Kim et al., 2023). Figure 1 displays example explanations generated by these two explanation methods.

To conduct the study, we design and build a web-based platform where participants can remotely finish the whole procedure of the experiment. After users log into the

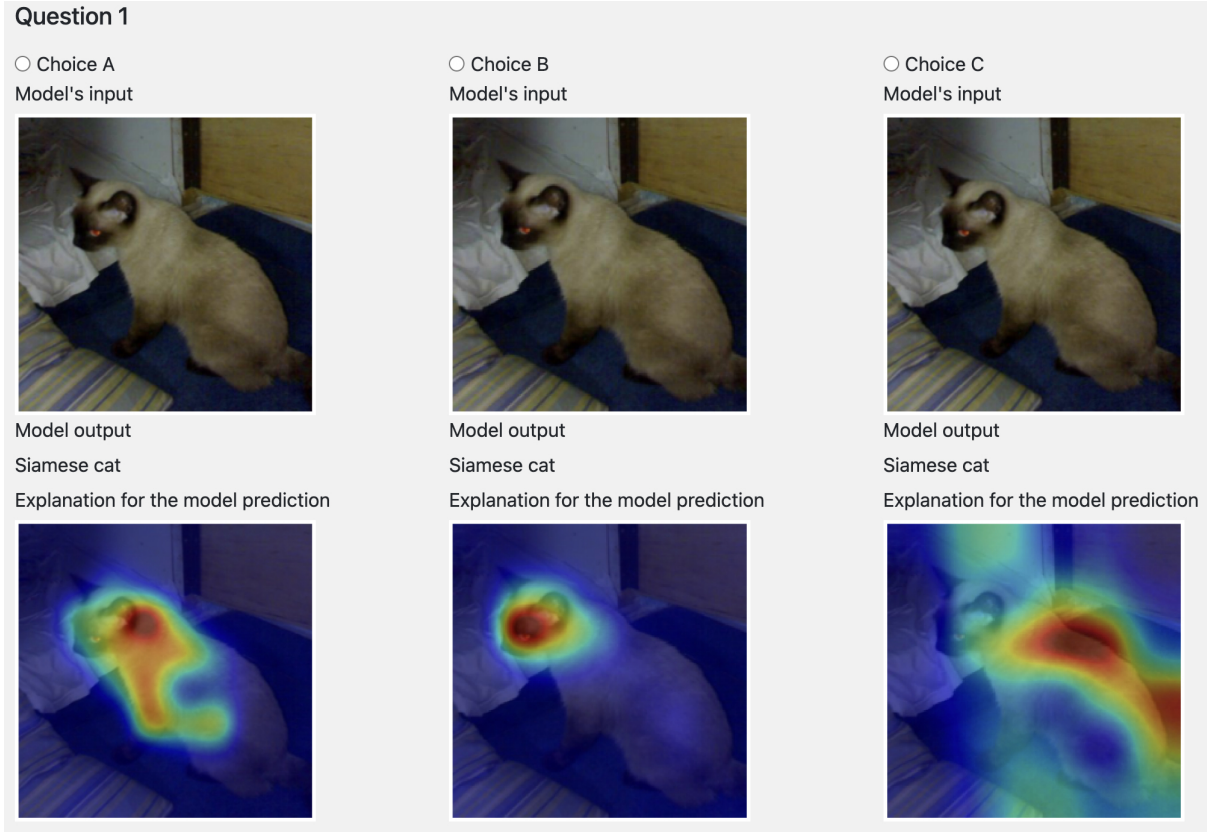


Figure 2. An example of the objective evaluation. The objective evaluation aims to objectively measure participants' comprehension of static explanations. Each choice contains a prediction from a different classification model, paired with its respective static explanation. Participants need to choose the best model based on the explanations.

1 platform, we first evaluate their objective and subjective understanding of static explana-
 2 tions. The objective explanations require participants to choose, from three classification
 3 models, the most accurate on unobserved test data. The three classification models yield
 4 identical decisions on 5 images. The only differences between the three networks lie in
 5 their explanations. Hence, to select the best model, the participants must rely on the
 6 explanations. Figure 2 presents an example question, including the original image, the
 7 model outputs, and the explanations. The full set of questions used in the study can be
 8 found in Appendix A.

TABLE 2: *DETAILED QUESTIONS IN THE SUBJECTIVE EVALUATION. THE USER WILL RESPOND TO EACH QUESTION USING A 7-POINT LIKERT SCALE.*

Aspect	Question
Comprehension	How much do you think you understand the explanations provided for predictions of deep learning models?
Perceived Usefulness	Using explanations would improve my understanding of deep learning models' predictions.
	Using explanations would enhance my effectiveness in understanding predictions of deep learning models.
	I would find explanations useful in understanding predictions of deep learning models.
Perceived Ease-of-Use	I become confused when I use the explanation information.
	It is easy to use explanation information to understand predictions of deep learning models.
	Overall, I would find explanation information easy to use.
Behavioral Intention	I would prefer getting explanation information as long as it is available when getting predictions from deep learning models.
	I would recommend others use explanation information to understand predictions of deep learning models.
Trust	How would you rate the competence of the explanation method? - i.e. to what extent does the explanation method perform its function properly?
	How would you rate the dependability of the explanation method? - i.e. to what extent can you count on the explanation method to explain predictions of deep learning models?
	How would you rate your degree of faith that the explanation method will be able to explain predictions of deep learning models in the future?
	How would you rate your overall trust in the explanation method?

1 The subjective evaluation measures participants' self-reported perception of the
2 static explanations, including their comprehension (Cheng et al., 2019; Hoffman, Mueller,

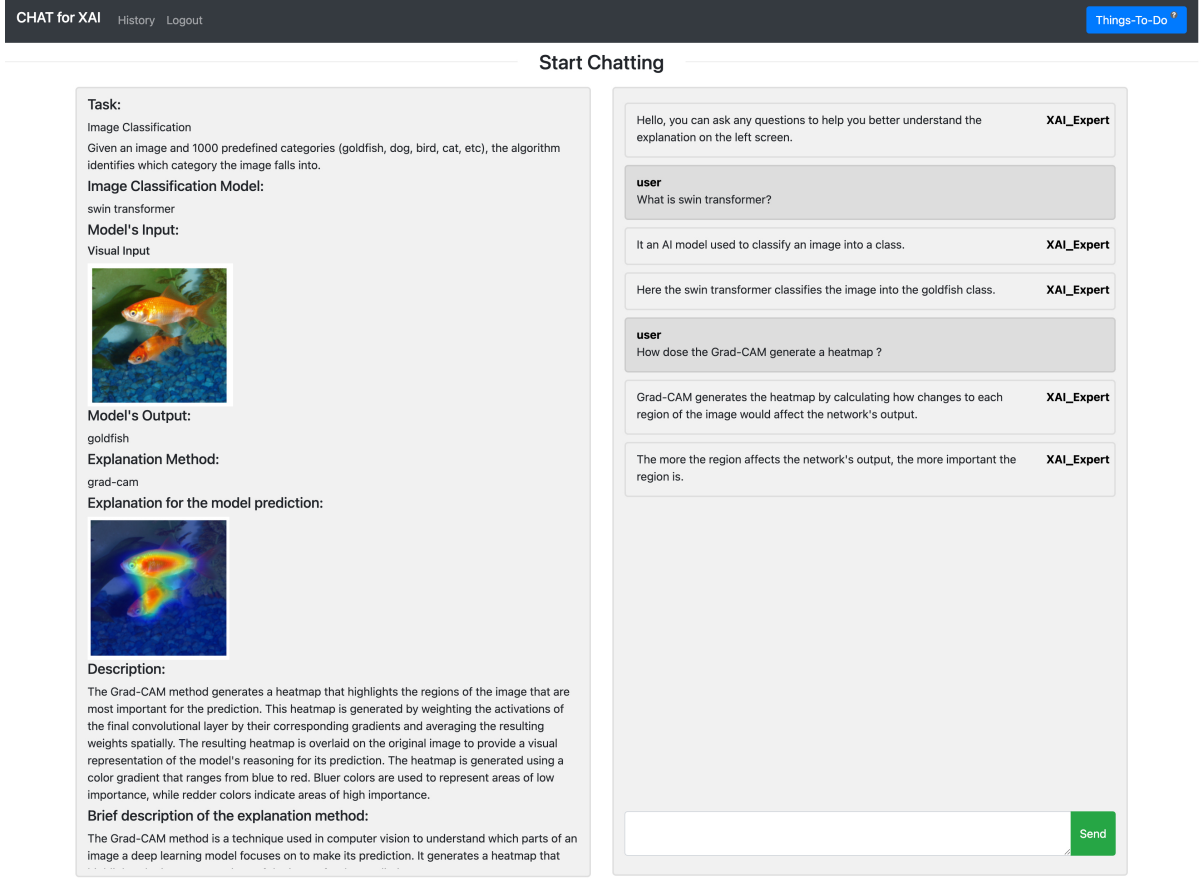


Figure 3. The web page where users can discuss static explanations with an expert.

1 Klein, & Litman, 2018), acceptance (Davis, 1989; Davis, Bagozzi, & Warshaw, 1989; Diop,
 2 Zhao, & Duy, 2019; Flathmann et al., 2023), and trust (Davis, 1989; Davis et al., 1989;
 3 Diop et al., 2019; Guo, Yang, & Shi, 2023; X. J. Yang, Unhelkar, Li, & Shah, 2017).
 4 Based on an in-depth review of existing literature, we chose the questions from those
 5 that have been validated in prior research. The subjective evaluation contains a total
 6 of 13 questions, each utilizing a 7-point Likert scale for responses. Table 2 lists all the
 7 questions we used. Labels of the 7-point Likert scale are listed in Appendix A.

8 After these two evaluations, participants are divided into two groups, i.e., the con-
 9 trol group and the experimental group. Participants in the control group read static
 10 explanations for 15 minutes. Participants in the experimental group conduct conversa-
 11 tional explanations with participants in the Wizard-of-Oz (WoZ) setting (Kelley, 1984).
 12 They interact with a dialogue system that they believe to be autonomous but is actually
 13 operated by a human expert on machine learning.

To support the WoZ experiment, we built a conversation page with a two-section structure, as depicted in Figure 3. On the left, the page shows a task description, a textual description of the prediction model, a textual description of the explanation technique, an example input image, the model prediction on the input image, a static explanation for the prediction, and a textual description of the explanation. On the right, the interface enables users to converse with XAI experts, seeking clarifications and posing questions about the explanation. For the users in the control group, we replace the textual chat user interface with a 15-minute timer. Once the timer reaches zero, users are allowed to proceed to the post-evaluations. Users from both groups receive the same post-evaluations, which are identical to the pre-evaluations. We discuss the evaluations below.

