THE UNIVERSITY OF CHICAGO

EMPOWERING HUMANS WITH MACHINES: FROM EXPLANATIONS TO TEACHING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF COMPUTER SCIENCE

BY
HAN LIU

CHICAGO, ILLINOIS
AUGUST 2024

To my grandfather, Wenzhang Zhou,

who taught me to strive rather than merely aim for excellence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to extend my heartfelt gratitude to my advisor, Prof. Chenhao Tan, for his guidance and support throughout my Ph.D. studies. His dedication to research and his ability to think critically about problems have been a constant source of inspiration. His question-driven approach to research has encouraged me to think deeply about the issues I work on. Prof. Tan's invaluable feedback has shaped my research ideas and significantly improved my writing skills. His mentorship has been instrumental in my growth as both a researcher and a person. I am grateful for the opportunity to work with him on many exciting research projects.

I would also like to thank the members of my dissertation committee, Prof. Yuxin Chen and Prof. Aritrick Chatterjee. Prof. Chen's insightful comments and suggestions have been invaluable. He encouraged me to explore different research directions and provided valuable feedback on my work. Prof. Chen introduced me to new research areas and helped me broaden my research horizons. He connected me with other cohorts in the department and provided opportunities for collaboration and communication with many other brilliant minds. I am grateful for his support and guidance throughout my Ph.D. studies.

Prof. Chatterjee's feedback on my research has been invaluable, and his guidance has helped me grow as a researcher. His expertise in radiology and medical imaging has been a valuable resource for my research on AI-driven tutorials. His advice on how to design and evaluate AI assistance for radiology tasks has been instrumental in shaping my research ideas. His optimism and enthusiasm for research have inspired me to work harder and think more creatively about the problems I work on.

I would like to thank my collaborators, cohorts in the department, and labmates at the Chicago Human+AI Lab for their support and encouragement: Vivian Lai, Chao-Chun (Joe) Hsu, Yangqiaoyu (Rosa) Zhou, Chacha Chen, Mourad Heddaya, Karen Zhou, Samuel Carton, Shi Feng, Yizhou (Harry) Tian, Yiming Zhang, Jiamin Yang, Shuyuan (Lily) Wang, Dang

Nguyen, Yixuan (Tom) Wang, Haokun Liu, Chenghao Yang, Chaoqi Wang, Yibo Jiang, Jibang Wu, Renyu Zhang, Dr. Aytek Oto, Dr. Maryellen L. Giger, and many others. Their feedback on my research ideas and our discussions on various topics have been invaluable. The days and nights we spent together in Crerar, Searle, and other places on campus have been some of the most memorable moments of my Ph.D. studies. I am grateful for their friendship and support throughout my time at the University of Chicago.

I would like to thank my kind and caring friends both near and far from Hyde Park for their support and encouragement: Yang Yang, Lili Zhou, Kaiyu Yang, Yang Li, Ruo Yang, Chao Guo, Ruojing Jiang, Fang Wu, Xiaoyu Gao, Zongyi Li, Jiaye Wu, Xingjian Ju, Chenwei Wang, Jingyi Yan, Tianci Hu, Ji Li, Rui Pan, Yichi Zhang, Yu-Chen (Jaky) Lee, and many others. Their friendship and support have been a source of strength and comfort during the challenging times of my Ph.D. studies. I am also grateful for the many fun and memorable moments we shared together.

I would like to thank all of my meowy friends for their unconditional love and companionship: Coal, Oreo, Shige, Shimei, Rourou, Mimi, Dingding, Huihui, and many others. Shige and Shimei have been with me closely through the COVID-19 pandemic and have been a source of faith during challenging times.

I would also like to thank my grandfather, Wenzhang Zhou, for his wisdom and guidance throughout my life. His teachings have inspired me to strive for excellence and make the most of my abilities. I am grateful for his love and support throughout my life. I would like to thank my grandmother, Shuyan Wang, for her love and care throughout my life. Her kindness and generosity shaped my values and beliefs and have been the lighthouse guiding me through the storms of life.

Finally, I would like to thank my parents, Xinyu Zhou and Qiang Liu, for their love and support throughout my Ph.D. studies and my life. Their curiosity and enthusiasm for learning about my research, despite the lack of a technical background, have pushed me to think outside the box and to communicate my research ideas more clearly and effectively to a broader audience.

ABSTRACT

Artificial intelligence (AI) has demonstrated increasing potential to assist humans in decision-making tasks. However, AI systems solve problems differently than humans, leading to varying performance in human-AI teams. An ideal human-AI collaboration system should leverage the strengths of both humans and AI while mitigating their weaknesses. In this dissertation, I will discuss how to empower humans with AI systems in decision-making tasks. First, I will explain how to understand human-AI teams under different distribution types and interactive interfaces (Chapter 2). Then, I will describe how to align AI models with human perceptions for better decision support (Chapter 3). Third, I will explore how to build AI-driven tutorials using selected concepts and examples to assist and teach humans in fine-grained image classification tasks (Chapter 4). Finally, I will discuss future directions for human-AI collaboration and how to enable better two-way communication between humans and AI systems (Chapter 5).

CHAPTER 1

INTRODUCTION

In the recent decades, there has been huge advances in the research and development of Artifical Intelligence (AI) systems. With greater ability and capacity to process data, AI systems have achieved remarkable performance in various tasks, such as image classification, speech recognition, and natural language processing. Such advances have invited more and more AI systems to be deployed in real-world applications in high-stakes domains such as healthcare, criminal justice, and finance. As they are increasingly used to assist humans in different kinds of tasks and domains, researchers start to wonder how to effectively communicate and collaborate with AI systems. Although AI systems have shown remarkable performance in various tasks, they solve problems in different ways than humans do, resulting in different success and failure modes from humans. An ideal human-machine collaboration system should leverage the strengths of both humans and machines, and mitigate their weaknesses. In this way, the system can make better predictions than either humans or machines alone, obtaining complementary performance [Bansal et al., 2019]. To achieve this goal, we need to understand how machines solve problems. What are their strengths and weaknesses? How do they make decisions? How to understand their predictions? How can we better interact with them? These questions are especially important in high-stakes domains where the decisions cannot made by AI systems alone. Therefore, we need more knowledge and tools to empower humans to better understand and collaborate with AI systems.

In this dissertation, we focus on how to enable a two-way communication between humans and AI systems. The goal is to enable humans to understand how AI systems make decisions and to enable AI systems to make predictions in a way that is compatible with human intuition or understandable to humans. We investigate how to achieve this goal in three different aspects: understanding machines through data and explanations, learning from humans for better decision support, and teaching humans with concepts and examples.

1.1 Understanding Machines Through Data and Explanations

In this part, we focus on understanding machines in challenging prediction tasks. When machines make predictions, they often rely on knowledge learned from data. However, the data machines are trained on may not be representative of the real world, or may be biased in some way. When the machine encounters unseen data, it may make mistakes. To achieve complementary performance, we need to understand what performance humans may achieve when their machine teammate are making predictions for out-of-distribution data. On the other hand, when machines make predictions, they often do not provide explanations for their decisions. We also need to understand how machines make their decisions in order to better interact with them. Therefore, we also explore how to understand machine predictions through interactive explanations. We design a novel interactive explanation system that allows users to interact with the machine to obtain explanations for its predictions. We also discuss how human understandings change when they interact with the machine. We emphasize the importance of helping humans to understand the machine's predictions in order to better collaborate with them.

1.2 Learning from Humans for Better Decision Support

In this part, we focus on learning from humans for better decision support. When humans make decisions, they often rely on their intuition and past experience. For example, when they predict the outcomes of a new case, they often refer to similar cases they have seen before. Such case-based decision making is common in many domains, such as healthcare where radiologists refer to similar cases to make diagnoses and criminal justice where judges refer to similar cases to make sentencing decisions. To achieve satisfactory performance, decision makers relies on the similarity between the new case and the cases they have seen before. Therefore, similarity judgments that are more compatible with human intuition can lead to better cases-based decision support. However, the perception of similarity between cases may vary between AI systems and humans. To achieve better decision support, we need to build AI systems that can capture the similarity between cases as perceived by humans. In this part, we investigate how to learn human-compatible representations in AI systems to retrieve decision support cases that are more effective for human decision makers. We emphasize the importance of building AI systems that are aligned with human intuition to better support human decision making.

1.3 Teaching Humans with Concepts and Examples

In this part, we focus on teaching image classification tasks with distinctive concepts and informative examples. When making predictions on images, humans often rely on their knowledge of the world and the concepts they have learned. For example, when they classify images of birds, they often refer to specific concepts such as the shape of the beak, the color of the feathers, and the size of the body. To build effective AI-driven tutorials, we need to identify key concepts for the tasks and important examples associated with the concepts. We propose a novel teaching paradigm with concept and example selection algorithm to teach simulated human learners to do fine-grained image classification tasks. We evaluate the effectiveness of the teaching on a number of fine-grained natural image classification tasks and discuss how different selection methods can affect learner performance. We emphasize the importance of finding informative concepts and examples as common ground between humans and AI systems to transfer knowledge effectively and improve human performance.

1.4 Organization and Contributions

The rest of the thesis is organized as follows.

In Chapter 2, we investigate how to understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. We show that human-AI teams have different interactions when making predictions on in-distribution and out-of-distribution examples. We also show that interactive explanations may not always improve human performance and may reinforce human biases.

In Chapter 3, we investigate how to learn human-compatible representations for case-based decision support. We show that AI systems that are trained to capture human similarity judgments can produce human-compatible representations. With different decision support selection policies, we show that human-compatible representations can lead to better decision support and improve human performance.

In Chapter 4, we investigate how to teach fine-grained image classification with concept and example selection. We show that teaching with informative concepts and associated examples can improve human performance.

In Chapter 5, we discuss future directions for human-AI collaboration and how to enable better two-way communication between humans and AI systems.

CHAPTER 2
UNDERSTANDING THE EFFECT OF OUT-OF-DISTRIBUTION EXAMPLES AND INTERACTIVE EXPLANATIONS

2.1 Overview

Although AI holds promise for improving human decision making in societally critical domains, it remains an open question how human-AI teams can reliably outperform AI alone and human alone in *challenging* prediction tasks (also known as *complementary performance*). We explore two directions to understand the gaps in achieving complementary performance. First, we argue that the typical experimental setup limits the potential of human-AI teams. To account for lower AI performance out-of-distribution than in-distribution because of distribution shift, we design experiments with different distribution types and investigate human performance for both in-distribution and out-of-distribution examples. Second, we develop novel interfaces to support interactive explanations so that humans can actively engage with AI assistance. Using virtual pilot studies and large-scale randomized experiments across three tasks, we demonstrate a clear difference between in-distribution and out-of-distribution, and observe mixed results for interactive explanations: while interactive explanations improve human perception of AI assistance's usefulness, they may reinforce human biases and lead to limited performance improvement. Overall, our work points out critical challenges and future directions towards enhancing human performance with AI assistance.

In this chapter, we start with explanations derived from AI systems that help humans understand the decisions made by AI systems. Most of the work in this chapter is published in Liu et al. [2021]. This is a joint work with Vivian Lai and Chenhao Tan.

2.2 Introduction

As AI performance grows rapidly and often surpasses humans in constrained tasks [Kleinberg et al., 2018, He et al., 2015, McKinney et al., 2020, Silver et al., 2018, Brown and Sandholm, 2019], a critical challenge to enable social good is to understand how AI assistance can be used to enhance *human performance.* AI assistance has been shown to improve people's efficiency in tasks such as transcription by enhancing their computational capacity [Lasecki et al., 2017, Gaur et al., 2016], support creativity in producing music [Louie et al., 2020, McCormack et al., 2019, Frid et al., 2020], and even allow the visually impaired to "see" images [Wu et al., 2017, Gurari et al., 2018]. However, it remains difficult to enhance human decision making in *challenging* prediction tasks [Kleinberg et al., 2015]. Ideally, with AI assistance, human-AI teams should outperform AI alone and human alone (e.g., in accuracy; also known as *complementary performance* [Bansal et al., 2021]). Instead, researchers have found that while AI assistance improves human performance compared to human alone, human-AI teams seldom outperform AI alone in a wide variety of tasks, including recidivism prediction, deceptive review detection, and hypoxemia prediction [Lai and Tan, 2019, Lai et al., 2020, Green and Chen, 2019b,a, Zhang et al., 2020, Poursabzi-Sangdeh et al., 2021, Carton et al., 2020, Lin et al., 2020, Weerts et al., 2019, Beede et al., 2020, Wang and Yin, 2021, Lundberg et al., 2018].

To address the elusiveness of complementary performance, we study two factors: 1) an overlooked factor in the experimental setup that may over-estimate AI performance; 2) the lack of two-way conversations between humans and AI, which may limit human understanding of AI predictions. First, we argue that prior work adopts a best-case scenario for AI. Namely, these experiments randomly split a dataset into a training set and a test set (Fig. 2.1). The training set is used to train the AI, and the test set is used to evaluate AI performance and human performance (with AI assistance). We hypothesize that this evaluation scheme is too optimistic for AI performance and provide limited opportunities for humans to contribute insights because the test set follows the same distribution as the training set (in-distribution). In practice, examples during testing may differ substantially from the training set, and AI performance can significantly drop for these out-of-distribution examples [McCoy et al., 2019, Clark et al., 2019, Jia and Liang, 2017]. Furthermore, humans are better equipped to detect problematic patterns in AI predictions and offer complementary insights in out-of-distribution examples. Thus, we propose to develop experimental designs with both out-of-distribution examples and in-distribution examples in the test set.

Second, although explaining AI predictions has been hypothesized to help humans understand AI predictions and thus improve human performance [Doshi-Velez and Kim, 2017], static explanations, such as highlighting important features and showing AI confidence, have been mainly explored so far [Green and Chen, 2019b, Lai and Tan, 2019, Bansal et al., 2021]. Static explanations represent a one-way conversation from AI to humans and may be insufficient for humans to understand AI predictions. In fact, psychology literature suggests that interactivity is a crucial component in explanations [Lombrozo, 2006, Miller, 2018]. Therefore, we develop interactive interfaces to enable a two-way conversation between decision makers and AI. For instance, we allow humans to change the input and observe how AI predictions would have changed in these counterfactual scenarios (Fig. 2.6). We hypothesize that interactive explanations improve the performance of humans and their subjective perception of AI assistance's usefulness. Although out-of-distribution examples and interactive explanations are relatively separate research questions, we study them together in this work as we hypothesize that they are critical missing ingredients towards complementary performance.

To investigate the effect of out-of-distribution examples and interactive explanations on human-AI decision making, we choose three datasets spanning two tasks informed by prior work: 1) recidivism prediction (COMPAS and ICPSR) (a canonical task that has received much attention due to its importance; COMPAS became popular because of the ProPublica

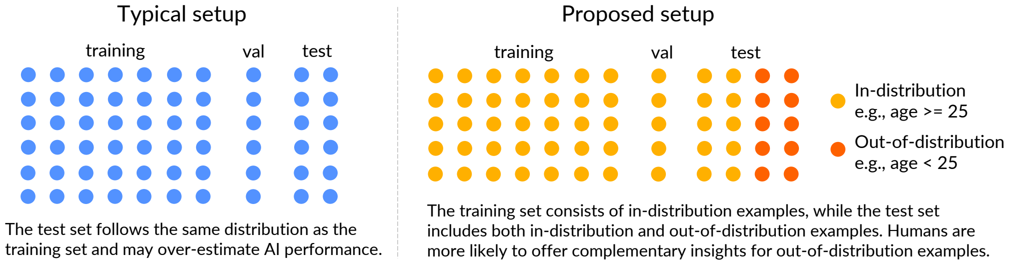

Typical setup
training
val
test
The test set follows the same distribution as the training set and may over-estimate AI performance.
Proposed setup
training
val
test
In-distribution e.g., age >= 25
Out-of-distribution e.g., age < 25
The training set consists of in-distribution examples, while the test set includes both in-distribution and out-of-distribution examples. Humans are more likely to offer complementary insights for out-of-distribution examples.

Figure 2.1: An illustration of the typical setup and our proposed setup that takes into account distribution types. For instance, in the recidivism prediction task we can use defendants of younger ages to simulate out-of-distribution examples, assuming our training set only contains older defendants referred as in-distribution examples. The fractions of data are only for illustrative purposes. See details of in-distribution vs. out-of-distribution setup in §2.4.2.

article on machine bias [Angwin et al., 2016], and ICPSR was recently introduced to the human-AI interaction community by Green and Chen [2019b,a], so it would be useful to see whether same results hold in both datasets); 2) profession detection (BIOS) (the task is to predict a person's profession based on a short biography; this task is substantially easier than recidivism prediction and other text-based tasks such as deceptive review detection, so crowdworkers may have more useful insights to offer for this task). We investigate human-AI decision making in these tasks through both virtual pilot studies and large-scale randomized experiments. We focus on the following three research questions:

- **RQ1**: how do distribution types affect the performance of human-AI teams, compared to AI alone?
- **RQ2**: how do distribution types affect human agreement with AI predictions?
- **RQ3**: how do interactive explanations affect human-AI decision making?

Our results demonstrate a clear difference between in-distribution and out-of-distribution. Consistent with prior work, we find that human-AI teams tend to underperform AI alone in in-distribution examples in all tasks. In comparison, human-AI teams can occasionally outperform AI in out-of-distribution examples in recidivism prediction (although the difference is small). It follows that the performance gap between human-AI teams and AI is smaller out-of-distribution than in-distribution, confirming that humans are more likely to achieve complementary performance out-of-distribution.

Distribution types also affect human agreement with AI predictions. In recidivism prediction (COMPAS and ICPSR), humans are more likely to agree with AI predictions in-distribution than out-of-distribution, suggesting that humans behave differently depending on the distribution type. Moreover, in recidivism prediction, human agreement with *wrong AI predictions* is lower out-of-distribution than in-distribution, suggesting that humans may be better at providing complementary insights into AI mistakes out-of-distribution. However, in BIOS, where humans may have more intuitions for detecting professions, humans are less likely to agree with AI predictions in-distribution than out-of-distribution. This observation also explains the relatively low in-distribution performance of human-AI teams in BIOS compared to AI alone.

Finally, although we do not find that interactive explanations lead to improved performance for human-AI teams, they significantly increase human perception of AI assistance's usefulness. Participants with interactive explanations are more likely to find real-time assistance useful in ICPSR and COMPAS, and training more useful in COMPAS. To better understand the limited utility of interactive explanations, we conduct an exploratory study on what features participants find important in recidivism prediction. We find that participants with interactive explanations are more likely to fixate on demographic features such as age and race, and less likely to identify the computationally important features based on Spearman correlation. Meanwhile, they make more mistakes when they disagree with AI. These observations suggest that interactive explanations might reinforce existing human biases and lead to suboptimal decisions.

Overall, we believe that our work adds value to the community in the emerging field of human-AI collaborative decision making in *challenging* prediction tasks. Our work points out an important direction in designing future experimental studies on human-AI decision making:

it is critical to think about the concept of out-of-distribution examples and evaluate the performance of human-AI teams both in-distribution and out-of-distribution. The implications for interactive explanations are mixed. On the one hand, interactive explanations improve human perception of AI usefulness, despite not reliably improving their performance. On the other hand, similar to ethical concerns about static explanations raised in prior work [Green and Chen, 2019a,b, Bansal et al., 2021], interactive explanations might reinforce existing human biases. It is critical to take these factors into account when developing and deploying improved interactive explanations. Our results also highlight the important role that task properties may play in shaping human-AI collaborative decision making and provide valuable samples for exploring the vast space of tasks.

2.3 Related Work and Research Questions

In this section, we review related work and formulate our research questions.

2.3.1 Performance of Human-AI Teams in Prediction Tasks

With a growing interest in understanding human-AI interaction, many recent studies have worked on enhancing human performance with AI assistance in decision making. Typically, these decisions are formulated as prediction tasks where AI can predict the outcome and may offer explanations, e.g., by highlighting important features. For instance, the bailing decision (whether a defendant should be bailed) can be formulated as a prediction problem of whether a defendant will violate pretrial terms in two years [Kleinberg et al., 2018]. Most studies have reported results aligning with the following proposition:

Proposition 1. *AI assistance improves human performance compared to without any assistance; however, the performance of human-AI teams seldom surpasses AI alone in challenging prediction tasks [Lai and Tan, 2019, Lai et al., 2020, Green and Chen, 2019b,a, Zhang et al., 2020, Poursabzi-Sangdeh et al., 2021, Carton et al., 2020, Lin et al., 2020, Weerts et al.,* *2019, Beede et al., 2020, Lundberg et al., 2018, Wang and Yin, 2021, Buçinca et al., 2020].*[1]

This proposition is supported in a wide variety of tasks, including recidivism prediction [Green and Chen, 2019b,a, Lin et al., 2020], deceptive review detection [Lai and Tan, 2019, Lai et al., 2020], income prediction [Poursabzi-Sangdeh et al., 2021], and hypoxemia prediction [Lundberg et al., 2018], despite different forms of AI assistance. To understand this observation, we point out that Proposition 1 entails that AI alone outperforms humans alone in these tasks (human $<$ human $+$ AI $<$ AI). Lai et al. [2020] conjectures that the tasks where humans need AI assistance typically fall into the *discovering* mode, where the groundtruth is determined by (future) external events (e.g., a defendant's future behavior) rather than human decision makers, instead of the *emulating* mode, where humans (e.g., crowdworkers) ultimately define the groundtruth.[2] We refer to prediction tasks in the discovering mode as *challenging prediction tasks*. Example tasks include the aforementioned recidivism prediction, deception detection, hypoxemia prediction, etc. These tasks are non-trivial to humans and two corollaries follow: 1) human performance tend to be far from perfect; 2) the groundtruth labels cannot be crowdsourced.[3] In such tasks, AI can identify non-trivial and even counterintuitive patterns to humans. These patterns can be hard for humans to digest and leverage when they team up with AI. As such, it is difficult for human-AI teams to achieve complementary performance.

A notable exception is Bansal et al. [2021], which shows that human-AI team performance surpasses AI performance in sentiment classification (beer reviews and Amazon reviews) and

[1] Our focus in this work is on understanding the performance of human-AI teams compared to AI performance and do not recommend AI to replace humans in any means. In fact, many studies have argued that humans should be the final decision makers in societally critical domains for ethical and legal reasons such as recidivism prediction and medical diagnosis [Green and Chen, 2019b, Lai and Tan, 2019, Liptak, 2017, Supreme Court of Wisconsin, 2016, Supreme Court of the United States, 1993].

[2] In fact, it is unclear what complementary performance means in the *emulating* mode if humans define the groundtruth as human performance is by definition 100%. A more subtle discussion can be found in footnote 4.

[3] Whether a task is challenging (in the discovering mode) also depends on characteristics of humans. For instance, sentiment analysis of English reviews might not be challenging for native speakers, but could remain challenging for non-native speakers.

Table 2.1: Definitions of complementary performance and comparable performance.

Complementary performance. An ideal outcome of human-AI collaborative decision making: the performance of human-AI teams is better than AI alone and human alone.
Comparable performance. The performance of human alone is *similar* to AI alone, yielding more potential for complementary performance as hypothesized in Bansal et al. [2021]. There lacks a quantitative definition of what performance gap counts as comparable. We explore different ranges in this work.

LSAT question answering. Their key hypothesis is that human-AI teams are likely to excel when human performance and AI performance are comparable, while prior studies tend to look at situations where the performance gap is substantial. It naturally begs the question of what size of performance gap counts as comparable performance, whether comparable performance alone is sufficient for complementary performance, and whether other factors are associated with the observed complementary performance (we summarize the definitions of complementary performance and comparable performance in Table 2.1 to help readers understand these concepts). For instance, it is useful to point out that sentiment analysis is closer to the emulating mode.[4] We will provide a more in-depth discussion in §2.8.

Our core hypothesis is that a standard setup in current experimental studies on human-AI interaction might limit the potential of human-AI teams. Namely, researchers typically follow standard machine learning setup in evaluating classifiers by randomly splitting the dataset into a training set and a test set, and using the test set to evaluate the performance of human-AI teams and AI alone. It follows that the data distribution in the test set is similar to the training set by design. Therefore, this setup is designed for AI to best leverage the patterns learned from the training set and provide a strong performance. In practice, a critical growing concern is distribution shift [Goodfellow et al., 2016, Quionero-Candela et al., 2009, Sugiyama and Kawanabe, 2012]. In other words, the test set may differ from the training set,

[4] Although labels in sentiment analysis are determined by the original author, sentiment analysis is generally viewed as a natural language understanding task that humans are capable of. AI is thus designed to emulate human capability. In the emulating mode, improving human performance is essentially aligning single-person decisions with the majority of a handful of annotators. We argue that data annotation is qualitatively different from decision making in challenging tasks such as recidivism prediction.

so the patterns that AI identifies can fail during testing, leading to a substantial drop in AI performance [McCoy et al., 2019, Clark et al., 2019, Jia and Liang, 2017]. Throughout this paper, we refer to testing examples that follow the same distribution as the training set as in-distribution (IND) examples and that follow a different distribution as out-of-distribution (OOD) examples.

Thus, our first research question (**RQ1**) examines how distribution types affect the performance of human-AI teams, compared to AI alone. We expect our results in in-distribution examples to replicate previous findings and be consistent with Proposition 1. In comparison, we hypothesize that humans are more capable of spotting problematic patterns and mistakes in AI predictions when examples are not similar to the training set (out-of-distribution), as humans might be robust against distribution shift. Even if human-AI teams do not outperform AI alone in out-of-distribution examples, we expect the performance gap between human-AI teams and AI alone to be smaller out-of-distribution than in-distribution. Inspired by the above insights on comparable performance, we choose three tasks where humans and AI have performance gaps of different sizes so that we can investigate the effect of distribution type across tasks.

2.3.2 Agreement with AI

In addition to human performance, human agreement with AI predictions is critical for understanding human-AI interaction, especially in tasks where humans are the final decision makers. When AI predictions are explicitly shown, this agreement can also be interpreted as the trust that humans place in AI. Prior work has found that in general, the more information about AI predictions is given, the more likely humans are going to agree with AI predictions [Lai and Tan, 2019, Feng and Boyd-Graber, 2019, Bansal et al., 2021, Ghai et al., 2020]. For instance, explanations, presented along with AI predictions, increase the likelihood that humans agree with AI [Lai et al., 2020, Bansal et al., 2021, Ghai et al., 2020].

Confidence levels have also been shown to help humans calibrate whether to agree with AI [Zhang et al., 2020, Bansal et al., 2021]. In a similar vein, Yin et al. [2019] investigate the effect of observed and stated accuracy on humans' trust in AI and find that both stated and observed accuracy can affect human trust in AI. Finally, expertise may shape humans' trust in AI: Feng and Boyd-Graber [2019] find that novices in Quiz Bowl trust the AI more than experts when visualizations are enabled.

However, little is known about the effect of distribution types as it has not been examined in prior work. Our second research question (**RQ2**) inquires into the effect of distribution types on human agreement with AI predictions. We hypothesize that humans are more likely to agree with AI in-distribution than out-of-distribution because the patterns that AI learns from in-distribution examples may not apply out-of-distribution and AI performance is worse out-of-distribution than in-distribution. Furthermore, given prior results that humans are more likely to agree with correct AI predictions than wrong AI predictions [Lai and Tan, 2019, Bansal et al., 2021], it would be interesting to see whether that trend is different out-of-distribution from in-distribution.

Additionally, we are interested in having a closer look at the effect of distribution types on human agreement by zooming in on the correctness of AI predictions. Prior work has introduced three terms to address these different cases of agreement [Wang and Yin, 2021]: appropriate trust [McBride and Morgan, 2010, McGuirl and Sarter, 2006, Merritt et al., 2015, Muir, 1987] (the fraction of instances where humans agree with correct AI predictions and disagree with wrong AI predictions; this is equivalent to human-AI team accuracy in binary classification tasks), overtrust [Parasuraman and Riley, 1997, de Visser et al., 2014] (the fraction of instances where humans agree with wrong AI predictions), and undertrust [Parasuraman and Riley, 1997, de Visser et al., 2014] (the fraction of instances where humans disagree with correct AI predictions). To simplify the measurement, we only consider agreement with AI predictions in this work because disagreement and agreement add

Table 2.2: Definition of human agreement based on the correctness of AI predictions.

	Correct AI predictions	**Wrong AI predictions**
Humans agree with	Appropriate agreement	Overtrust
Humans disagree with	Undertrust	Appropriate disagreement

up to 1. We define the fraction of instances where humans agree with *correct* AI predictions as *appropriate agreement* and the fraction of instances where humans agree with *incorrect* AI predictions as *overtrust*, and similarly the counterparts in disagreement as *undertrust* and *appropriate disagreement*. Table 2.2 shows the full combinations of human agreement and AI correctness. The term *appropriate trust* then is the sum of *appropriate agreement* and *appropriate disagreement*. We hypothesize that patterns embedded in the AI model may not apply to out-of-distribution examples, humans can thus better identify wrong AI predictions in out-of-distribution examples (i.e., overtrust is lower out-of-distribution). Similarly, our intuition is that *appropriate agreement* is also likely lower out-of-distribution as AI may make correct predictions based on non-sensible patterns. While we focus on how distribution types affect *appropriate agreement* and *overtrust*, it also entails how distribution types affect *undertrust* and *appropriate disagreement*.

2.3.3 Interactive Explanations

A key element in developing AI assistance are explanations of AI predictions, which have attracted a lot of interest from the research community [Lipton, 2016, Doshi-Velez and Kim, 2017, Ribeiro et al., 2016, Lundberg and Lee, 2017, Koh and Liang, 2017, Lakkaraju et al., 2016, Gilpin et al., 2018]. Experimental studies in human-AI decision making have so far employed static explanations such as highlighting important features and showing similar examples, a few studies have also investigated the effect of explanations with an interactive interface. However, literature in social sciences has argued that explanations should be interactive. For instance, Lombrozo [2006] suggests that an explanation is a byproduct of an interaction process between an explainer and an explainee, and Miller [2018] says that explanations are social in that they are transferable knowledge that is passed from one person to the other in a conversation. We hypothesize that the one-way conversation in static explanations is insufficient for humans to understand AI predictions, contributing to the proposition that human-AI teams have yet to outperform AI alone.

It is worth pointing out that industry practitioners have worked towards developing interactive interfaces to take advantage of deep learning models' superior predictive power. For instance, Tenney et al. [2020] develop an interative interpretability tool that provide insightful visualizations for NLP tasks. Similar interactive tools have been used to support data scientists in debugging machine learning models and improving model performance [Kaur et al., 2020, Hohman et al., 2019, Wu et al., 2019]. While data scientists are familiar with machine learning, laypeople may not have the basic knowledge of machine learning. We thus focus on developing an interface that enables meaningful interactive explanations for laypeople to support decision making rather than debugging. Our ultimate goal is to improve human performance instead of model performance. In addition, there have been interactive systems that provide AI assistance for complicated tasks beyond constrained prediction tasks [Cai et al., 2019b, Xie et al., 2020, Yang et al., 2019]. Our scope in this work is limited to explanations of AI predictions where the human task is to make a simple categorical prediction. Most similar to our work is Cheng et al. [2019], which examines the effect of different explanation interfaces on user understanding of a model and shows improved understandings with interactive explanations, whereas our work focuses on the effect of interactive explanations on human-AI decision making.

As such, our final research question (**RQ3**) investigates the effect of interactive explanations on human-AI decision making. We hypothesize that interactive explanations lead to better human-AI performance, compared to static explanations. We further examine the effect of interactive explanations on human agreement with AI predictions. If interactive explanations enable humans to better critique incorrect AI predictions, then humans may become less reliant on the incorrect predicted labels (i.e., lower overtrust). Finally, we expect interactive explanations to improve subjective perception of usefulness over static explanations because interactive explanations enable users to have two-way conversations with the model.

2.3.4 Differences from Interactive Machine Learning and Transfer Learning

It is important to note that our focus in this work is on how distribution types and interactive explanations affect human performance in decision making and our ultimate goal is to enhance human performance. While other areas such as transfer learning and interactive machine learning have conducted user studies where people interact with machine learning models, the goal is usually to improve model performance. Specifically, interactive machine learning tends to involve machine learning practitioners, while our work considers the population that does not have a machine learning background [Hohman et al., 2019, Krause et al., 2016, Tenney et al., 2020, Wexler et al., 2019]. Similarly, transfer learning focuses on improving models that would generalize well on other domains (distributions), whereas our work investigates how examples in different distributions affect *human performance* [Zhuang et al., 2020, Liang and Zheng, 2020, Torrey and Shavlik, 2010]. Although improving AI will likely improve human performance in the long run, we focus on the effect of AI assistance on human decision making where the AI is not updated.

2.4 Methods

In order to evaluate the performance of human-AI teams, we consider three important ingredients in this work: 1) Prediction tasks: we consider three prediction tasks that include both tabular and text datasets as well as varying performance gaps between human alone and AI alone (§2.4.1); 2) In-distribution (IND) vs. out-of-distribution (OOD): a key contribution of our work is to highlight the importance of distribution shift and explore ways to design human-AI experimental studies with considerations of in-distribution and out-of-distribution examples (§2.4.2); 3) Explanation type: another contribution of our work is to design novel interactive explanations for both tabular data and text data (§2.4.3). We further use virtual pilot studies to gather qualitative insights and validate our interface design (§2.4.4), and then conduct large-scale experiments with crowdworkers on Mechanical Turk (§2.4.5).

2.4.1 Prediction Tasks

We use two types of tasks, recidivism prediction, and profession prediction. Recidivism prediction is based on tabular datasets, while profession prediction is based on text datasets.

- ICPSR [United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics., 2014]. This dataset was collected by the U.S. Department of Justice. It contains defendants who were arrested between 1990 and 2009, and the task is to predict if a defendant will violate the terms of pretrial release. Violating terms of pretrial release means that the defendant is rearrested before trial, or fails to appear in court for trial, or both. We clean the dataset to remove incomplete rows, restrict the analysis to defendants who were at least 18 years old, and consider only defendants who were released before trial as we only have ground truth for this group. We consider seven attributes as features in this dataset: Gender, Age, Race, Prior Arrests, Prior Convictions, Prior Failure to Appear, and Offense Type (e.g., drug, violent). To protect defendant privacy, we only selected defendants whose features are identical to at least two other defendants in the dataset. This yielded a dataset of 40,551 defendants.

- COMPAS [Angwin et al., 2016]. The task is to predict if the defendant will recidivate in two years. The features in this dataset are Sex, Age, Race, Prior Crimes, Charge Degree, Juvenile Felony Count, and Juvenile Misdemeanor Count. Both datasets have overlapping features such as Age and Race. There are 7,214 defendants in this dataset.

- BIOS [De-Arteaga et al., 2019]. This dataset contains hundreds of thousands of online biographies from the Common Crawl corpus. The task is to predict a person's profession given a biography. The original dataset consists of 29 professions, and we narrow it down to five professions to make the task feasible for humans, namely, psychologist, physician, surgeon, teacher, and professor.[5] This yielded a dataset of 205,360 biographies.

As Bansal et al. [2021] hypothesize that comparable performance between humans and AI is critical for complementary performance, our tasks cover varying performance gaps. The in-distribution performance gap between AI alone and human alone in-distribution is relatively small (∼7%) in recidivism prediction (68.4% vs. 60.9% in ICPSR and 65.5% vs. 60.0% in COMPAS), but large (∼20%) in profession prediction (see Table 2.3 and §2.5 for a more detailed discussion on performance gap). Note that human performance in ICPSR and COMPAS is derived from our experiments with crowdworkers. Although they are not representative of judges (see more discussion in §2.8), they outperform random baselines and can potentially be improved with AI assistance. In fact, human performance in LSAT is also ∼60% in Bansal et al. [2021], and crowdworkers were able to achieve complementary performance. Finally, we include gender and race for recidivism prediction to understand how humans might use the information, but they should not be included in AI for deployment.

2.4.2 In-distribution vs. Out-of-distribution Setup

As argued in §2.3, prior work randomly split a dataset to evaluate the performance of human-AI teams. This setup constitutes a best-case scenario for AI performance and may have contributed to the elusiveness of complementary performance. We expect humans to be more capable of providing complementary insights (e.g., recognizing that AI falsely generalizes a pattern) on examples following different distributions from the training data

[5] To choose these five professions, we built maximum spanning trees with 4, 5, 6 nodes from a graph based on the confusion matrix of a classifier trained with all biographies. Thus, the maximum spanning tree identifies the most confusing professions for the AI.

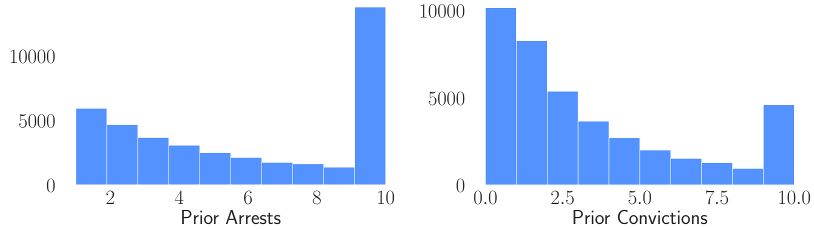


Figure 2.2: Histograms of numbers of instances for "Prior Arrests" and "Prior Convictions" in ICPSR.

(out-of-distribution). Therefore, it is crucial to evaluate the performance of human-AI teams on out-of-distribution examples. We thus provide the first attempt to incorporate distribution shift into experimental studies in the context of human-AI decision making.

Designing In-distribution vs. Out-of-distribution

To simulate the differences between in-distribution and out-of-distribution examples, our strategy is to split the dataset into an in-distribution (IND) subset and an out-of-distribution (OOD) subset based on a single attribute (e.g., age ≥ 25 as in-distribution and age < 25 as out-of-distribution to simulate a scenario where young adults are not presented in the training set). We develop the following desiderata for selecting an attribute to split the dataset: 1) splitting by this attribute is sensible and interpretable to human (e.g., it makes little sense to split biographies based on the number of punctuation marks); 2) splitting by this attribute could yield a difference in AI performance between in-distribution and out-of-distribution so that we might expect different human behavior in different distribution types; 3) this attribute is "smoothly" distributed in the dataset to avoid extreme distributions that can limit plausible ways to simulate IND and OOD examples (see the supplementary materials for details). Now we discuss the attribute selected for each dataset and present rationales for not using other attributes.

- ICPSR. We choose the age of the defendant as the attribute. We also tried Gender, but it failed desiderata 2 due to a small AI performance difference (1%) between in-distribution and out-of-distribution. Other features such as Prior Arrests and Prior Convictions do not satisfy desiderata 3, because they have a huge spike towards the end (see Fig. 2.2) and thus limit possible IND/OOD splits.

- COMPAS. We choose the age of the defendant as the attribute. We also tried Sex and Prior Crimes, but they failed desiderata 2 and 3 respectively as Gender and Prior Convictions did in ICPSR.

- BIOS. We choose the length of the biography (i.e., the total number of characters) as the attribute. Note that our dataset contains biographies from the web, a dataset created by De-Arteaga et al. [2019]. Although one may think that professor, surgeon, psychologist, and physician require more education than teacher and thus resulting in longer biographies, the average biography length of a teacher's biography is not the shortest in our dataset. Interestingly, physicians have the shortest biographies with 348 characters and teachers have an average biography length of 367 characters. We also experimented with gender but it does not satisfy desiderata 2 since we observed a small AI performance difference (3%) between in-distribution and out-of-distribution.

Given the selected attribute, for each dataset, we split the data into 10 bins of equal size based on the attribute of choice. Then, we investigate which bins to use as in-distribution and out-of-distribution. Our goal in this step is to maximize the AI performance gap between in-distribution and out-of-distribution so that we can observe whether humans would behave differently with AI assistance depending on distribution types (see supplementary materials). The chosen splits for each dataset are: 1) age ≥ 25 as IND and age < 25 as OOD in ICPSR, 2) age ≥ 26 as IND and age < 26 as OOD in COMPAS, and 3) length ≥ 281 characters as IND and length < 281 characters as OOD in BIOS. For each potential split, we use 70% of the data in the IND bins for training and 10% of the data in the IND bins for validation. Our test set includes two subsets: 1) the remaining 20% of the data in the IND bins, and 2) the data in the OOD bins. We also balance the labels in each bin of our test set for performance evaluation.

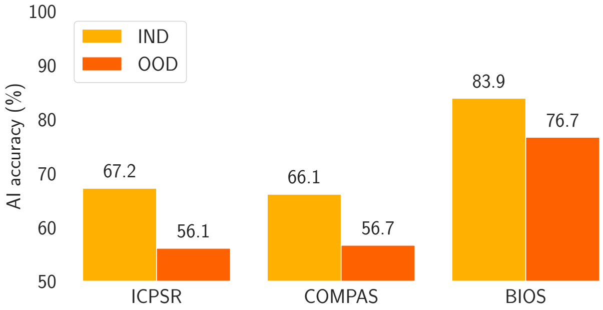


Figure 2.3: Accuracy of machine learning models on the in-distribution and out-of-distribution test set for the user study. Since the test set is balanced, the baseline in ICPSR and COMPAS is 50%. AI outperforms the random baseline even out-of-distribution in ICPSR and COMPAS despite that its performance is lower out-of-distribution than in-distribution. AI performance drops by about 10% in recidivism prediction and about 7% in BIOS for out-of-distribution examples compared to in-distribution examples.

AI Performance in-distribution and out-of-distribution.

Following prior work [Lai et al., 2020, De-Arteaga et al., 2019], we use a linear SVM classifier with unigram bag-of-words for BIOS and with one-hot encoded features for recidivism prediction tasks. The standard procedure of hyperparameter selection (a logarithmic scale between 10^{-4} and 10^{4} for the inverse of regularization strength) is done with the validation set. We focus on linear models in this work for three reasons: 1) linear models are easier to explain than deep models and are a good starting point to develop interactive explanations [Feng and Boyd-Graber, 2019, Poursabzi-Sangdeh et al., 2021]; 2) prior work has shown that human performance is better when explanations from simple models are shown [Lai et al., 2020]; 3) there is a sizable performance gap between humans and AI even with a linear model, although smaller than the case of deception detection [Lai and Tan, 2019, Lai et al., 2020].

Finally, to reduce the variance of human performance so that each example receives multiple human evaluations, we randomly sample 180 IND examples and 180 OOD examples from the test set to create a balanced pool for our final user study.[6] Fig. 2.3 shows AI performance on these samples: the IND-OOD gap is about 10% in recidivism prediction and

[6] We choose from five random seeds the one that leads to the greatest AI performance difference between in-distribution samples and out-of-distribution samples.

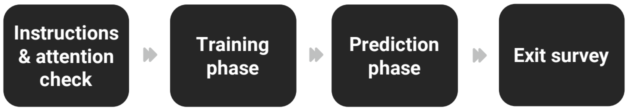


Figure 2.4: The workflow of our experiments. In the training phase, we introduce a novel feature quiz where users choose one positive and one negative feature after each example. Human decisions in the prediction phase are used to study human-AI decision making.

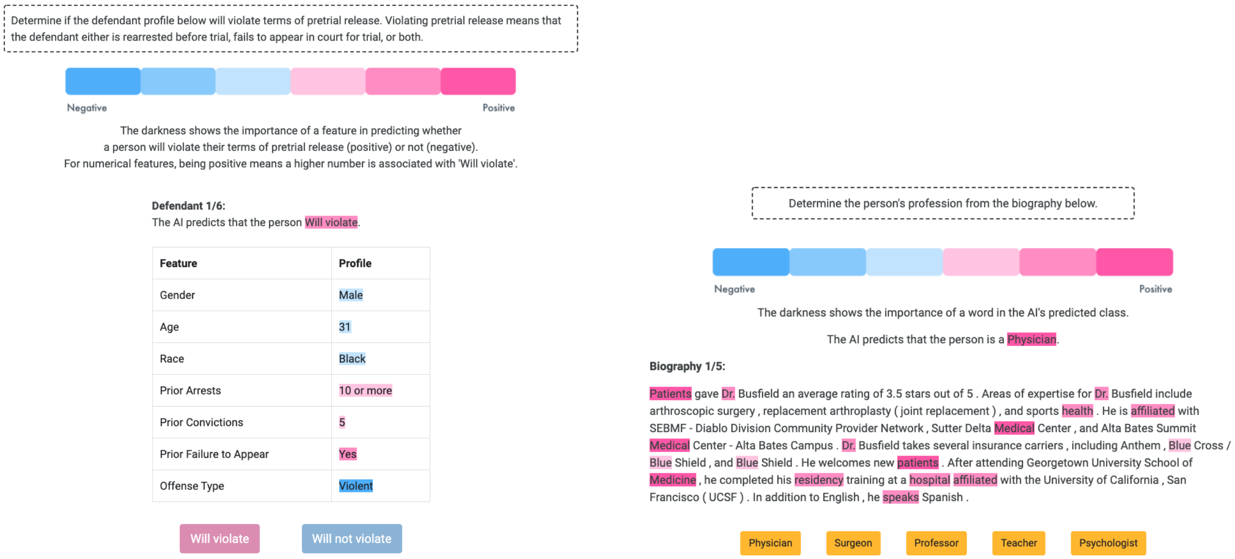
7% in BIOS. It entails that the absolute performance necessary to achieve complementary performance is lower OOD than IND. Because of this AI performance gap in-distribution and out-of-distribution, we will focus on understanding the performance difference between human-AI teams and AI alone (*accuracy gain*). As discussed in §2.3, we hypothesize that the accuracy gain is greater out-of-distribution than in-distribution.

2.4.3 Interactive Explanations and Explanation Type

To help users understand the patterns embedded in machine learning models, following Lai et al. [2020], our experiments include two phases: 1) a training phase where users are shown no more than six representative examples and the associated explanations; and 2) a prediction phrase that is used to evaluate the performance of human-AI teams with 10 random in-distribution examples and 10 random out-of-distribution examples. Fig. 2.4 shows the workflow of our experiments. Our contribution is to develop interactive explanations to enable a two-way conversation between humans and AI and examine the effect of interactive explanations. We also consider a static version of AI assistance in each phase for comparison. We refer to AI assistance during the prediction phase as *real-time assistance*.

Static Assistance

Our static assistance for an AI prediction includes two components (see Fig. 2.5). First, we highlight important features based on the absolute value of feature coefficients to help users understand what factors determine the AI prediction. We color all seven features in ICPSR and COMPAS to indicate whether a feature contributes positively or negatively to the



Determine if the defendant profile below will violate terms of pretrial release. Violating pretrial release means that the defendant either is rearrested before trial, fails to appear in court for trial, or both.

The darkness shows the importance of a feature in predicting whether a person will violate their terms of pretrial release (positive) or not (negative). For numerical features, being positive means a higher number is associated with 'Will violate'.

Defendant 1/6:
The AI predicts that the person Will violate.

Feature	Profile
Gender	Male
Age	31
Race	Black
Prior Arrests	10 or more
Prior Convictions	5
Prior Failure to Appear	Yes
Offense Type	Violent

(a) Static assistance for ICPSR.

Determine the person's profession from the biography below.

The darkness shows the importance of a word in the AI's predicted class.

The AI predicts that the person is a Physician.

Biography 1/5:

Patients gave Dr. Busfield an average rating of 3.5 stars out of 5 . Areas of expertise for Dr. Busfield include arthroscopic surgery , replacement arthroplasty (joint replacement) , and sports health . He is affiliated with SEBMF - Diablo Division Community Provider Network , Sutter Delta Medical Center , and Alta Bates Summit Medical Center - Alta Bates Campus . Dr. Busfield takes several insurance carriers , including Anthem , Blue Cross / Blue Shield , and Blue Shield . He welcomes new patients . After attending Georgetown University School of Medicine , he completed his residency training at a hospital affiliated with the University of California , San Francisco (UCSF) . In addition to English , he speaks Spanish .

(b) Static assistance for BIOS.

Figure 2.5: Screenshots for static assistance in ICPSR and BIOS. The interface for COMPAS is similar to ICPSR (see Fig. 6.8.

prediction (Fig. 2.5a). As BIOS has many words as features, we highlight the top 10 most important words. We only show the colors but hide the feature coefficient numbers because 1) we have not introduced the notion of prediction score; 2) showing numerical values without interaction may increase the cognitive burden without much gain. Second, we also show the AI predicted label along with the highlights. In the training phase, following Lai et al. [2020], the actual label is revealed after users make their predictions so that they can reflect on their decisions and actively think about the task at hand.

The purpose of the training examples is to allow participants to familiarize themselves with the task, extract useful and insightful patterns, and apply them during the prediction phase. We use SP-LIME [Ribeiro et al., 2016, Lai et al., 2020] to identify 5-6 representative training examples that capture important features (6 in ICPSR and COMPAS and 5 in BIOS).[7] We make sure the selected examples are balanced across classes. For the control condition, we simply include the first two examples. Finally, during training, to ensure that users

[7] We include 10 examples in the pilot studies, but mechanical turkers commented that the experiment took too long.

understand the highlighted important features, we add a feature quiz after each example where users are required to choose a positive and a negative feature (see Fig. 6.11).

Interactive Explanations

To help humans better understand how AI makes a prediction and the potential fallacies in AI reasoning, we develop a suite of interactive experiences. There are two important components. First, we enable users to experiment with counterfactual examples of a given instance. This allows participants to interact with each feature and observe changes in AI predictions. Second, we make the feature highlights dynamic, especially for BIOS where there are many features. Specifically, our designs are as follows:

- Interactive explanations for tabular-data classification (ICPSR and COMPAS; Fig. 2.6a gives a screenshot for ICPSR). We present the original profile of the defendant and the counterfactual ("What-if scenario profile") on the left of the screen (Fig. 2.6a(1)). Users can adjust features to change the counterfactual profile (Fig. 2.6a(2)) via sliders, radio buttons, and select lists (Fig. 2.6a(3-5)). For instance, users can investigate how a younger or older age affects the prediction by adjusting a defendant's age using the slider. In addition, we show all the features and their associated weight on the right, sorted in descending order (Fig. 2.6a(6)).

- Interactive explanations for text classification (BIOS; see Fig. 2.6b). To enable the counterfactuals, users can delete any word in the text and see how the prediction would change (removal can be undone by clicking the same word again). For dynamic highlight, a slider is available for users to adjust the number of highlighted words (Fig. 2.6b(1)). In addition, we provide a searchable table to display all words presented in the text and their associated feature importance, sorted in descending order (Fig. 2.6b(2)).

The searchable table allows users to the explore the high-dimensional feature space in BIOS, a text classification task. While it may seem that showing coefficients in recidivism prediction is not as useful, we highlight that these numerical values make little sense on their own. The counterfactual profile enables users to examine how these numerical values affect prediction outcomes.

2.4.4 Virtual Pilot Studies

We conducted virtual pilot studies to obtain a qualitative understanding of human interaction with interactive explanations. The pilot studies allow us to gather insights on how humans use interactive explanations in their decision-making process, as well as feedback on the web application before conducting large-scale randomized experiments.

Experimental design. We employed a concurrent think-aloud process with participants [Nielsen et al., 2002]. Participants are told to verbalize the factors they considered behind a prediction. During the user study session, participants first read instructions for the task and proceed to answer a couple of attention-check questions (see Fig. 6.10), which ensure that they understand the purpose of the user study. Upon passing the attention-check stage, they undergo a training phase before proceeding to the prediction phase. Finally, they answer an exit survey (see Fig. 6.13) that asks for demographic information and semi-structured questions on the web application and interactive explanations. A participant works on ICPSR and BIOS in a random order.

We recruited 15 participants through mailing lists at the University of Colorado Boulder: 7 were female and 8 were male, with ages ranging from 18 to 40.[8] To understand the general population that does not have a machine learning background, we sent out emails to computer science and interdisciplinary programs. Participants included both undergraduate and graduate students with and without machine learning background. The user study is conducted on Zoom due to the pandemic. The user study sessions were recorded with the participants' consent. Participants were compensated for $10 for every 30 minutes. A typical

[8] Note that the wide range in age is due to the available choices in our exit survey. Namely, the first option is 18-25 and the second option is 26-40.

user study session lasted between an hour to an hour and a half. Participants were assigned in a round-robin manner to interactive and static explanations. For instance, if a participant was assigned to static explanations in BIOS, the participant would be assigned to interactive explanations in ICPSR. As the user study sessions were recorded on Zoom cloud, we used the first-hand transcription provided by Zoom and did a second round of transcribing to correct any mistranscriptions. Subsequently, thematic analysis was conducted to identify common themes in the think-aloud processes, and thematic codes were collectively coded by two researchers.

Next, we summarize the key themes from the pilot studies and the changes to our interface.

Disagreement with AI predictions. Participants tend to disagree with AI predictions when the explanations provided by the AI contradict their intuitions. For instance, although AI suggests that the drug offense type is correlated with "Will violate", P4 thinks that "drug offense is not something serious, a minor offense" and thus disagrees with AI and chooses "Will not violate". With a similar train of thought, P7 asks why AI suggests the violent offense type to be correlated with "Will not violate" and thinks that it should be the other way around. A potential reason is that people are more likely to restrain themselves after serious crimes as the consequence can be dire, but it seemed difficult for the participants to reason about this counterintuitive pattern. The above comments suggest that some patterns that AI identifies can be counterintuitive and thus challenging for humans to make sense of.

Furthermore, participants disagree with AI predictions due to focusing too much on a few patterns they learned from AI. For instance, if a participant learns that *Prior Failure to Appear* positively relates to "Will violate", they will apply the same logic on future examples and disagree with the AI when the pattern and prediction disagrees. Quoting from P9, "The current example has no for *Prior Failure to Appear* and drug offense but the previous examples had yes for *Prior Failure to Appear* and drug offense". P9 then chooses "Will not violate" because of these two features. This observation highlights the importance of paying attention to features globally, which can be challenging for humans.

Finally, participants are more confident in BIOS than in ICPSR as they are able to relate to the task better and understand the explanations provided by the AI better. They believe that the biography text is sufficient to detect the profession, but much of the crucial information is missing in ICPSR. P9 said, "there was more background on what they did in their lives, and how they got there and whatnot, so it helped me make a more educated decision". This observation also extends to their evaluation of AI predictions, quoting from P12, "the AI would be more capable of predicting based on a short snippet about someone than predicting something that hasn't happened".

Strategies in different tasks. Different strategies are employed in different tasks. Since BIOS is a task requiring participants to read a text, most participants look for (highlighted) keywords that distinguish similar professions. For instance, while both professor and teacher teach, participants look for keywords such as "phd" to distinguish them. Similarly, in the case of surgeon and physician, participants look for keywords such as "practice" and "surgery". In ICPSR, as there are only seven features, most participants pay extra attention to a few of them, including *Prior Failure to Appear*, *Prior Convictions*, *Prior Arrest*, and *Offense Type*. We also noticed during the interview that most participants tend to avoid discussing or mentioning sensitive features such as *Race*. In §2.8, we elaborate and discuss findings on an exploratory study on important features identified by participants.

The effect of interactive explanations. Participants could be categorized into two groups according to their use of the interactive console, either they do not experiment with it, or they play with it excessively. Participants in the former group interact with the console only when prompted, while the latter group result in a prolonged user study session. Some participants find the additional value of interactive console limited as compared to static explanations such as highlights. They are unsure of the 'right' way to use it as P12 commented, "I know how it works, but I don't know what I should do. Maybe a few use cases can be helpful.

Like examples of how to use them”. Other participants do not interact much with it, but still think it is helpful. With reference to P6, “I only played with it in the first few examples. I just use them to see the AI’s decision boundaries. Once I get it in training, I don’t need them when I predict.”

Another interesting finding was that while some participants make decisions due to visual factors, others make decisions due to numerical factors. P2 said, “the color and different darkness were really helpful instead of just having numbers”. In contrast, P4, who often made decisions by looking at the numbers, commented on one of the many justifications that the defendant “will not violate because the numbers are low.” This observation suggests that our dynamic highlights may provide extra benefits to static highlights.

Web application feedback. As some participants were unsure of how to use the interactive console and make the most out of it, we added an animated video that showcased an example of using the interactive console on top of the walk-through tutorial that guides a user through each interactive element (see the supplementary materials). We also added a nudging component describing how many instances they have used interactive explanations with to remind participants of using the interactive console (see Fig. 2.6).

In addition to Zoom sessions, we conducted pilot studies on Mechanical Turk before deploying them in large-scale tasks. Since some Zoom sessions took longer than we expected, we wanted to investigate the total time taken for completing 10 training and 20 test instances. We noted from the feedback collected from exit surveys of pilot studies that the training was too time consuming and difficult. We thus reduced the number of training instances and improved the attention check questions and instruction interfaces. See the supplementary materials for details.

2.4.5 Large-scale Experiments with Crowdworkers

Finally, we discuss our setup for the large-scale experiments on Amazon Mechanical Turk. First, in order to understand the effect of out-of-distribution examples, we consider the performance of humans without any assistance as our control setting. Second, another focus of our study is on interactive explanations, we thus compares interactive explanations and static explanations.[9]

Specifically, participants first go through a training phase to understand the patterned embedded in machine learning models, and then enter the prediction phase where we evaluate the performance of human-AI teams. We allow different interfaces in the training phase and in the prediction phase because the ideal outcome is that participants can achieve complementary performance without real-time assistance after the training phase. To avoid scenarios where users experience a completely new interface during prediction, we consider situations where the assistance in training is more elaborate than the real-time assistance in prediction. Therefore, we consider the following six conditions to understand the effect of explanation types during training and prediction (the word before and after "/" refers to the assistance type during training and prediction respectively):

- **None/None.** Participants are not given any form of AI assistance in either the training phase or the prediction phase. In the training phase, there are only two examples instead of 5-6 in other conditions to help participants understand the task. In other words, this condition is a *human-only* condition.

- **Static/None.** Participants are provided static assistance in the training phase. Important features are highlighted in shades of pink/blue and AI predictions are provided. Participants are *not* provided any assistance in the prediction phase.

- **Static/Static.** Participants are provided static assistance in both training and prediction.

[9] A natural question is about the effect of explanations vs. AI assistance without explanations. We refer readers to prior work on this question [Lai et al., 2020, Lai and Tan, 2019, Green and Chen, 2019b].

- **Interactive/None.** Participants are provided interactive explanations during the training phase, and no assistance in the prediction phase.
- **Interactive/Static.** Participants are provided interactive explanations in the training phase and static assistance in the prediction phase.
- **Interactive/Interactive.** Participants are provided interactive explanations in both training and prediction.

We refer to these different conditions as ***explanation type*** in the rest of this paper. The representative examples are the same during training in Interactive and Static. Participants are recruited via Amazon Mechanical Turk and must satisfy three criteria to work on the task: 1) residing in the United States, 2) have completed at least 50 Human Intelligence Tasks (HITs), and 3) have been approved for 99% of the HITs completed. Following the evaluation protocol in prior work [Green and Chen, 2019a,b], each participant is randomly assigned to one of the explanation types, and their performance is evaluated on 10 random in-distribution examples and 10 random out-of-distribution examples. We do not allow any repeated participation. We used the software program G*Power to conduct a power analysis. Our goal was to obtain .95 power to detect a small effect size of .1 at the standard .01 alpha error probability using F-tests. As such, we employed 216 participants for each explanation type, which adds up to 1,296 participants per task. Note that our setup allows us to examine human performance on random samples beyond a fixed set of 20 examples, which alleviates the concern that our findings only hold on a dataset of 20 instances.

The median time taken to complete a HIT is 9 minutes and 22 seconds. Participants exposed to interactive conditions took 12 minutes, while participants exposed to non-interactive conditions took 7 minutes (see Fig. 6.12). Our focus in this work is on human performance, so we did not limit the amount of time in the experiments. Participants were allowed to spend as much time as they needed so that they were able to explore the full capacities of our interface. Participants were paid an average wage of $11.31 per hour. We leave consideration

Table 2.3: Performance comparison between human alone and AI alone. We also add numbers from prior work to contextualize these numbers. Note that AI performance here is slightly different ($\leq$1.2%) from that in Fig. 2.3, because AI performance in this table is calculated from a subset of examples shown in None/None (human alone) while the AI performance in Fig. 2.3 is calculated from the out-of-distribution test set of 180 examples.

Task	IND (typical setup)			OOD (proposed setup)		
	Human	AI	Difference between humans and AI	Human	AI	Difference between humans and AI
ICPSR	60.9	68.4	−7.5	55.9	55.0	0.9
COMPAS	60.0	65.5	−5.5	54.5	56.1	−1.6
BIOS	63.5	84.1	−20.6	68.4	76.6	−8.2
Deception detection [Lai et al., 2020] [Lai and Tan, 2019]	∼51	∼87.0	∼ −36	—	—	—
LSAT [Bansal et al., 2021]	∼58	65	∼ −7	—	—	—
Beer reviews [Bansal et al., 2021]	∼82	84	∼ −2	—	—	—

of efficiency (i.e., maintaining good performance while reducing duration of interactions) for future work.

2.5 RQ1: The Effect of In-distribution and Out-of-distribution Examples on Human Performance

Our first research question examines how in-distribution and out-of-distribution examples affect the performance of human-AI teams. Recall that Bansal et al. [2021] hypothesize that comparable performance is important to achieve complementary performance. Table 2.3 compares the performance of human alone and AI alone in the three prediction tasks both in-distribution and out-of-distribution (we also add tasks from other papers to illustrate the ranges in prior work). The performance gap between human alone and AI alone in ICPSR and COMPAS is similar to tasks considered in Bansal et al. [2021]. In BIOS, the in-distribution performance gap between human alone and AI alone is greater than the tasks in Bansal et al. [2021] but much smaller than deception detection, and the out-of-distribution performance gap between human alone and AI alone becomes similar to LSAT in Bansal et al. [2021]. As a result, we believe that our chosen tasks somewhat satisfy the condition of "comparable performance" and allow us to study human-AI decision making over a variety of performance gaps between human alone and AI alone.

Note that AI performance here is calculated from the random samples shown in None/None (human alone), and is thus slightly different ($\leq$1.2%) from AI performance in Fig. 2.3, which is calculated from the in-distribution and out-of-distribution test set of 180 examples each. To account for this sample randomness and compare human performance in different explanation types for these two distribution types, we need to establish a baseline given the random samples (we show absolute accuracy in the supplementary material as the performance difference without accounting for the baseline is misleading; see Fig. 6.1). Therefore, we calculate the accuracy difference on the same examples between a human-AI team and AI, and use *accuracy gain* as our main metric. Accuracy gain is positive if a human-AI team outperforms AI. In the rest of this paper, we will use *human performance* and *the performance of human-AI teams* interchangeably. Since the results are similar between ICPSR and COMPAS, we show the results for ICPSR in the main paper and include the figures for COMPAS in the supplementary materials (see Fig. 6.2-Fig. 6.6).

Preview of results. To facilitate the understanding of our complex results across tasks, we provide a preview of results before unpacking the details of each analysis. Our results indeed replicate existing findings that AI performs better than human-AI teams in in-distribution examples. However, human-AI teams fail to outperform AI in out-of-distribution examples. The silver lining is that the performance gap between human-AI teams and AI is smaller out-of-distribution than in-distribution. These results are robust across tasks (see Table 2.4 for a summary).

Table 2.4: Summary of results on human-AI team performance.

	IND (typical setup)			OOD (proposed setup)		
	ICPSR	COMPAS	BIOS	ICPSR	COMPAS	BIOS
AI performs better than human-AI teams in in-distribution examples.	✓	✓	✓	—	—	—
Human-AI teams perform better than AI in out-of-distribution examples.	—	—	—	✗	✗	✗
The performance difference between human-AI teams and AI is smaller out-of-distribution than in-distribution.	see the OOD columns			☑	✓	✓

✓: holds ☑: holds in at least half of the explanation types

✗: rejected ☒: rejected in all except one explanation type

Human-AI teams underperform AI in in-distribution examples (see Fig. 2.7). We use t-tests with Bonferroni correction to determine whether the accuracy gain for in-distribution examples is statistically significant. Consistent with Proposition 1, our results show that accuracy gain is negative across all explanation types ($p < 0.001$). In other words, the performance of human-AI teams is lower than AI performance for in-distribution examples. This observation also holds across all tasks, which means that AI may have an advantage in both challenging (ICPSR and COMPAS) and relatively simple tasks (BIOS) for humans if the test set follows a similar distribution as the training set (in-distribution).

Human-AI teams do not outperform AI in out-of-distribution examples, although the accuracy gain out-of-distribution is sometimes positive (see Fig. 2.7). Similarly, we use t-tests with Bonferroni correction to determine whether the accuracy gain for out-of-distribution examples is statistically significant. The results are different than what we expected: humans seldom outperform AI in out-of-distribution examples. Interestingly, we observe quite different results across different tasks. In BIOS, accuracy gain is significantly below 0 across all explanation types ($p < 0.001$). In ICPSR and COMPAS, accuracy gain is occasionally positive, including None/None, Static/Static, Interactive/None, Interactive/Static in ICPSR, and Interactive/None in COMPAS, although none of them is statistically significant. The negative accuracy gain (Static/None) in ICPSR is not significant either. These results suggest that although AI performs worse out-of-distribution than in-distribution, it remains challenging for human-AI teams to outperform AI alone out-of-distribution. The performance of human-AI teams, however, becomes comparable to AI performance in challenging tasks such as recidivism prediction, partly because the performance of AI alone is more comparable to human alone out-of-distribution (e.g., 0.9% in ICPSR vs. -8.2% in BIOS in None/None (human alone) in Fig. 2.7).

Interestingly, Interactive/None leads to the highest accuracy gain in ICPSR, while Interactive/Interactive leads to a tiny negative gain, suggesting interactive explanations as real-time assistance might hurt human performance in ICPSR. We will elaborate on this observation in §2.7.

The performance gap between human-AI teams and AI is smaller in out-of-distribution examples than in in-distribution examples (see Fig. 2.7). We finally examine the difference between in-distribution and out-of-distribution examples. We use two approaches to determine whether there exists a significant difference. First, for each explanation type in each task, we test whether the accuracy gain in out-of-distribution examples is significantly different from that in in-distribution examples with t-tests after Bonferroni correction. In both BIOS and COMPAS, accuracy gain is significantly greater in out-of-distribution examples than in in-distribution examples across all explanation types ($p < 0.001$). In ICPSR, accuracy gain is significantly greater in out-of-distribution examples than in in-distribution examples in all explanation types ($p < 0.001$) except Static/None. Second, we conduct two-way ANOVA based on distribution types and explanation types. We focus on the effect of distribution types here and discuss the effect of explanation types in §2.7. We observe a strong effect of distribution type across all tasks ($p < 0.001$), suggesting a clear difference between in-distribution and out-of-distribution. Note that this reduced performance gap does not necessarily suggest that humans behave differently out-of-distribution from in-distribution, as it is possible that human performance stays the same and the reduced performance gap is simply due to a drop in AI performance. We further examine human agreement with AI predictions to shed light on the reasons behind this reduced performance gap.