Experimental Design

There are two independent variables and two categories of dependent variables. The independent variable in the experiments is the explanation method: LIME or Grad-CAM and the method of understanding static explanations: conversation with human experts or reading static explanations. As we devise both subjective and objective evaluations before and after conversations or readings, two categories of dependent variables were collected in the experiment: the model selection accuracy and the self-reported perception scores.

Objective Evaluation – Selection of Classification Models. The evaluation aims to objectively evaluate participants’ understanding of the static explanations. Participants are presented with 5 input images, on which the three neural networks make the same decisions. The only differences between the three networks lie in their explanations. Participants need to choose the one that would be the most accurate on unobserved test data. Hence, to make the correct selection, the participants must understand the explanations. We use the accuracy of selecting the correct model to measure participants’ objective understanding of static explanations.

We recognize that existing explanation techniques are not always faithful to the underlying model (Adebayo et al., 2018; Jacovi & Goldberg, 2020; Kindermans et al., 2019)

and do not always provide actionable information for model selection. As our goal is to test if the users can understand the static explanations *when* they do provide actionable information, rather than evaluating the static explanations themselves, we selected input images where better classification models indeed have more reasonable and intuitive explanations. This approach allows users to easily pick the best classification models if they understand the static explanations well. We deem an explanation more reasonable when it focuses more on discriminative features that are unique to the predicted class and less on spurious features that are irrelevant to the class. In addition, good models should have explanations that rely on multiple types of discriminative features. This is because a model relying on multiple features is robust and makes the correct decision even if some discriminative features are missing or occluded. In the example in Figure 2, Model B is better than Model A or Model C as Model B utilizes both the head and the body of the cat for classification. In addition, unlike Model A, Model B does not focus on the background, which is irrelevant to the predicted class, Siamese Cat.

Subjective Evaluation. We also measure participants’ subjective perception of static explanations, including their comprehension, acceptance, and trust. The subjective evaluation contains a total of 13 questions listed in table 2. All questions utilize a 7-point Likert scale for responses.

- Comprehension (Cheng et al., 2019; Hoffman et al., 2018): Participants’ subjective perceptions of their understanding of explanations. It complements the objective evaluation, providing a holistic perspective on participants’ understanding of static explanations.
- Perceived Usefulness (Davis, 1989; Davis et al., 1989; Diop et al., 2019): The degree to which participants feel that the explanations enhance their experience with deep learning models. Along with *perceived ease of use* and *behavioral intention*, these three aspects measure participants’ acceptance of static explanations. They are derived from the Technology Acceptance Model (TAM) (Davis, 1989; Davis et al., 1989; Diop et al., 2019), a widely applied theory for understanding individual acceptance and usage of information systems. As the explanations are used by end-

users, investigating their acceptance of the explanations is very important.

- Perceived Ease of Use (Davis, 1989; Davis et al., 1989; Diop et al., 2019): Participants’ assessment of the simplicity and clarity of the explanations.
- Behavioral Intention (Davis, 1989; Davis et al., 1989; Diop et al., 2019): The tendency of participants to utilize the explanation information in the future.
- Trust (Bach, Khan, Hallock, Beltrão, & Sousa, 2022; Muir & Moray, 1996): Participants’ confidence in the explanation methods keeping functioning as intended. Trust has been recognized as an important factor in human-AI collaboration as it mediates the human’s reliance on AI models, thus directly affecting the effectiveness of the human-AI team (Doshi-Velez & Kim, 2017; Seaborn, Miyake, Pennefather, & Otake-Matsuura, 2021; Sebo et al., 2020; Silva, Schrum, Hedlund-Botti, Gopalan, & Gombolay, 2023; Vorm & Combs, 2022).

The literature demonstrated that static explanations have inconsistent effects on users’ trust in AI systems. On one hand, several studies have demonstrated that detailed explanations (Glass, McGuinness, & Wolverton, 2008; Ha & Kim, 2023; Silva et al., 2023), contrastive explanations (Larasati, Liddo, & Motta, 2020), and example-based explanations (F. Yang, Huang, Scholtz, & Arendt, 2020) can enhance user trust in systems. On the other hand, studies showed that static explanations do not have strong effects on user trust in AI systems (Cheng et al., 2019; Kunkel, Donkers, Michael, Barbu, & Ziegler, 2019; Wang & Yin, 2021; Y. Zhang et al., 2020).

One main reason for these inconsistent reports is that trust is mediated by the users’ understanding of the static explanations (Kunkel et al., 2019; Wang & Yin, 2021; Y. Zhang et al., 2020), and such understanding is often absent. According to theories of trust (Hoffman et al., 2018; Lim, Dey, & Avrahami, 2009; McKnight, Cummings, & Chervany, 1998), the ability to build a mental model of AI systems is the key for user trust in AI. Unsurprisingly, studies on the effects of static explanations for laypersons show that users with limited knowledge of machine learning

struggle to understand static explanations and the decision-making processes they are supposed to explain. Consequently, these users do not exhibit increased trust in AI systems after receiving static explanations (Wang & Yin, 2021; Y. Zhang et al., 2020).

With this paper, we quantitatively investigate whether customized conversations about static model explanations can enhance user understanding and improve trust. The conversational approach toward explanations has been advocated by previous studies (Feldhus et al., 2022; Glass et al., 2008; Lakkaraju et al., 2022; Pieters, 2011; Schaffer, O'Donovan, Michaelis, Raglin, & Höllerer, 2019) but never experimentally verified. For example, through interviews with decision-makers, Lakkaraju et al. (2022) found that decision-makers strongly prefer conversational explanations that allow them to ask follow-up questions.

Detailed Study Procedure

Before participation, individuals are required to sign an informed consent form that outlines the objectives and procedures of the study. The form also clarifies compensation details and assures both the anonymity and confidentiality of data collected during the study. Upon signing the consent, participants receive an email that guides them to access the study platform.

After logging in, a pop-up prompt provides an overview of the tasks ahead. Participants then complete pre-experiment objective and subjective evaluations of the static explanations. The objective evaluation measures participants' understanding of static explanations by letting them choose, from three classification models, the most accurate on unobserved test data. There are 5 explanation examples in the objective evaluation. The subjective evaluation, with 13 self-reporting questions, probes the perceived comprehension, acceptance, and trust towards the static explanations. Following these evaluations, participants in the experimental group engage in a WoZ discussion about static explanations. During the conversation, one example image is displayed on the screen. The example image is different from those used in the evaluations; however, the explanation

1 methods remain the same. Participants are motivated to understand the explanations
 2 as they need to select the best-performing classification model using explanations only
 3 when doing objective evaluation. Our XAI experts faithfully answer the user’s questions
 4 based on their knowledge, trying to help the user gradually understand the explanation.
 5 For participants in the control group, they read the static explanation for 15 minutes
 6 which is the average conversation time of the experimental group. After the conversation
 7 or 15-minute reading, participants complete the same set of evaluations as before. All
 8 evaluation outcomes and conversation records are documented. Upon study completion,
 9 each participant receives a \$10 reward.

TABLE 3: *RESULTS OF THE EXPERIMENTAL GROUP BEFORE AND AFTER CONVERSATIONS, AND THE CONTROL GROUP BEFORE AND AFTER 15-MINUTE READING. EACH SCORE IS PRESENTED AS MEAN \pm STANDARD DEVIATION AND THE CHANGE δ BEFORE AND AFTER. * $p < 0.001$*

Explanation Methods	Group	Evaluation Timing	Objective Understanding (Decision-Making Accuracy)	Subjective Understanding	Perceived Usefulness	Perceived Ease of Use	Behavioral Intention	Trust
LIME	experimental	before	0.38 ± 0.20	4.03 ± 1.35	5.09 ± 1.07	4.48 ± 0.94	5.25 ± 0.95	4.15 ± 0.88
		after	$0.53^* \pm 0.16$	$5.30^* \pm 0.88$	$5.92^* \pm 0.66$	$5.28^* \pm 0.84$	$5.83^* \pm 0.81$	$4.92^* \pm 0.73$
	control	before	0.37 ± 0.17	4.57 ± 1.43	5.67 ± 0.95	4.87 ± 1.26	5.73 ± 0.69	4.37 ± 0.90
		after	0.40 ± 0.20	4.60 ± 1.16	5.33 ± 0.96	4.48 ± 1.26	5.27 ± 1.08	4.36 ± 1.05
Grad-CAM	experimental	before	0.82 ± 0.21	4.17 ± 0.91	5.49 ± 0.97	4.71 ± 0.95	5.52 ± 0.65	4.40 ± 1.00
		after	$0.92^* \pm 0.11$	$5.43^* \pm 0.97$	$6.12^* \pm 0.60$	$5.58^* \pm 0.82$	$6.08^* \pm 0.79$	$5.19^* \pm 0.80$
	control	before	0.81 ± 0.20	4.07 ± 1.34	5.58 ± 0.59	4.36 ± 1.15	5.45 ± 0.71	4.22 ± 0.96
		after	0.79 ± 0.19	4.40 ± 1.28	5.46 ± 0.69	4.70 ± 1.21	5.33 ± 0.83	4.42 ± 0.87

Results & Discussion

11 Table 3 tabulates the mean and standard deviation (SD) for all the measures. As
 12 explanation methods (LIME vs. Grad-CAM) and group (experimental vs. control) are
 13 between-subjects variables and time (before vs. after) is a within-subject variable, we
 14 conduct a three-way Analysis of Variance (ANOVAs).

Effects of explanations on objective decision accuracy and subjective measures

Results show significant main effects of group ($F(1, 116) = 5.60, p = .02$), method ($F(1, 116) = 218, p < .001$) and time ($F(1, 116) = 12.51, p < .001$). The experimental group, the Grad-CAM method, and the after-conversation condition display a higher objective decision accuracy. We also find a significant interaction effect between group and time ($F(1, 116) = 11.3, p = .01$), as displayed in the figure 4. In the participant's initial decision, there were no significant differences between the experimental and control conditions. During participants' final decision, those who interact with the XAI expert (i.e., experimental condition) have better decision accuracy. This phenomenon highlights the effectiveness of conversational explanations in enhancing the objective understanding of static explanations of users.

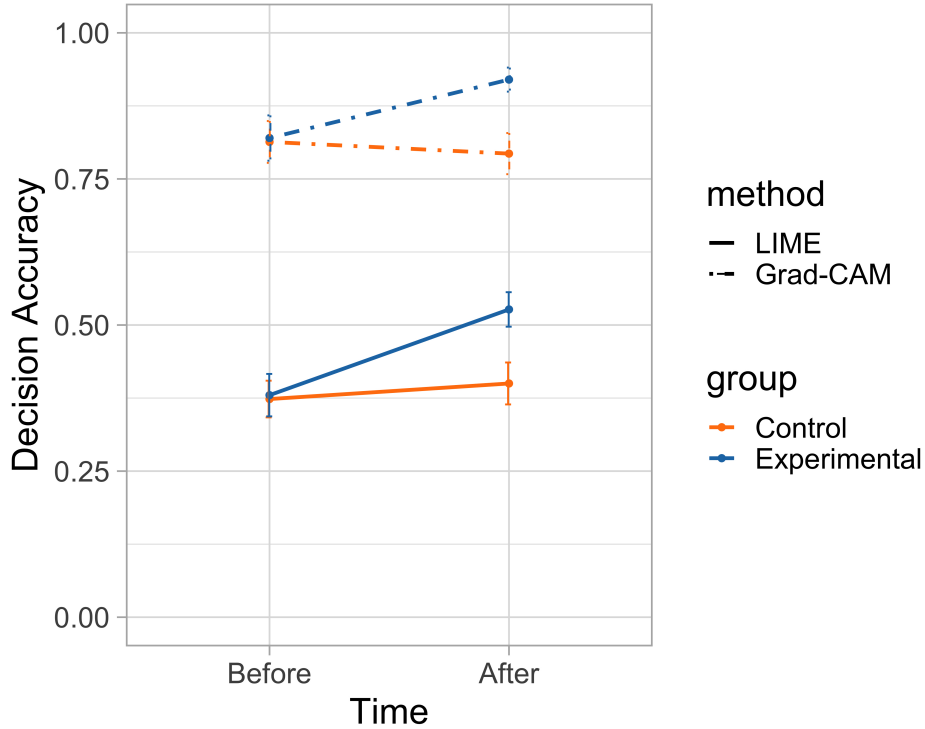


Figure 4. Objective decision accuracy for different groups before and after conditions.

We observe varied objective performance between LIME and Grad-CAM ($F(1, 116) = 218, p < .001$). Grad-CAM has a higher accuracy of objective decision accuracy compared to LIME. A potential reason might be the inherently intuitive nature of the explanations produced by Grad-CAM compared to LIME.

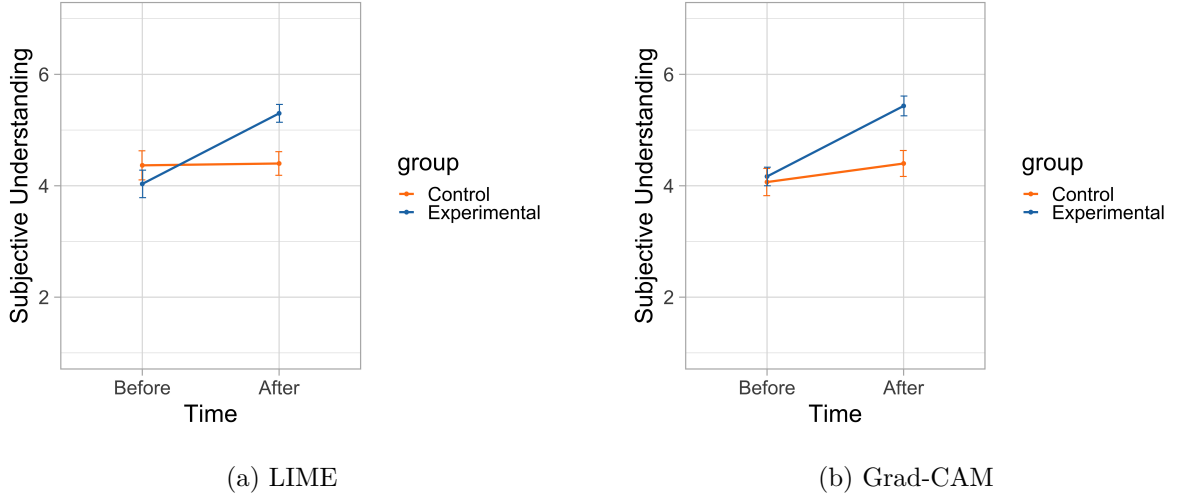


Figure 5. Subjective understanding score for (a) LIME and (b) Grad-CAM before and after conditions.

In terms of participants’ subjective understanding, we find a significant main effect of the evaluation timing ($F(1, 116) = 4.08, p < .001$). Participants receiving conversational explanations have a significantly larger improvement in subjective understanding. We also observe a significant interaction effect between group and time ($F(1, 116) = 37.3, p < .001$), shown in figure 5. Initially, there is no significant difference in the participants’ self-reported understanding of static explanations between the experimental and control groups. After the conditions, participants in the experimental group demonstrate a higher self-report understanding compared to those in the control group.

The main effect of the explanation method ($F(1, 116) = .72, p = .40$) is not significant for participants’ subjective understanding, contrasting with its effect on objective understanding. Even though participants can intuitively choose the best classification model based on the heatmap in the objective evaluation, participants’ initial self-reporting understanding score of Grad-CAM is just slightly larger than 4 (average understanding). This shows that participants still feel confused about how Grad-CAM works and how it explains models’ predictions, even though they can perform well in the objective evaluation. This also demonstrates that subjective and objective evaluations measure participants’ understanding of static explanations from complementary aspects. Self-reporting scores can be influenced by personal biases, while the objective evaluation might not capture

users' feelings about understanding. Combining both methods can provide a comprehensive view of participants' understanding of static explanations.

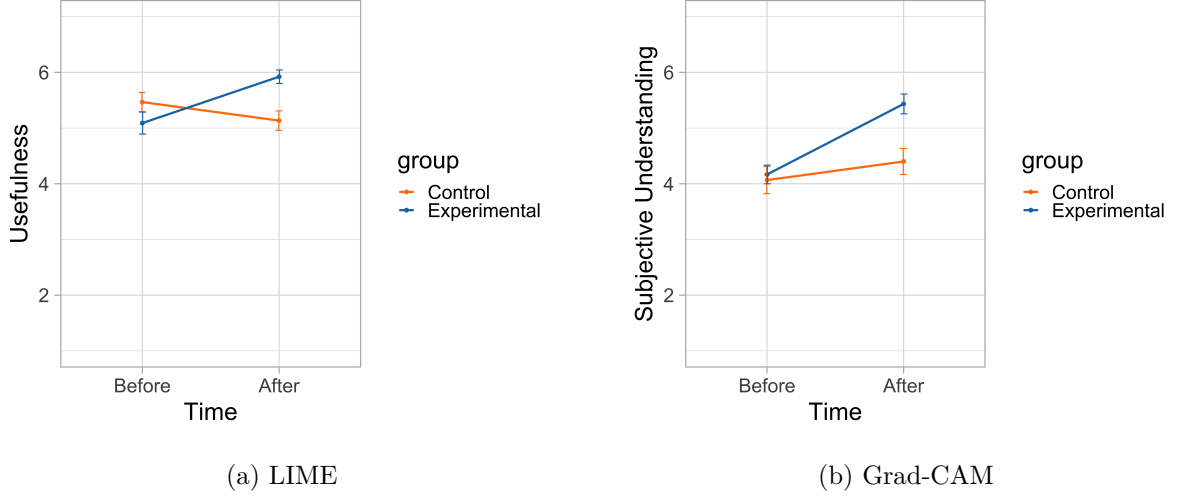


Figure 6. Participants' self-report usefulness score for (a) LIME and (b) Grad-CAM before and after conditions.

For the perceived usefulness, results show a significant main effect of time ($F(1, 116) = 14.6, p < .001$), as well as a significant interaction effect between group and time ($F(1, 116) = 52.9, p < .001$), as depicted in figure 6. The experiment group (i.e., receiving conversational explanation) results in a larger increment of perceived usefulness. For the control group, the Grad-CAM method increases perceived ease of use when participants are given more time to view the static explanation. However, a reversed trend is observed for the LIME method in the control group – the perceived ease of use drops after additional time is provided.

Similar results are observed for participants' perceived ease of use. There are significant main effects of group ($F(1, 116) = 5.19, p = .002$) and of time ($F(1, 116) = 30.3, p < .001$), as well as a significant interaction effect between group and time ($F(1, 116) = 33.7, p < .001$). The perceived ease of use increases largely for the experiment group after interacting with XAI experts. For the control group, the Grad-CAM method increases perceived ease of use while LIME methods decrease it when giving participants more time to view the static explanation.

For the behavioral intention, results show a significant main effect of the time

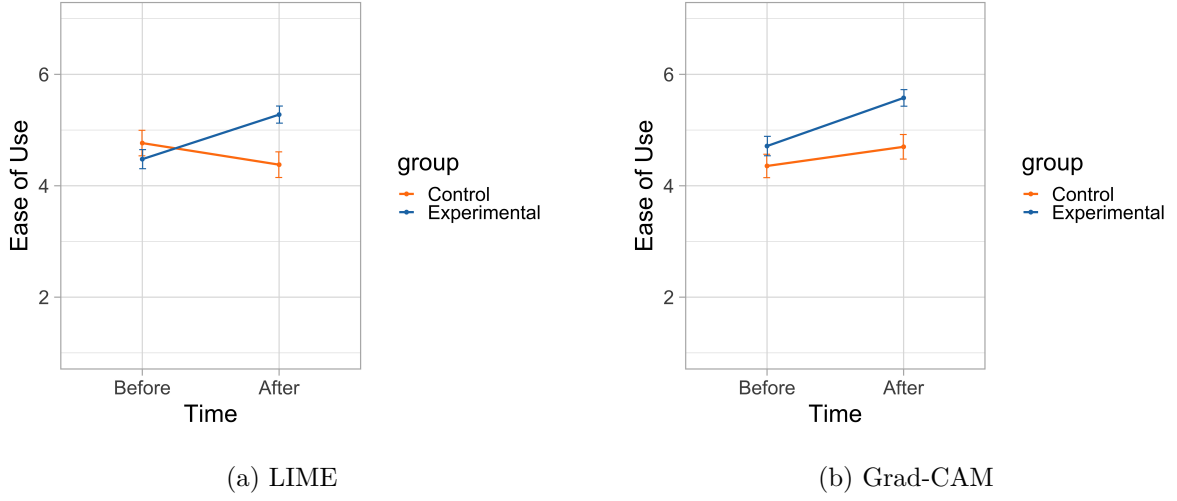


Figure 7. Participants' self-report ease of use score for (a) LIME and (b) Grad-CAM before and after conditions.