In short, our results suggest a significant difference between in-distribution and out-of-distribution, and human-AI teams are more likely to perform well in comparison with AI out-of-distribution. These results are robust across different explanation types. In general, the accuracy gain is greater in recidivism prediction than in BIOS. After all, the in-distribution AI performance in BIOS is much stronger than humans without any assistance. This observation resonates with the hypothesis in Bansal et al. [2021] that comparable performance between humans and AI is related to complementary performance. However, we do not observe complementary performance in our experiments, which suggests that comparable performance between humans and AI alone is insufficient for complementary performance.

2.6 RQ2: Agreement/Trust of Humans with AI

Our second research question examines how well human predictions agree with AI predictions depending on the distribution type. Agreement is defined as the percentage of examples where the human gives the same prediction as AI. Humans have access to AI predictions in Static/Static, Interactive/Static, Interactive/Interactive, so agreement in these explanation types may be interpreted as how much *trust* humans place in AI predictions (we use *overtrust* to refer to agreement with incorrect predictions in all explanation types). Since both ICPSR and COMPAS yield similar results, we show ICPSR results in the main paper and COMPAS in the supplementary materials (see Fig. 6.2-Fig. 6.6).

Preview of results. Different from results in performance, we observe intriguing differences across tasks. Our results show that humans tend to show higher agreement with AI predictions

Table 2.5: Summary of results on agreement with AI. Recall that appropriate agreement refers to humans agreeing with correct AI predictions, and overtrust refers to humans agreeing with incorrect AI predictions.

	IND (typical setup)			OOD (proposed setup)		
	ICPSR	COMPAS	BIOS	ICPSR	COMPAS	BIOS
Agreement is higher in-distribution than out-of-distribution.	see the OOD columns			✓	☑	$\boxed{!}$
Agreement is higher when AI predictions are correct (appropriate agreement) than when AI predictions are wrong (overtrust).	$\boxed{!}$	✗	✓	☑	☑	✓
When AI predictions are correct, agreement (appropriate agreement) is higher in-distribution than out-of-distribution.	see the OOD columns			☑	✗	!
When AI predictions are wrong, agreement (overtrust) is higher in-distribution than out-of-distribution.	see the OOD columns			✓	☑	☒

✓: holds
✗: rejected
!: mostly supported in the reverse direction except one explanation type

☑: holds in at least half of the explanation types
☒: rejected in all except one explanation type
$\boxed{!}$: reversed only in one explanation type

in in-distribution examples than out-of-distribution examples in ICPSR and COMPAS, but not in BIOS. When it comes to appropriate agreement vs. overtrust, the results depend on distribution types. We first compare the extent of appropriate agreement and overtrust in the same distribution type. In out-of-distribution examples, human agreement with AI predictions is higher when AI predictions are correct than when AI predictions are wrong (appropriate agreement exceeds overtrust). But for in-distribution examples, this is only true for BIOS, but false in ICPSR and COMPAS. To further understand these results, we compare appropriate agreement and overtrust in-distribution to out-of-distribution. We find that both appropriate agreement and overtrust are stronger in-distribution than out-of-distribution in ICPSR, but in BIOS, the main statistical significant results are that appropriate agreement is stronger out-of-distribution than in-distribution. See Table 2.5 for a summary.

Humans are more likely to agree with AI on in-distribution examples than out-of-distribution examples in ICPSR and COMPAS, but not in BIOS (see Fig. 2.8). As AI performance is typically better in-distribution than out-of-distribution, we expect humans to agree with AI predictions more often in-distribution than out-of-distribution. To determine whether the difference is significant, we use t-test with Bonferroni correction for each explanation type in Fig. 2.8. In ICPSR, agreement is indeed significantly greater in-distribution than out-of-distribution in all explanation types ($p < 0.001$). In COMPAS, in-distribution agreement is significantly higher in all explanation types ($p < 0.05$ in None/None, $p < 0.01$ in Static/None and Interactive/Interactive, $p < 0.001$ in Interactive/None) except Static/Static and Interactive/Static (see Fig. 6.3). These results suggest that in ICPSR and COMPAS, humans indeed behave more differently from AI out-of-distribution. However, in BIOS, we find the agreement is generally higher for out-of-distribution examples than for in-distribution examples, and the difference is statistically significant in Static/Static ($p < 0.05$). Note that the agreement difference between in-distribution and out-of-distribution is much smaller in BIOS (<4%, usually within 2%) than in ICPSR and COMPAS (~10%).

These results echo observations in our virtual pilot studies that humans are more confident in themselves when detecting professions and are less affected by in-distribution vs. out-of-distribution differences, and may turn to AI predictions out-of-distribution because the text is too short for them to determine the label confidently. In comparison, the fact that humans agree with AI predictions less out-of-distribution than in-distribution in recidivism prediction suggests that humans seem to recognize that AI predictions are more likely to be wrong out-of-distribution than in-distribution in ICPSR and COMPAS. To further unpack this observation, we analyze human agreement with correct AI predictions vs. incorrect AI predictions.

Out-of-distribution appropriate agreement *mostly exceeds* out-of-distribution

overtrust in all of the three tasks; in-distribution appropriate agreement exceeds in-distribution overtrust *only* in BIOS (see Fig. 2.9). We next examine the role of distribution type in whether humans can somehow distinguish when AI is correct from when AI is wrong. First, for each distribution type, we use t-test with Bonferroni correction to determine if humans agree with AI more when AI predictions are correct. Consistent with prior work [Lai and Tan, 2019, Bansal et al., 2021], we find that human-AI teams are more likely to agree with AI when AI predictions are correct than when AI predictions are wrong in most explanation types. This is true both in-distribution and out-of-distribution in BIOS ($p < 0.001$): the agreement gap between correct and incorrect AI predictions is close to 20%, and even reaches 30%-40% out-of-distribution with some explanation types (Fig. 2.9b). In ICPSR and COMPAS, we mostly find significantly greater appropriate agreement than overtrust out-of-distribution. In fact, IND appropriate agreement tends to be lower than IND overtrust, though only significantly in Interactive/Interactive ($p < 0.05$) in ICPSR. In comparison, for out-of-distribution examples, appropriate agreement is significantly higher than overtrust in three explanation types in ICPSR ($p < 0.01$ in None/None, Interactive/None, and Interactive/Static). In COMPAS, appropriate agreement is also significantly higher than overtrust in out-of-distribution examples ($p < 0.05$ in None/None and Interactive/Static, $p < 0.01$ in Static/None and Interactive/None) except Static/Static and Interactive/Interactive (see Fig. 6.4). These results are especially intriguing as they suggest that although the performance of human alone and AI alone is worse out-of-distribution than in-distribution in recidivism prediction, humans can more accurately detect AI mistakes, which explains the small positive accuracy gain in Fig. 2.7.

In-distribution and out-of-distribution appropriate agreement comparison shows different results in each of the three tasks (see Fig. 2.9). We further compare human agreement between in-distribution and out-of-distribution when AI is correct. Similarly, we use t-tests with Bonferroni corrections for each explanation type. Different from our expectation, appropriate agreement is significantly higher out-of-distribution than in-distribution in all explanation types in BIOS except Interactive/Static ($p < 0.001$ in None/None and Static/None; $p < 0.01$ in Static/Static, Interactive/None, and Interactive/Interactive). This is consistent with the observation of higher overall agreement out-of-distribution than in-distribution in BIOS in Fig. 2.8. In ICPSR, appropriate agreement for in-distribution examples is significantly higher than for out-of-distribution examples in all explanation types except None/None ($p < 0.01$ in Interactive/None, Interactive/Static, and Interactive/Interactive, $p < 0.05$ in Static/None and Static/Static). In COMPAS, no significant difference is found between in-distribution and out-of-distribution.

These results suggest that appropriate agreement is stronger out-of-distribution than in-distribution in BIOS. In other words, humans can recognize correct AI predictions better out-of-distribution than in-distribution. This could relate to that humans have higher confidence in their own predictions when the text is longer. As a result, they are more likely to overrule correct AI predictions. However, appropriate agreement is stronger in-distribution than out-of-distribution in ICPSR, which relatively weakens the performance of human-AI teams compared to AI alone out-of-distribution, and suggests that a reduced overtrust is the main contributor to the aforementioned reduced performance gap. In comparison, it seems that in COMPAS, humans simply tend to agree with AI predictions more in-distribution than out-of-distribution, without the ability to recognize when AI predictions are correct.

Overtrust is lower out-of-distribution than in-distribution in ICPSR and COMPAS, but not in BIOS (see Fig. 2.9). In comparison, when AI predictions are wrong, human agreement is significantly lower for out-of-distribution examples than in-distribution examples in all explanation types ($p < 0.001$) in ICPSR. This also holds for some explanation types ($p < 0.01$ in Static/None, Interactive/None, and Interactive/Static) in COMPAS. However, overtrust in in-distribution examples has no significant difference from out-of-distribution examples in BIOS except for None/None ($p < 0.01$). These results suggest that in recidivism prediction, human decisions contradict wrong AI predictions out-of-distribution more accurately than in-distribution, but it is not the case in BIOS.

In summary, the contrast between appropriate agreement and overtrust is interesting as it explains the different stories behind the reduced performance gap out-of-distribution compared to in-distribution in ICPSR and in BIOS: the reduced performance gap in BIOS is mainly attributed to the higher appropriate agreement out-of-distribution, while the reduced performance gap in ICPSR is driven by the lower overtrust out-of-distribution. These results may relate to the task difficulty for humans. Recidivism prediction is more challenging for humans and the advantage of humans may lie in the ability to recognize obvious AI mistakes. In constrast, as humans are more confident in their predictions in BIOS, it is useful that they avoid overruling correct AI predictions. Such asymmetric shifts in agreement rates highlight the complementary insights that humans can offer when working with AI assistance and suggest interesting design opportunities to leverage human expertise in detecting AI mistakes.

2.7 RQ3: The Effect of Interactive Explanations

In this section, we focus on the effect of interactive explanations in human decision making. We revisit human performance and human agreement and then examine human perception of AI assistance's usefulness collected in our exit survey. Finally, for ICPSR and COMPAS, we take a deep look at the most important features reported by humans in the exit survey to understand the limited improvement in the performance of human-AI teams.

Preview of results. In general, we do not find significant impact from interactive explanations with respect to the performance of human-AI team or human agreement with wrong AI predictions, compared to static explanations. However, humans are more likely to find AI assistance useful with interactive explanations than static explanations in ICPSR and COMPAS, but not in BIOS. Table 2.6 summarizes the results.

Real-time assistance leads to better performance than no assistance in BIOS,

Table 2.6: Summary of results on the effect of interactive explanations.

	IND (typical setup)			OOD (proposed setup)		
	ICPSR	COMPAS	BIOS	ICPSR	COMPAS	BIOS
Interactive explanations lead to better human-AI team performance.	✗	✗	✗	✗	✗	✗
Interactive explanations lead to lower human agreement with wrong AI predictions (overtrust).	✗	✗	✗	✗	✗	✗
Human-AI teams are more likely to find AI assistance useful with interactive explanations.	see the OOD columns			☑	☑	✗
✓: holds ✗: rejected	☑: holds in at least half of the explanation types ☒: rejected in all except one explanation type					

but interactive explanations do not lead to better human-AI performance than AI alone (see Fig. 2.7). We conduct one-way ANOVA on explanation type for in-distribution and out-of-distribution separately on human performance due to the clear difference between in-distribution and out-of-distribution. We find that explanation type affects human performance in both distribution types significantly in BIOS ($p < 0.001$), but not in ICPSR ($p = 0.432$ IND, $p = 0.184$ OOD) nor in COMPAS ($p = 0.274$ IND, $p = 0.430$ OOD). We further use Tukey's HSD test to see if differences between explanation types are significant. In BIOS, we find Static/Static, Interactive/Static, and Interactive/Interactive have significantly better performance than None/None, Static/None, and Interactive/None for in-distribution examples ($p < 0.001$). For out-of-distribution examples, we have almost the same observation ($p < 0.05$) except that the difference between Interactive/Static and None/None is no longer significant. These results suggest that real-time assistance in the prediction phase improves human performance in BIOS, consistent with [Lai and Tan, 2019, Bansal et al., 2021], although there is no significant difference between static and interactive explanations. In ICPSR and COMPAS, no significant difference exists between any pair of explanation types. In other words, no explanation type leads to better nor worse human-AI team performance in recidivism prediction.

Interactive explanations do not lead to significantly lower overtrust (see Fig. 2.9). We use one-way ANOVA to determine whether significant differences in overtrust exist between different explanation types. We also do this separately for in-distribution and out-of-distribution examples. We observe a strong effect in all tasks in both distributions ($p < 0.001$). However, Tukey's HSD test shows overtrust in Interactive/Interactive is not statistically different from Static/Static; similarly, Interactive/None is not statistically different from Static/None either. The strong effect comes from the significant differences between explanation types with real-time assistance and those without, likely because predicted labels are shown in real-time assistance. For example, in out-of-distribution examples in BIOS, three explanation types without real-time assistance (None/None, Static/None, Interactive/None) have significantly lower overtrust than the three with real-time assistance (Static/Static, Interactive/Static, Interactive/Interactive) ($p < 0.001$ for most pairs; $p < 0.01$ for Interactive/None vs. Static/Static and Interactive/None vs. Interactive/Static). Similarly, in out-of-distribution examples in ICPSR, None/None and Interactive/None has significantly lower overtrust than Static/Static, Interactive/Static, and Interactive/Interactive ($p < 0.001$ for most pairs; $p < 0.01$ for None/None vs. Interstatic/Static, Interactive/None vs. Static/Static, and Interactive/None vs. Interactive/Static). In fact, Interactive/Interactive has the highest overtrust in both in-distribution and out-of-distribution examples in ICPSR. Results in COMPAS are qualitatively similar (see Fig. 6.3).[10]

These results are contrary to our expectation: interactive explanations do not lead to lower overtrust. In fact, they lead to the highest overtrust in ICPSR, so they may not encourage users to critique incorrect AI predictions. Our observations also resonate with prior work

[10] For in-distribution overtrust, None/None is significantly lower than explanation types with real-time assistance ($p < 0.05$ in Static/Static; $p < 0.001$ in Interactive/Static and Interactive/Interactive). For out-of-distribution overtrust, all explanation types without real-time assistance are significantly lower than Static/Static ($p < 0.05$) and Interactive/Interactive ($p < 0.01$). However, similarly to ICPSR, we do not see significantly lower overtrust in interactive explanations than in static explanations either in-distribution or out-of-distribution.

that shows higher overall agreement with AI predictions when predicted labels are shown [Lai et al., 2020, Lai and Tan, 2019].

Human-AI teams are more likely to find AI assistance useful with interactive explanations in ICPSR and COMPAS, but not in BIOS (see Fig. 2.10). We ask participants whether they find training and real-time assistance useful when applicable. Since only Static/Static, Interactive/Static, and Interactive/Interactive have real-time assistance, we focus our analysis here on these three explanation types. We use one-way ANOVA to test the effect of explanation type for the usefulness of training and real-time AI assistance separately. For training, the effect of explanation type is significant only in COMPAS ($p < 0.05$). With Tukey's HSD test, we find the perception of training usefulness is significantly higher in Interactive/Interactive than in Static/Static ($p < 0.05$). These results show that human-AI team with interactive explanations are more likely to find training useful in COMPAS.

For perception of real-time assistance, explanation type has a significant effect in COMPAS ($p < 0.001$) and ICPSR ($p < 0.001$), but not in BIOS ($p = 0.6$). We also use Tukey's HSD test to determine whether there is a pairwise difference among explanation types. In COMPAS, Interactive/Interactive achieves a significantly higher human perception of real-time assistance usefulness than both Static/Static ($p < 0.001$) and Interactive/Static ($p < 0.05$) (see Fig. 6.5). Perception of Interactive/Static is also significantly higher than that of Static/Static ($p < 0.001$). We find similar results in ICPSR except that the difference between Static/Static and Interactive/Static is not significant. In BIOS, Interactive/Interactive has the highest human perception of AI assistance usefulness, but no significant difference is found. These results suggest that with interactive explanations, human-AI teams perceive real-time assistance as more useful, especially in recidivism prediction. A possible reason is that human perception of usefulness depends on the difficulty of tasks. COMPAS is more challenging than BIOS to humans as recidivism prediction is not an average person's experience, thus interactive explanations may have decreased the difficulty of the task in perception.

Exploratory study on important features. Finally, since there are only seven features in ICPSR and COMPAS, we asked participants to identify the top three most important features that made the biggest influence on their own predictions in the exit survey (see Fig. 6.13 for the wording of all survey questions). We also identify important features based on Spearman correlation as a comparison point. The top three are ("Prior Failure to Appear", "Prior Arrests", "Prior Convictions") in ICPSR, and ("Prior Crimes", "Age", and "Race") in COMPAS. By comparing these computationally important features with human-perceived important features, we can identify potential biases in human perception to better understand the limited performance improvement.

Fig. 2.11a shows the percentage of participants that choose each feature as an important feature for their decisions in ICPSR. We group participants based on explanation types: 1) without interactions (Static/None and Static/Static) and 2) with interactions (Interactive/None, Interactive/Static, and Interactive/Interactive). Humans largely choose the top computationally important features in both groups in ICPSR. We use t-test with Bonferroni correction to test whether there is a difference between the two groups. In ICPSR, we find participants with interaction choose significantly more "Age" and "Offense Type", but less "Prior Convictions" (all $p < 0.01$). In fact, participants with interaction are less likely to choose all of the top three features than those without. In COMPAS (see Fig. 6.6), we find participants with interaction choose significantly more "Race" and "Sex", but less "Charge Degree" ($p < 0.001$ in "Race", $p < 0.05$ in "Sex" and "Charge Degree"). These results suggest that participants with interaction are more likely to fixate on demographic features and potentially reinforce human biases,[11] but are less likely to identify computationally important features in ICPSR and COMPAS.

This observation may also relate to why interactive explanations do not lead to better performance of human-AI teams. We thus hypothesize that participants with interaction make

[11] Race is indeed important in COMPAS, so this might be justified to a certain extent.

more mistakes when they disagree with AI predictions, which can explain the performance difference between Interactive/None and Interactive/Interactive in Fig. 2.7. Fig. 2.8 shows that users disagree with AI predictions less frequently in Interactive/Interactive than in Interactive/None, and Fig. 2.11b further shows that they are indeed more likely to be wrong when they disagree (not statistically significant).

2.8 Discussion

In this work, we investigate the effect of out-of-distribution examples and interactive explanations on human-AI decision making through both virtual pilot studies and large-scale, randomized human subject experiments. Consistent with prior work, our results show that the performance of human-AI teams is lower than AI alone in-distribution. This performance gap becomes smaller out-of-distribution, suggesting a clear difference between in-distribution and out-of-distribution, although complementary performance is not yet achieved. We also observe intriguing differences between tasks with respect to human agreement with AI predictions. For instance, participants in ICPSR and COMPAS agree with AI predictions more in-distribution than out-of-distribution, which is consistent with AI performance differences in-distribution and out-of-distribution, but it is not the case in BIOS. As for the effect of interactive explanations, although they fail to improve the performance of human-AI teams, they tend to improve human perception of AI assistance's usefulness, with an important caveat of potentially reinforcing human biases.

Our work highlights the promise and importance of exploring out-of-distribution examples. The performance gap between human-AI teams and AI alone is smaller out-of-distribution than in-distribution both in recidivism prediction, where the task is challenging and humans show comparable performance with AI, and in BIOS, where the task is easier for both humans and AI but AI demonstrates a bigger advantage than humans. However, complementary performance is not achieved in our experiments, suggesting that out-of-distribution examples and interactive explanations (as we approach them) are not the only missing ingredients. Similarly, comparable performance alone might not be a sufficient condition for complementary performance. While results with respect to human-AI team performance and the effect of interactive explanations are relatively stable across tasks, the intriguing differences in human agreement with AI predictions between tasks demonstrate the important role of tasks and the complexity of interpreting findings in this area. We group our discussion of implications by out-of-distribution experiment design, interactive explanations, and choice of tasks, and then conclude with other limitations.

Out-of-distribution experimental design. The clear differences between in-distribution and out-of-distribution suggest that distribution type should be an important factor when designing experimental studies on human-AI decision making. Our results also indicate that it is promising to reduce the performance gap between human-AI teams and AI for out-of-distribution examples, as AI is more likely to suffer from distribution shift. Out-of-distribution examples, together with typical in-distribution examples, provide a more realistic examination of human-AI decision making and represent an important direction to examine how humans and AI complement each other.

However, it remains an open question of what the best practice is for evaluating the performance of human-AI teams out-of-distribution.[12] To simulate out-of-distribution examples, we use separate bins based on an attribute (age for ICPSR and COMPAS; length for BIOS). Our setup is realistic in the sense that it is possible that age distribution in the training data differs from the testing data and leads to worse generalization performance in out-of-distribution examples in recidivism prediction. Similarly, length is a sensible dimension for distributon mistach in text classification. That said, our choice of separate bins leads to non-overlapping out-of-distribution and in-distribution examples. In practice, the difference between out-of-distribution and in-distribution can be continuous and subtle to

[12] Concurrently with this work, Chiang and Yin [2021] investigates human reliance on machine predictions when humans are aware of distribution shifts.

quantify [Koh et al., 2021]. From an experimental point of view, it is challenging to investiage the effect of out-of-distribution examples on a continuous spectrum, and out-of-distribution examples that are very close to in-distribution examples may not be interesting to study. As a result, it makes sense to zoom in on the challenging out-of-distribution examples and have a clear separation between in-distribution and out-of-distribution. We believe that our design represents a reasonable first attempt in understanding the effect of out-of-distribution examples and future work is required to address the spectrum of out-of-distribution.

Notably, a side effect of our split is that out-of-distribution examples are more difficult than in-distribution examples for humans in recidivism prediction (but not in BIOS; see Fig. 6.1). We encourage future work to examine to what extent this is true in practice and how this shift affects human decision making. Furthermore, out-of-distribution examples might benefit from new feature representations, which humans can extract, pointing to novel interaction with AI. Overall, many research questions emerge in designing experiments and interfaces to effectively integrate humans and AI under distribution shift.

Interactive explanations and appropriate trust in AI predictions. We find that interactive explanations improve human perception of AI assistance but fail to improve the performance of human-AI teams. While the idea of interactive explanations is exciting, our implementation of interactive explanations seems insufficient. That said, our results suggest future directions for interactive explanations: 1) detecting out-of-distribution examples and helping users calibrate their trust in-distribution and out-of-distribution (e.g., by suggesting how similar an example is to the training set); 2) automatic counterfactual suggestions [Wachter et al., 2017] to help users navigate the decision boundary as it might be difficult for decision makers to come up with counterfactuals on their own; 3) disagreement-driven assistance that frames the decision as to whether to agree with AI predictions or not and help decision makers explore features accordingly.

Meanwhile, we show that interactive explanations may reinforce human biases. While this observation is preliminary and further work is required to understand the effect of interactive explanations on human biases, this concern is consistent with prior work showing that explanations, including random ones, may improve people's trust in AI predictions [Lai and Tan, 2019, Bansal et al., 2021, Green and Chen, 2019a,b]. Therefore, it is important to stay cautious about the potential drawback of interactive explanations and help humans not only detect issues in AI predictions but also reflect biases from themselves. Future work is required to justify these interactive explanations to be deployed to support human decision making.

Choice of tasks and the complexity of interpreting findings in human-AI decision making. Our work suggests tasks can play an important role and it can be challenging to understand the generalizability of findings across tasks. We observe intriguing differences with respect to human agreement with AI predictions between recidivism prediction and BIOS. A surprising finding is that humans agree with AI predictions more out-of-distribution than in-distribution in BIOS, despite that AI performs worse out-of-distribution than in-distribution. Furthermore, there exists an asymmetry of human agreement with AI predictions when comparing OOD with IND: the reduced performance gap out-of-distribution in recidivism prediction is because humans are less likely to agree with **incorrect** predictions OOD than IND, but the reduced performance gap in BIOS is due to that humans are more likely to agree with **correct** AI predictions OOD than IND. This asymmetry indicates that humans perform better relatively with AI OOD than IND for different reasons in different tasks. One possible interpretation of this observation is that humans can complement AI in different ways in different tasks. To best leverage human insights, it may be useful to design appropriate interfaces that guide humans to find reasons to respectively reject AI predictions or accept AI predictions.

Moreover, by exploring tasks with different performance gaps, our results suggest that comparable performance alone might not be sufficient for complementary performance, echoing the discussion in Bansal et al. [2021]. These differences could be driven by many possible factors related to tasks, including difficulty levels, performance gap, and human expertise/confidence. Although these factors render it difficult to assess the generalizability of findings across tasks, it is important to explore the diverse space and understand how the choice of tasks may induce different results in the emerging area of human-AI interaction. We hope that our experiments provide valuable samples for future studies to explore the question of what tasks should be used and how findings would generalize in the context of human-AI decision making.

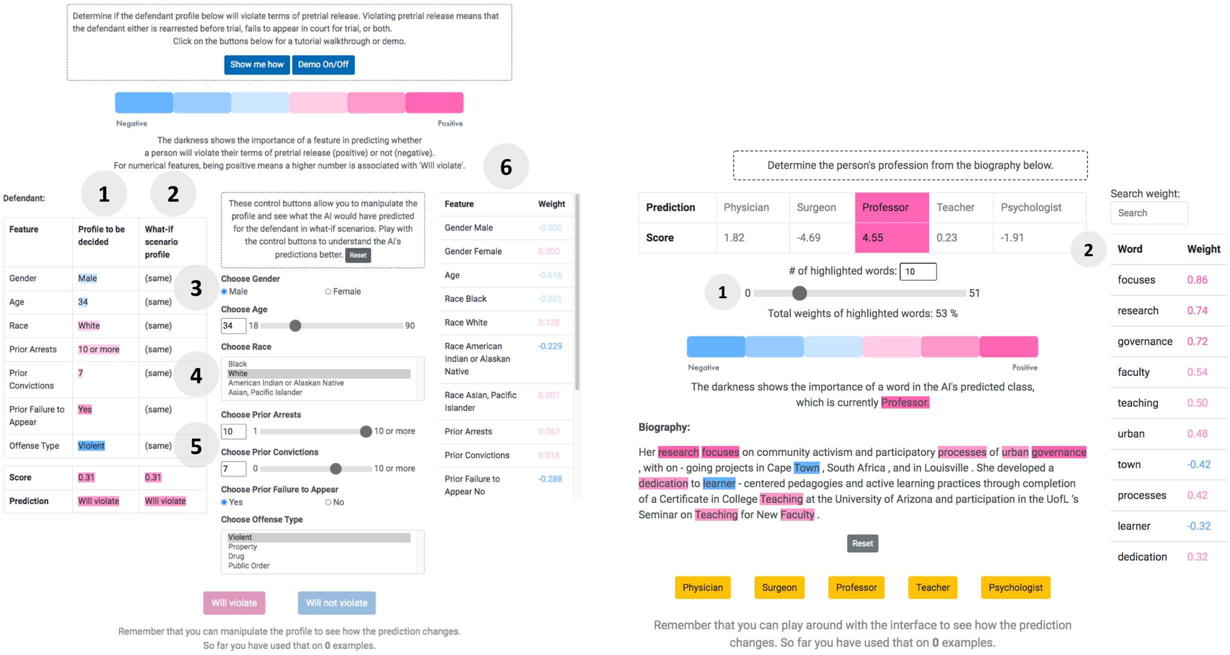
Our choice of tasks is aligned with the discovering mode proposed in Lai et al. [2020], where AI can identify counterintuitive patterns and humans may benefit from AI assistance beyond efficiency. In contrast, humans define the labels in tasks such as question answering and object recognition in the emulating mode, in which case improving performance is essentially improving the quality of data annotation. We argue that improvement in these two cases can be qualitatively different.

We include recidivism prediction because of its societal importance. One might argue that complementary performance is not achieved because crowdworkers are not representative of decision makers in this task (i.e., judges) and recidivism prediction might be too difficult for humans. Indeed, crowdworkers are not the best demographic for recidivism prediction and lack relevant experieince compared to judges. That said, we hypothesized that complementary performance is possible in recidivism prediction because 1) humans and AI show comparable performance, in fact <1% out-of-distribution (as a result, the bar to exceed AI performance out-of-distribution is quite low and the absolute performance is similar to LSAT in Bansal et al. [2021]); 2) prior studies have developed valuable insights on this task with mechanical turkers [Green and Chen, 2019a,b] and mechanical turkers outperform random guessing, indicating that they can potentially offer valuable insights, despite their lack of experience compared to judges. Therefore, we believe that this was a reasonable attempt, although it is possible that the performance of judge-AI teams would differ. As for the difficulty of this task, it is useful to note that this task is challenging for judges as well. This difficulty might have contribued to the elusiveness of complementary performance, but is also why it is especially important to improve human performance in these challenging tasks where human performance is low, ideally while preserving human agency.

To complement recidivism prediction, we chose BIOS because humans including mechanical turkers have strong intuitions about this task and can potentially provide complementary insights from AI. Indeed, mechanical turkers are more likely to override wrong AI predictions in BIOS than in recidivism prediction. However, the performance gap between AI and humans in BIOS might be too big to count as "comparable". As "comparable performance" is a new term, it is difficult to quantify and decide what performance gap constitutes comparable performance.

Model complexity and other limitations. In this work, we have focused on linear models because they are relatively simple to "explain". However, a growing body of work has shown that "explaining" linear models is non-trivial in a wide variety of tasks [Lai et al., 2020, Poursabzi-Sangdeh et al., 2021]. We speculate that the reason is that the relatively simple patterns in linear models are still challenging for humans to make sense of, e.g., why violent crimes are associated with "will not violate pretrial terms". Humans need to infer the reason might be that the consequence is substantial in that scenario. We expect such challenges to be even more salient for complex deep learning models. We leave it to future work for examining the role of model complexity in human-AI decision making.

Our limitations in samples of human subjects also apply to our virtual pilot studies. University students are not necessarily representative of decision makers for each task. Our findings may depend on the sample population, although it is reassuring that both virtual pilot studies and large-scale, randomized experiments show that humans may not identify important features or effectively use patterns identified by AI.



Determine if the defendant profile below will violate terms of pretrial release. Violating pretrial release means that the defendant either is rearrested before trial, fails to appear in court for trial, or both.
Click on the buttons below for a tutorial walkthrough or demo.

The darkness shows the importance of a feature in predicting whether a person will violate their terms of pretrial release (positive) or not (negative).
For numerical features, being positive means a higher number is associated with 'Will violate'.

Defendant:

Feature	Profile to be decided	What-if scenario profile
Gender	Male	(same)
Age	34	(same)
Race	White	(same)
Prior Arrests	10 or more	(same)
Prior Convictions	7	(same)
Prior Failure to Appear	Yes	(same)
Offense Type	Violent	(same)

Score	0.31	0.31
Prediction	Will violate	Will violate

These control buttons allow you to manipulate the profile and see what the AI would have predicted for the defendant in what-if scenarios. Play with the control buttons to understand the AI's predictions better.

Choose Gender
● Male ○ Female

Choose Age
34 18 90

Choose Race
Black
White
American Indian or Alaskan Native
Asian, Pacific Islander

Choose Prior Arrests
10 1 10 or more

Choose Prior Convictions
7 0 10 or more

Choose Prior Failure to Appear
● Yes ○ No

Choose Offense Type
Violent
Property
Drug
Public Order

Feature	Weight
Gender Male	-0.000
Gender Female	0.000
Age	-0.016
Race Black	-0.001
Race White	0.128
Race American Indian or Alaskan Native	-0.229
Race Asian, Pacific Islander	0.007
Prior Arrests	0.067
Prior Convictions	0.018
Prior Failure to Appear No	-0.288

Remember that you can manipulate the profile to see how the prediction changes.
So far you have used that on **0** examples.

(a) Interactive explanation for ICPSR.

Determine the person's profession from the biography below.

Prediction	Physician	Surgeon	Professor	Teacher	Psychologist
Score	1.82	-4.69	4.55	0.23	-1.91

# of highlighted words: 10
0 51
Total weights of highlighted words: 53 %

The darkness shows the importance of a word in the AI's predicted class, which is currently Professor.

Biography:

Her research focuses on community activism and participatory processes of urban governance , with on - going projects in Cape Town , South Africa , and in Louisville . She developed a dedication to learner - centered pedagogies and active learning practices through completion of a Certificate in College Teaching at the University of Arizona and participation in the UofL 's Seminar on Teaching for New Faculty .