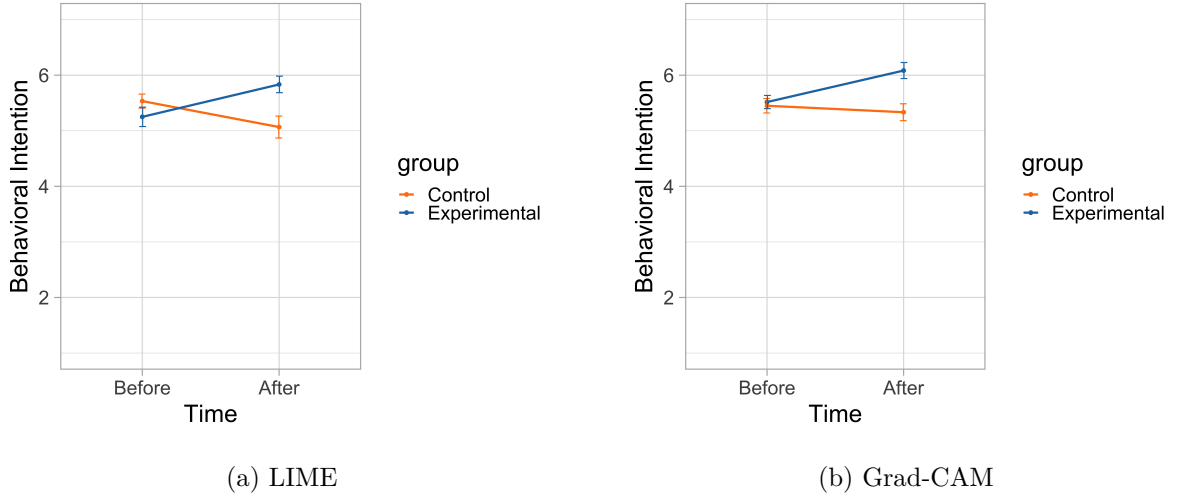


Figure 8. Participants' self-report behavioral intention score for (a) LIME and (b) Grad-CAM before and after conditions.

1 ($F(1, 116) = 3.92, p = .005$) and a significant interaction effect between group and time
 2 ($F(1, 116) = 3.92, p < .001$) as shown in figure 8. Participants increase their behavioral
 3 intention and are more inclined to use explanations in future scenarios after receiving
 4 conversational explanations. On the contrary, the behavioral intention of the control
 5 group decrease for both Grad-CAM and LIME.

6 The boost in usefulness, ease of use, and behavioral intention for the experimental

group can be attributed to the increased understanding of static explanations. Prior to the expert interactions, participants might have had limited knowledge or even misconceptions about the explanation methods. Experiment results show that participants gain a clearer understanding of how the XAI methods function, after the participants' questions are addressed in the conversations. Consequently, they report perceiving the static explanations as more useful and easier to use, and report higher inclination to use the static explanations in future tasks.

The perceived usefulness, ease of use, and behavioral intention of the control group all decrease after reading static explanations for a longer time. This trend suggests a decreased willingness to utilize explanations in future scenarios. The reluctance may be attributed to the frustration the control group faced in attempting to comprehend the static explanations on their own. Research by Carolin Ebermann and Weibelzahl (2023) on the impact of cognitive fit and misfit in the acceptance of AI system usage highlights this phenomenon. They found that users experiencing a cognitive misfit with the AI system often report negative moods, which in turn, reduce their perceived usefulness, ease of use, and behavioral intention of the AI systems. The contrary results of the control group and the experimental group also underscore the importance and effectiveness of conversations in enhancing user behavioral intentions of static explanations.

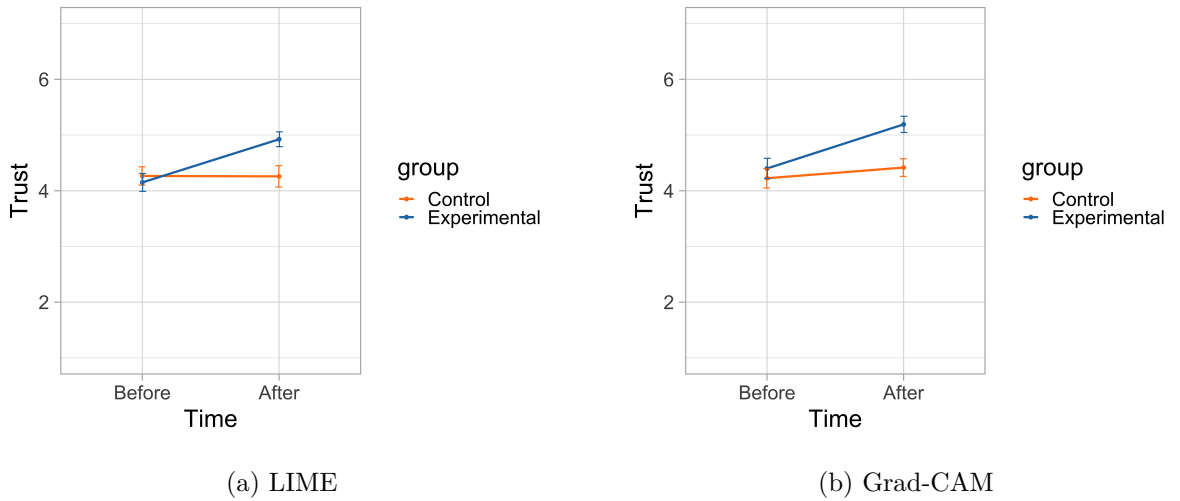


Figure 9. Participants' trust for (a) LIME and (b) Grad-CAM before and after conditions.

For the trust, results show significant main effects of group ($F(1, 116) = 4.31$, $p = .04$) and time ($F(1, 116) = 70.0$, $p < .001$). The experimental group and the after condition display a higher trust score of participants. We also find a significant interaction effect between group and time ($F(1, 116) = 43.7$, $p < .001$), as displayed in the figure 9. Initially, there were no significant differences in trust scores between the experimental and control conditions. During participants' final decision, those who interact with the XAI expert (i.e., experimental condition) report a higher trust score. The enhancements of the experimental group, contrasted with the unchanged trust score of the control group indicate that informativeness and clarity through conversations can help static explanations gain more trust from users. While there exist numerous studies on how explanations of AI predictions can influence users' trust in AI predictions (Cheng et al., 2019; S. S. Y. Kim et al., 2023; Kunkel et al., 2019; Ma et al., 2023; Yu, Berkovsky, Taib, Zhou, & Chen, 2019; Y. Zhang et al., 2020), to our knowledge, this is the first experiment designed explicitly to gauge the impact of conversations on enhancing participants' trust in explanations.

Analysis of Collected Conversations

We collect 60 free-form conversations between XAI experts and participants from 4 different discipline groups. On average, each conversation had 27.4 turns, with each turn comprising approximately 14.4 tokens. By analyzing the users' questions, we divide them into six categories:

- Basic concepts in machine learning: Questions about basic terms and concepts in machine learning that lay people may not know, e.g., what is a deep learning model, what is accuracy, the model structure, and the training data, etc.
- Application and performance of machine learning models: Questions about the ability, accuracy, and limitations of machine learning.
- Diagram reading: Questions about the explanation diagram generated by Grad-CAM or LIME, e.g., what different colors represent in the heatmap.

- 1 • Basic concepts in explainable AI: Questions about basic concepts of explanation
2 methods, e.g., what are explanation methods?
- 3 • Mechanism of explanation methods: Questions about how explanation methods
4 work and how the provided explanation is generated.
- 5 • Other explanations: Questions that require the generation of other types of expla-
6 nations on the current predictions, explanations for different predictions, or com-
7 parisons between the provided explanation and other explanation methods.

8 Based on this categorization, we build a repository for questions that could occur
9 in the conversations. In total, we collected 397 questions from the four different cat-
10 egories. Table 4 contains examples and the number of questions in each category. As
11 observed in Table 4, the questions of participants mainly revolve around basic concepts in
12 machine learning, the fundamentals of explanation methods, and their underlying mech-
13 anisms. This trend might be attributed to the multi-disciplinarity of the participants.
14 It suggests that many participants may not be familiar with machine learning models
15 and explanation methods, which is aligned with the real application of explanation meth-
16 ods. Therefore, it’s crucial to tailor responses to these questions to help users better
17 understand explanations. Furthermore, we note a marked interest in new explanations.
18 This could indicate that as users become more familiar with provided explanation ex-
19 amples, they exhibit curiosity about alternative explanation methods and how models
20 might behave under specific scenarios. Concurrently, the diagram reading category con-
21 tains only 16 questions, implying that explanations generated by Grad-CAM and LIME
22 were relatively straightforward and easy to understand. The diverse range of questions
23 sourced from our conversations underscores that static, one-off explanations are often
24 insufficient for users to understand them. Engaging in dialogue can provide more dy-
25 namic and tailored explanations to users, hence deepening their understanding of static
26 explanations.

27 Having well internalized their knowledge, experts are often unable to estimate what
28 laypeople know (Wittwer et al., 2008). This phenomenon is also referred to as the “curse of

knowledge” (Camerer, Loewenstein, & Weber, 1989). As a result, experts tend to overlook potential areas of confusion or make unwarranted assumptions about what is “common knowledge”. While analyzing the collected conversations, we often find ourselves unable to anticipate the user questions, which corroborates the literature. We describe a few examples below.

Several participants misunderstood the idea of the heatmap produced by Grad-CAM as depicting literal heat dissipating from objects. They infer that the model uses the temperature of objects to perform classification. In reality, a heatmap is just a metaphor that visualizes numerical values distributed spatially, which refers to the feature importance in our case. This misconception leads to questions about how the heat of objects is measured and why non-living objects are warmer than their environment. Some example utterances from participants include: *“So the Grad-cam method basically just refers to the usage of generating a heatmap to capture living matters correct? ... based on the parts of the image that generate more heat?”* –P36, *“basically using heat to predict what is the input right?...how will we know what is the animal or input simply based on heat?”* – P47, *“if these are pictures, how do they figure out the heat since the animal isn’t generating heat”* – P49, *“So a heat sensor is not required? A heatmap is automatically generated from each photo and analyzed using the model.”* – P52.

A second common misconception is the conflation between the post-hoc explanation technique and the classification models. Some example user questions include: *“is the explanation method what the model uses to classify & predict what the image is supposed to be?”* – P6, *“Swin transformer uses LIME model? ... what are the differences between lime model and Swin transformer?”* – P8. Furthermore, participants face challenges in understanding certain terms commonly used in AI and XAI, even though these terms are frequently used and understood within academic communities. Many participants asked questions about basic concepts in machine learning, such as: *“what is the explanation method?”* – P7, *“how do you classify the image?”* – P17, *“what is the algorithm? does it mean lime? what are deep neural networks?”* – P32, *“How would you explain the term “perturbations of images” to a five-year-old?”* – P46.

1 The observations from the interactions between XAI experts and layperson users
2 demonstrate the importance of conversations for users to understand static explanations
3 as they bridge the knowledge gap between the two groups. Conversations can reveal the
4 specific areas of misunderstanding, such as incorrect implicit assumptions the users make
5 and knowledge they lack. Hence, conversational explanations may help the AI system
6 communicate with and bring genuine understanding to the users.

TABLE 4: *OVERVIEW OF COLLECTED QUESTIONS. INCLUDING CATEGORIES OF QUESTIONS, EXAMPLES, AND THE COUNT OF QUESTIONS IN EACH CATEGORY.*

Question Category	Question Examples	Num
Basic concepts in machine learning	<ul style="list-style-type: none"> • What is a deep learning model? • What is the image classification task? • How does the model know what features to extract? 	85
Application, performance, and limitations of machine learning models	<ul style="list-style-type: none"> • How about the precision of the classification model? • Where has this Swin Transformer classification method been used in practical applications? • Will the different species of an animal affect the classification model categorizing the animal? 	68
Diagram reading	<ul style="list-style-type: none"> • Are regions colored in red areas that have been identified as containing key features for the animal? • What are the yellow line spots for (in LIME explanations)? • What do the red and blue colors mean (in Grad-CAM explanations)? 	16
Basic concepts of explanation methods	<ul style="list-style-type: none"> • What is the explanation model used for? • Can LIME be used without the internet? • What are some limitations of the Grad-CAM (LIME) method? 	95
Mechanism of explanation methods	<ul style="list-style-type: none"> • Why does the (LIME) explanation not highlight all the parts of the leopard? • How LIME model recognize the most important parts for the model prediction? • Seems like the Classification Model and the Explanation Model are trained separately - how can we be sure that the underlying logic of making a prediction is the same for both models? 	91
Other explanations	<ul style="list-style-type: none"> • Can you list other visualization methods? • Is there anything special about the Grad-CAM (or LIME) method that is different from others? • What if there are both fishes and humans in an image? How should this image be classified, and can you provide such explanations? 	42

1 Implications for building dialogue systems to explain static explanations

2 Our study indicates the impact of conversational explanations on user comprehension, acceptance, and trust of static explanations. Static explanations, while informative, 3 may not cater to users with varied backgrounds and expertise. Engaging in conversational 4 explanations provides a dynamic and interactive medium for users to seek clarifications, 5 ask questions, and thereby facilitate a deeper and more personalized understanding. 6

7 The emergence of advanced conversational agents (Ni et al., 2023; Shuster et al., 8 2022; T. Zhang et al., 2022), especially knowledge-based question-answering (Lan et al., 9 2021; M. Luo et al., 2023; L. Zhang et al., 2023) powered by large language models 10 (Ouyang et al., 2022; Touvron et al., 2023; Zhao et al., 2023) paves the way toward 11 conversational agents that can explain model decisions and discuss static explanations. 12 Our study suggests the following desiderata for such agents.

- 13 • *Extensive knowledge of AI and XAI.* As observed in our study, a large portion of 14 user questions are related to core concepts of machine learning models and explanation 15 methods. To answer those questions, conversational agents need to be 16 trained on a comprehensive corpus encompassing AI and XAI concepts. Besides, 17 in our study, participants also are curious about the applications, performances, 18 and limitations of machine learning models and explanation methods. Therefore, 19 besides answering abstract questions, dialogue systems also should relate them to 20 real-world applications and limitations.
- 21 • *Capability to generate new explanations as needed.* As an improved understanding 22 of the provided explanations, participants in our study exhibit curiosity about 23 alternative explanation methods and explaining different predictions. Dialogue systems 24 should provide new explanations to users when requested. For instance, if a 25 user is curious about how changing a feature would affect the model output, the 26 system should generate a new explanation with the new feature, which showcases 27 the effect.
- 28 • *Capability to interpret scientific diagrams and visualizations.* A significant portion

of AI and XAI explanations often comes in the form of diagrams (Ribeiro et al., 2016; Selvaraju et al., 2017), such as heatmaps or feature importance visualizations. Our study reveals that users have questions related to understanding these diagrams. Answering these questions usually requires an understanding of specific regions of the diagrams, such as answering what parts of the object are highlighted by the yellow line in LIME explanations. Therefore, future dialogue systems should have visual processing capabilities, understanding and interpreting diagrams contextually. For instance, they should be able to recognize colors, patterns, and other graphical elements in heatmaps or charts and relate them to users' questions. The recent development in multimodal large language models (Driess et al., 2023; Gong et al., 2023; Zhu, Chen, Shen, Li, & Elhoseiny, 2023) is a promising direction to achieve this goal.

Limitations

Despite the insights gained, there are several limitations that should be acknowledged. First, the static explanations used in our study are limited. Our experiments focused on feature attribution explanation methods. The applicability of our findings to other explanation methods, such as example-based explanation methods, remains an open question. Second, as our main objective was to discern the effects of free-form conversational explanations, we did not delve into the comparative performance of different explanation methods. In our experiments, we intentionally selected explanation examples where the best classification model yielded the most reasonable explanations. The explanation examples discussed by participants and XAI experts were chosen such that they reasonably explain the predictions of the classification model. Future work would be to extend these conversations to include explanations that might be less reliable. Third, we explore how conversations foster user trust in explanations in our study. Nevertheless, previous studies (Ha & Kim, 2023; Wang & Yin, 2021; Y. Zhang et al., 2020) have shown that humans may trust AI models even if they make wrong decisions. We do not explore whether users' trust in our study is misplaced, which we leave for future work. Fourth, we use AI to classify the images. Previous studies (Bankins, Formosa, Griep, & Richards,

2022; Formosa, Rogers, Griep, Bankins, & Richards, 2022) found that participants favor humans over AI decision-makers when their decisions directly affect participant welfare. In our study, AI decisions do not directly affect participant welfare. We also did not investigate if the participants preferred conversations with humans or AI chatbots or if their trust in the explanations was affected by that variable. Finally, our research is confined to one geographical region and includes only students and staff from the university. Factors such as cultural backgrounds and age-related differences could potentially influence user interactions with XAI and how they seek to clarify confusion. Future studies could involve recruiting participants from diverse countries, regions, and age groups.

Conclusion

In our work, we conduct Wizard-of-Oz experiments to investigate how free-form conversations assist users in understanding static explanations, promoting trust, and making informed decisions about AI models. Participants engage in conversational explanations with XAI experts to understand how the provided static explanation explains the model decision. To evaluate the effects of conversations, we design objective and subjective measurements. We observe a notable improvement in users' comprehension, acceptance, trust, and collaboration after conversations. From collected conversations, we find that participants' questions and confusions are diverse and unanticipated. Our findings advocate for the integration of dialogue systems in future XAI designs to ensure more personalized explanations.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 chi conference on human factors in computing systems* (p. 1–18).
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 52138–52160.
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in neural information processing systems* (p. 9525–9536).
- Adebayo, J., Muelly, M., Liccardi, I., & Kim, B. (2020). Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*.
- Alkan, O., Wei, D., Mattetti, M., Nair, R., Daly, E., & Saha, D. (2022). FROTE: Feedback rule-driven oversampling for editing models. *Proceedings of Machine Learning and Systems*, 4, 276–301.
- Alvarez-Melis, D., & Jaakkola, T. (2017). A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 412–421). doi: 10.18653/v1/D17-1042
- Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4), 105–120. doi: 10.1609/aimag.v35i4.2513
- Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *Springer-Verlag*, 63, 512.
- Ashktorab, Z., Liao, Q. V., Dugan, C., Johnson, J., Pan, Q., Zhang, W., . . . Campbell, M. (2020). Human-AI collaboration in a cooperative game setting: Measuring social perception and outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 4.
- Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2022). A system-

atic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human-Computer Interaction*, 0(0), 1-16. doi: 10.1080/10447318.2022.2138826

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bankins, S., Formosa, P., Griep, Y., & Richards, D. (2022). AI decision making with dignity? contrasting workers' justice perceptions of human and AI decision making in a human resource management context. *Information Systems Frontiers*, 24(3), 857–875.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., ... Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–16).

Bhat, S., Lyons, J. B., Shi, C., & Yang, X. J. (2024). Evaluating the impact of personalized value alignment in human-robot interaction: Insights into trust and team performance outcomes. In *Proceedings of the 2024 acm/ieee international conference on human-robot interaction* (pp. 32–41). doi: 10.1145/3610977.3634921

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5(4), 2403–2424.

Biswas, A., & Parikh, D. (2013). Simultaneous active learning of classifiers & attributes via relative feedback. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 644–651).

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., & Rinzivillo, S. (2023, Sep 01). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5). doi: 10.1007/s10618-023-00933-9

Cai, C. J., Winter, S., Steiner, D., Wilcox, L., & Terry, M. (2019). "Hello AI": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. In *Proceedings of the acm on human-computer interaction* (Vol. 3). doi: 10.1145/3359206

- 1 Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic
2 settings: An experimental analysis. *Journal of Political Economy*, 97(5), 1232–
3 1254.
- 4 Carissoli, C., Negri, L., Bassi, M., Storm, F. A., & Fave, A. D. (2023). Mental workload
5 and human-robot interaction in collaborative tasks: A scoping review. *International*
6 *Journal of Human–Computer Interaction*, 0(0), 1-20. doi: 10.1080/10447318.2023
7 .2254639
- 8 Carolin Ebermann, M. S., & Weibelzahl, S. (2023). Explainable AI: The effect
9 of contradictory decisions and explanations on users’ acceptance of AI systems.
10 *International Journal of Human–Computer Interaction*, 39(9), 1807-1826. doi:
11 10.1080/10447318.2022.2126812
- 12 Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible
13 models for healthcare: Predicting pneumonia risk and hospital 30-day readmission.
14 In *Proceedings of the 21th acm sigkdd international conference on knowledge dis-*
15 *covery and data mining* (p. 1721–1730). doi: 10.1145/2783258.2788613
- 16 Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., & Rudin, C. (2019). This looks like
17 that: Deep learning for interpretable image recognition. In *Proceedings of the 33rd*
18 *international conference on neural information processing systems*.
- 19 Chen, Y., Li, B., Yu, H., Wu, P., & Miao, C. (2021). Hydra: Hypergradient data
20 relevance analysis for interpreting deep neural networks. In *Proceedings of the aaai*
21 *conference on artificial intelligence* (Vol. 35, pp. 7081–7089).
- 22 Cheng, H.-F., Wang, R., Zhang, Z., O’connell, F., Gray, T., Harper, F. M., & Zhu, H.
23 (2019). Explaining decision-making algorithms through UI: Strategies to help non-
24 expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in*
25 *computing systems* (pp. 1–12).
- 26 Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In *Perspectives on*
27 *socially shared cognition* (p. 127-149). doi: 10.1037/10096-006
- 28 Clark, H. H., & Marshall, C. R. (1981). Definite knowledge and mutual knowledge. In
29 *Elements of discourse understanding* (pp. 10–63).