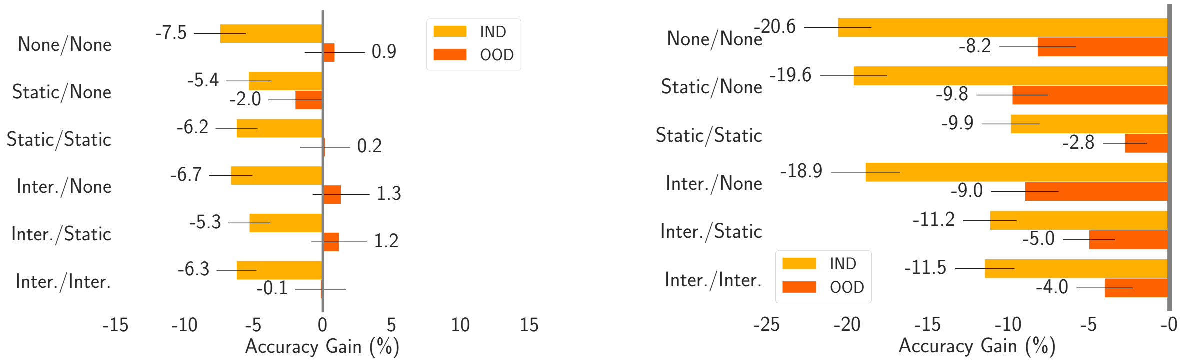
Search weight:

Word	Weight
focuses	0.86
research	0.74
governance	0.72
faculty	0.54
teaching	0.50
urban	0.48
town	-0.42
processes	0.42
learner	-0.32
dedication	0.32

Remember that you can play around with the interface to see how the prediction changes. So far you have used that on **0** examples.

(b) Interactive explanation for BIOS.

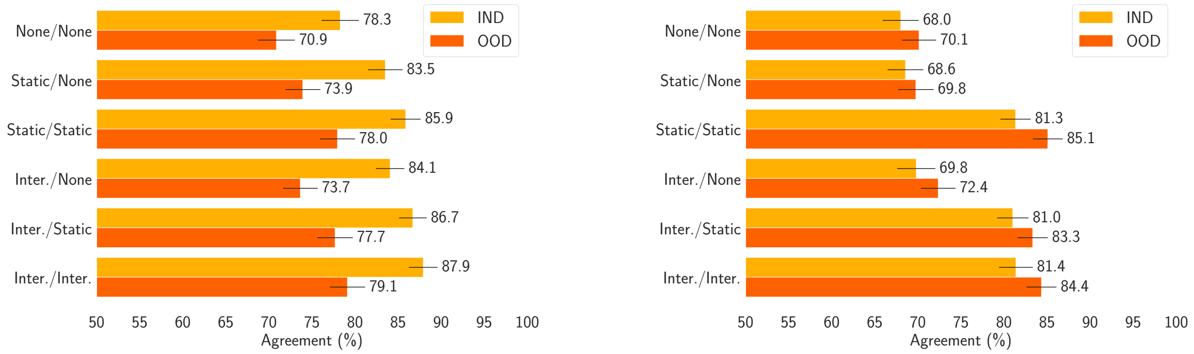
Figure 2.6: Screenshots for interactive explanations in ICPSR and BIOS. In addition to static assistance such as feature highlights and showing AI predictions, users are able to manipulate the features of a defendant's profile to see any changes in the AI prediction in ICPSR. The interactive console for ICPSR includes: 1) the actual defendant's profile; 2) the edited defendant's profile if user manipulates any features; 3) users are able to edit the value of *Gender* and *Prior Failure to Appear* with radio buttons; 4) users are able to edit the value of *Race* and *Offense Type* with dropdown; 5) users are able to edit the value *Age*, *Prior Arrests*, and *Prior Convictions* with sliders; 6) a table displaying features and coefficients, the color and darkness of the color shows the feature importance in predicting whether a person will violate their terms of pretrial release or not. In BIOS, users are able to remove any words from the biography to see any changes in the AI prediction. The interactive console for BIOS includes: 1) user is able to edit the number of highlighted words with a slider; 2) a table displaying features and respective coefficients, the color and darkness of the color shows the importance of a word in the AI's predicted class. The interface for COMPAS is similar to ICPSR (see Fig. 6.9).



(a) Accuracy gain in ICPSR.

(b) Accuracy gain in BIOS.

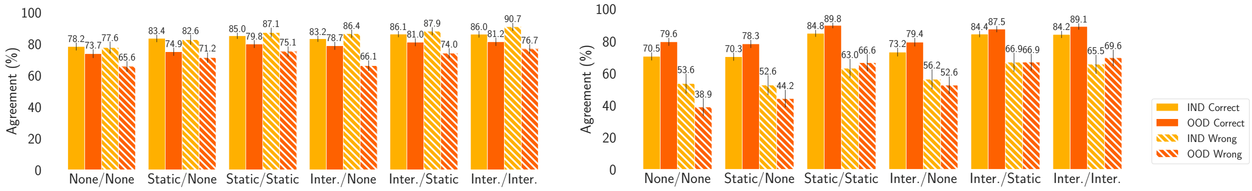
Figure 2.7: Accuracy gain in ICPSR and BIOS. Distribution types are indicated by the color of the bar and error bars represent 95% confidence intervals. All accuracy gains are statistically significantly negative in-distribution, indicating that Human-AI teams underperform AI based on typical random split of training/test sets. However, results are mixed for out-of-distribution examples. While accuracy gain in BIOS is always negative, accuracy gain in ICPSR is sometimes positive (although not statistically significant). The performance gap between human-AI teams and AI is generally smaller out-of-distribution than in-distribution, suggesting that humans may have more complementary insights to offer out-of-distribution. Results in COMPAS are similar to ICPSR and can be found in the supplementary materials.



(a) Agreement with AI predictions in ICPSR.

(b) Agreement with AI predictions in BIOS.

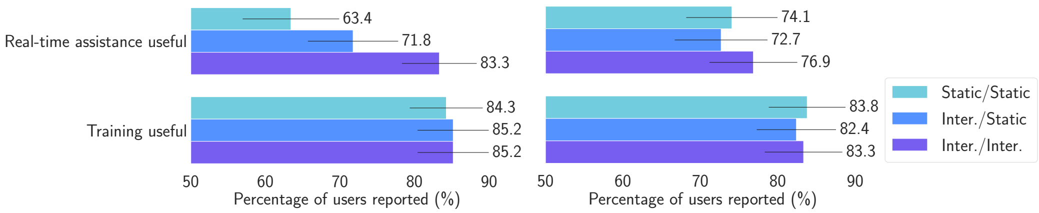
Figure 2.8: Agreement with AI predictions in ICPSR and BIOS. Distribution types are indicated by the color of the bar and error bars represent 95% confidence intervals. In ICPSR and COMPAS, agreement with AI predictions is much higher in-distribution than out-of-distribution. However, this trend is reversed in BIOS. In BIOS, agreement is generally higher in Static/Static, Interactive/Static, and Interactive/Interactive, where AI predictions and explanations are shown. We will discuss the effect of explanation type in §2.7.



(a) ICPSR agreement by correctness.

(b) BIOS agreement by correctness.

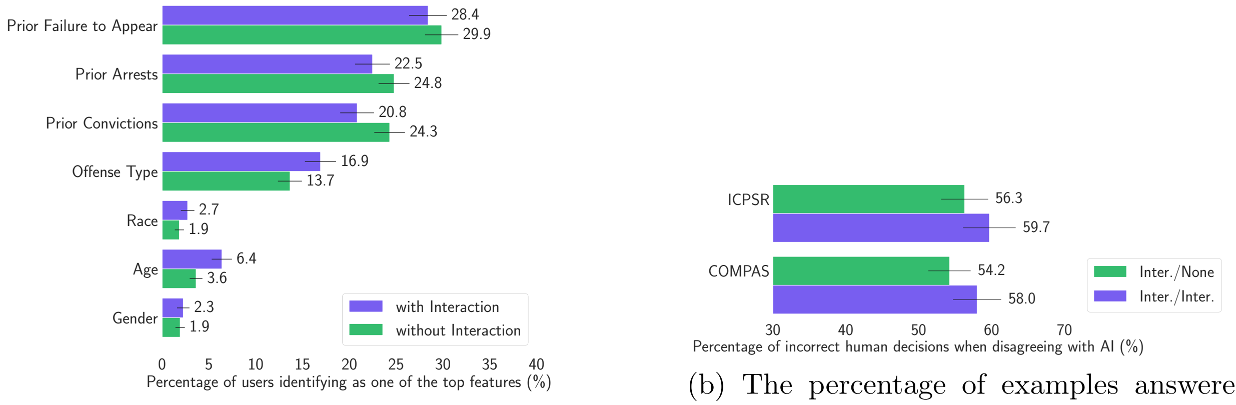
Figure 2.9: Agreement with AI grouped by distribution type and whether AI predictions are correct. Distribution types are indicated by the color of the bar, bars with stripes represent wrong AI predictions, and error bars represent 95% confidence intervals. A notable observation is that when AI is wrong, humans are significantly less likely to agree with AI predictions out-of-distribution than in-distribution in ICPSR and COMPAS, but it is not the case in BIOS.



(a) ICPSR.

(b) BIOS.

Figure 2.10: Human perception on whether real-time assistance is useful and whether training is useful. x-axis shows the percentage of users that answered affirmatively. Error bars represent 95% confidence interval.



(a) Percentage of participants finding a feature important in ICPSR.

(b) The percentage of examples answered wrongly by participants when they disagree with AI predictions.

Figure 2.11: The features in 2.11a are sorted in descending order from top to bottom by their Spearman correlation with groundtruth labels.

CHAPTER 3

LEARNING HUMAN-COMPATIBLE REPRESENTATIONS FOR CASE-BASED DECISION SUPPORT

3.1 Overview

Algorithmic case-based decision support provides examples to aid people in decision making tasks by providing contexts for a test case. Despite the promising performance of supervised learning, representations learned by supervised models may not align well with human intuitions: what models consider similar examples can be perceived as distinct by humans. As a result, they have limited effectiveness in case-based decision support. In this work, we incorporate ideas from metric learning with supervised learning to examine the importance of alignment for effective decision support. In addition to instance-level labels, we use human-provided triplet judgments to learn human-compatible decision-focused representations. Using both synthetic data and human subject experiments in multiple classification tasks, we demonstrate that such representation is better aligned with human perception than representation solely optimized for classification. Human-compatible representations identify nearest neighbors that are perceived as more similar by humans and allow humans to make more accurate predictions, leading to substantial improvements in human decision accuracies (17.8% in butterfly vs. moth classification and 13.2% in pneumonia classification).

Unlike the previous chapter, which focuses helping humans understand AI systems, this chapter focuses on helping AI systems understand humans. Most of the work in this chapter is published in Liu et al. [2023]. This is a joint work with Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan.

3.2 Introduction

Despite the impressive performance of machine learning (ML) models, humans are often the final decision maker in high-stake domains due to ethical and legal concerns [Lai and Tan, 2019, Green and Chen, 2019b], so ML models as decision support is preferred over full automation. In order to provide meaningful information to human decision makers, the model cannot be illiterate in the underlying problem, e.g., a model for assisting breast cancer radiologists should have a high diagnostic accuracy by itself. However, a model with high *autonomous* performance may not provide the most effective decision support, because it could solve the problem in a way that is not comprehensible or even perceptible to humans, e.g., AlphaGo's famous move 37 [Silver et al., 2016, 2017, Metz et al., 2016]. Our work studies the relation between these two objectives that effective decision support must balance: achieving high autonomous performance and aligning with human intuitions.

We focus on case-based decision support for classification problems [Kolodneer, 1991, Begum et al., 2009, Liao, 2000, Lai and Tan, 2019]. For each test example, in addition to showing the model's predicted label, case-based decision support shows one or more related examples retrieved from the training set. These examples can be used to justify the model's prediction, e.g., by showing similar-looking examples with the predicted label, or to help human decision makers calibrate its uncertainty, e.g., by showing similar-looking examples from other classes. Both use cases require the model to know what is similiar-looking to the human decision maker. In other words, an important consideration in aligning with human intuition is approximating human judgment of similarity.

Figure 3.1 illustrates the importance of such alignment on a classification problem of distinguishing butterfly from moth. A high-accuracy ResNet [He et al., 2016] produces a highly linearly-separable representation space, which leads to high classification accuracy. But the nearest neighbor cannot provide effective justification for model prediction because it looks dissimilar to the test example for humans. The similarity measured in model representation

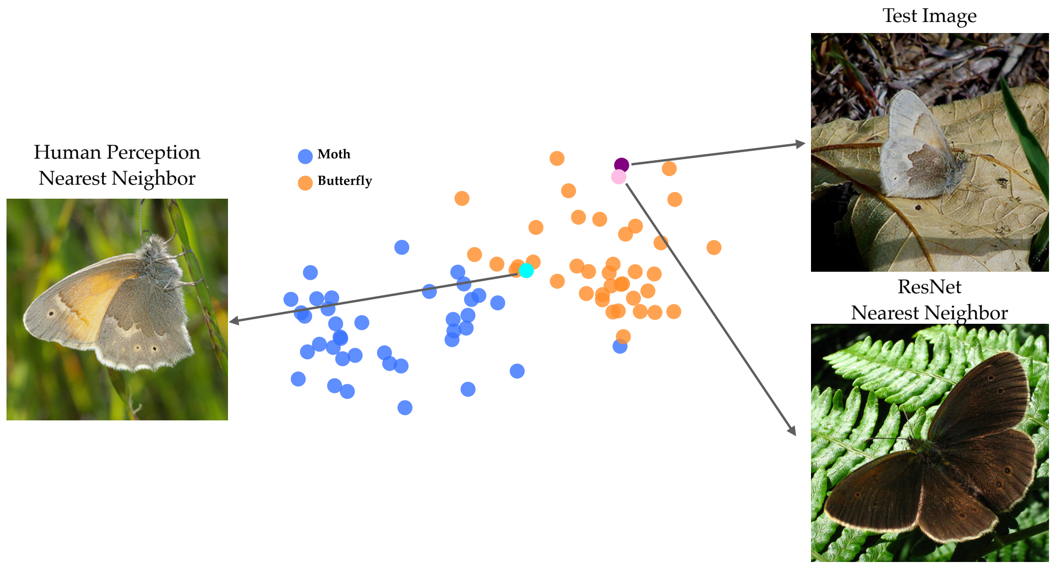


Figure 3.1: Nearest neighbor retrieved by the model representation might not align with human similarity judgment. The MLE representations (512-dim) are visualized using t-SNE [Van der Maaten and Hinton, 2008]. The purple circle represents a specific test instance. The nearest neighbor found by MLE representations (pink circle) is not as visually similar as the instance in cyan circle found by optimizing a metric learning objective.

space does not align with human visual similarity. If we instead use representations from a second model trained specifically to mimic human visual similarity rather than to classify images, the nearest neighbor would provide strong justification for the model prediction. However, using the second model for decision support has the risk of misleading or even deceiving the human decision maker because the "justification" is generated based on a representation space that is different from the model used to predict the label; it becomes persuasion rather than justification.

The goal of this work is to learn a *single* representation space that satisfies two properties: (i) producing easily separable representations for different classes to support accurate classification, and (ii) constituting a metric space that is aligned with human perception of similarity between examples. Simultaneously matching the best model on classification accuracy and achieving perfect approximation of human similarity might not be possible, but we hypothesize that a good trade-off between the two would benefit decision support. We propose a novel multi-task learning method that combines supervised learning and metric learning. We supplement the standard maximum likelihood objective with a triplet margin loss function from Balntas et al. [2016]. Our method learns from human annotations of similarity judgments among data instances in the triplet form.

We validate our approach with both synthetic data and user study. We show that representations learned from our framework identify nearest neighbors that are perceived as more similar by the synthetic human than that based on supervised classification (henceforth *MLE representations*, see §3.3 for details), and are therefore more suitable to provide decision support. We further demonstrate that the advantage of human-compatible representations indeed derives from human perception rather than data augmentation.

We further conduct human subject experiments using two classification tasks: (i) butterfly vs. moth classification from ImageNet [Krizhevsky et al., 2012], and (ii) pneumonia classification based on chest X-rays [Kermany et al., 2018]. Our results show that human-compatible representations provide more effective decision support than MLE representations. In particular, human-compatible representations allow laypeople to achieve an accuracy of 79.1% in pneumonia classification, 15.3% higher than MLE representations. A similar improvement has been observed on the butterfly vs. moth classification task (34.8% over MLE representations and 17.8% over random).

To summarize, our main contributions include:

- We highlight the importance of alignment in learning human-compatible representations for case-based decision support.
- We propose a multi-task learning framework that combines supervised learning and metric learning to simultaneously learn classification and human visual similarity.
- We design a novel evaluation framework for comparing representations in decision support.
- Empirical results with synthetic data and human subject experiments demonstrate the effectiveness of our approach.

3.3 Case-Based Decision Support

Consider the problem of using a classification model $h : \mathcal{X} \to \mathcal{Y}$ as decision support for humans. Simply showing the predicted label from the model provides limited information. Explanations are commonly hypothesized to improve human performance by providing additional information [Doshi-Velez and Kim, 2017]. We focus on information presented in the form of examples from the training data, also known as case-based decision support [Kolodneer, 1991, Begum et al., 2009, Liao, 2000, Lai and Tan, 2019]. Case-based decision support can have diverse use cases and goals. Given a test example (x) and its predicted label ($\hat{y}$), two common use cases are:

- Presenting the nearest neighbor of x along with label $\hat{y}$ as a justification of the predicted label. We refer to this scenario as *justification* [Kolodneer, 1991].
- Presenting the nearest neighbor in each class without presenting $\hat{y}$. This approach makes a best-effort attempt to provide evidence and leaves the final decision to humans, without biasing humans with the predicted label. We refer to this scenario as *neutral decision support* [Lai and Tan, 2019].

Formulation. Building on Kolodneer [1991], we formalize the problem of case-based decision support in the context of representation learning. The goal is to assist humans on a classification problem with groundtruth $f : \mathcal{X} \to \mathcal{Y}$. We assume access to a representation model g, which takes an input $x \in \mathcal{X}$ and generates an m-dimensional representation $g(x) \in \mathbb{R}^m$. For each test instance x, an example selection policy π chooses k labeled examples from the training set D^{train} and shows them to the human (optionally along with the labels); the human then makes a prediction by choosing a label from $\mathcal{Y}$. As discussed in the two common use cases, we consider nearest-neighbor-based selection policies in this work. The focus of this work is thus on the effectiveness of g for case-based decision support.

Given a neural classification model $h : \mathcal{X} \to \mathcal{Y}$, the representation model is the last layer before the classification head, which is a byproduct derived from h. We refer to this model as $e(h)$.[1] In justification, the example selection policy is $\pi = \mathrm{NN}(x, e(h), D^{\text{train}}_{\hat{y}})$, where $\hat{y} = h(x)$, $D^{\text{train}}_{\hat{y}}$ refers to the subset of training data with label $\hat{y}$ (i.e., $\{(x, y) \in D^{\text{train}} \mid y = \hat{y}\}$), and NN finds the nearest neighbor of x using representations from $e(h)$ among the subset of examples with label $\hat{y}$. In decision support, the example selection policy is $\{\mathrm{NN}(x, e(h), D^{\text{train}}_{y}),\ \forall y \in \mathcal{Y}\}$.

Misalignment with human similarity metric is detrimental. We argue that aligning model representations with human similarity metric is crucial for case-based decision support; we refer to it as the *metric alignment problem.* To illustrate the importance of alignment, we need to reason about the goal of case-based decision support. Let us start with justification, which is a relatively easy case. To justify a predicted label, the chosen example should ideally *appear similar* to the test image. Crucially, this similarity is perceived by humans (i.e., interpretable), and the example selection policy identifies the nearest neighbor based on model representation (i.e., faithful). The gap between human representation and model representation (Fig. 3.1) leads to undesirable justification.

Neutral decision support, however, represents a more complicated scenario. We start by emphasizing that the goal is not simply to maximize human decision accuracy, because one may use policies that intentionally show distant examples to nudge or manipulate humans towards making a particular decision.[2] Choosing the nearest neighbors in each class is thus an attempt to present *faithful* and *neutral* evidence from the representation space so that humans can make their own decisions, hence preserving their agency. Therefore, the chosen nearest neighbors should be visually similar to the test instance by human perception, again highlighting the potential gap between model representation and human representation.

[1] In general, we can use the representation in any layer, but in preliminary experiments, we find representation from the last layer is most effective.

[2] We will consider one such policy for the sake of evaluating the quality of representations in §3.4.

Assuming that humans follow the natural strategy by picking the presented instance that's most *similar* to the test instance and answering with the corresponding label, then ideally, nearest neighbors in each class retain key information useful for classification so that they can reveal the separation learned in the model.

It is unlikely that we get high alignment by solely optimizing classification even when the model's classification accuracy is comparable to the human's. Models trained with supervised learning almost always exploit patterns in the training data that are (i) not robust to distribution shifts, and (ii) counterintuitive or even unobservable for humans [Ilyas et al., 2019, Xiao et al., 2020].

Combining metric learning on human triplets with supervised classification. We propose to address the metric alignment problem with additional supervision on the human similarity metric. We collect data in the form of human similarity judgment triplets (or *triplets* for short). Each triplet is an ordered tuple: (x^r, x^+, x^-), which indicates x^+ is judged by humans as being closer to the reference x^r than x^- [Balntas et al., 2016]. Given a triplet dataset T and labeled classification dataset D, we learn a model θ using triplet margin loss [Balntas et al., 2016] in conjunction with cross-entropy loss, controlled by a hyperparameter λ:

$$\underbrace{\lambda\left[-\sum_{(x,y)\sim D}\log\left(p_\theta(y|x)\right)\right]}_{\text{Cross-entropy loss}}+(1-\lambda)\underbrace{\left[\sum_{(x^r,x^+,x^-)\sim T}\max\left(d_\theta(x^r,x^+)-d_\theta(x^r,x^-)+1,0\right)\right]}_{\text{Triplet margin loss}} \tag{3.1}$$

where $d_\theta(\cdot,\cdot)$ is the similarity metric based on model representations; we use Euclidean distance. In this work, we initialize θ with a pretrained ResNet [He et al., 2016]. When $\lambda = 1$ and the triplet margin loss is turned off, the model reduces to a finetuned ResNet. When $\lambda = 0$ and the cross-entropy loss is turned off, the model reduces to the triplet based-learning model of Balntas et al. [2016]; we call it `TMLModel` and will use it to simulate humans in some synthetic experiments in the appendix. Our work is concerned with the representations learned by these models. Our approach uses the representations learned with $\lambda = 0.5$ (henceforth *human-compatible representations* and `HC` for short). We refer to the representations fine-tuning ResNet with the cross-entropy loss as *MLE representations* (`MLE` for short) and the representations from `TMLModel` as `TML`.

3.4 Experimental Setup

In this section, we provide the specific model instantiation and detailed experiment setup.

Models. All models and baselines use ResNet-18 [He et al., 2016] pretrained on ImageNet as the backbone image encoder. Following Chen et al. [2020], we take the output of the average pooling layer and feed it into an MLP projection head with desired embedding dimension. We use the output of the projection head as our final embeddings (i.e., representations), where we add task-specific head and loss for training and evaluation. We use Euclidean distance as the similarity metric for both loss calculation and distance measurement during example selection in decision support.

Our first baseline uses representations from ResNet finetuned with classification labels using cross-entropy loss (i.e., `MLE`). ResNet typically achieves high classification accuracy but does not necessarily produce human-aligned representations. Our second baseline uses representations from the same pretrained model finetuned with human triplets using triplet margin loss [Balntas et al., 2016] (i.e., `TML`). We expect `TML` to produce more aligned representations but achieve lower classification accuracy than `MLE` and may provide limited effectiveness in decision support.

Our representations, `HC`, are learned by combining the two loss terms following Equation 3.1. The hyperparameter λ controls the trade-off between metric alignment and classification accuracy: with higher λ we expect `HC` to be more similar to `MLE`, while lower λ steers `HC` towards `TML`. Empirically tuning λ confirms this hypothesis. For the main paper, we present results with $\lambda = 0.5$. More details about model specification and hyperparameter tuning can be found in the appendix.

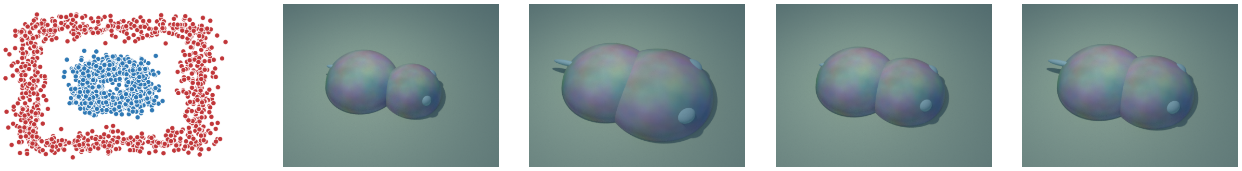
Filtering classification-inconsistent triplets. Human triplets may not always align with classification: triplet annotators may choose the candidate from the incorrect class over the one from the correct class. We refer to these data points as *classification-inconsistent triplets*. We consider a variant of human-compatible representations where we isolate human intuition that's compatible with classification and remove these classification-inconsistent triplets from the training set; we refer to this condition as `HC-filtered`. Filtering is yet another way to strike a balance between human intuition and classification. We leave further details on filtering in the appendix.

Evaluation metrics. Our method is designed to align representations with human similarity metrics and at the same time retain the representations' predictive power for classification. We can evaluate these representations with classification and triplet accuracy using existing data, but our main evaluation is designed to simulate case-based decision support scenarios.

- **Head-to-head comparisons** ("**H2H**"). To evaluate justification, we set up head-to-head comparisons between two representations (R_1 vs. R_2) and ask: given a test instance and two justifications retrieved by R_1 and R_2, which justification do humans consider as closer to the test instance? We report the fraction of rounds that R_1 is preferable. In addition to the typical justification for the predicted label, we also examine that for classes other than the predicted class, as those examples will be used in decision support for users to examine the plausibility of each class. We refer to the nearest example in the *predicted* class as *NI*, and the nearest example in the other class as *NO*.

- **Neutral decision support**. Following §3.3, we retrieve the nearest neighbors from each class. We use the accuracy of humans as the measure of effective decision support.

- **Persuasive decision support**. We retrieve the nearest example with the predicted label and the furthest example from the other class. If the representation is aligned with human similarity metric, this approach encourages people to follow the predicted label, which



(a) Decision boundary (b) Vespula 1 (c) Vespula 2 (d) Weevil 1 (e) Weevil 2

Figure 3.2: VW dataset. (a) shows the dataset where labels are determined (non-linearly) by two features: the head and the body size of the fictional insects. (b)-(d) show samples of the two classes; the Weevil has a mid-sized body and mid-sized head, while the Vespula does not. Tail length and texture are two non-informative features.

likely leads to over-reliance and may be unethical in practice. Here, we use this scenario as a surrogate to evaluate the quality of the learned representations.

Note that we do not show model predictions so that humans focus on the similarity between examples.

3.5 Synthetic Experiment

To understand the strengths and limitations of our method, we first experiment with synthetic datasets. Using simulated human similarity metrics, we control and vary the level of disagreement between the classification groundtruth and the synthetic human's knowledge.

3.5.1 Synthetic Dataset and Simulated Human Similarity Metrics

We use the synthetic dataset "Vespula vs Weevil" (VW) from Chen et al. [2018b]. It is a binary image classification dataset of two fictional species of insects. Each example contains four features, two of them—head and body size—are predictive of the label, and the other two—tail length and texture—are completely non-predictive. We generate 2000 images and randomly split the dataset into training, validation, and testing sets in a 60%:20%:20% ratio. The labels are determined by various synthetic decision boundaries, such as the one shown in Fig. 3.2a.

To generate triplets data, we define simulated human similarity metrics as a weighted Euclidean distance over the visual features: for any instance a and b, $d(a, b) = \sqrt{\sum_i w_i(a_i - b_i)^2}$, where i refers to the i-th feature. By changing the weight of each feature, we can control the level of disagreement between a synthetic human and the groundtruth. All procedures that involve humans (i.e., triplet data collection and evaluation) are done by the synthetic human in this section.

To quantify the disagreement, we use 1-NN classification accuracy following the synthetic human similarity metric; we refer to it as the *task alignment score*. Note that this is different from our main alignment problem, which is about the representations. The task alignment score ranges from 50% (setting the informative features' weights to 0 and distractor weights to 1) to 100%. See the appendix for more details on how we generate these weights. In each setting, we generate 40,000 triplets.

3.5.2 Results

We compare `HC`, `MLE`, `TML` on classification accuracy, triplet accuracy, and decision support performance for the synthetic human. We train all three representations with a large dimension of 512 and a small dimension of 50 and observe that the 512-dimension representation is preferable based on most metrics. We also train `HC` on filtered vs. unfiltered triplets as well as with different values λ. For our main results, we report the performance with $\lambda = 0.5$ and filtered triplets for the decision boundary in Fig. 3.2a. We will discuss the effect of filtering later in this section. λ's role is relatively limited and we will discuss its effect and other decision boundaries in the appendix.

In synthetic experiments, `HC` achieves the same perfect classification accuracy as `MLE` (100%), and a triplet accuracy of 96.8%, which is comparable to `TML` (97.3%). This shows that `HC` indeed learns both the classification task and human similarity prediction task. We next present the evaluation of case-based decision support with the synthetic human, which is the key goal of this work.

`HC` significantly outperforms `MLE` in H2H. If there is no difference between `HC` and `MLE`,

Table 3.1: Experiment results on VW with H2H comparison and decision support evaluations.

Task alignment	50%	80%	83%	92%	92.5%	100%
Weights	[0,0,1,1]	[1,0,1,1]	[0,1,1,1]	[1,256,256,256]	[256,1,256,256]	[1,1,1,1]
NI-H2H						
HC vs. MLE	0.917	0.914	0.903	0.880	0.872	0.808
NO-H2H						
HC vs. MLE	0.916	0.968	0.946	0.958	0.962	0.970
Neutral decision support						
MLE	0.753	0.899	0.896	0.897	0.901	0.929
TML	0.568	0.775	0.807	0.868	0.877	**1.000**
HC	**0.759**	**0.901**	**0.928**	**0.949**	**0.955**	**1.000**
Persuasive decision support						
MLE	0.704	0.900	0.903	0.903	0.901	0.919
TML	0.906	0.881	0.863	0.876	0.877	**1.000**
HC	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

the synthetic human should prefer HC about 50% of times. However, as shown in Table 3.1, our synthetic human prefer HC over MLE by a large margin (about 90% of times) as justifications for both nearest in-class examples and nearest out-of-class examples, indicating the NIs and NOs selected based on the HC representations are more aligned with the synthetic human than MLE. For NI H2H, the preference towards HC declines as the task alignment improves, because if alignment between human similarity and classification increases, MLE can capture human similarity as a byproduct of classification.

HC **provides the best decision support.** Table 3.1 shows that HC achieves the highest neutral and persuasive decision support accuracies in all task alignments. In neutral decision support, MLE consistently outperforms TML, highlighting that representation solely learned for metric learning is ineffective for decision support. For all models, the decision support performance improves as the task alignment increases, suggesting that decision support is easier when human similarity judgment is aligned with the classification task. MLE and TML are more comparable in persuasive decision support, while HC consistently achieves 100%. The fact that MLE shows comparable performance between neutral and persuasive decision support further confirms that `MLE` does not capture human similarity for examples from different classes.

Filtering triplets leads to better decision support. Fig. 3.3 shows that filtering class-inconsistent triplets improves `HC`'s decision support performance across all alignments. Further details in the appendix show that filtering slightly hurts H2H performance. This suggests that in terms of decision support, the benefit of filtering out human noise may outweigh the loss of some similarity judgment.

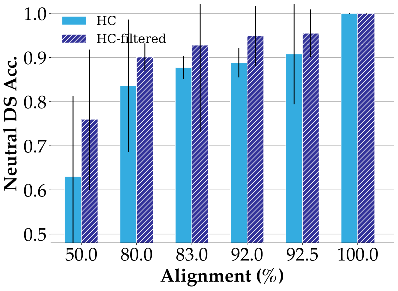


Figure 3.3: Neutral decision support with `HC` and `HC-filtered`. `HC-filtered` leads to improved performance.

The importance of human perception. One may question whether filtering class-inconsistent triplets essentially provides additional label supervision in the form of triplets. We show this is not the case by experimenting with `HC` trained on *label-derived triplets*. Assuming that an instance is more similar to another instance with the same label than one with a different label, we derive label-derived triplets directly from groundtruth labels (x^+ from the same class as x^r and x^- from the other class), containing no human perception information. Table 3.2 shows decision support results for this setting: `HC` label-derived triplets show worse performance than `HC-filtered`. In fact, `HC` label-derived triplets show even worse neutral and persuasive decision support than `MLE`, which may be due to label-derived triplets causing overfitting. This suggests that triplets without human perception do not lead to human-compatible representations.

We also experiment with `HC` trained on same-class triplets, human-triplets but only those where the non-reference cases (x^+, x^-) are from the same class; that is, the triplets cannot provide any label supervision. We observe from Table 3.2 that `HC` trained on these triplets show similar results to `HC-filtered` across all decision support evaluations. This suggests that human perception is the main factor in driving human-compatible representations' high

Table 3.2: Experiment results on VW using synthetic human with 92% alignment. Comparing MLE representations and `HC-filtered` with `HC` trained on label-derived triplets and `HC` trained on same-class triplets. 40,000 new triplets were generated for each condition.

Evaluations	`MLE`	`HC` label-derived triplets	`HC` same-class triplets	`HC-filtered`
NI-H2H with `MLE`	N/A	0.509	0.890	0.889
NO-H2H with `MLE`	N/A	0.607	0.970	0.958
Neutral DS	0.897	0.723	0.960	0.949
Persuasive DS	0.903	0.803	0.998	1.000

decision support performance.