- 1 Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization
2 techniques to open black box data mining models. *Information Sciences*, 225, 1–
3 17.
- 4 Croce, D., Rossini, D., & Basili, R. (2019). Auditing deep learning processes through
5 kernel-based explanatory models. In *Proceedings of the 2019 conference on empirical*
6 *methods in natural language processing and the 9th international joint conference*
7 *on natural language processing (emnlp-ijcnlp)* (pp. 4037–4046).
- 8 Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A
9 survey of the state of explainable AI for natural language processing. In *Proceedings*
10 *of the 1st conference of the asia-pacific chapter of the association for computational*
11 *linguistics and the 10th international joint conference on natural language process-*
12 *ing* (pp. 447–459).
- 13 Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of
14 information technology. *MIS Quarterly*, 13(3), 319–340.
- 15 Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer
16 technology: A comparison of two theoretical models. *Management science*, 35(8),
17 982–1003.
- 18 Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A
19 large-scale hierarchical image database. In *Proceedings of the 2009 ieee conference*
20 *on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009
21 .5206848
- 22 Diop, E. B., Zhao, S., & Duy, T. V. (2019). An extension of the technology acceptance
23 model for understanding travelers’ adoption of variable message signs. *PLoS one*,
24 14(4).
- 25 Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine
26 learning. *arXiv preprint arXiv:1702.08608*.
- 27 Doshi-Velez, F., Wallace, B. C., & Adams, R. (2015). Graph-Sparse LDA: A topic model
28 with structured sparsity. In *Twenty-ninth aaai conference on artificial intelligence*.
- 29 Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., ... Florence,

P. (2023). PALM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

D’Avella, S., Camacho-Gonzalez, G., & Tripicchio, P. (2022). On multi-agent cognitive cooperation: Can virtual agents behave like humans? *Neurocomputing*, 480(C), 27–38. doi: 10.1016/j.neucom.2022.01.025

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I., Muller, M., ... others (2021). The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509*.

Fails, J. A., & Olsen Jr, D. R. (2003). Interactive machine learning. In *Proceedings of the 8th international conference on intelligent user interfaces* (pp. 39–45).

Feldhus, N., Ravichandran, A. M., & Möller, S. (2022). Mediators: Conversational agents explaining NLP model behavior. *arXiv preprint arXiv:2206.06029*.

Feng, S., & Boyd-Graber, J. (2019). What can AI do for me? Evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th international conference on intelligent user interfaces* (p. 229–239). doi: 10.1145/3301275.3302265

Flathmann, C., Schelble, B. G., McNeese, N. J., Knijnenburg, B., Gramopadhye, A. K., & Madathil, K. C. (2023). The purposeful presentation of AI teammates: Impacts on human acceptance and perception. *International Journal of Human–Computer Interaction*, 0(0), 1–18. doi: 10.1080/10447318.2023.2254984

Formosa, P., Rogers, W., Griep, Y., Bankins, S., & Richards, D. (2022). Medical AI and human dignity: Contrasting perceptions of human and artificially intelligent (AI) decision making in diagnostic and medical resource allocation contexts. *Computers in Human Behavior*, 133, 107296. doi: <https://doi.org/10.1016/j.chb.2022.107296>

Gero, K. I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., ... Zhang, W. (2020). Mental models of AI agents in a cooperative game setting. In (p. 1–12). doi: 10.1145/3313831.3376316

Glass, A., McGuinness, D. L., & Wolverton, M. (2008). Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on intelligent user interfaces* (p. 227–236). doi: 10.1145/1378773.1378804

- Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., ... Chen, K. (2023). *MultiModal-GPT: A vision and language model for dialogue with humans*.
- González, A. V., Bansal, G., Fan, A., Mehdad, Y., Jia, R., & Iyer, S. (2021). Do explanations help users detect errors in open-domain QA? An evaluation of spoken vs. visual explanations. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 1103–1116). doi: 10.18653/v1/2021.findings-acl.95
- Guesmi, M., Chatti, M. A., Joarder, S., Ain, Q. U., Alatrash, R., Siepmann, C., & Vahidi, T. (2023). Interactive explanation with varying level of details in an explainable scientific literature recommender system. *International Journal of Human-Computer Interaction*, 0(0), 1-22. doi: 10.1080/10447318.2023.2262797
- Guo, Y., Yang, X. J., & Shi, C. (2023). Enabling team of teams: A trust inference and propagation (TIP) model in multi-human multi-robot teams. In *Robotics: Science and Systems XIX*. doi: 10.15607/RSS.2023.XIX.003
- Ha, T., & Kim, S. (2023). Improving trust in AI with mitigating confirmation bias: Effects of explanation type and debiasing strategy for decision-making with explainable AI. *International Journal of Human-Computer Interaction*, 0(0), 1-12. doi: 10.1080/10447318.2023.2285640
- Häuslschmid, R., von Bülow, M., Pfleging, B., & Butz, A. (2017). Supporting trust in autonomous driving. In (p. 319–329). doi: 10.1145/3025171.3025198
- He, X., Hong, Y., Zheng, X., & Zhang, Y. (2023). What are the users' needs? Design of a user-centered explainable artificial intelligence diagnostic system. *International Journal of Human-Computer Interaction*, 39(7), 1519-1542. doi: 10.1080/10447318.2022.2095093
- Herse, S., Vitale, J., & Williams, M.-A. (2023). Using agent features to influence user trust, decision making and task outcome during human-agent collaboration. *International Journal of Human-Computer Interaction*, 39(9), 1740-1761. doi: 10.1080/10447318.2022.2150691
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

- Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems* (p. 1–13). doi: 10.1145/3290605.3300809
- Hu, L., Chen, J., Nair, V. N., & Sudjianto, A. (2018). Locally interpretable models and effects based on supervised partitioning (LIME-SUP). *arXiv preprint arXiv:1806.00663*.
- Idahl, M., Lyu, L., Gadiraju, U., & Anand, A. (2021). Towards benchmarking the utility of explanations for model debugging. *arXiv preprint arXiv:2105.04505*.
- Ignatiev, A., Narodytska, N., & Marques-Silva, J. (2019). Abduction-based explanations for machine learning models. In *Proceedings of the aaai conference on artificial intelligence* (pp. 1511–1519).
- Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4198–4205).
- Jeyakumar, J. V., Noor, J., Cheng, Y.-H., Garcia, L., & Srivastava, M. (2020). How can I explain this to you? An empirical study of deep neural network explanation methods. In *Advances in neural information processing systems* (Vol. 33, pp. 4211–4222).
- Karimi, A.-H., Barthe, G., Balle, B., & Valera, I. (2020). Model-agnostic counterfactual explanations for consequential decisions. In *International conference on artificial intelligence and statistics* (pp. 895–905).
- Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, 2(1), 26–41. doi: 10.1145/357417.357420
- Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th international conference on neural information processing systems* (p. 2288–2296).
- Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A.

(2023). "help me help the ai": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 chi conference on human factors in computing systems*. doi: 10.1145/3544548.3581001

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., . . . Kim, B. (2019). The (un)reliability of saliency methods. In *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 267–280).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (Vol. 25).

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces* (p. 126–137). doi: 10.1145/2678025.2701399

Kunkel, J., Donkers, T., Michael, L., Barbu, C.-M., & Ziegler, J. (2019). Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 chi conference on human factors in computing systems* (p. 1–12). doi: 10.1145/3290605.3300717

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency* (p. 29–38). doi: 10.1145/3287560.3287590

Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 1675–1684).

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking explainability as a dialogue: A practitioner’s perspective. *arXiv preprint arXiv:2202.01875*.

Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W. X., & Wen, J.-R. (2021). A survey on complex knowledge base question answering: Methods, challenges and solutions. In *Proceedings of the thirtieth international joint conference on artificial intelligence*,

IJCAI-21 (pp. 4483–4491). International Joint Conferences on Artificial Intelligence Organization. doi: 10.24963/ijcai.2021/611

Larasati, R., Liddo, A. D., & Motta, E. (2020). The effect of explanation styles on user’s trust. In *2020 workshop on explainable smart systems for algorithmic transparency in emerging technologies*.

Lertvittayakumjorn, P., Specia, L., & Toni, F. (2020). FIND: Human-in-the-Loop Debugging Deep Text Classifiers. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 332–348). doi: 10.18653/v1/2020.emnlp-main.24

Liang, W., Zou, J., & Yu, Z. (2020). ALICE: Active learning with contrastive natural language explanations. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 4380–4391). doi: 10.18653/v1/2020.emnlp-main.355

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 chi conference on human factors in computing systems* (p. 1–15).

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2119–2128).

Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1–45.

Liu, L., Guo, F., Zou, Z., & Duffy, V. G. (2022). Application, development and future opportunities of collaborative robots (cobots) in manufacturing: A literature review. *International Journal of Human-Computer Interaction*, 0(0), 1-18. doi: 10.1080/10447318.2022.2041907

Liu, N., Huang, X., Li, J., & Hu, X. (2018). On interpretation of network embedding via

- taxonomy induction. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1812–1820).
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 10012–10022).
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, 10(10), 464–470.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Luo, M., Fang, Z., Gokhale, T., Yang, Y., & Baral, C. (2023). End-to-end knowledge retrieval with multi-modal queries. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8573–8589). doi: 10.18653/v1/2023.acl-long.478
- Luo, R., Du, N., & Yang, X. J. (2022). Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time. *International Journal of Human–Computer Interaction*, 38(18-20), 1962–1971. doi: 10.1080/10447318.2022.2097602
- Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. In *Proceedings of the 2023 chi conference on human factors in computing systems*. doi: 10.1145/3544548.3581058
- McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management review*, 23(3), 473–490.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1–38.
- Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers

through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 607–617).

Muir, B. M., & Moray, N. (1996). Trust in automation. part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460.

Nguyen, G., Kim, D., & Nguyen, A. (2021). The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, *34*, 26422–26436.

Nguyen, G., Taesiri, M. R., & Nguyen, A. (2022). Visual correspondence-based explanations improve AI robustness and human-ai team accuracy. *Neural Information Processing Systems (NeurIPS)*.

Ni, J., Young, T., Pandelea, V., Xue, F., & Cambria, E. (2023). Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, *56*(4), 3055–3155.