3.6 Human Subject Experiments

We conduct human subject experiments on two image classification datasets: a natural image dataset, Butterflies v.s. Moths (BM) and a medical image dataset of chest X-rays (CXR). For BM, we followed Singla et al. [2014] and acquired 200 images from ImageNet [Krizhevsky et al., 2012]. BM is a binary classification problem and each class contains two species. CXR is a balanced binary classification subset taken from Kermany et al. [2018] with 3,166 chest X-ray images that are labeled with either normal or pneumonia. We randomly split the datasets following 60%:20%:20% ratio. The classification accuracy with our base supervised learning models are 97.5% for BM and 97.3% for CXR. We only present results with human subjects in the main paper, but results from simulation experiments with `TML` as a synthetic agent, such as filtering triplets providing better results, are qualitatively consistent. See §3.13 and §3.14 in the appendix for more details.

3.6.1 Triplet Annotation

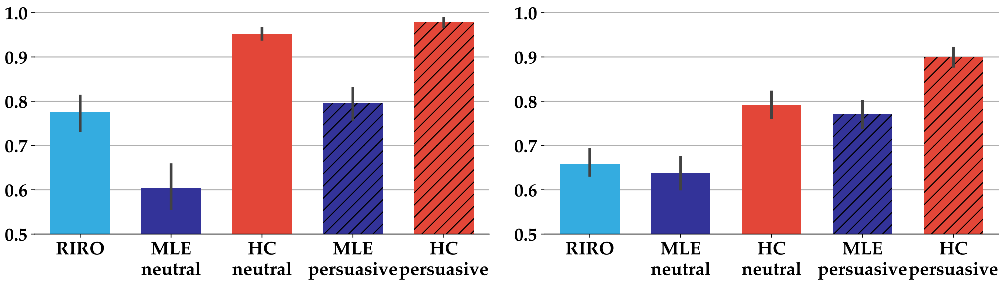
We recruit crowdworkers on Prolific to acquire visual similarity triplets. In each question, we show a reference image on top and two candidate images below, and ask a 2-Alternative-Forced-Choice (2AFC) question: which candidate image looks more similar to the reference image? A screenshot of the interface can be found in the appendix. To generate triplets for annotation, we first sample the reference image from either the training, the validation, or the test set. Then for each reference image, we sample two candidates from the training set. We sample the candidates only from the training set because in decision support, the selected examples should always come from the training set, and thus we only need to validate and test triplet accuracies with candidates from the training set.

For BM we recruit 80 crowdworkers, each completing 50 questions, giving us 4000 triplets. For CXR we recruit 100 crowdworkers, each answering 20 questions, yielding 2000 triplets. Our pilot study suggests that visual similarity judgment on chest X-rays is a more mentally demanding task, so we decrease the number of questions for each CXR survey.

3.6.2 Results on Butterflies v.s. Moths

We recruit crowdworkers on Prolific to evaluate representations produced by our models by doing decision support tasks. We acquire examples with different example selection policies from `HC` and `MLE`. We choose the dimension and training triplets of the representation based on the models' classification accuracy, triplet accuracy, and decision support simulation results based on synthetic agents. See more details in the appendix. We do not include `TML` in human studies, because in practice, `TML` models cannot make predictions on class labels, therefore are unable to distinguish and select in-class and out-of-class examples and thus cannot be used for decision support.

H2H comparison results show `HC` NI examples are slightly but significantly preferred over `MLE` NI examples according to human visual similarity. We recruit 30 Prolific workers to make H2H comparisons between `HC` NI examples and `MLE` NI examples over the entire test set. The mean preference for `HC` over `MLE` is 0.5316 with a 95% confidence interval of ±0.0302 ($p = 0.0413$ with one-sample t-test). This means the `HC` NI examples are closer to the test images than `MLE` NI examples with statistical significance according to human visual similarity.



(a) Butterfly vs. Moth.

(b) Pneumonia classification.

Figure 3.4: Decision support accuracy with human subject studies. Error bars show 95% confidence intervals. HC dominates MLE in both neutral and persuasive decision support.

Decision support results show HC is significantly better than MLE both in neutral and persuasive decision support. Combining two example selection policies with two representations, we have four conditions: HC neutral, HC persuasive, MLE neutral, MLE persuasive. We also add a baseline condition with random supporting examples, which we call random in-class random out-of-class (RIRO). We recruit 30 Prolific workers for each condition and ask them to go through the images in the test set with supporting examples from each class in the training set. Both the order of the test images and the order of the supporting images within each test question are randomly shuffled.

Figure 3.4a shows the human classification accuracies with different decision support scenarios and different representations. In neutral decision support, we observe that HC achieves much higher accuracy than MLE (95.3% vs. 60.5%, $p = 4\text{e}{-19}$ with two-sample t-test). In fact, even RIRO provides better decision support than MLE representations, suggesting that the supporting images based on MLE are confusing and hurt human decision making (77.5% vs. 60.5%, $p = 3\text{e}{-6}$). As expected, the accuracies are generally higher in persuasive decision support. HC enables an accuracy of 97.8%, which is much better than MLE at 79.5% ($p = 2\text{e}{-13}$). HC in neutral decision support already outperforms MLE in persuasive decision support. These findings confirm our results with VW synthetic experiments that human-compatible representations provide much better decision support than MLE representations.

3.6.3 Results on Chest X-rays

We use the same experimental setup as BM to evaluate HC and MLE representations in CXR. **H2H comparison results show HC NI examples are slightly preferred over MLE NI examples but the difference is not statistically significant.** We recruit 50 Prolific workers to each make 20 H2H comparisons between HC NI examples and MLE NI examples. The mean preference for HC over MLE is 0.516 with a 95% confidence interval of ± 0.0725 ($p = 0.379$ with one-sample t-test). H2H comparison in CXR is especially challenging as laypeople need to differentiate between two chest X-rays in the same class, hence the slightly worse performance in H2H compared to BM.

Similar to BM, HC outperforms MLE in both neutral and persuasive decision support in CXR. As expected, Fig. 3.4b shows that pneumonia classification is a much harder task than butterfly vs. moth classification, indicated by the lower accuracies across all conditions. In neutral decision support, HC enables much better accuracy than MLE (79.1% vs. 63.8%, $p = 2\text{e}{-8}$ with two-sample t-test). In fact, similar to the BM setting, MLE provides similar performance with RIRO (63.8% vs. 65.9%, $p = 0.390$), suggesting that MLE representations are no different from random representations for selecting nearest neighbors within a class. To contextualize our results, we would like to highlight that our crowdworkers are laypeople and have no medical training. It is thus impressive that human-compatible representations enable an accuracy of almost 80% in neutral decision support, which demonstrates the potential of human-compatible representations.

In persuasive decision support, HC provides the highest decision support accuracy at 90.0%, also much higher than MLE at 77.0% ($p = 2\text{e}{-10}$). Again, while we do not recommend persuasive decision support as a policy for decision support in practice, these results show that our human-compatible representations are indeed more compatible with humans than MLE representations.

3.7 Related Work

Ordinal embedding. The ordinal embedding problem [Ghosh et al., 2019, Van Der Maaten and Weinberger, 2012, Kleindessner and von Luxburg, 2017, Kleindessner and Luxburg, 2014, Terada and Luxburg, 2014, Park et al., 2015] seeks to find low-dimensional representations that respect ordinal feedback. Currently, there exist several techniques for learning ordinal embeddings. Generalized Non-metric Multidimensional Scaling [Agarwal et al., 2007] takes a max-margin approach by minimizing hinge loss. Stochastic Triplet Embedding [Van Der Maaten and Weinberger, 2012] assumes the Bradley-Terry-Luce noise model [Bradley and Terry, 1952, Luce, 1959] and minimizes logistic loss. The Crowd Kernel [Tamuz et al., 2011] and t-STE [Van Der Maaten and Weinberger, 2012] propose alternative non-convex loss measures based on probabilistic generative models. These results are primarily empirical and focus on minimizing prediction error on unobserved triplets. In principle, one can plugin these approaches in our framework as alternatives to the triplet margin loss in Eq. 3.1.

AI explanations and AI-assisted decision making. Various explanation methods have been developed to explain black-box AI models [Guidotti et al., 2018], such as feature importance [Ribeiro et al., 2016, Shrikumar et al., 2017], saliency map [Zhou et al., 2016, Selvaraju et al., 2017], and decision rules [Ribeiro et al., 2018]. Example-based explanations are also a type of common explanation methods that use examples to explain AI models. Nearest-neighbor examples can explain a model's local decision [Wang and Yin, 2021, Nguyen et al., 2021, Taesiri et al., 2022, Lai and Tan, 2019]. To the best of our knowledge, there has been no prior work that examines the role of representations in choosing the nearest neighbors in the context of AI explanations. Meanwhile, global example-based explanations such as prototypes can explain a model's global behavior or a model's understanding of the data distribution [Kim et al., 2016, Chen et al., 2018a, Cai et al., 2019a, Lai et al., 2020]. Explaining a model's global behavior is also closely related to machine teaching [Zhu et al., 2018].

Many of these explanation methods have been used in AI-assisted decision making to explain AI predictions or inform users about the AI model or training data [Lai et al., 2021]. Among them, example-based explanations have shown be useful in many high-stake domains where full AI automation is often not desired, such as recidivism prediction [Hayashi and Wakabayashi, 2017] and medical diagnosis [Cai et al., 2019c, Rajpurkar et al., 2020, Tschandl et al., 2020]. While many of the current literature in AI-assisted decision making focus on generating explanations of AI without considering human feedback, our decision support methods offer assistance by learning from human perceptions and provide examples from human-compatible representations.

3.8 Conclusion

Our work formulates the novel problem of learning human-compatible representations for case-based decision support. As we identify in this paper, the key to providing effective case-based support with a model is the alignment between the model and the human in terms of similarity metrics: two examples that appear similar to the model should also appear similar to the human. But models trained to perform classification do not automatically produce representations that satisfy this property. To address this issue, we propose a multi-task learning method to combine two sources of supervision: labeled examples for classification and triplets of human similarity judgments. With synthetic experiments and user studies, we validate that human-compatible representations (i) consistently get the best of both worlds in classification accuracy and triplet accuracy, (ii) select visually more similar examples in head-to-head comparisons, (iii) and provide better decision support.

3.9 Ethics Statement

Although coming from a genuine goal to improve human-AI collaboration by aligning AI models with human intuition, our work may have potential negative impacts for the society. We discuss these negative impacts from two perspectives: the multi-task learning framework and the decision support policies.

Multi-task learning framework

Our human-compatible representations models are trained with two sources of data. The first source of data is classification annotations where groundtruth maybe be derived from scientific evidence or crowdsourcing with objective rules or guidelines. The second source of data is human judgment annotations where groundtruth is probably always acquired from crowdworkers with subjective perceptions. When our data is determined with subjective perceptions, the model that learns from it may inevitably develop bias based on the sampled population. If not carefully designed, the human judgment dataset may contain bias against certain minority group depending on the domain and the task of the dataset. For example, similarity judgment based on chest X-rays of patients in one gender group or racial group may affect the generalizability of the representations learned from it, and may lead to fairness problems in downstream tasks. It is important for researchers to audit the data collection process and make efforts to avoid such potential problems.

Decision support policies

Among a wide variety of example selection policies, our policies to choose the decision support examples are only attempts at leveraging AI model representations to increase human performance. We believe that they are reasonable strategies for evaluating representations learned by a model, but future work is required to establish their use in practice.

The neutral decision support policy aims to select the nearest examples in each class, therefore limiting the decision problem to a small region around the test example. We hope this policy allow human users to zoom in the local neighborhood and scrutinize the difference between the relatively close examples. In other words, neutral decision support help human users develop a local decision boundary with the smallest possible margin. This could be useful for confusing test cases that usually require careful examinations. However, the neutral decision support policy adopts an intervention to present a small region in the dataset and may downplay the importance of global distribution in human users' decision making process.

The persuasive decision support policy aims to select the nearest in-class examples but the furthest out-of-class examples. It aims to maximize the visual difference between examples in opposite class, thus require less effort for human users to adopt case-based reasoning for classification. It also helps human users to develop a local decision boundary with the largest possible margin. However, when model prediction is incorrect, the policy end up selecting the furthest in-class examples with the nearest out-of-class examples, completely contrary to what it is design to do, may lead to even over-reliance or even adversarial supports.

In general, decision support policies aim to choose a number of supporting examples without considering some global properties such as representativeness and diversity. While aiming to reduce humans' effort required in task by encouraging them to make decision in a local region, the decision support examples do not serve as a representative view of the whole dataset, and may bias human users to have a distorted impression of the data distribution. It remains an open question that how to ameliorate these negative influence when designing decision support interactions with case-based reasoning.

3.10 Code and Data

Our code and data are available at `https://github.com/ChicagoHAI/learning-human-compatible-representations`.

3.11 Limitations

We discuss some of the limitations in our work.

Limitations of decision support policies. Our decision support policies are simple first steps towards a more general example selection policy for decision support. There are certain limitations of our selection policies. For example in this work, we only look at selecting two examples from the two classes in binary image classifcation tasks. We encourage future work to explore more selection methods towards effective decision-support.

In addition to the ethical concerns discussed in the main paper and the ethics statement, our neutral decision support and persuasive decision support policies have different limitations and use cases. Neutral decision support selects the nearest example from each class. Therefore when a test example lies too close to the decision boundary, the test example, in-class example, and out-of-class example may appear too similar to be distinguished by humans. This is where we may need to select examples further away with different features so that users are more likely to spot the distinction. Persuasive decision support selects the most similar example in the predicted class and the least similar example in the other class, the latter of which has a risk of being an outlier. This may invite biases about the data distribution of the other class and degrade effectiveness of decision support.

Limitations of experimenting with crowdworkers. There are several limitations of experimenting with crowdworkers. First, crowdworkers may not invest as much time as domain experts in the tasks. Therefore, collected triplets may come from superficial or the salient features among the images. Second, crowdworkers or in general lay people have limited domain knowledge such as basic anatomy of body parts when working with medical image. Therefore it is less likely for them to notice the most important feature in the images. In our CXR task, we mitigate this limitation by providing an instruction and quiz section before our main study that provides basic information about how to examine chest X-rays. However, in other tasks, we may need to provide more detailed instructions and quizzes to help crowdworkers understand the task and in this way polish collected triplets.

As the expertise level of the end users increases, HC should be able to learn a high-quality representation. The effectiveness of our decision support methods may vary due to experts strong domain knowledge, but we would still expect our human-compatible representation to provide more effective decision support than MLE representations.

Our ultimate goal is to apply our method to domain experts. We start with crowdworkers and the positive results are encouraging. We hope these results could be used to convince and invite more domain experts to get involved and work towards an applicable system together in the future.

Limitations of design choices in the algorithm. A number of decision choices were made in the algorithm. For example, we use Euclidean distance as the distance metric to be learned for the representation space. Experimenting with different kinds of metrics (e.g., in the psychology literature) and exploring the effectiveness of their respective representations in decision support would be an interesting future direction.

We used ResNet as the backbone network for feature extraction of images due to its competitiveness and popularity. Although model architecture is not the main concern of this paper, one could also plug in other common backbones such as DenseNet [Huang et al., 2017] and ViT [Dosovitskiy et al., 2021] into our representation learning algorithm. We leave the exploration of additional architecture and the effectiveness of their learned representation on decision support to future work.

3.12 Synthetic Experiment Results

3.12.1 Hyperparameters

For our `MLE` backbone we use We use different controlling strength between classification and human judgment prediction, including λs at 0.2, 0.5, and 0.8, and discuss the effect of λ in

Table 3.3: Classification and triplet accuracy of human-compatible representations with different λ. `TMLModel` has no classfication head and no classification accuray.

Model	**Classification accuracy**	**Triplet accuracy**
`MLE`	0.998 ± 0.003	0.673 ± 0.014
`HC` $\lambda = 0.8$	0.998 ± 0.032	0.970 ± 0.024
`HC` $\lambda = 0.5$	0.995 ± 0.000	0.972 ± 0.004
`HC` $\lambda = 0.2$	0.996 ± 0.016	0.973 ± 0.039
`TML`	N/A	0.973 ± 0.016

the next section. In contrast to the experiments on BM, we observe that human-compatible representations with 512-dimension embedding shows overall better performance than human-compatible representations with 50-dimension embedding and show results for the latter in the next section. We use the Adam optimizer [Kingma and Ba, 2014] with learning rate $1e-4$. We use a training batch size of 40 for triplet prediction, and 30 for classification.

3.12.2 Additional Results

Classification and triplet accuracy. Table 3.3 shows how tuning λ affects human-compatible representations's classification and triplet accuracy. Higher λ drives human-compatible representations to behave more simlar to MLE representations while lower human-compatible representations is more similar to `TMLModel`.

Experiment results on VW with confidence intervals. Table 3.4 presents results on VW with human-compatible representations $\lambda = 0.5$. This is is simply Table 1 in the main paper with 0.95 confidence intervals.

Results for different λ. In Table 3.5 and Table 3.6 we show experiment results with human-compatible representations using $\lambda = 0.2$ and $\lambda = 0.8$. We do not observe a clear trend between λ and evaluation metric performances. In the main paper we present human-compatible representations with $\lambda = 0.5$ as it shows best overall performance.

Number of triplets. We examine the effect of the number of triplets, showing the results

Table 3.4: Experiment results on VW. Models use 512-dimension embeddings; HC uses $\lambda = 0.5$ and filtered triplets. This is the same table as Table 3.1 and adds confidence intervals.

Alignments	50%	80%	83%	92%	92.5%	100%
Weights	[0,0,1,1]	[1,0,1,1]	[0,1,1,1]	[1,256,256,256]	[256,1,256,256]	[1,1,1,1]
			NI-H2H			
HC vs. MLE	0.917 ± 0.064	0.914 ± 0.007	0.903 ± 0.016	0.880 ± 0.022	0.872 ± 0.020	0.808 ± 0.017
			NO-H2H			
HC vs. MLE	0.916 ± 0.093	0.968 ± 0.011	0.946 ± 0.009	0.958 ± 0.031	0.962 ± 0.008	0.970 ± 0.008
			Neutral decision support			
MLE	0.753 ± 0.056	0.899 ± 0.025	0.896 ±0.044	0.897 ± 0.045	0.901 ± 0.025	0.929 ± 0.028
TML	0.568 ± 0.049	0.775 ± 0.084	0.807 ± 0.038	0.868 ± 0.012	0.877 ± 0.025	**1.000** ± 0.000
HC	**0.759** ± 0.080	**0.901** ± 0.016	**0.928** ± 0.099	**0.949** ± 0.034	**0.955** ± 0.027	**1.000** ± 0.00
			Persuasive decision support			
MLE	0.704 ± 0.028	0.900 ± 0.017	0.903 ±0.017	0.903 ± 0.017	0.901 ± 0.017	0.919 ± 0.016
TML	0.906 ± 0.011	0.881 ± 0.043	0.863 ± 0.044	0.876 ± 0.027	0.877 ± 0.076	**1.000** ± 0.000
HC	**1.000** ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000

in Fig. 3.5. We decrease number of triplets by powers of 2 and find that H2H preference towards human-compatible representations indeed declines as HC is less human-compatible with fewer training data. As for decision support, in neutral decision support HC performance declines and eventually approaches MLE representations except an outlier in the end, while in persuasive decision support HC performance is able to stay 100% even as the number of triplets declines.

Additional details on weight generation. We generate alignment scores by searching through weight combinations of the simulated human visual similarity metrics. We search the weights in powers of 2, from 0 to 2^{10}, producing a sparse distribution of alignments (Fig. 3.6). Increasing search range to powers of 10 produces smoother distribution, but the weights are also more extreme and unrealistic. We note that the alignment distribution may vary across different datasets. In our experiments we choose weights and alignments to be as representative to the distribution as possible.

Table 3.5: Experiment results on VW. Models using 512-dimension embeddings; HC uses $\lambda = 0.2$ and filtered triplets.

Alignments	50%	80%	83%	92%	92.5%	100%
Weights	[0,0,1,1]	[1,0,1,1]	[0,1,1,1]	[1,256,256,256]	[256,1,256,256]	[1,1,1,1]
			NI-H2H			
HC vs. MLE	0.920 ± 0.005	0.890 ± 0.032	0.906 ± 0.053	0.895 ± 0.016	0.862 ± 0.254	0.832 ± 0.058
			NO-H2H			
HC vs. MLE	0.901 ± 0.439	0.948 ± 0.095	0.970 ± 0.019	0.972 ± 0.095	0.933 ± 0.154	0.981 ± 0.040
			Neutral decision support			
MLE	**0.753** ± 0.056	0.899 ± 0.025	0.896 ±0.044	0.897 ± 0.045	0.901 ± 0.025	0.929 ± 0.028
TML	0.568 ± 0.049	0.775 ± 0.084	0.807 ± 0.038	0.868 ± 0.012	0.877 ± 0.025	**1.000** ± 0.000
HC	0.740 ± 0.540	**0.925** ± 0.127	**0.933** ± 0.064	**0.935** ± 0.000	**0.945** ± 0.349	**1.000** ± 0.000
			Persuasive decision support			
MLE	0.704 ± 0.028	0.900 ± 0.017	0.903 ±0.017	0.903 ± 0.017	0.901 ± 0.017	0.919 ± 0.016
TML	0.906 ± 0.011	0.881 ± 0.043	0.863 ± 0.044	0.876 ± 0.027	0.877 ± 0.076	**1.000** ± 0.000
HC	**0.996** ± 0.016	**0.995** ± 0.000	**0.998** ± 0.000	**0.996** ± 0.016	**0.995** ± 0.000	0.995 ± 0.032

3.12.3 Additional Decision Boundaries

We create a variant of the VW dataset where the labels are populated by a linear separator. We refer to this dataset as VW-Linear (Fig. 3.7). We find the results are overall similar to the original VW data.

Classification and triplet accuracy. Table 3.9 shows classification and triplet accuracy of tuning λ, showing a similar trend to the previous experiment.

H2H and decision support results In Table 3.10 we present results with the best set of hyperparameter: filtered triplets, 512-dimension embedding, $\lambda = 0.5$. We show results for $\lambda = 0.2$ in Table 3.5 and $\lambda = 0.8$ in Table 3.6.

Similar to the experiment on VW square decision boundary, we see no clear relation between λ, embedding dimension and our evaluation metrics.

Table 3.6: Experiment results on VW. Models using 512-dimension embeddings; HC uses $\lambda = 0.8$ and filtered triplets.

Alignments	50%	80%	83%	92%	92.5%	100%
Weights	[0,0,1,1]	[1,0,1,1]	[0,1,1,1]	[1,256,256,256]	[256,1,256,256]	[1,1,1,1]
NI-H2H						
HC vs. MLE	0.916 ± 0.082	0.869 ± 0.217	0.891 ± 0.029	0.879 ± 0.066	0.853 ± 0.164	0.828 ± 0.138
NO-H2H						
HC vs. MLE	0.902 ± 0.193	0.944 ± 0.093	0.959 ± 0.005	0.956 ± 0.090	0.942 ± 0.026	0.969 ± 0.034
Neutral decision support						
MLE	**0.753** ± 0.056	**0.899** ± 0.025	0.896 ±0.044	0.897 ± 0.045	0.901 ± 0.025	0.929 ± 0.028
TML	0.568 ± 0.049	0.775 ± 0.084	0.807 ± 0.038	0.868 ± 0.012	0.877 ± 0.025	**1.000** ± 0.000
HC	0.740 ± 0.095	0.894 ± 0.111	**0.929** ± 0.079	**0.960** ± 0.032	**0.923** ± 0.127	**1.000** ± 0.000
Persuasive decision support						
MLE	0.704 ± 0.028	0.900 ± 0.017	0.903 ±0.017	0.903 ± 0.017	0.901 ± 0.017	0.919 ± 0.016
TML	0.906 ± 0.011	0.881 ± 0.043	0.863 ± 0.044	0.876 ± 0.027	0.877 ± 0.076	**1.000** ± 0.000
HC	**0.998** ± 0.032	**0.995** ± 0.000	**0.998** ± 0.000	**0.998** ± 0.032	**0.995** ± 0.000	**0.999** ± 0.016

3.13 Human Subject Study on Butterflies v.s. Moths

3.13.1 Dataset

Our BM dataset include four species of butterflies and moths including: Peacock Butterfly, Ringlet Butterfly, Caterpiller Moth, and Tiger Moth. An example of each species is shown in Fig 3.8.

3.13.2 Hyperparameters

We use different controlling strength between classification and human judgment prediction, including λs at 0.2, 0.5, and 0.8. We use the Adam optimizer [Kingma and Ba, 2014] with learning rate $1e-4$. Our training batch size is 120 for triplet prediction, and 30 for classification. All models are trained for 50 epoches. The checkpoint with the lowest validation total loss in each run is selected for evaluations and applications.

Table 3.7: Experiment results on VW. Models use 512-dimension embeddings; HC uses $\lambda = 0.5$ and unfiltered triplets.

Alignments	50%	80%	83%	92%	92.5%	100%
Weights	[0,0,1,1]	[1,0,1,1]	[0,1,1,1]	[1,256,256,256]	[256,1,256,256]	[1,1,1,1]
NI-H2H						
HC vs. MLE	0.921 ± 0.015	0.900 ± 0.035	0.920 ± 0.023	0.895 ± 0.008	0.867 ± 0.034	0.846 ± 0.016
NO-H2H						
HC vs. MLE	0.951 ± 0.034	0.969 ± 0.024	0.991 ± 0.002	0.991 ± 0.004	0.958 ± 0.010	0.980 ± 0.023
Neutral decision support						
MLE	**0.753** ± 0.056	**0.899** ± 0.025	**0.896** ±0.044	**0.897** ± 0.045	**0.901** ± 0.025	0.929 ± 0.028
TML	0.568 ± 0.049	0.775 ± 0.084	0.807 ± 0.038	0.868 ± 0.012	0.877 ± 0.025	**1.000** ± 0.000
HC	0.603 ± 0.051	0.801 ± 0.025	0.848 ± 0.053	0.880 ± 0.000	0.880 ± 0.081	**1.000** ± 0.000
Persuasive decision support						
MLE	0.704 ± 0.028	0.900 ± 0.017	0.903 ±0.017	0.903 ± 0.017	0.901 ± 0.017	0.919 ± 0.016
TML	0.906 ± 0.011	0.881 ± 0.043	0.863 ± 0.044	0.876 ± 0.027	0.877 ± 0.076	**1.000** ± 0.000
HC	**0.996** ± 0.004	**0.999** ± 0.004	**0.996** ± 0.004	**0.996** ± 0.004	**0.996** ± 0.004	0.997 ± 0.004

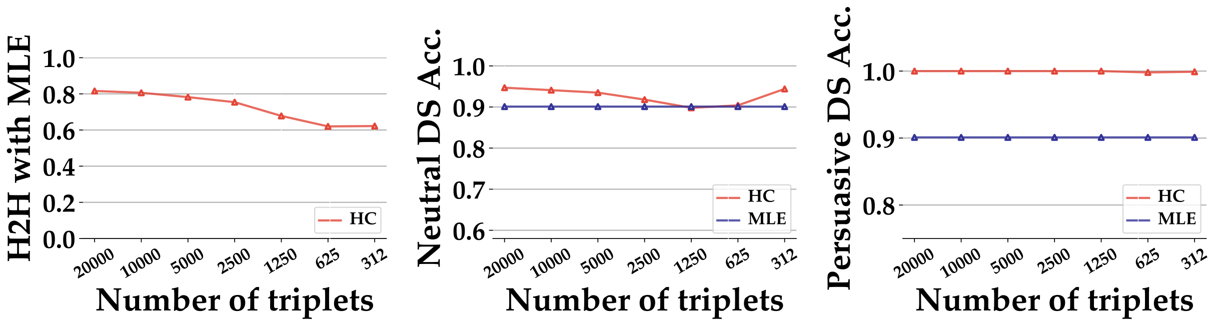


Figure 3.5: HC performance declines as the number of triplets decreases, but shows strong persuasive decision support accuracy even with very few triplets.

3.13.3 Classification and Triplet learning/Accuracy

We present the test-time classification and triplet accuracy of our models in Table 3.13. Both MLE and HC achieve above 97.5% classification accuracy. HC in the 512-dimension unfiltered setting achieve 100.0% classification accuracy. Both TML and HC achieve above 70.7% triplet accuracy. Both TML and HC achieve the highest triplet accuracy in the 50-dimension unfiltered setting with triplet accuracy at 75.9% and 76.2% respectively. Filtering out class-inconsistent triplets removes 15.75% of the triplet annotations in this dataset.

We also evaluate the pretrained LPIPS metric [Zhang et al., 2018] on our triplet test set

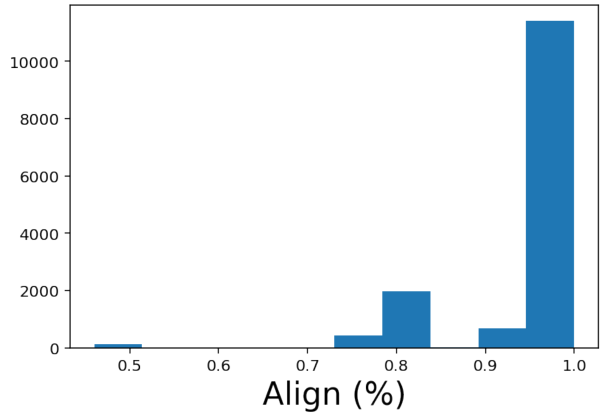


Figure 3.6: Histogram of alignments generated by searching informative weights in powers of 2.

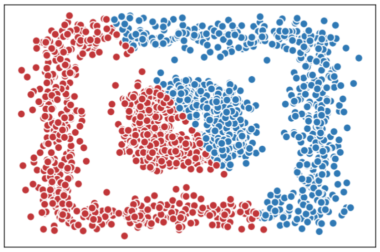


Figure 3.7: VW-Linear

Table 3.9: HC performance with different λ on VW linear decision boundary data.

Model	**Classification accuracy**	**Triplet accuracy**
MLE	0.993 ± 0.003	0.673 ± 0.014
HC $\lambda = 0.8$	0.988 ± 0.032	0.968 ± 0.030
HC $\lambda = 0.5$	0.978 ± 0.013	0.966 ± 0.007
HC $\lambda = 0.2$	0.978 ± 0.032	0.970 ± 0.010
TML	N/A	0.976 ± 0.012

as baselines for learning perceptual similarity. Results with AlexNet backbone and VGG backbone are at 54.5% and 55.0% triplet accuracy respectively, suggesting that TML and HC provides much better triplet accuracy in this task.

3.13.4 Effect of Triplet Amount and Type

We evaluate the effect of the number of triplets on our models in Fig. 3.9. Similar to the VW experiments, H2H preference towards human-compatible representations and neutral decision support performance decrease as the number of triplets decreases. Human-compatible representations achieve strong persuasive decision support performance even with very few triplets.

Table 3.10: Experiment results on VW-Linear. Models use 512-dimension embeddings; HC uses $\lambda = 0.5$ and filtered triplets.

Alignments	56%	84%	95%	98.5%
Weights	[0,1,1,1]	[1,0,1,1]	[1,1,1,1]	[32,256,1,1]
		NI-H2H		
HC vs. MLE	0.913 ± 0.023	0.922 ± 0.008	0.899 ± 0.020	0.848 ± 0.055
		NO-H2H		
HC vs. MLE	0.932 ± 0.034	0.960 ± 0.027	0.921 ± 0.013	0.928 ± 0.034
		Neutral decision support		
MLE	0.778 ± 0.084	0.792 ± 0.144	0.839 ± 0.130	0.927 ± 0.019
TML	0.554 ± 0.175	0.770 ± 0.318	0.950 ± 0.095	0.914 ± 0.075
HC	**0.841** ± 0.053	**0.911** ± 0.053	**0.967** ± 0.009	**0.961** ± 0.014
		Persuasive decision support		
MLE	0.802 ± 0.249	0.815 ± 0.151	0.848 ± 0.188	0.953 ± 0.051
TML	0.473 ± 1.016	0.653 ± 1.747	0.441 ± 0.016	0.381 ± 0.474
HC	**0.979** ± 0.014	**0.977** ± 0.009	**0.977** ± 0.009	**0.978** ± 0.013

3.13.5 Model Evaluation with Synthetic Agent

We trained models with different configurations. We mainly discuss two factors: 1) filtering out class-inconsistent triplets or not; 2) a large dimension at 512 vs. a small dimension at 50 for the output representations. We also tried different hyperparameters such as different λs that control the strength of the classification loss and triplet margin loss as well as different random seeds. We select the best TML / HC / MLE in each filtering-dimension configuration with the highest average of test classification accuracy and test triplet accuracies.

Label accuracy and triplet accuracy. As this task is relatively simple, both MLE and HC achieves test accuracy of above 97.5%. In fact, HC without filtering out class-inconsistent triplets achieved 100%. Note that TML cannot classify alone. As for triplet accuracy, as expected, both HC and TML outperform MLE. Dimensionality does not affect triplet accuracy, but filtering out class-inconsistent triplets decrease triplet accuracy (76.2% vs. 70.7% with 50 dimensions, 74.1% vs. 70.9% with 512 dimensions). This is because filtering creates a distribution shift of the triplet annotations, and limits the models' ability to learn general

Table 3.11: Experiment results on VW-Linear. Models use 512-dimension embeddings; `HC` uses $\lambda = 0.2$ and filtered triplets.