Numata, T., Sato, H., Asa, Y., Koike, T., Miyata, K., Nakagawa, E., ... Sadato, N. (2020). Achieving affective human–virtual agent communication by enabling virtual agents to imitate positive expressions. *Scientific reports*, *10*(1), 5977.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pieters, W. (2011). Explanation and trust: what to tell the user in security and ai? *Ethics and information technology*, *13*, 53–64.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 chi conference on human factors in computing systems*. doi: 10.1145/3411764.3445315

Powles, J., & Hodson, H. (2017). Google deepmind and healthcare in an age of algorithms. *Health and Technology*, *7*(4), 351–367. doi: 10.1007/s12553-017-0179-1

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE:

Feasible and actionable counterfactual explanations. In *Proceedings of the aaai/acm conference on ai, ethics, and society* (pp. 344–350).

Quinn, T. P., Senadeera, M., Jacobs, S., Coghlan, S., & Le, V. (2021). Trust and medical AI: The challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association*, 28(4), 890–894.

Ribeiro, M. T., Singh, S., & 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* (pp. 1135–1144).

Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., . . . others (2020). Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Transactions on Cognitive and Developmental Systems*, 13(3), 717–728.

Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17* (pp. 2662–2670). doi: 10.24963/ijcai.2017/371

Rudziński, F. (2016). A multi-objective genetic optimization of interpretability-oriented fuzzy rule-based classifiers. *Applied Soft Computing*, 38, 118–133.

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th international conference on intelligent user interfaces* (p. 240–251). doi: 10.1145/3301275.3302308

Schmid, U., & Wrede, B. (2022). What is missing in XAI so far? An interdisciplinary perspective. *KI-Künstliche Intelligenz*, 36(3-4), 303–315.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., . . . Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8), 476–486.

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2021). Voice in human–agent interaction: A survey. *ACM Computing Surveys*, 54(4). doi:

10.1145/3386867

2 Sebo, S., Dong, L. L., Chang, N., Lewkowicz, M., Schutzman, M., & Scassellati, B.
3 (2020). The influence of robot verbal support on human team members: Encourag-
4 ing outgroup contributions and suppressing ingroup supportive behavior. *Frontiers*
5 *in Psychology*, 3584.

6 Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017).
7 Grad-CAM: Visual explanations from deep networks via gradient-based localiza-
8 tion. In *Proceedings of the ieee international conference on computer vision* (pp.
9 618–626).

10 Sharma, S., Henderson, J., & Ghosh, J. (2019). CERTIFAI: Counterfactual explanations
11 for robustness, transparency, interpretability, and fairness of artificial intelligence
12 models. *arXiv preprint arXiv:1905.07857*.

13 Shen, H., Huang, C.-Y., Wu, T., & Huang, T.-H. K. (2023). ConvXAI: Delivering hetero-
14 geneous AI explanations via conversations to support human-ai scientific writing.
15 In *Companion publication of the 2023 conference on computer supported cooperative*
16 *work and social computing* (p. 384–387). doi: 10.1145/3584931.3607492

17 Shih, A., Choi, A., & Darwiche, A. (2018). A symbolic approach to explaining bayesian
18 network classifiers. *arXiv preprint arXiv:1805.03364*.

19 Shuster, K., Xu, J., Komeili, M., Ju, D., Smith, E. M., Roller, S., ... others (2022).
20 BlenderBot 3: a deployed conversational agent that continually learns to responsibly
21 engage. *arXiv preprint arXiv:2208.03188*.

22 Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Ex-
23 plainable artificial intelligence: Evaluating the objective and subjective impacts of
24 XAI on human-agent interaction. *International Journal of Human-Computer In-*
25 *teraction*, 39(7), 1390-1404. doi: 10.1080/10447318.2022.2101698

26 Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional net-
27 works: Visualising image classification models and saliency maps. *arXiv preprint*
28 *arXiv:1312.6034*.

29 Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-

scale image recognition. In *Proceedings of 3rd international conference on learning representations*.

Slack, D., Krishna, S., Lakkaraju, H., & Singh, S. (2023). Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence*, 5(8), 873-883. doi: 10.1038/s42256-023-00692-8

Smith-Renner, A., Fan, R., Birchfield, M., Wu, T., Boyd-Graber, J., Weld, D. S., & Findlater, L. (2020). No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–13).

Springer, A., & Whittaker, S. (2019). Progressive disclosure: Empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces* (p. 107–120). doi: 10.1145/3301275.3302322

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319–3328).

Taesiri, M. R., Nguyen, G., & Nguyen, A. (2022). Visual correspondence-based explanations improve AI robustness and human-ai team accuracy. *Advances in Neural Information Processing Systems*, 35, 34287–34301.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., . . . Yuan, A. (2020). The language interpretability tool: Extensible, interactive visualizations and analysis for NLP models. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 107–118). doi: 10.18653/v1/2020.emnlp-demos.15

Teso, S., Bontempelli, A., Giunchiglia, F., & Passerini, A. (2021). Interactive label cleaning with example-based explanations. *Advances in Neural Information Processing Systems*, 34, 12966–12977.

Teso, S., & Kersting, K. (2019). Explanatory interactive machine learning. In *Proceedings of the 2019 aaai/acm conference on ai, ethics, and society* (p. 239–245). doi: 10.1145/3306618.3314293

- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., . . . others (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Tran, K. H., Ghazimatin, A., & Saha Roy, R. (2021). Counterfactual explanations for neural recommenders. In *Proceedings of the 44th international acm sigir conference on research and development in information retrieval* (pp. 1627–1631).
- Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.
- Vorm, E. S., & Combs, D. J. Y. (2022). Integrating transparency, trust, and acceptance: The intelligent systems technology acceptance model (istam). *International Journal of Human–Computer Interaction*, 38(18-20), 1828-1845. doi: 10.1080/10447318.2022.2070107
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31, 841.
- Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th international conference on intelligent user interfaces* (p. 318–328). doi: 10.1145/3397481.3450650
- Wilkesmann, M., & Wilkesmann, U. (2011). Knowledge transfer as interaction between experts and novices supported by technology. *Vine*, 41(2), 96–112.
- Wittwer, J., Nückles, M., & Renkl, A. (2008). Is underestimation less detrimental than overestimation? The impact of experts’ beliefs about a layperson’s knowledge on learning and question asking. *Instructional Science*, 36, 27–52.
- Xu, W., Dainoff, M. J., Ge, L., & Gao, Z. (2023). Transitioning to human interaction with AI systems: New challenges and opportunities for hci professionals to enable human-centered ai. *International Journal of Human–Computer Interaction*, 39(3), 494-518. doi: 10.1080/10447318.2022.2041900
- Yang, F., Du, M., & Hu, X. (2019). Evaluating explanation without ground truth in interpretable machine learning. *arXiv preprint arXiv:1907.06831*.

- 1 Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations
2 foster end users' appropriate trust in machine learning? In *Proceedings of the 25th*
3 *international conference on intelligent user interfaces* (p. 189–201). doi: 10.1145/
4 3377325.3377480
- 5 Yang, H., Rudin, C., & Seltzer, M. (2017). Scalable bayesian rule lists. In *International*
6 *conference on machine learning* (pp. 3921–3930).
- 7 Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user
8 experience and system transparency on trust in automation. In *Proceedings of the*
9 *2017 acm/ieee international conference on human-robot interaction* (pp. 408–416).
10 doi: 10.1145/2909824.3020230
- 11 Yoon, J., Arik, S., & Pfister, T. (2020). Data valuation using reinforcement learning. In
12 *International conference on machine learning* (pp. 10842–10851).
- 13 Yu, K., Berkovsky, S., Taib, R., Zhou, J., & Chen, F. (2019). Do I trust my machine
14 teammate? An investigation from perception to decision. In *Proceedings of the 24th*
15 *international conference on intelligent user interfaces* (p. 460–468). doi: 10.1145/
16 3301275.3302277
- 17 Zhang, L., Zhang, J., Wang, Y., Cao, S., Huang, X., Li, C., . . . Li, J. (2023). FC-KBQA:
18 A fine-to-coarse composition framework for knowledge base question answering. In
19 *Proceedings of the 61st annual meeting of the association for computational linguis-*
20 *tics (volume 1: Long papers)* (pp. 1002–1017). doi: 10.18653/v1/2023.acl-long.57
- 21 Zhang, T., Liu, Y., Li, B., Zeng, Z., Wang, P., You, Y., . . . Cui, L. (2022, December).
22 History-aware hierarchical transformer for multi-session open-domain dialogue sys-
23 tem. In *Findings of the association for computational linguistics: Emnlp 2022* (pp.
24 3395–3407). doi: 10.18653/v1/2022.findings-emnlp.247
- 25 Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation
26 on accuracy and trust calibration in AI-assisted decision making. In *Proceedings*
27 *of the 2020 conference on fairness, accountability, and transparency* (p. 295–305).
28 doi: 10.1145/3351095.3372852
- 29 Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., . . . Wen, J.-R. (2023). A

1 survey of large language models. *arXiv preprint arXiv:2303.18223*.

2 Zheng, Y., Rowell, B., Chen, Q., Kim, J. Y., Kontar, R. A., Yang, X. J., & Lester, C. A.

3 (2023). Designing human-centered AI to prevent medication dispensing errors:

4 Focus group study with pharmacists. *JMIR Formative Research*, 7(1), e51921. doi:

5 10.2196/51921

6 Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing

7 vision-language understanding with advanced large language models. *arXiv preprint*

8 *arXiv:2304.10592*.

Biographies

Zhang Tong is a research assistant and Ph.D. student in the School of Computer Science and Engineering at Nanyang Technological University, Singapore. He received a B.E. degree (2020) in Computer Science and Technology from Shandong University.

X. Jessie Yang is an Assistant Professor in the Department of Industrial and Operations Engineering at the University of Michigan Ann Arbor. She obtained a PhD (2014) and a MEng (2009) in Mechanical and Aerospace Engineering (Human Factors), and a BEng (2006) in Electrical and Electronic Engineering, all from Nanyang Technological University.

Boyang Li is a Nanyang Associate Professor at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He was a Senior Research Scientist at Baidu Research USA, and a Research Scientist and Group Leader at Disney Research. He received his Ph.D. degree (2015) from Georgia Institute of Technology.

Appendix

Objective Evaluation – Selection of classification models.

1 The evaluation aims to objectively evaluate participants' understanding of static
2 explanations. We ask participants to choose, from three classification models, the most
3 accurate on unobserved test data. All three classification models make the same decisions
4 on 5 images, accompanied by static explanations from the same explanation method. The
5 only differences between the three networks lie in their explanations. Hence, to make the
6 correct selection, the participants must understand the explanations.

7 Figure A1 presents the full set of images listed in the objective evaluation for Grad-
8 CAM, while Figure A2 showcases the same for LIME.

Questionnaire 1

Questionnaire Description

The questionnaire consists of questions that each offer three choices. Each choice contains an input image, the prediction from a deep learning model for that input, and an explanation of how the model arrived at its prediction. The deep learning model is designed to classify images into specific categories, such as Goldfish or Siberian Husky.

It is important to note that while the deep learning models in different choices have differing levels of accuracy, the explanation method remains consistent.

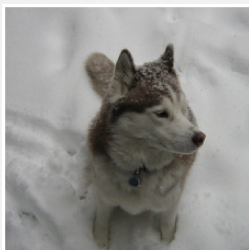
Your responsibility is to assess and compare the explanations provided for different deep learning models and choose the deep learning model that you believe best explains its prediction.

We greatly value your participation, and please rest assured that all responses will be kept anonymous and confidential.

Question 1

☐ Choice A

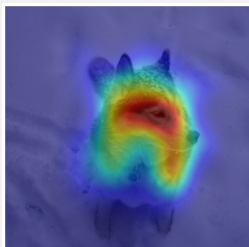
Model's input



Model output

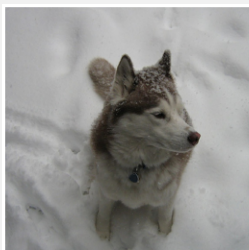
Siberian husky

Explanation for the model prediction



☐ Choice B

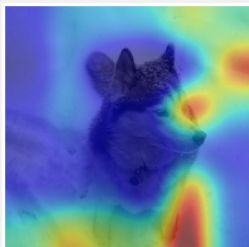
Model's input



Model output

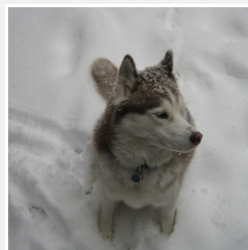
Siberian husky

Explanation for the model prediction



☐ Choice C

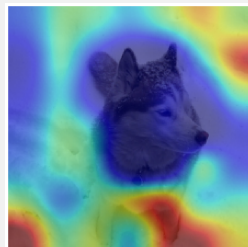
Model's input



Model output

Siberian husky

Explanation for the model prediction



Question 2

☐ Choice A

Model's input



Model output

Goldfish

☐ Choice B

Model's input

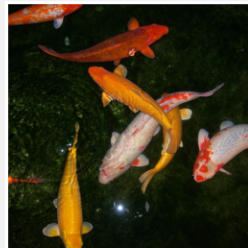


Model output

Goldfish

☐ Choice C

Model's input

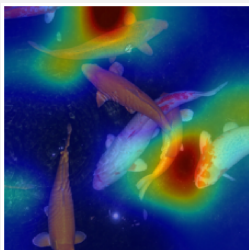


Model output

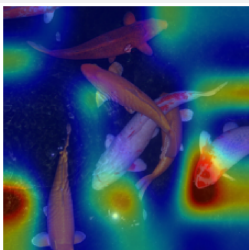
Goldfish

Figure A1. Objective evaluation questions used for Grad-CAM.

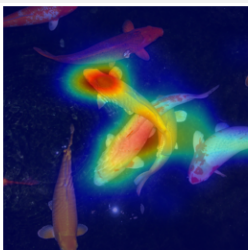
Explanation for the model prediction



Explanation for the model prediction




Explanation for the model prediction



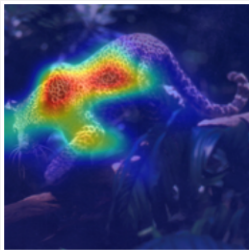
Question 3

☐ Choice A
Model's input




Model output
Leopard

Explanation for the model prediction

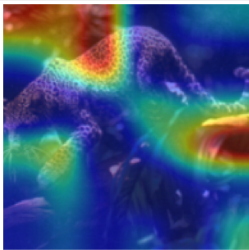


☐ Choice B
Model's input




Model output
Leopard

Explanation for the model prediction

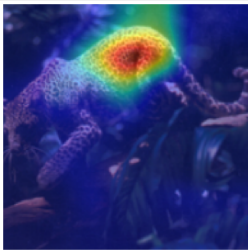


☐ Choice C
Model's input




Model output
Leopard

Explanation for the model prediction




Question 4

☐ Choice A
Model's input




Model output
Bee

Explanation for the model prediction




☐ Choice B
Model's input




Model output
Bee

Explanation for the model prediction



☐ Choice C
Model's input



Model output
Bee

Explanation for the model prediction




Figure A1. Objective evaluation questions used for Grad-CAM.

Question 5

☐ Choice A

Model's input

Model output

Siamese cat

Explanation for the model prediction

☐ Choice B

Model's input

Model output

Siamese cat

Explanation for the model prediction

☐ Choice C

Model's input

Model output

Siamese cat

Explanation for the model prediction

Submit

Cancel

Figure A1. Objective evaluation questions used for Grad-CAM.

Pre Questionnaire 1

Questionnaire Description

The questionnaire consists of questions that each offer three choices. Each choice contains an input image, the prediction from a deep learning model for that input, and an explanation of how the model arrived at its prediction. The deep learning model is designed to classify images into specific categories, such as Goldfish or Siberian Husky.

It is important to note that while the deep learning models in different choices have differing levels of accuracy, the explanation method remains consistent.

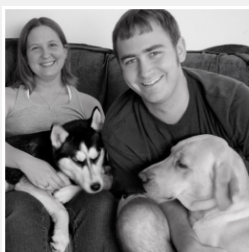
Your responsibility is to assess and compare the explanations provided for different deep learning models and choose the deep learning model that you believe best explains its prediction.

We greatly value your participation, and please rest assured that all responses will be kept anonymous and confidential.

Question 1

☐ Choice A

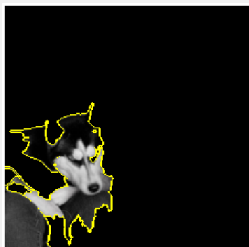
Model's input



Model output

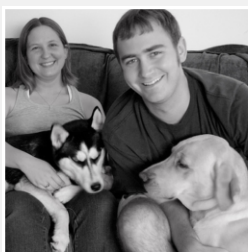
Siberian husky

Explanation for the model prediction



☐ Choice B

Model's input



Model output

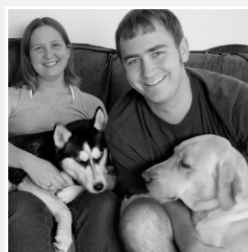
Siberian husky

Explanation for the model prediction



☐ Choice C

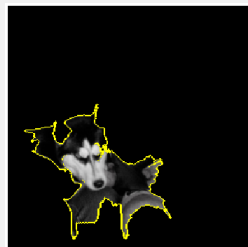
Model's input



Model output

Siberian husky

Explanation for the model prediction



Question 2

☐ Choice A

Model's input



Model output

Goldfish

☐ Choice B

Model's input

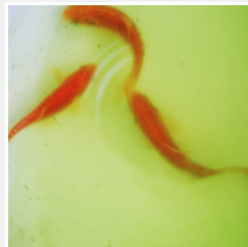


Model output

Goldfish

☐ Choice C

Model's input

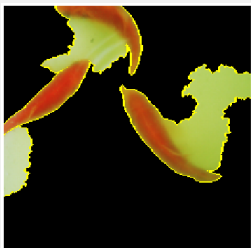


Model output

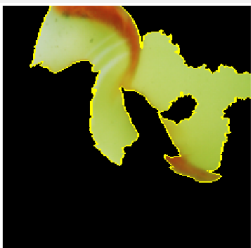
Goldfish

Figure A2. Objective evaluation questions used for LIME.


Explanation for the model prediction



Explanation for the model prediction

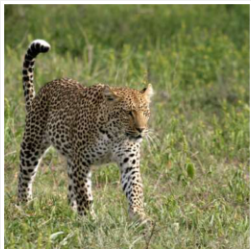


Explanation for the model prediction



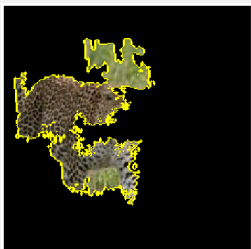
Question 3

☐ Choice A
Model's input

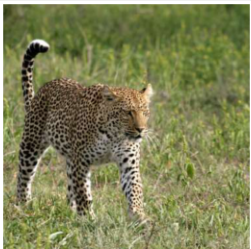


Model output
Leopard

Explanation for the model prediction

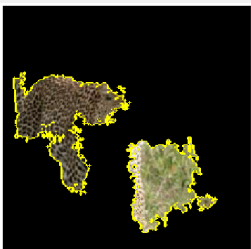


☐ Choice B
Model's input

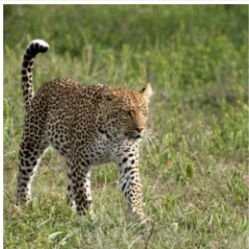


Model output
Leopard

Explanation for the model prediction




☐ Choice C
Model's input




Model output
Leopard

Explanation for the model prediction



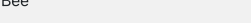
Question 4

☐ Choice A
Model's input




Model output
Bee

Explanation for the model prediction

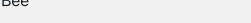


☐ Choice B
Model's input




Model output
Bee

Explanation for the model prediction



☐ Choice C
Model's input



Model output
Bee

Explanation for the model prediction

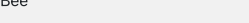
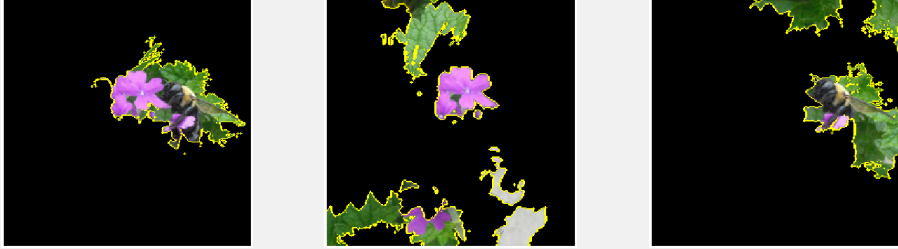

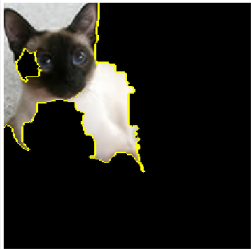



Figure A2. Objective evaluation questions used for LIME.

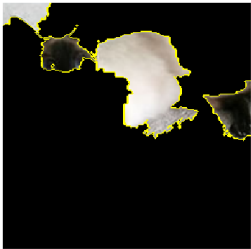



Question 5

☐ Choice A
Model's input


Model output
Siamese cat
Explanation for the model prediction


☐ Choice B
Model's input


Model output
Siamese cat
Explanation for the model prediction


☐ Choice C
Model's input


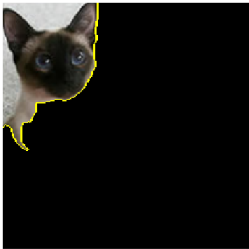
Model output
Siamese cat
Explanation for the model prediction


Figure A2. Objective evaluation questions used for LIME.