Alignments	56%	84%	95%	98.5%
Weights	[0,1,1,1]	[1,0,1,1]	[1,1,1,1]	[32,256,1,1]
		NI-H2H		
`HC` vs. `MLE`	0.936 ± 0.024	0.921 ± 0.008	0.912 ± 0.074	0.856 ± 0.034
		NO-H2H		
`HC` vs. `MLE`	0.946 ± 0.032	0.974 ± 0.032	0.949 ± 0.003	0.934 ± 0.029
		Neutral decision support		
`MLE`	0.778 ± 0.084	0.792 ± 0.144	0.839 ± 0.130	0.927 ± 0.019
`TML`	0.554 ± 0.175	0.770 ± 0.318	0.950 ± 0.095	0.914 ± 0.075
`HC`	**0.845** ± 0.127	**0.880** ± 0.127	**0.956** ± 0.016	**0.956** ± 0.111
		Persuasive decision support		
`MLE`	0.802 ± 0.249	0.815 ± 0.151	0.848 ± 0.188	0.953 ± 0.051
`TML`	0.473 ± 1.016	0.653 ± 1.747	0.441 ± 0.016	0.381 ± 0.474
`HC`	**0.974** ± 0.016	**0.970** ± 0.064	**0.968** ± 0.064	**0.988** ± 0.032

human visual similarity.

To run synthetic experiments for case-based decision support, we select the `TML` with the best test triplet accuracy as our synthetic agent, and then evaluate the examples produced by all representations. We do not show results of `TML` as we use it as the synthetic agent.

Human-compatible representations is prefered over MLE representations in H2H. We compare examples selected from different models in different configurations to examples selected by the `MLE` baseline with the same dimensionality.

Table 3.15 shows how often the synthetic agent prefers the tested model examples to baseline `MLE` examples. In all settings, the preference towards `HC` is above 50%, but not as high as those in our synthetic experiments with the VW dataset. Filtering out class-inconsistent triplets improves the preference for the nearest example with the predicted label, while hurting the preference for the nearest out-of-class example.

Decision support simulations shows a large dimension benefits MLE representations but hurts unfiltered human-compatible representations in neutral decision

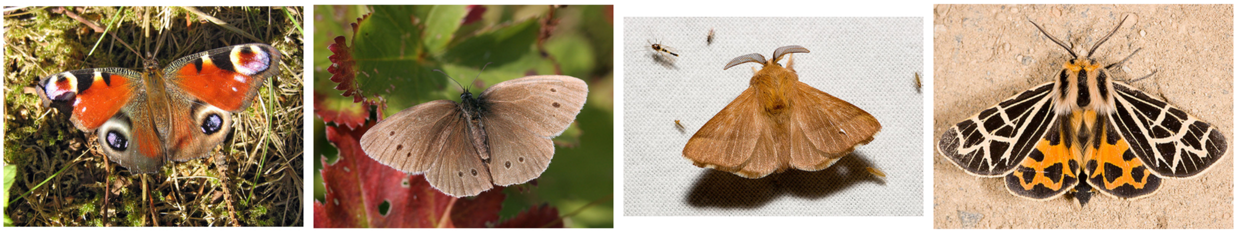
Table 3.12: Experiment results on VW-Linear. Models use 512-dimension embeddings; HC uses $\lambda = 0.8$ and filtered triplets.

Alignments	56%	84%	95%	98.5%
Weights	[0,1,1,1]	[1,0,1,1]	[1,1,1,1]	[32,256,1,1]
NI-H2H				
HC vs. MLE	0.906 ± 0.122	0.909 ± 0.111	0.882 ± 0.135	0.848 ± 0.050
NO-H2H				
HC vs. MLE	0.926 ± 0.021	0.955 ± 0.199	0.936 ± 0.053	0.912 ± 0.095
Neutral decision support				
MLE	0.778 ± 0.084	0.792 ± 0.144	0.839 ± 0.130	0.927 ± 0.019
TML	0.554 ± 0.175	0.770 ± 0.318	**0.950** ± 0.095	0.914 ± 0.075
HC	**0.824** ± 0.175	**0.895** ± 0.159	**0.950** ± 0.032	**0.969** ± 0.016
Persuasive decision support				
MLE	0.802 ± 0.249	0.815 ± 0.151	0.848 ± 0.188	0.953 ± 0.051
TML	0.473 ± 1.016	0.653 ± 1.747	0.441 ± 0.016	0.381 ± 0.474
HC	**0.981** ± 0.048	**0.964** ± 0.206	**0.961** ± 0.175	**0.978** ± 0.064

support. We also run simulated decision support with the TML synthetic agent. Table 3.16 shows decision support accuracy for different settings. MLE have both higher neutral decision support accuracy and persuasive decision support scores when we use a large dimension at 512. We hypothesize that for MLE, reducing dimension may force the network to discard dimensions useful for human judgments but keep dimensions useful for classification. We then use the 512-dimension MLE with the highest intrinsic evaluation scores as our MLE baseline in later studies.

For HC, neutral decision support accuracy are in general comparable to 87.5% score of the 512-dimension MLE baseline except unfiltered 512-dimension HC which has only 80%. We hypothesize that representations of large dimension may struggle more with contradicting signals between metric learning and supervised classification in the unfiltered settings. For persuasive decision support, HC achieves perfect scores in all settings.

Overall, to proceed with our human-subject experiments, we choose HC filtered with 50 dimensions as our best HC as it achieves a good balance between H2H and neutral decision



(a) Ringlet Butterfly (b) Peacock Butterfly (c) Caterpiller Moth (d) Tiger Moth

Figure 3.8: An example of each species in the BM dataset.

Table 3.13: Classification and triplet accuracy of BM models.

Model	**Classification accuracy**	**Triplet accuracy**
	Dimension 50	
MLE	0.975	0.610
HC	0.975	0.762
HC-filtered	0.975	0.707
TML	N/A	0.759
TML-filtered	N/A	0.721
	Dimension 512	
MLE	0.975	0.631
HC	1.000	0.741
HC-filtered	0.975	0.709
TML	N/A	0.748
TML-filtered	N/A	0.732

support. For MLE, we choose the representation with 512 dimensions. We conduct head-to-head comparison between these two representations. Our synthetic agent prefers HC in 70% of the nearest in-class examples and in 97.5% of the nearest out-of-class examples.

3.13.6 Interface

We present the screenshots of our interface at the end of the appendix. Our interface consists of four stages. Participants will see the consent page at the beginning, as shown in Fig 3.12. After consent page, participants will see task specific instructions, as shown in Fig 3.14. After entering the task, partipants will see the questions, as shown in Fig 3.15. We also

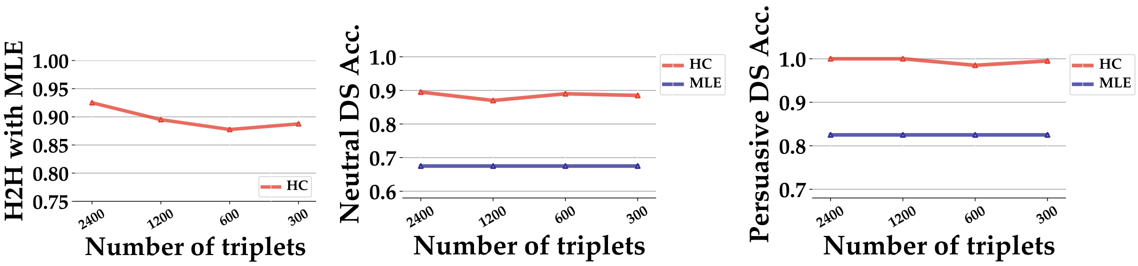


Figure 3.9: HC performance declines as the number of triplets decreases, but shows strong persuasive decision support accuracy even with very few triplets.

Table 3.15: BM H2H preference results with synthetic agent.

Dimensions	50	512
NI H2H with MLE		
HC	0.838	0.575
HC filtered	0.863	0.725
NO H2H with MLE		
HC	0.775	0.925
HC filtered	0.700	0.775

Table 3.16: BM decision support accuracy with synthetic agent.

Dimensions	50	512
Neutral Decision Support		
MLE	0.675	0.875
HC	0.900	0.800
HC filtered	0.875	0.900
Persuasive Decision Support		
MLE	0.825	0.875
HC	1.000	1.000
HC filtered	1.000	1.000

include two attention check questions in all studies to check whether participants are paying attention to the questions. Following suggestions on Prolific, we design the attention check with explicit instructions, as shown in Fig 3.17. After finishing all questions, participants will reach the end page and return to Prolific, as shown in Fig 3.19. Our study is reviewed by the Institutional Review Board (IRB) at our institution (IRB22-0388).

3.13.7 Crowdsourcing

We recruit our participants on a crowdsourcing platform: Prolific (www.prolific.co) [April-May 2022]. We conduct three total studies: an annotation study, a decision support study, and a head-to-head comparison study. We use the default standard sampling on Prolific for participant recruitment. Eligible participants are limited to those reside in United States. Participants are not allowed to attempt the same study more than once.

Triplet annotation study We recruit 90 participants in total. We conduct a pilot study with 7 participants to test the interface, and recruit 83 participants for the actual collection of annotations. 3 participants fail the attention check questions and their responses are excluded in the results. We spend in total $76.01 with an average pay at $10.63 per hour. The median time taken to complete the study is 3'22".

Decision support study We recruit 161 participants in total. 3 participants fail the attention check questions and their responses are excluded in the results. We take the first 30 responses in each conditon to compile the results. We spend in total $126.40 with an average pay at $9.32 per hour. The median time taken to complete the study is 3'53".

Head-to-head comparison study We recruit 31 participants in total, where 1 participant fail the attention check questions and their responses are excluded in the results. We spend in total $24.00 with an average pay at $9.40 per hour. The median time taken to complete the study is 3'43".

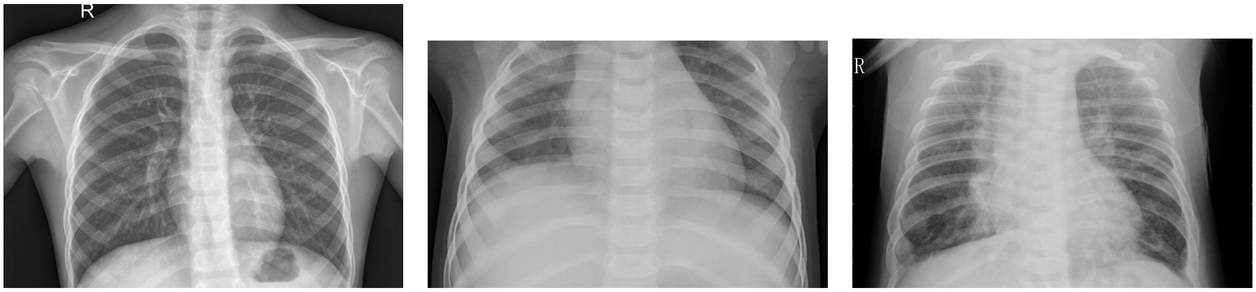
3.14 Human Subject Study on Chest X-rays

3.14.1 Dataset

Our CXR dataset is constructed from a subset of the chest X-ray dataset used by Kermany et al. [2018], which had 5,232 images. We take a balanced subset of 3,166 images, 1,583 characterized as depicting pneumonia and 1,583 normal. The pneumonia class contains bacterial pneumonia and viral pneumoia images, but we do not differentiate them for this study. An example of each image class is shown in Fig 3.10.

3.14.2 Hyperparameters

For CXR experiment, instead of ResNet-18 pretrained from ImageNet, we use a ResNet-18 finetuned on CXR classifcation as our CNN backbone, as we observe it provides better



(a) Normal (b) Bacterial pneumonia (c) Viral pneumonia

Figure 3.10: An example of each image class in the CXR dataset.

Table 3.17: Classification and triplet accuracy of CXR models.

Model	**Classification accuracy**	**Triplet accuracy**
	Dimension 50	
MLE	0.973	0.571
HC	0.954	0.576
HC-filtered	0.955	0.574
TML	N/A	0.602
TML-filtered	N/A	0.587
	Dimension 512	
MLE	0.973	0.588
HC	0.968	0.602
HC-filtered	0.971	0.561
TML	N/A	0.618
TML-filtered	N/A	0.591

decision support simulation results. For training our HC model we use λ of 0.5. We use the Adam optimizer [Kingma and Ba, 2014] with learning rate $1e-4$. Our training batch size is 16 for triplet prediction, and 30 for classification. All models are trained for 10 epoches. The checkpoint with the lowest validation total loss in each run is selected for evaluations and applications.

3.14.3 Classification and Triplet Accuracy

We present the test-time classification and triplet accuracy of our models in Table 3.17. Both MLE and HC achieve above 95% classification accuracy. Both TML and HC achieve above above 65% triplet accuracy. Both TML model and HC achieve the highest triplet accuracy in

Table 3.18: CXR H2H preference results with synthetic agent.

Dimensions	50	512
NI H2H with MLE		
HC	0.536	0.675
HC filtered	0.472	0.599
NO H2H with MLE		
HC	0.535	0.635
HC filtered	0.487	0.494

Table 3.19: CXR decision support accuracy with synthetic agent.

Dimensions	50	512
Neutral Decision Support		
MLE	0.711	0.726
HC	0.742	0.779
HC-filtered	0.732	0.804
Persuasive Decision Support		
MLE	0.881	0.882
HC	0.949	0.966
HC-filtered	0.948	0.946

the 512-dimension unfiltered setting with triplet accuracy at 69.1% and 72.2% respectively. Filtering out class-inconsistent triplets removes 20.69% of the triplet annotations in this dataset.

3.14.4 Model Evaluation with Synthetic Agent

Similar to the BM setting, we select the TML with the best test triplet accuracy as our synthetic agent, and then evaluate the examples produced by all representations. As table 3.18 shows, preference for HC over MLE in H2H is less significant compared to BM, likely due to the challenging nature of the CXR dataset. We still observe the patten that filtering improves H2H performance.

Table 3.19 shows decision support accuracy for different settings. All models benefit from a large dimension at 512. We observe consistent patterns such as filtering leading to better decision support.

3.14.5 Effect of Triplet Amount and Type

We evaluate the effect of the number of triplets on our models in Fig. 3.11. Similar to the BM experiments, H2H preference towards human-compatible representations and neutral decision support performance decrease as the number of triplets decreases. Human-compatible

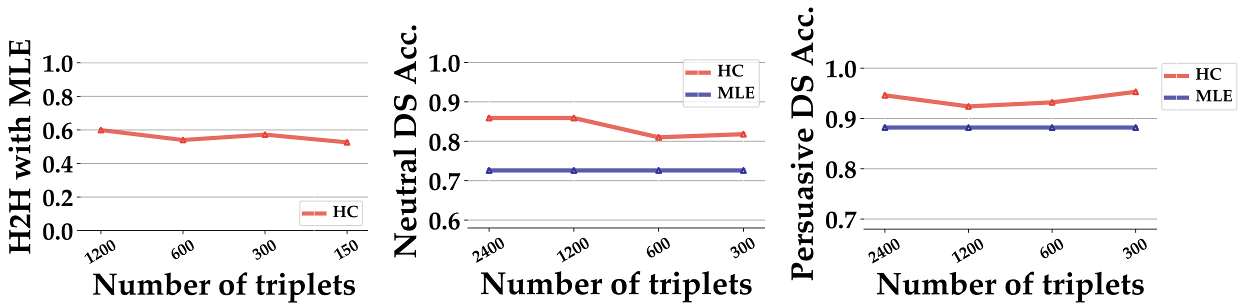


Figure 3.11: HC performance declines as the number of triplets decreases, but shows strong persuasive decision support accuracy even with very few triplets.

representations achieve strong persuasive decision support performance even with very few triplets.

3.14.6 Interface

Our CXR interface is mostly the same as our BM interface, except that we add basic chest X-ray instructions as participants may not be familiar with medical images. After the consent page at the beginning, participants will see basic chest X-ray instructions showing where the lungs and hearts. Then, they enter an multiple-choice attention check, as shown in Fig 3.13. The correct answer in “lungs and adjacent interfaces”. Failing the attention check will disqualify the participant. After correctly answering the pre-task attention check, participants will see the same task specific instructions as in the BM studies, as shown in Fig 3.14. Screenshots of questions are shown in Fig 3.16. We also include two in-task attention check questions simlar to the BM study. Our study is reviewed by the Institutional Review Board (IRB) at our institution with study number that we will release upon acceptance to preserve anonymity.

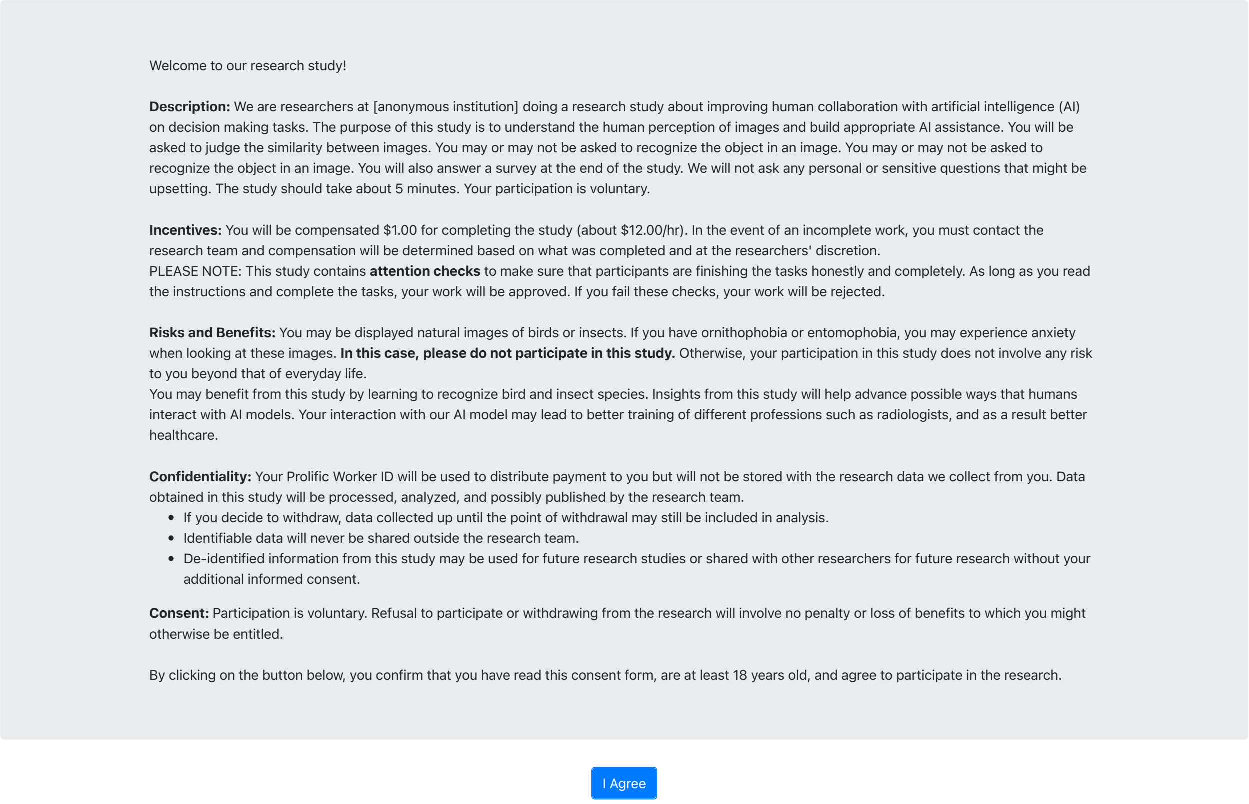
3.14.7 Crowdsourcing

We recruit our participants on Prolific (www.prolific.co) [September 2022]. We conduct three total studies: an annotation study, a decision support study, and a head-to-head comparison study. We use the default standard sampling on Prolific for participant recruitment. Eligible participants are limited to those reside in United States. Participants are not allowed to attempt the same study more than once.

Triplet annotation study We recruit 123 participants in total. 20 partipants fail the pre-task attention check question and 3 participants fail the in-task attention check questions; their responses are excluded in the results. We spend in total $80.00 with an average pay at $10.70 per hour. The median time taken to complete the study is 3'22".

Decision support study We recruit 296 participants in total. 34 partipants fail the pre-task attention check question and 10 participants fail the in-task attention check questions; their responses are excluded in the results. We spend in total $221.67 with an average pay at $11.00 per hour. The median time taken to complete the study is 3'40".

Head-to-head comparison study We recruit 57 participants in total. 6 partipants fail the pre-task attention check question and 1 participants fail the in-task attention check questions; their responses are excluded in the results. We spend in total $40.00 with an average pay at $10.54 per hour. The median time taken to complete the study is 3'25".



Welcome to our research study!

Description: We are researchers at [anonymous institution] doing a research study about improving human collaboration with artificial intelligence (AI) on decision making tasks. The purpose of this study is to understand the human perception of images and build appropriate AI assistance. You will be asked to judge the similarity between images. You may or may not be asked to recognize the object in an image. You may or may not be asked to recognize the object in an image. You will also answer a survey at the end of the study. We will not ask any personal or sensitive questions that might be upsetting. The study should take about 5 minutes. Your participation is voluntary.

Incentives: You will be compensated $1.00 for completing the study (about $12.00/hr). In the event of an incomplete work, you must contact the research team and compensation will be determined based on what was completed and at the researchers' discretion.
PLEASE NOTE: This study contains **attention checks** to make sure that participants are finishing the tasks honestly and completely. As long as you read the instructions and complete the tasks, your work will be approved. If you fail these checks, your work will be rejected.

Risks and Benefits: You may be displayed natural images of birds or insects. If you have ornithophobia or entomophobia, you may experience anxiety when looking at these images. **In this case, please do not participate in this study.** Otherwise, your participation in this study does not involve any risk to you beyond that of everyday life.
You may benefit from this study by learning to recognize bird and insect species. Insights from this study will help advance possible ways that humans interact with AI models. Your interaction with our AI model may lead to better training of different professions such as radiologists, and as a result better healthcare.

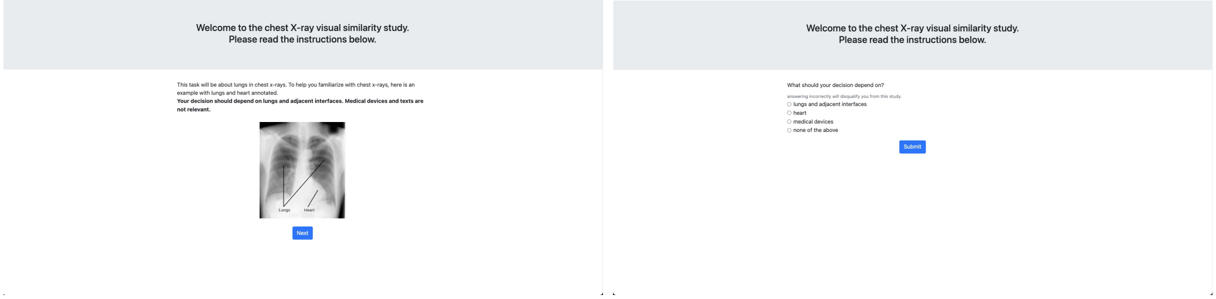
Confidentiality: Your Prolific Worker ID will be used to distribute payment to you but will not be stored with the research data we collect from you. Data obtained in this study will be processed, analyzed, and possibly published by the research team.

- If you decide to withdraw, data collected up until the point of withdrawal may still be included in analysis.
- Identifiable data will never be shared outside the research team.
- De-identified information from this study may be used for future research studies or shared with other researchers for future research without your additional informed consent.

Consent: Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled.

By clicking on the button below, you confirm that you have read this consent form, are at least 18 years old, and agree to participate in the research.

Figure 3.12: The consent form page on our interface.



Welcome to the chest X-ray visual similarity study.
Please read the instructions below.

This task will be about lungs in chest x-rays. To help you familiarize with chest x-rays, here is an example with lungs and heart annotated.
Your decision should depend on lungs and adjacent interfaces. Medical devices and texts are not relevant.

(a) Basic instructions about chest X-rays.

Welcome to the chest X-ray visual similarity study.
Please read the instructions below.

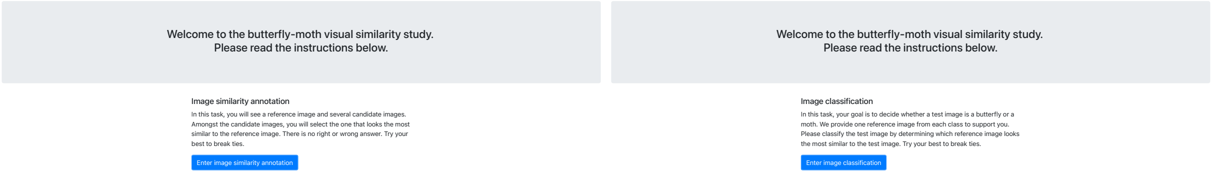
What should your decision depend on?

answering incorrectly will disqualify you from this study.

○ lungs and adjacent interfaces
○ heart
○ medical devices
○ none of the above

(b) Multiple-choice attention check for CXR tasks. The correct answer is "lungs and adjacent interfaces".

Figure 3.13: Pre-task instructions and attentions check for CXR tasks



Welcome to the butterfly-moth visual similarity study.
Please read the instructions below.

Image similarity annotation

In this task, you will see a reference image and several candidate images. Amongst the candidate images, you will select the one that looks the most similar to the reference image. There is no right or wrong answer. Try your best to break ties.

Welcome to the butterfly-moth visual similarity study.
Please read the instructions below.

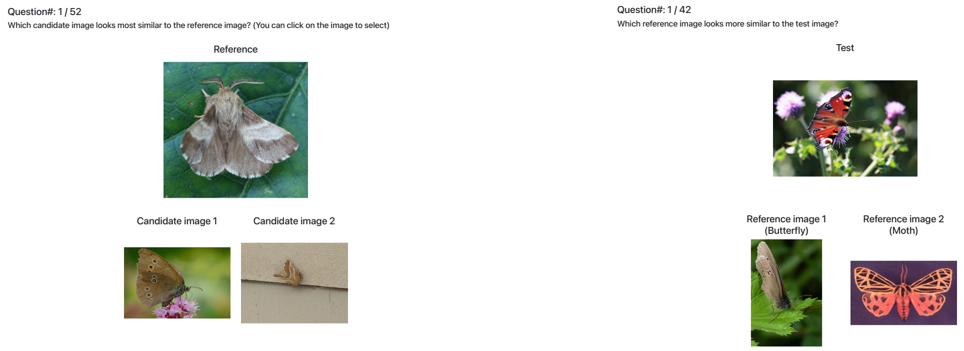
Image classification

In this task, your goal is to decide whether a test image is a butterfly or a moth. We provide one reference image from each class to support you. Please classify the test image by determining which reference image looks the most similar to the test image. Try your best to break ties.

(a) The annotation and head-to-head comparision task instructions.

(b) The decision support task instructions.

Figure 3.14: The task-specific instruction page on our interface.



Question#: 1 / 52

Which candidate image looks most similar to the reference image? (You can click on the image to select)

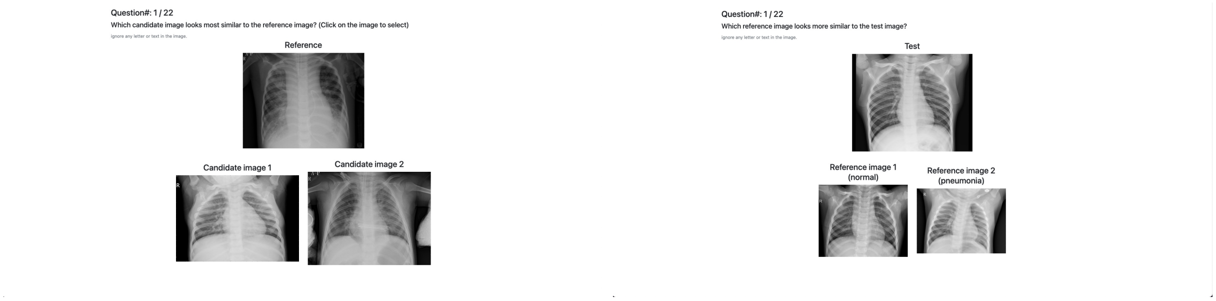
Question#: 1 / 42

Which reference image looks more similar to the test image?

(a) The annotation and head-to-head comparision task questions.

(b) The decision support task questions.

Figure 3.15: The task-specific questions for BM.



Question#: 1 / 22

Which candidate image looks most similar to the reference image? (Click on the image to select)

ignore any letter or text in the image.

Question#: 1 / 22

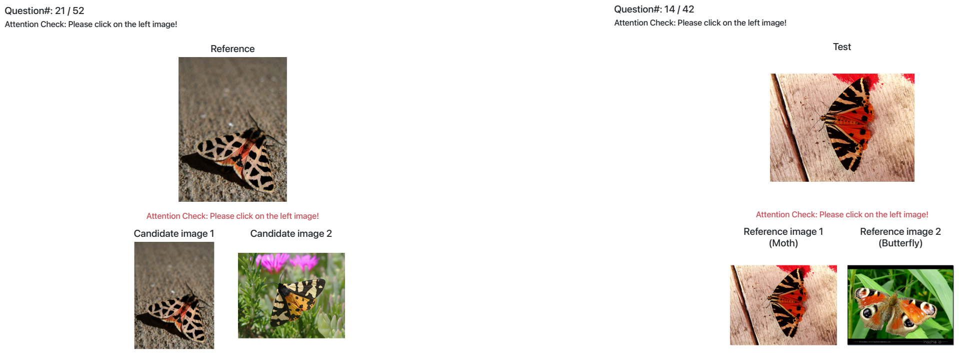
Which reference image looks more similar to the test image?

ignore any letter or text in the image.

(a) The annotation and head-to-head comparision task questions.

(b) The decision support task questions.

Figure 3.16: The task-specific questions for CXR.



Question#: 21 / 52

Attention Check: Please click on the left image!

Attention Check: Please click on the left image!

Question#: 14 / 42

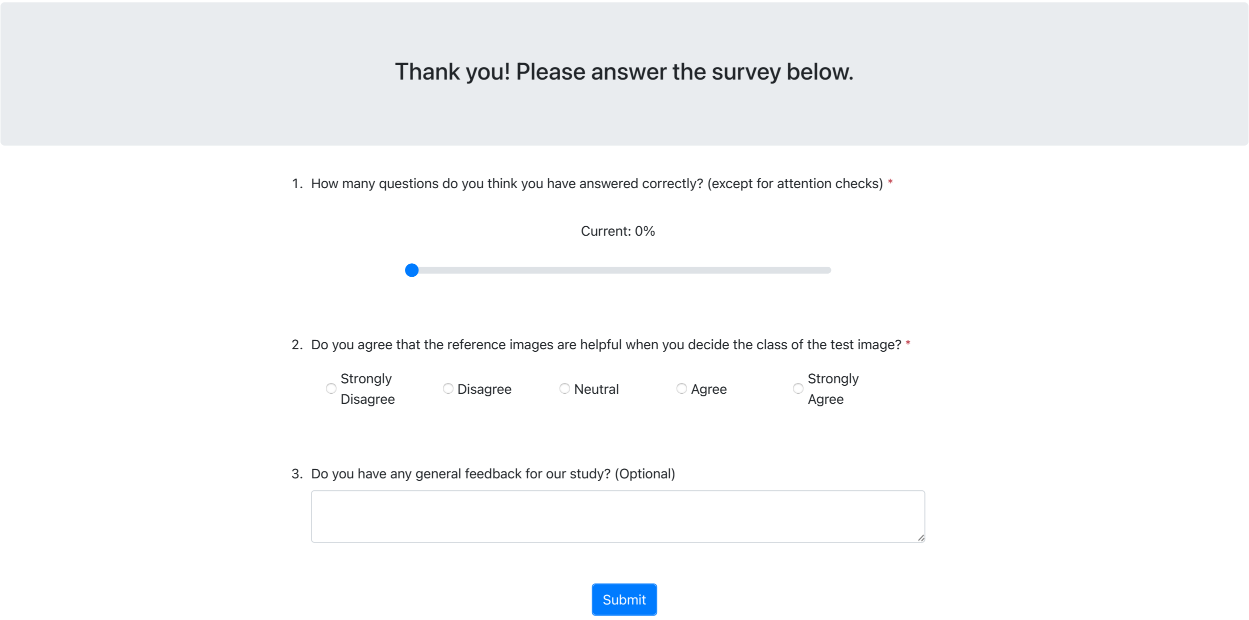
Attention Check: Please click on the left image!

Attention Check: Please click on the left image!

(a) The annotation and head-to-head comparison task attention check questions.

(b) The decision support task attention check questions.

Figure 3.17: The task-specific attention check questions for BM.



Thank you! Please answer the survey below.

1. How many questions do you think you have answered correctly? (except for attention checks) *

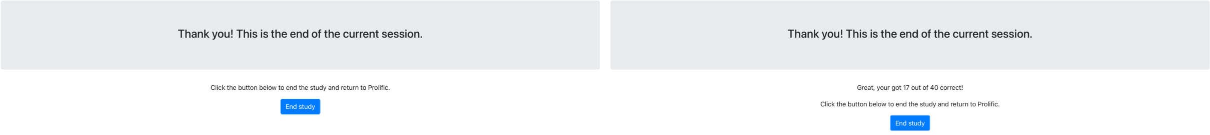
Current: 0%

2. Do you agree that the reference images are helpful when you decide the class of the test image? *

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

3. Do you have any general feedback for our study? (Optional)

Figure 3.18: The survey page of the decision support task on our interface.



Thank you! This is the end of the current session.

Click the button below to end the study and return to Prolific.

Thank you! This is the end of the current session.

Great, your got 17 out of 40 correct!

Click the button below to end the study and return to Prolific.

(a) The annotation and head-to-head comparision task end page.

(b) The decision support task end page.

Figure 3.19: The task-specific end page on our interface.

CHAPTER 4

TEACHING HUMANS WITH CONCEPTS AND EXAMPLES

4.1 Overview

In this chapter, we present a novel teaching paradigm that combines concepts and examples to teach simulated human learners to do fine-grained image classification tasks. We propose a method to select concepts and examples that are informative for the task to teach a concept bottleneck learner. We evaluate the effectiveness of the teaching on 200 bird species identification tasks with different levels of granularity and in both in-distribution and out-of-distribution examples from two different datasets. Results show that the teaching is more effective on tasks with the highest level of granularity, which suggests concepts and examples selected at the proper level of granularity are more informative for the task and have more potential to improve the performance of human learners.

Unlike the previous chapters, which focuses explanations of tasks and predictions, this chapter focuses on the transfer of knowledge from AI systems to humans, namely, teaching humans to perform tasks. This work is exploratory in nature and aims to understand the effectiveness of the teaching with concepts and examples paradigm on image classification tasks.

4.2 Introduction

Just like communication among people and each other cancels ambiguity and promotes agreement, communication between human and AI should also aim for mutual understanding. In this chapter we seek to find common ground between humans and AI, and discuss specifically how to teach simulated human learners to do fine-grained image classification through concepts and examples.

We argue that the key to mutual understanding lies in the ability to communicate effectively through distinctive concepts and representative examples. AI models powered by deep learning have shown remarkable performance in various tasks and demonstrated the ability to learn from large-scale data. However, the lack of interpretability and explainability of these models has been a major obstacle to their adoption in high-stakes domains such as healthcare, criminal justice, and finance. Recent advances in explainable AI (XAI) have made significant progress in building models that make predictions based on concepts transferred from natural language descriptions via vision language models [Oikarinen et al., 2023, Yuksekgonul et al., 2022, Yan et al., 2023, Yang et al., 2023].

In this chapter, we propose to investigate how to leverage these models in building AI-driven tutorials that can teach learners with concepts and examples. We will also investigate how to build AI-driven tutorials by identifying key concepts for the tasks and important examples associated with the concepts. We will evaluate the effectiveness of these tutorials in improving learner performance and understanding of the underlying concepts on a number of fine-grained natural image classification tasks.

4.2.1 Recent Advances on Vision Language Models and Explainable AI

Visual perception and language understanding, two of the most important functions of human brains, have been proved to be well emulated, sometimes even "outperformed," by machine-learning models on curated datasets and in specific domains. While human brains are able to often effortlessly perform and make connections between the two through multiple levels of abstractions, machine learning models struggles to manage both at the same time. However, recent advances in contrastive representation learning sheds new light on how to effectively learn transferable knowledge from supervision signals with a large scale of data in both the text and the image format [Radford et al., 2021].

Self-supervised deep learning models learned with framework like CLIP [Radford et al., 2021] trained from a large scale are black-box models where model behaviors, for example predictions in classification tasks, are not transparent and interpretable to there human users. Recent work in explainable AI (XAI) have come up with various techniques to produce global and local explanations, we argue that these explanations are, on a spectrum of interpretability, too close to machines but too far from humans.

Example-based explanation techniques usually either select prototypical images from training set as global explanation, or select nearest neighbors of the input images as local explanations. These examples can serve as probing to the black-box model's input/output space, but hardly an explanation for the decision making process.

Local feature-based explanations like GradCam [Selvaraju et al., 2017] or LIME [Ribeiro et al., 2016] use gradient-based or perturbation-based methods to identify pixel regions contributing to the prediction and outputs heatmaps for the input images. These methods aim to select in the causal chain of the prediction the important parts of the input, but may sometimes still be incomprehensible should the causal relationship between the highlighted part and the prediction remained obscure. For example, a heatmap highlighting regions outside a patient's liver could be produced by such algorithms, while the task is to predict from a series of MRI images the malignancy of a lesion inside the patient's liver.

Another important breakthrough in building explainable AI models is the use of concepts as intermediate representations. Concept bottleneck models [Koh et al., 2020] have been proposed to learn interpretable representations by mapping the input data to a set of predefined concepts. These models can provide explanations in terms of concepts, which are more interpretable to humans than raw pixels or features. However, these models are still limited because they rely on predefined concepts, which have to be annotated by humans for each example in the training data. However, recent work has shown that it is possible to learn concepts from data without human supervision: [Oikarinen et al., 2023] and [Yuksekgonul et al., 2022] have proposed models that can learn concepts from natural language descriptions through multimodal models and use them to make predictions on images. [Yan et al., 2023] and [Yang et al., 2023] have proposed models where concepts queried from a large language model are used to train concept-bottleneck model. These models can leverage concepts for prediction without any human supervision in the pipeline yet still yield good performance on image classification tasks.

4.3 Methods

In this work, we propose to investigate how to leverage vision-language models and concept bottleneck models in building AI-driven tutorials that can teach humans in a more interpretable way. Given a set of concepts and examples, we aim to build a system that can generate tutorials that identify the most important concepts and examples for a given task, and present them to the user in an interpretable way. The user can then interact with the system to learn the concepts and examples that are most relevant to the task.

4.3.1 Problem Formulation

We formulate the teaching problem as follows. Given a labeled image classification dataset $D = \{(i, y)\}$, where i is an image and $y \in \mathcal{Y}$ is the label, we aim to select teaching examples $\mathcal{M}$ from all of the N images for training the learner to predict the label from the image. To simulate the thought process of human learners, we assume that the learner is a concept bottleneck model that can predict the label from the image using a set of concept $\mathcal{C}$, which are extracted from natural language descriptions.

The goal is to build the concept bottleneck with the most important concepts $\mathcal{C}^*$, and select the most effective teaching examples $\mathcal{M}^*$ from the dataset D for the learner to learn to predict the label from the image through the concepts bottleneck. Suppose we have a vision-language model such as CLIP ([Radford et al., 2021]) that is pretrained with image-text pair to align images and texts in a shared representation space. We can use the image encoder

$\mathcal{I}$ to encode the image dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ where $\mathbf{x}_i \in \mathbb{R}^d$ is the extracted feature of image i with dimension d, and the text encoder $\mathcal{T}$ to encode the concepts $\mathcal{C}$ into $\mathbf{E} \in \mathbb{R}^{N_C \times d}$ from natural language descriptions, where N_C is the number of concept, i.e. the largest possible size of the concept bottleneck. $\mathbf{E} = \left[\mathcal{T}(c_1); \ldots; \mathcal{T}(c_{N_C})\right]$.

To make a prediction with a concept-bottleneck model, we can then compute the concept activation $\mathbf{a}_i \in \mathbb{R}^{N_C}$ of any image i on each concept c_j in the bottleneck with a function g:

$$a_{ij} = g(\mathbf{x}_i, \mathcal{T}(c_j)), j = 1 \ldots N_C \tag{4.1}$$

$$\mathbf{a}_i = \left(a_{i1} \ldots a_{iN_C}\right)^\top \tag{4.2}$$

where g is a function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that computes the activation score from image feature $\mathbf{x}_i$ of any given image and any concept feature $\mathcal{T}(c_j)$ from all N_C concepts in the bottleneck. In our experiments, we use cosine similarity as the function g. Since our image features and concept features are both normalized, g is equivalent to taking the inner product of $\mathbf{x}_i$ and $\mathcal{T}(c_j)$.

Now we simulate the human learner as a linear classifier that can predict the label y_i of any image i with the concept activation $\mathbf{a}_i$. The label prediction function f: $\hat{y}_i = f(\mathbf{a}_i)$ where f is a linear function $f : \mathbb{R}^{N_C} \to \mathcal{Y}$ that predicts the label from the concept activation.

4.3.2 Concept Selection

Identifying the most important concepts to form the concept bottleneck is crucial for the learner to predict the label from the image. Given a set of concepts $\mathcal{S}$, we aim to select the most important concepts $\mathcal{C}^*$ for the learner to make predictions on the image dataset $\mathcal{D}$. We need to design a scoring function $\mathcal{F}_{\mathcal{C}} : 2^{\mathcal{S}} \to \mathbb{R}$ that scores the importance of a set of concepts $\mathcal{C}$. We adapt the concept selection objectives from [Yang et al., 2023] to select the most important concepts for the task. Specifically, we can define $D(c)$ as the discriminability score of concept c on the image dataset $\mathcal{D}$, which measures how well the concept can distinguish the images in the dataset. Concepts with higher discriminability scores are associated with fewer classes. We can calculate the discriminability score of a concept c by first measuring the similarity score between a class y and a concept c:

$$\operatorname{Sim}(y, c) = \frac{1}{|\mathcal{X}_y|} \sum_{\boldsymbol{x} \in \mathcal{X}_y} \boldsymbol{x} \cdot \mathcal{T}(c)^\top \tag{4.3}$$

where $\mathcal{X}_y$ is the set of images in class y, and $\mathcal{T}(c)$ is the text embedding of concept c. We can normalize the class similarity of a concept as $\overline{\operatorname{Sim}}(y \mid c) = \operatorname{Sim}(y, c) / \sum_{y' \in \mathcal{Y}} \operatorname{Sim}\left(y', c\right)$. The normalized similarity score $\overline{\operatorname{Sim}}(y \mid c)$ measures the conditional likelihood of class y given concept c. Now we can compute the negative entropy of the normalized similarity score as the discriminability score of concept c:

$$D(c) = \sum_{y' \in Y} \overline{\operatorname{Sim}}\left(y' \mid c\right) \cdot \log\left(\overline{\operatorname{Sim}}\left(y' \mid c\right)\right) \tag{4.4}$$

A high discriminability score $D(c)$ indicates the appearance of concept c over examples from different classes has a skewer distribution, which means the concept c is associated with fewer classes and can distinguish the images from different classes better. We can rank the concepts by their discriminability scores and identify the most associated class of each concept by comparing the class similarity scores:

$$\operatorname{Class}(c) = \arg\max_{y \in \mathcal{Y}} \operatorname{Sim}(y, c) \tag{4.5}$$

We can then select the most important concepts $\mathcal{C}^*$ by picking the top concepts for each class and maximizing the total discriminability score of the given set of selected concepts $\mathcal{C}$:

$$\mathcal{C}^* = \bigcup_{y \in \mathcal{Y}} \arg\max_{\mathcal{C} \subseteq \mathcal{S}_y} \mathcal{F}_{\mathcal{C}}(\mathcal{C}) \tag{4.6}$$

where $\mathcal{S}_y = \{c \in \mathcal{S} \mid \mathrm{Class}(c) = y\}$ are all the concepts in $\mathcal{S}$ associated with class y, and $\mathcal{C}_y^*$ is the selected set associated with class y. The size of the selected set is balanced for each class and are denoted by $\left|\mathcal{C}_y^*\right| = K_y$ where $K_y \cdot |\mathcal{Y}| = K = |\mathcal{C}^*|$ is the size of the bottleneck. For each class y we select the top K_y concepts and take the union as our final concept bottleneck $\mathcal{C}^*$.

4.3.3 Example Selection

In this section, we aim to select the most effective teaching examples $\mathcal{M}^*$ from the dataset $\mathcal{D}$ for teaching the learner to predict the label from the image through the selected concept bottleneck $\mathcal{C}^*$. We need to design a scoring function $\mathcal{F}_{\mathcal{M}} : 2^N \rightarrow \mathbb{R}$ that scores the effectiveness of the teaching set $\mathcal{M}$. Specifically, we can define $\mathrm{Sim}(i, C)$ as the saliency score of image i on the concept bottleneck $\mathcal{C}$, which measures how well the image can activate the concepts in the bottleneck. We first split the concepts $\mathcal{C}$ into two sets: the concepts that are associated with the true class of the image as $\mathcal{C}_y^+$ and those that are not as $\mathcal{C}_y^-$. We can calculate the saliency score of an image i by measuring the differences between the sum of the activations on associated concepts and the sum of the activations of non-associated concepts:

$$\begin{aligned} \mathrm{Sim}(i, C) &= \alpha \, \mathrm{Sim}(i, \mathcal{C}_y^+) - \beta \, \mathrm{Sim}(i, \mathcal{C}_y^-) && (4.7) \\ &= \alpha \sum_{c \in \mathcal{C}_y^+} g(\mathbf{x}_i, \mathcal{T}(c)) - \beta \sum_{c \in \mathcal{C}_y^-} g(\mathbf{x}_i, \mathcal{T}(c)) && (4.8) \end{aligned}$$

where α and β are hyperparameters that controls the strength of activation of associated concepts and non-associated concepts correspondingly. Finally we can calculate the scoring for any teaching set M as $\mathcal{F}_{\mathcal{M}} = \sum_{i \in \mathcal{M}} \text{Sim}(i, C^*)$ and select the optimal set by greedily adding examples with the highest saliency score $\text{Sim}(i, C^*)$.

4.4 Experimental Setup

4.4.1 Dataset

We conduct experiments on different granularities of binary classification tasks from two bird species image datasets. The iNaturalist dataset 2021 [Van Horn et al., 2018] is a large-scale natural image dataset with 2.7 million images of 10,000 classes of plants and animals. We use the bird species subset of the dataset, which contains 414,847 training images of 1,486 bird species. The CUB-200-2011 dataset [Wah et al., 2011] is a fine-grained dataset with 11,788 images of 200 bird species. The dataset contains 312 binary attributes for each image, which are annotated by crowd workers. We use these binary attributes as the concept pool for our concept selection experiments.

We look for the common species between the two datasets and select 175 species in total for our experiments. For each species in the iNaturalist dataset, we re-split the training images into 60% for training, 20% for validation, and 20% for testing randomly. We use the re-splitted training set to select concepts and examples, the validation set to do early stopping, and the testing set to evaluate the performance of the learner. Since the testing set comes from the same distribution as the training set, we refer to it as the in-distribution testing set. For the CUB-200-2011 dataset, we only use the default testing split for evaluation. Since the testing split comes from a different distribution than the training split in iNaturalist, we refer to it as the out-of-distribution testing set.

To investigate the effect of teaching with concepts and examples at different levels of granularity, we build a taxonomy tree for the bird species in the iNaturalist dataset. The taxonomy information provided by the dataset contains the scientific names, genus names, family names, order names, and class names of the species. We use the distance between the species in the taxonomy tree to define the level of granularity for the binary classification tasks. For example, two species with a distance of 2 in the taxonomy tree are considered to be in the same genus, while two species with a distance of 4 are considered to be in the same family. The binary classification tasks between species in the same genus are considered to be at the genus level, while the tasks between species in the same family are considered to be at the family level. Based on the taxonomy tree, we select binary classification tasks by randomly sampling two species from the dataset. There are four levels of granularity: genus, family, order, and class. We select 50 binary classification tasks for each level of granularity, resulting in 200 binary classification tasks in total.

4.4.2 Baseline

We compare our example selection method with two baselines: random selection and kmeans. The random selection baseline randomly selects examples from the training set to train the learner, while keeping examples selected balanced for each class. The kmeans baseline clusters the training examples into $|\mathcal{M}|/2$ clusters for each class using the kmeans algorithm on the image features. Then we select the images with the smallest Euclidean distance to each cluster center as the teaching examples.

4.4.3 Implementation Details

We use the CLIP model [Radford et al., 2021] as the vision-language model to encode images and concepts. We use the pretrained weights of the ViT-L/14 architecture with 768-dimensional embeddings. We use the default PyTorch implementation of linear models and the Adam optimizer with a learning rate of $1e-3$ to train the concept bottleneck model.

Table 4.1: Test accuracy of the learner with different concept bottleneck size K_y and number of examples M_y on the in-distribution and out-of-distribution testing sets. Accuracy is averaged over 50 binary classification tasks at each granularity level and each task over 5 runs with different random seeds.

Distribution		In-distribution				Out-of-distribution			
Level		Genus	Family	Order	Class	Genus	Family	Order	Class
K_y	M_y								
1	1	0.563	0.571	0.625	0.629	0.574	0.582	0.643	0.642
1	5	0.574	0.6	0.631	0.642	0.59	0.618	0.651	0.656
1	10	0.578	0.604	0.633	0.642	0.591	0.622	0.653	0.656
3	1	0.524	0.523	0.553	0.552	0.536	0.528	0.564	0.562
3	5	0.528	0.535	0.548	0.558	0.541	0.542	0.557	0.569
3	10	0.528	0.541	0.55	0.556	0.541	0.551	0.56	0.569
5	1	0.577	0.593	0.668	0.688	0.609	0.602	0.693	0.712
5	5	0.61	0.656	0.722	0.74	0.647	0.672	0.759	0.768
5	10	0.622	0.662	0.735	0.749	0.662	0.684	0.772	0.778

Learners can only predict the label from the selected concept activation, and do not take image features as inputs in the prediction. We use the cosine similarity as the activation function g to compute the concept activation, and cross-entropy loss to train the linear classifier f to predict the label from the concept activation. We use the early stopping strategy to prevent overfitting on the validation set. We train the model for 100 epochs and evaluate model checkpoints with the highest validation accuracy on the test set.

For each binary classification task, we experiment with different numbers of concepts and examples. We set the number of concepts per class K_y to be 1, 3, and 5, and the number of examples per class $M_y = |\mathcal{M}| \,/\, |\mathcal{Y}|$ to be 1, 5, and 10.

4.5 Main Results

4.5.1 Comparing Concept Bottleneck Size, Number of Examples, Granularity Levels, and Distribution Types

We investigate the effect of the concept bottleneck size and the number of examples on the performance of the learner. We report the test accuracy of the learner aggregated from the 50 binary classification tasks at each granularity level in Table 4.1. Each cell in the table shows the average accuracy over 5 runs with different random seeds. We report the accuracy on the in-distribution and out-of-distribution testing sets separately.

Larger bottleneck size does not always lead to better performance. We observe that the performance of the learner does not always improve with a larger bottleneck size. Across different levels of granularity and distribution types, the learner with a bottleneck size of 5 outperforms the learner with a bottleneck size of 3, but the learner with a bottleneck size of 3 does not always outperform the learner with a bottleneck size of 1.

More examples lead to better performance. We observe that the performance of the learner improves with more examples. The learner with 10 examples per class outperforms the learner with 5 examples per class, and the learner with 5 examples per class outperforms the learner with 1 example per class. This trend is consistent across different levels of granularity and distribution types.

Learners perform better on higher levels of granularity. We observe that the learner performs better on higher levels of granularity. The learner achieves the highest accuracy on the class level, followed by the order level, family level, and genus level. This suggest our method yield better performance when teaching concepts at higher levels of granularity.

Table 4.2: Win-rate of the learner over the baselines on the in-distribution and out-of-distribution testing sets at each granularity level. Number of concepts $K_y = 1$, number of examples $M_y = 1$. Win-rate greater or equal to 0.5 is highlighted in bold.

	In-distribution		Out-of-distribution	
Level	Random	KMeans	Random	KMeans
Genus	**0.5**	**0.625**	0.458	0.458
Family	0.417	0.417	0.354	0.292
Order	0.458	0.375	0.375	0.188
Class	**0.667**	**0.604**	**0.562**	0.417

Table 4.3: Win-rate of the learner over the baselines on the in-distribution and out-of-distribution testing sets at each granularity level. Number of concepts $K_y = 3$, number of examples $M_y = 5$. Win-rate greater or equal to 0.5 is highlighted in bold.

	In-distribution		Out-of-distribution	
Level	Random	KMeans	Random	KMeans
Genus	0.312	0.312	0.271	0.271
Family	0.375	0.396	0.375	0.458
Order	0.125	0.188	0.167	0.167
Class	0.229	0.229	0.229	0.208

Learners perform better on out-of-distribution testing sets. We observe that the learner performs better on out-of-distribution testing sets than on in-distribution testing sets. This suggests that our method is more robust to out-of-distribution data than in-distribution data. This may seem counterintuitive, but through further investigation, we find that the observation could be due to the fact that the out-of-distribution testing sets have better image quality than the in-distribution testing sets, which makes the task easier for the learner to predict the label from the image.

4.5.2 Comparing Example Selection Methods

We compare the performance of our concept selection method with two baselines: random selection and kmeans. We evaluate the performance of the learners on the in-distribution and out-of-distribution testing sets. For each binary classification task, we calculate accuracy of the learners on the testing sets over 5 runs with different random seeds. Our metric is how

Table 4.4: Win-rate of the learner over the baselines on the in-distribution and out-of-distribution testing sets at each granularity level. Number of concepts $K_y = 5$, number of examples $M_y = 10$. Win-rate greater or equal to 0.5 is highlighted in bold.

	In-distribution		Out-of-distribution	
Level	Random	KMeans	Random	KMeans
Genus	0.146	0.375	0.146	0.375
Family	0.188	0.188	0.167	0.25
Order	0.312	0.438	0.292	0.458
Class	0.438	**0.542**	0.417	**0.562**

many times the learner outperforms the baselines in all of the 50 tasks at each granularity level. We report results with three settings of bottleneck size K_y and number of examples M_y: $K_y = 1, M_y = 1$, $K_y = 3, M_y = 5$, and $K_y = 5, M_y = 10$ in Table 4.2, Table 4.3, and Table 4.4 respectively.

Ours outperforms random selection and kmeans on genus and class level across with small bottleneck size and number of examples. $(K_y = 1, M_y = 1)$ When the bottleneck size is 1 and the number of examples is 1, our method outperforms random selection and kmeans on both genus and class level for in-distribution examples and outperforms random selection on the class level for out-of-distribution examples. This suggests that our method is effective in selecting concepts and examples for teaching the learner to predict the label from the image when the bottleneck size is small and the number of examples is small.

Out method does not outperform random selection and kmeans with median bottleneck size and number of examples. $(K_y = 3, M_y = 5)$ When the bottleneck size is 3 and the number of examples is 5, our method does not outperform random selection and kmeans on different levels of granularity for either in-distribution examples and out-of-distribution examples.

Ours outperforms kmeans on class level with larger bottleneck size and number of examples. $(K_y = 5, M_y = 10)$ When the bottleneck size is 5 and the number of examples is 10, our method outperforms kmeans on the class level for out-of-distribution examples. This suggests that our method has more potential in selecting concepts and examples for higher levels of granularity when the bottleneck size is large and the number of examples is large.

4.6 Discussion

From the results we can see that the effectiveness of the teaching with concepts and examples paradigm is subject to several factors.

First, both the quality of the concept pool and the selected bottleneck is crucial for the effectiveness of the teaching. The concept pool could be improved by including more diverse concepts that capture the distinctive features of data from different classes. The bottleneck selection could be improved by selecting a bottleneck that is more informative for the task. In our experiemnts, we perform teaching on tasks with different levels of granularity, and the results show that the teaching has different effectiveness on different tasks. Tasks with different levels of granularity may require different levels of abstraction in the concepts and examples. When the task is too fine-grained, the concepts pool may not be detailed enough to capture the nuanced differences between classes. When the task is too coarse-grained, the concepts pool may not be diverse enough to capture the broad distribution of the data. In our experiments, we find that the teaching is more effective on tasks with the highest level of granularity, which suggests that our concept pool may be too general and not detailed enough to capture the differences between very similar looking classes. In future work, we could improve the concept pool by including more class-specific concepts that capture the distinctive features that are unique to one class or only appear in a few classes.

Second, the effectiveness of the teaching is also subject to the quality of the examples. In our experiments, we find that the teaching is more effective when the examples are more diverse and representative of the data distribution. This is observed from the performance of random and kmeans examples, which covers a broader distribution of the data. They are more effective than our concept selection method that are based on the concept bottleneck and does not consider coverage of the data distribution. The comparasion of the performance of different example selection methods in out-of-distribution examples also suggests that the examples should be more diverse and representative to teach more generalizable learners. In future work, we could improve the example selection by selecting examples that are more diverse and representative of the data distribution besides the objective of having stronge concept association.

4.7 Conclusion

In conclusion, we have presented a novel teaching paradigm that combines concepts and examples to teach simulated human learners to perform natural image classification tasks from four different levels of granularity and in both in-distribution and out-of-distribution examples. We find that the teaching is more effective on tasks with the highest level of granularity, which suggests that our concept pool may be too general and not detailed enough to capture the differences between very similar looking classes. These exploratory results suggest that the teaching with concepts and examples paradigm has the potential to improve the performance of human learners on image classification tasks when the concept pool and examples are carefully selected to be properly informative for the task.

CHAPTER 5
FUTURE WORK

In this dissertation, we have investigated how to enable a two-way communication between humans and AI systems. We have explored three different aspects: understanding machines through data and explanations, learning from humans for better decision support, and teaching humans with concepts and examples. In this chapter, we discuss future directions for human-AI collaboration in these three aspects.

From the perspective of understanding machines, we have explored how to understand machine predictions through data distribution and interactive explanations. We have shown that human-AI teams have different interactions when making predictions on in-distribution and out-of-distribution examples. One way to extend our current work is to investigate how to improve human-AI interactions on out-of-distribution examples. When human-AI teams make predictions on out-of-distribution examples, we should provide more informative explanations to help humans understand the machine's predictions. We need to inform humans about the distribution shift and the potential biases in the data and safeguard against over-reliance on the machine's predictions.

We have also shown that interactive explanations may not always improve human performance and may reinforce human biases. We should also investigate how to help humans understand the machine's predictions in a way that does not change humans' decision-making processes on the task.

From the perspective of learning from humans, we have explored how to learn human-compatible representations for case-based decision support. We have shown that AI systems that are trained to capture human similarity judgments can produce human-compatible representations. With different decision support policies, we have shown that human-compatible representations can lead to better decision support and improve human performance. One way to extend our current work is to investigate which decision support policies are more suitable for given tasks and human preferences. When AI systems have good performance on a task, we may consider policies that encourage humans to rely more on the AI system. When AI systems have poor performance on a task, we may consider policies that encourage humans to rely more on their own intuition. We can explore methods to adapt the decision support policies to the task and human preferences.

From the perspective of teaching humans, we have explored how to teach fine-grained image classification with concept and example selection. We have shown that teaching with informative concepts and examples can improve human performance. One way to extend our current work is to investigate how to incorporate coverage and diversity into the concept and example selection algorithm. When we select concepts and examples, we should consider the coverage of the concept pool and the diversity of the examples to help humans understand the task more comprehensively.

Another way to extend our current work is to enforce contrastive thinking during the human learning process. We should investigate how to teach humans in batches where we encourage them to think about the similarities and differences between examples. This could be implemented with careful design of an interactive interface that asks humans to reflect on the differences between examples in a batch. We hope to see better teaching performance when we encourage humans to think contrastively.

In conclusion, there are many opportunities for future work in encouringe two-way communication between humans and AI systems. We invite researchers to explore different aspects of human-AI collaboration and think about what is the best way to enable humans and AI systems to work together effectively. We hope that our work will inspire future research in this area and help to build AI systems that empower humans to make better decisions.

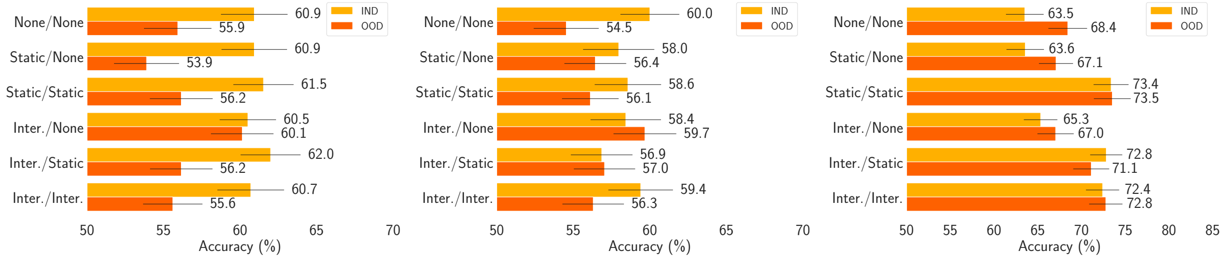
CHAPTER 6

APPENDIX

6.1 Appendix for Chapter 2

6.1.1 Human Performance in Absolute Accuracy

Fig. 6.1 shows human performance in absolute accuracy.



(a) ICPSR (b) COMPAS (c) BIOS

Figure 6.1: Human-AI team performance of different explanation types. Distribution types are indicated by color of the bar and error bars represent 95% confidence intervals. In ICPSR, human-AI team performance is significantly higher in-distribution than out-of-distribution in all explanation types ($p < 0.01$) except Interactive/None. In COMPAS, in-distribution performance is significantly higher only in None/None ($p < 0.005$). In BIOS, out-of-distribution performance is significantly higher only in None/None ($p < 0.01$).

6.1.2 COMPAS Figures

We also present the figures related to our hypotheses and results for COMPAS. The accuracy gain in COMPAS is shown in Fig. 6.2. The agreement and agreement by correctness are shown in Fig. 6.3 and Fig. 6.4. The subjective perception on whether real-time assistance is useful and whether training is useful is shown in Fig. 6.5. Fig. 6.6 shows the percentage of participants who rate a feature important.

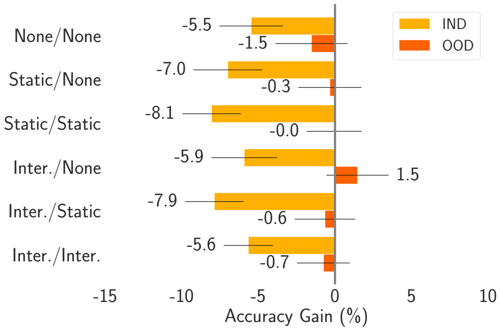


Figure 6.2: Accuracy gain of different conditions in COMPAS. Distribution types are indicated by color of the bar and error bars represent 95% confidence intervals. Accuracy gain is only sometimes positive (although not statistically significant). Performance gap between human-AI teams and AI is significantly smaller in all explanation types except None/None.

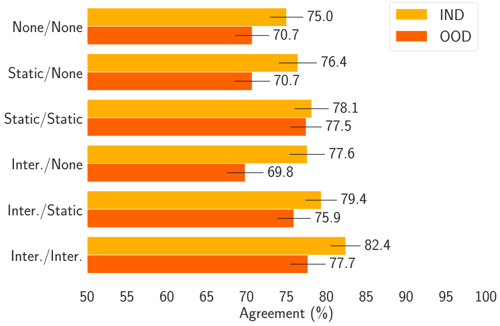


Figure 6.3: Agreement with AI predictions of different conditions in COMPAS. Distribution types are indicated by color of the bar and error bars represent 95% confidence intervals. As compared to BIOS, agreement with AI predictions is much higher in-distribution than out-of-distribution in all explanation types except Static/Static and Interactive/Static.

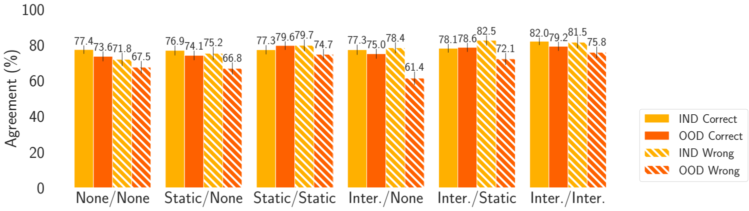


Figure 6.4: Agreement with AI grouped by distribution type and whether AI predictions are correct in COMPAS. Distribution types are indicated by color of the bar, bars with stripes represent wrong AI predictions, and error bars represent 95% confidence intervals. human-AI teams are only more likely to agree with correct AI predictions out-of-distribution for all explanation types except None/None.

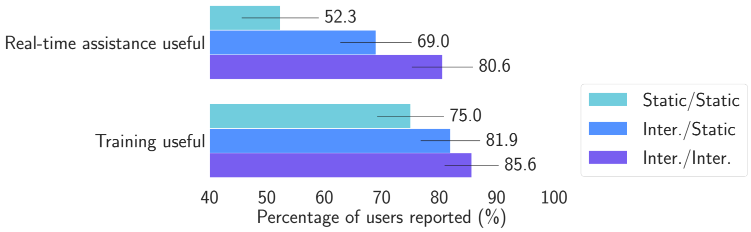


Figure 6.5: Subjective perception on whether real-time assistance is useful and whether training is useful. x-axis shows the percentage of users that answered affirmatively.

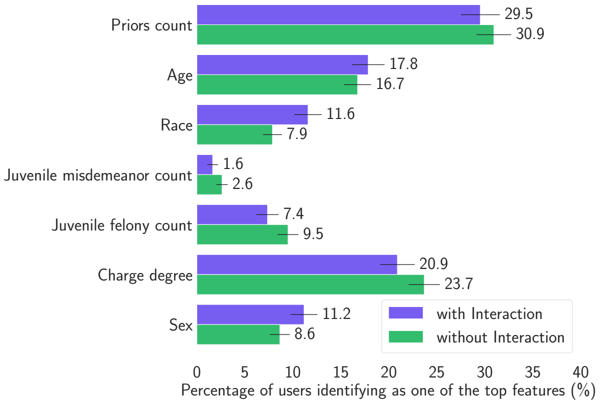
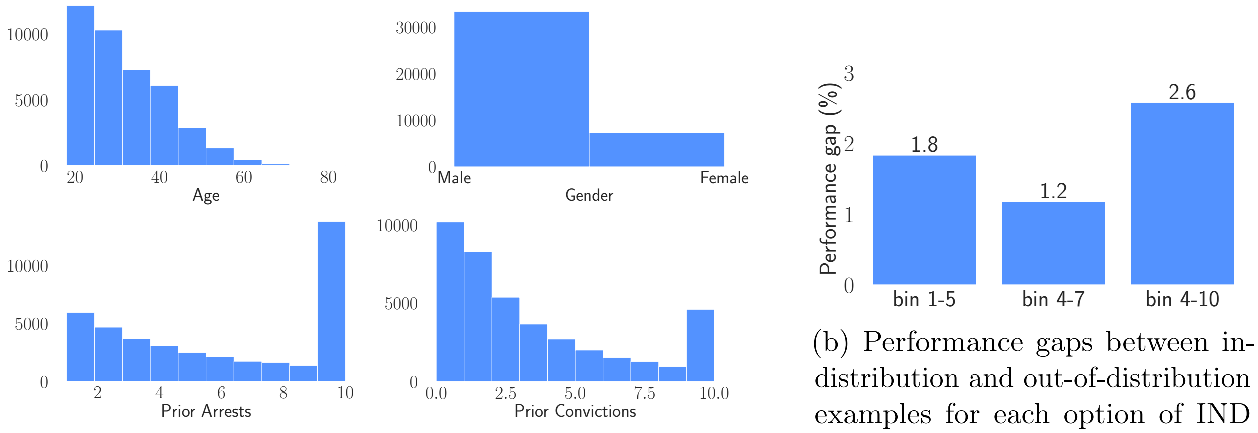


Figure 6.6: Percentage of users finding a feature important in COMPAS. The features are sorted in decreasing order from top to bottom by their Spearman correlation with groundtruth labels.

6.1.3 In-distribution vs. Out-of-distribution Setup

In this section, we will explain how we split in-distribution examples and out-of-distribution examples in ICPSR as an demonstration of the in-distribution vs. out-of-distribution setup procedures. First, we need to select an attribute for splitting. For each candidate attribute, we split the data into 10 bins of equal size based on this attribute. We do this because we want to explore different settings of splitting, e.g. different ranges of bins to use for training. In other words, we hope to have as much control as possible when we consider which bins are IND and which are OOD. For example in Fig. 6.7a we show the histogram of four candidate attributes that we can use to split the examples. The distribution is so extreme in Gender and Prior Arrests (too many "Male" in Gender and too many "10" in Prior Arrests) that if we choose any of these two attributes, we would have no choice but to use nearly half of our data as either IND or OOD, because we want to avoid having the same value in both distribution types. Similarly Prior Convictions also limits our choices of bins due to its extreme distribution. Since there are too many instances with value "0," bin 1 and bin 2 would both consist of defendants who have 0 prior convictions after binning. If we were to use a splitting where bin 1 is IND and bin 2 is OOD, then this splitting does not make sense (one distribution type falls into the other's distribution). Therefore we finally choose Age as the attribute. We also design desiderata 3) for the in-distribution vs. out-of-distribution setup to avoid these situations.

After selecting the attribute, we also need to decide which bins we use as in-distribution examples and which bins as out-of-distribution examples. In ICPSR, the options we explore are: 1) bin 1-5 as IND: age ≥ 30 as IND and age > 30 as OOD. 2) bin 4-7 as IND: age between 25-36 as IND and age < 25 or age > 36 as OOD; 3) bin 4-10 as IND: age ≥ 25 as IND and age < 25 as OOD; We finally settled on option 3) because it gives us the largest performance gap between in-distribution examples and out-of-distribution examples (Fig. 6.7b). Note that this performance gap looks different from what we present in Fig. 2 in the main paper



(a) Histograms of a subset of features in ICPSR. We choose Age as the attribute to split the distribution types because it has a relatively uniform distribution.

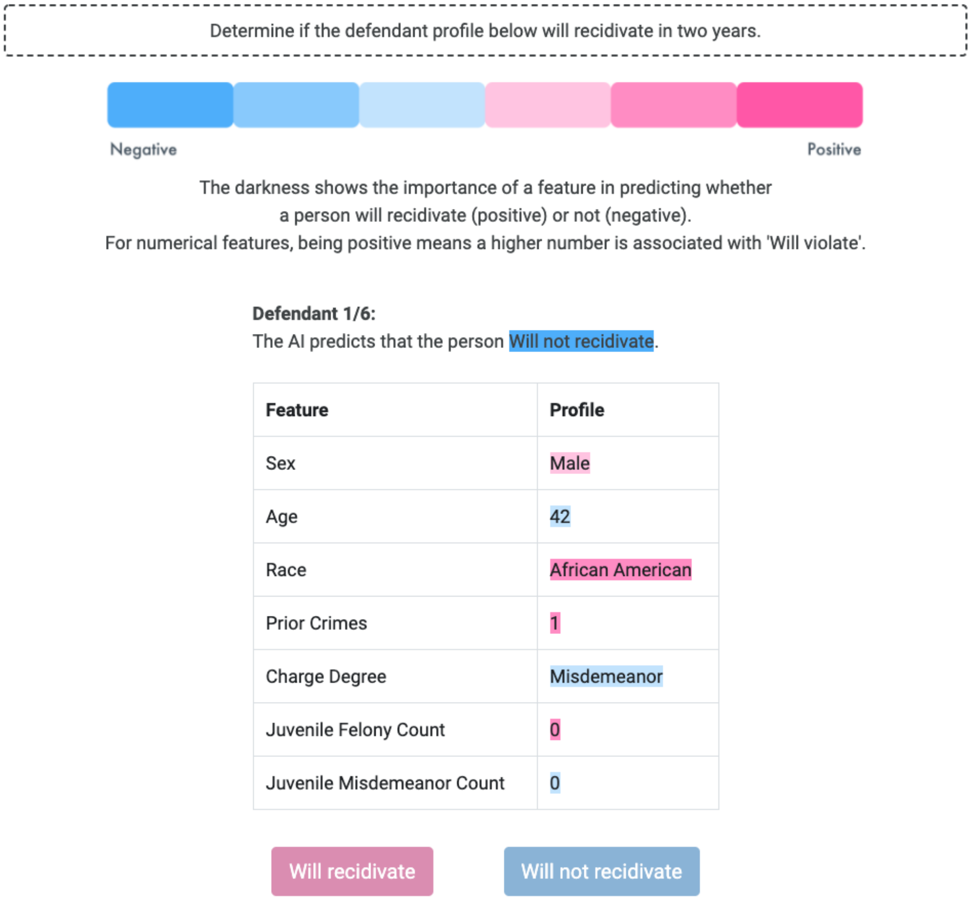
(b) Performance gaps between in-distribution and out-of-distribution examples for each option of IND bins in ICPSR. Using bin 4-10 as IND gives the largest performance gap.

Figure 6.7: Figures for ICPSR in-distribution vs. out-of-distribution setup.

because here we use the entire testset (after balancing labels) for evaluation, instead of the 360 randomly sampled examples we prepare for the user study. The in-distribution examples in the random samples are easier for AI, therefore giving us an even larger performance gap between in-distribution and out-of-distribution.

6.1.4 User Interface Designs

Screenshots for static assistance for COMPAS. Fig. 6.8 shows the static assistance for COMPAS.



Determine if the defendant profile below will recidivate in two years.

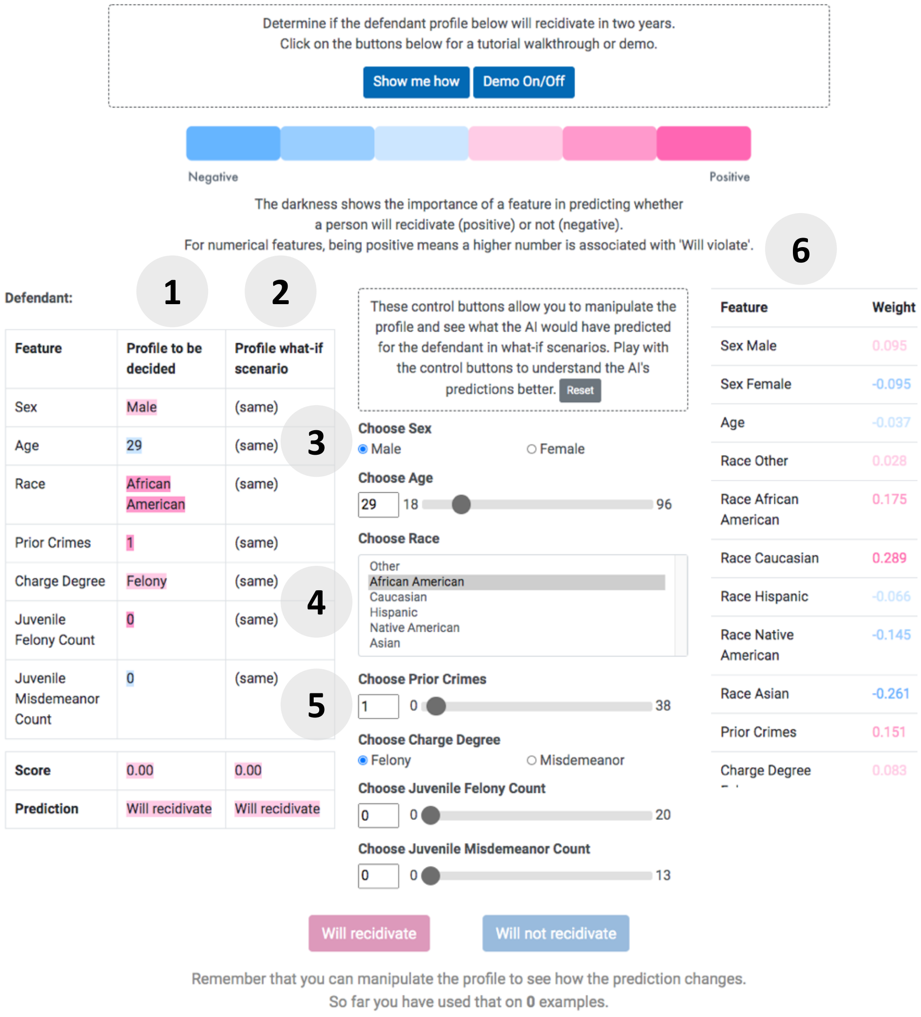
The darkness shows the importance of a feature in predicting whether
a person will recidivate (positive) or not (negative).
For numerical features, being positive means a higher number is associated with 'Will violate'.

Defendant 1/6:
The AI predicts that the person Will not recidivate.

Feature	Profile
Sex	Male
Age	42
Race	African American
Prior Crimes	1
Charge Degree	Misdemeanor
Juvenile Felony Count	0
Juvenile Misdemeanor Count	0

Figure 6.8: Static assistance for COMPAS.

Interactive interface for COMPAS. Fig. 6.9 shows the interactive interface for COMPAS.



Determine if the defendant profile below will recidivate in two years.
Click on the buttons below for a tutorial walkthrough or demo.

The darkness shows the importance of a feature in predicting whether a person will recidivate (positive) or not (negative).
For numerical features, being positive means a higher number is associated with 'Will violate'.

Defendant:

Feature	Profile to be decided	Profile what-if scenario
Sex	Male	(same)
Age	29	(same)
Race	African American	(same)
Prior Crimes	1	(same)
Charge Degree	Felony	(same)
Juvenile Felony Count	0	(same)
Juvenile Misdemeanor Count	0	(same)

Score	0.00	0.00
Prediction	Will recidivate	Will recidivate

These control buttons allow you to manipulate the profile and see what the AI would have predicted for the defendant in what-if scenarios. Play with the control buttons to understand the AI's predictions better. Reset

Choose Sex
● Male ○ Female

Choose Age
29 18 96

Choose Race
Other
African American
Caucasian
Hispanic
Native American
Asian

Choose Prior Crimes
1 0 38

Choose Charge Degree
● Felony ○ Misdemeanor

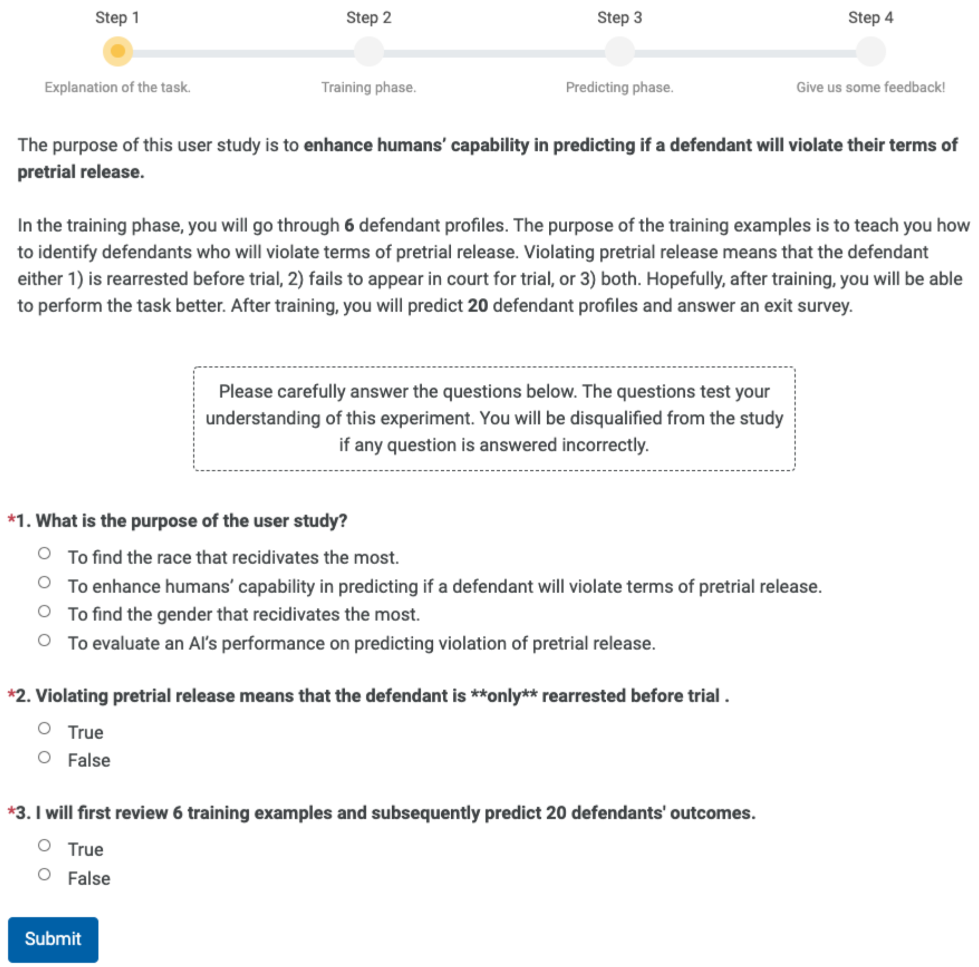
Choose Juvenile Felony Count
0 0 20

Choose Juvenile Misdemeanor Count
0 0 13

Feature	Weight
Sex Male	0.095
Sex Female	-0.095
Age	-0.037
Race Other	0.028
Race African American	0.175
Race Caucasian	0.289
Race Hispanic	-0.066
Race Native American	-0.145
Race Asian	-0.261
Prior Crimes	0.151
Charge Degree	0.083

Remember that you can manipulate the profile to see how the prediction changes.
So far you have used that on **0** examples.

Figure 6.9: In addition to static assistance such as feature highlights and showing AI predictions, users are able to manipulate the features of defendant's profile to see any changes in the AI prediction. Illustration of interactive console for COMPAS: 1) actual defendant's profile; 2) edited defendant's profile if user manipulates any features; 3) user is able to edit the value of *Sex* and *Charge Degree* with radio buttons; 4) user is able to edit the value of *Race* with dropdown; 5) user is able to edit the value *Age*, *Prior Crimes*, *Juvenile Felony Count*, and *Juvenile Misdemeanor Count* with sliders; 6) a table displaying features and respective coefficients, the color and darkness of the color shows the importance of a feature in predicting whether a person will recidivate or not.



The purpose of this user study is to **enhance humans' capability in predicting if a defendant will violate their terms of pretrial release.**

In the training phase, you will go through **6** defendant profiles. The purpose of the training examples is to teach you how to identify defendants who will violate terms of pretrial release. Violating pretrial release means that the defendant either 1) is rearrested before trial, 2) fails to appear in court for trial, or 3) both. Hopefully, after training, you will be able to perform the task better. After training, you will predict **20** defendant profiles and answer an exit survey.

Please carefully answer the questions below. The questions test your understanding of this experiment. You will be disqualified from the study if any question is answered incorrectly.

***1. What is the purpose of the user study?**

○ To find the race that recidivates the most.
○ To enhance humans' capability in predicting if a defendant will violate terms of pretrial release.
○ To find the gender that recidivates the most.
○ To evaluate an AI's performance on predicting violation of pretrial release.

***2. Violating pretrial release means that the defendant is **only** rearrested before trial .**

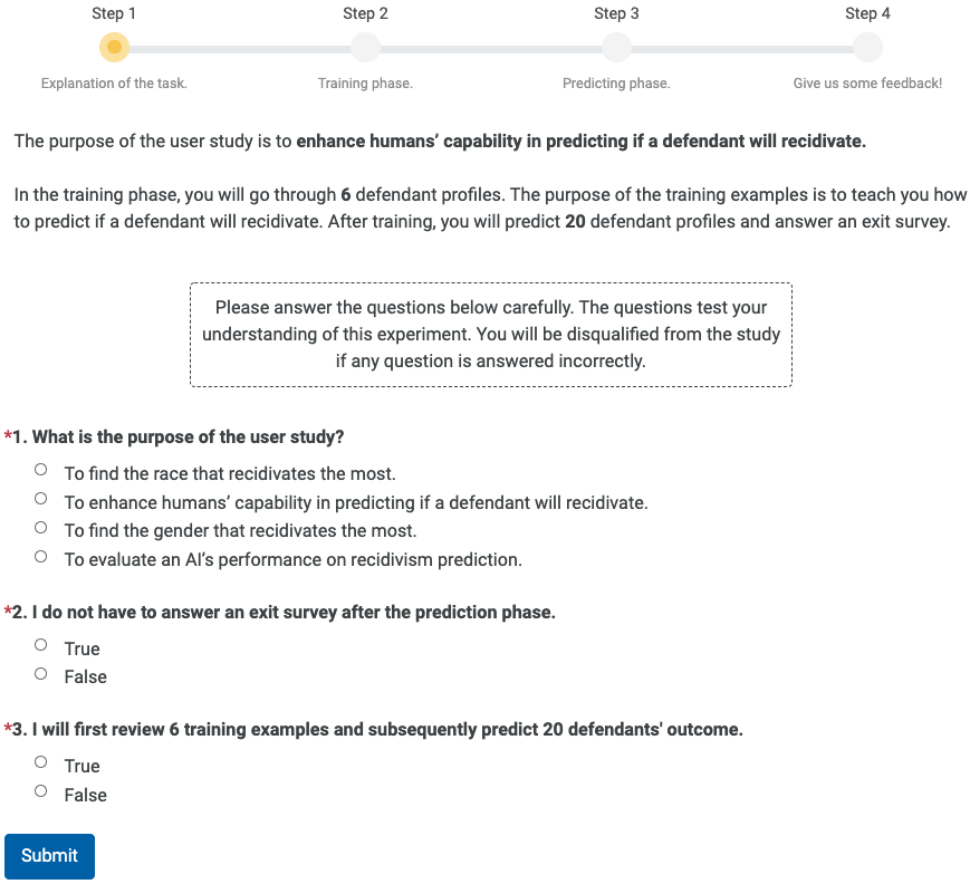
○ True
○ False

***3. I will first review 6 training examples and subsequently predict 20 defendants' outcomes.**

○ True
○ False

(a) Attention check for ICPSR. The user is required to select the correct answers before they are allowed to proceed to the training phase. The answers to the attention check questions can be found in the same page.

Attention check. In the recidivism prediction task, many participants found one of the attention-check questions to be very tricky. As the purpose of the attention-check questions was not to intentionally trick users into answering the wrong answer, we made edits to one of the attention-check questions to remove any confusion. In addition, many participants felt that it was better if they could refer to the definitions of certain terminology. As such, we combined the instructions and attention-check questions step in one page so participants are able to look up on the definitions if they had forgotten. Fig. 6.10 shows screenshots of attention check questions in all the three tasks.



The purpose of the user study is to **enhance humans' capability in predicting if a defendant will recidivate.**

In the training phase, you will go through **6** defendant profiles. The purpose of the training examples is to teach you how to predict if a defendant will recidivate. After training, you will predict **20** defendant profiles and answer an exit survey.

Please answer the questions below carefully. The questions test your understanding of this experiment. You will be disqualified from the study if any question is answered incorrectly.

***1. What is the purpose of the user study?**

○ To find the race that recidivates the most.
○ To enhance humans' capability in predicting if a defendant will recidivate.
○ To find the gender that recidivates the most.
○ To evaluate an AI's performance on recidivism prediction.

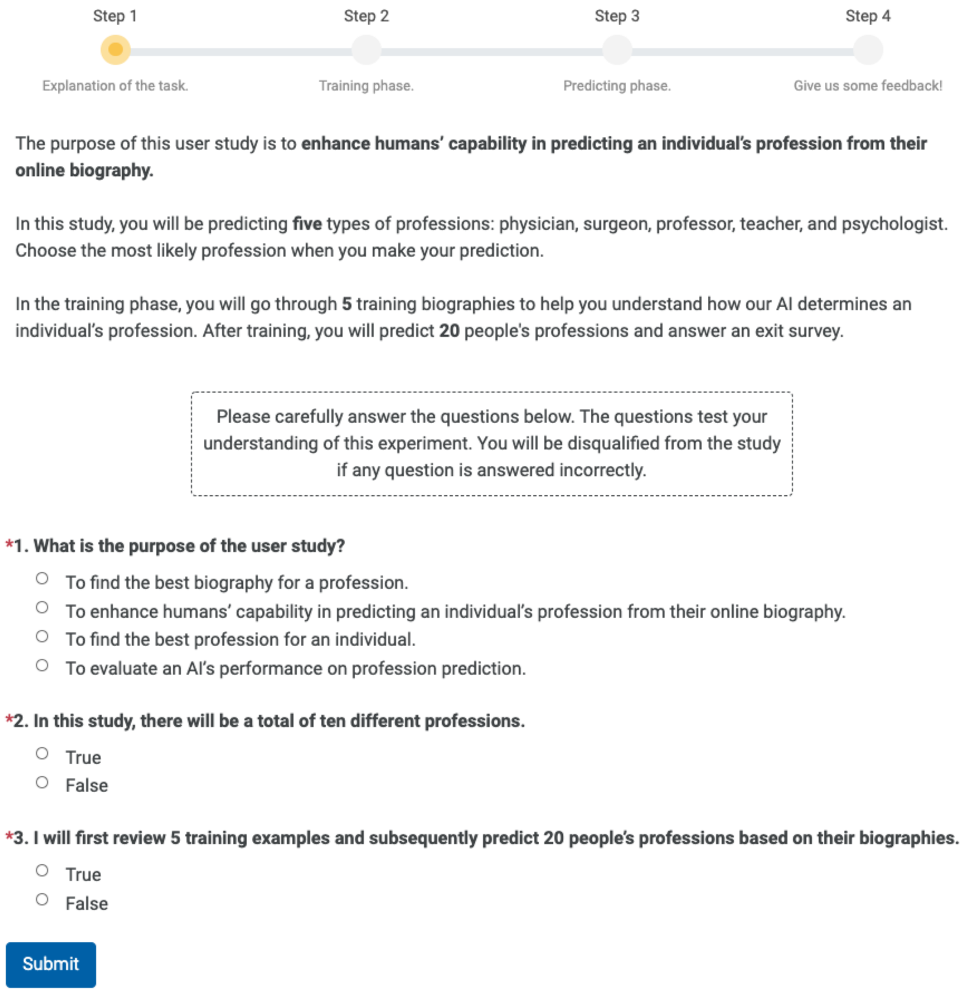
***2. I do not have to answer an exit survey after the prediction phase.**

○ True
○ False

***3. I will first review 6 training examples and subsequently predict 20 defendants' outcome.**

○ True
○ False

(b) Attention check for COMPAS. The user is required to select the correct answers before they are allowed to proceed to the training phase. The answers to the attention check questions can be found in the same page.



The purpose of this user study is to **enhance humans' capability in predicting an individual's profession from their online biography.**

In this study, you will be predicting **five** types of professions: physician, surgeon, professor, teacher, and psychologist. Choose the most likely profession when you make your prediction.

In the training phase, you will go through **5** training biographies to help you understand how our AI determines an individual's profession. After training, you will predict **20** people's professions and answer an exit survey.

Please carefully answer the questions below. The questions test your understanding of this experiment. You will be disqualified from the study if any question is answered incorrectly.

***1. What is the purpose of the user study?**

- ○ To find the best biography for a profession.
- ○ To enhance humans' capability in predicting an individual's profession from their online biography.
- ○ To find the best profession for an individual.
- ○ To evaluate an AI's performance on profession prediction.

***2. In this study, there will be a total of ten different professions.**

- ○ True
- ○ False

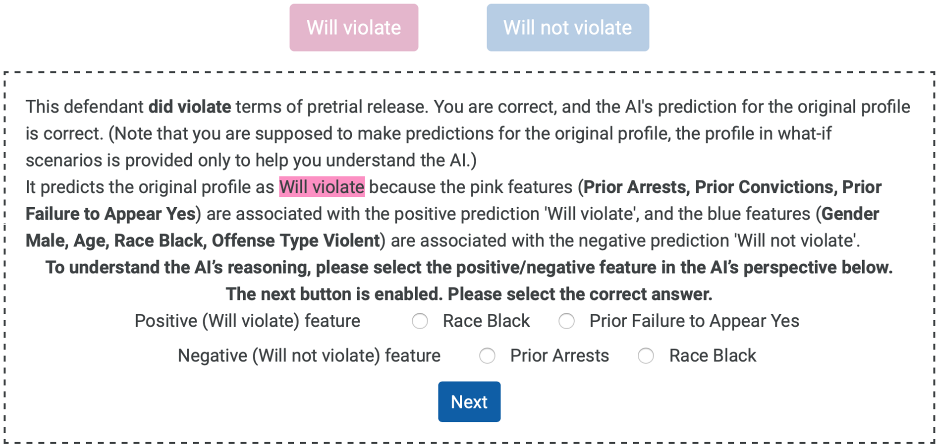
***3. I will first review 5 training examples and subsequently predict 20 people's professions based on their biographies.**

- ○ True
- ○ False

(c) Attention check for BIOS. The user is required to select the correct answers before they are allowed to proceed to the training phase. The answers to the attention check questions can be found in the same page.

Figure 6.10: Attention check questions.

Feature quiz. In the training phase of each task, for all explanation types except None/None, we also design a feature quiz to see if users understand the association between features and labels correctly. For each training instance in the training phase, we prompt users the quiz as in Fig. 6.11 after they make the prediction. We ask users to identify the positive and negative feature from two candidate features. The correct candidate is prepared by a random sampling from all the features that are currently shown in the interface, while the incorrect candidate is sampled from all features that do not have the correct polarity as prompted. The submit button is disabled for five seconds starting from the appearance of the check to refrain users from submitting a random answer.



This defendant **did violate** terms of pretrial release. You are correct, and the AI's prediction for the original profile is correct. (Note that you are supposed to make predictions for the original profile, the profile in what-if scenarios is provided only to help you understand the AI.)

It predicts the original profile as Will violate because the pink features (**Prior Arrests, Prior Convictions, Prior Failure to Appear Yes**) are associated with the positive prediction 'Will violate', and the blue features (**Gender Male, Age, Race Black, Offense Type Violent**) are associated with the negative prediction 'Will not violate'.

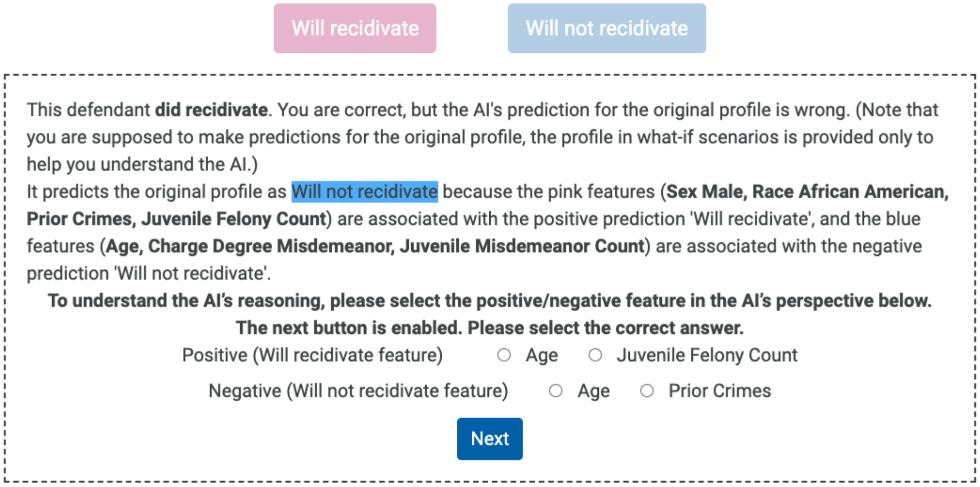
To understand the AI's reasoning, please select the positive/negative feature in the AI's perspective below.

The next button is enabled. Please select the correct answer.

Positive (Will violate) feature ○ Race Black ○ Prior Failure to Appear Yes

Negative (Will not violate) feature ○ Prior Arrests ○ Race Black

(a) Features quiz for ICPSR. The user is required to select the correct positive and negative feature before they are allowed to proceed to the next instance. In this example, the correct answer for positive feature is *Prior Failure to Appear Yes*, and the correct answer for negative feature is *Race Black*.



This defendant **did recidivate**. You are correct, but the AI's prediction for the original profile is wrong. (Note that you are supposed to make predictions for the original profile, the profile in what-if scenarios is provided only to help you understand the AI.)

It predicts the original profile as Will not recidivate because the pink features (**Sex Male, Race African American, Prior Crimes, Juvenile Felony Count**) are associated with the positive prediction 'Will recidivate', and the blue features (**Age, Charge Degree Misdemeanor, Juvenile Misdemeanor Count**) are associated with the negative prediction 'Will not recidivate'.

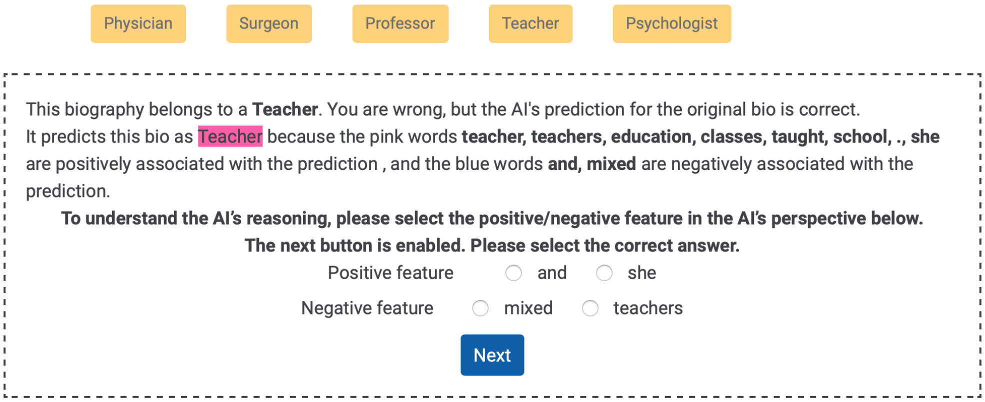
To understand the AI's reasoning, please select the positive/negative feature in the AI's perspective below.

The next button is enabled. Please select the correct answer.

Positive (Will recidivate feature) ○ Age ○ Juvenile Felony Count

Negative (Will not recidivate feature) ○ Age ○ Prior Crimes

(b) Features quiz for COMPAS. The user is required to select the correct positive and negative feature before they are allowed to proceed to the next instance. In this example, the correct answer for positive feature is *Juvenile Felony Count*, and the correct answer for negative feature is *Age*.



This biography belongs to a **Teacher**. You are wrong, but the AI's prediction for the original bio is correct.
It predicts this bio as Teacher because the pink words **teacher, teachers, education, classes, taught, school, ., she** are positively associated with the prediction , and the blue words **and, mixed** are negatively associated with the prediction.

To understand the AI's reasoning, please select the positive/negative feature in the AI's perspective below.

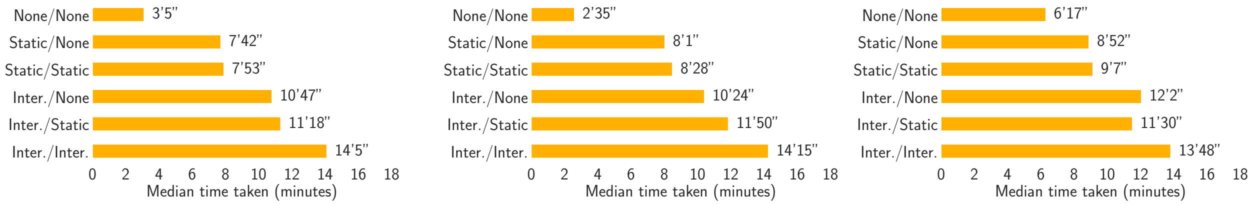
The next button is enabled. Please select the correct answer.

Positive feature ○ and ○ she

Negative feature ○ mixed ○ teachers

(c) Features quiz for BIOS. The user is required to select the correct positive and negative feature before they are allowed to proceed to the next instance. In this example, the correct answer for positive feature is *she*, and the correct answer for negative feature is *mixed*.

Figure 6.11: Feature quiz.



(a) Median time taken by users in ICPSR.

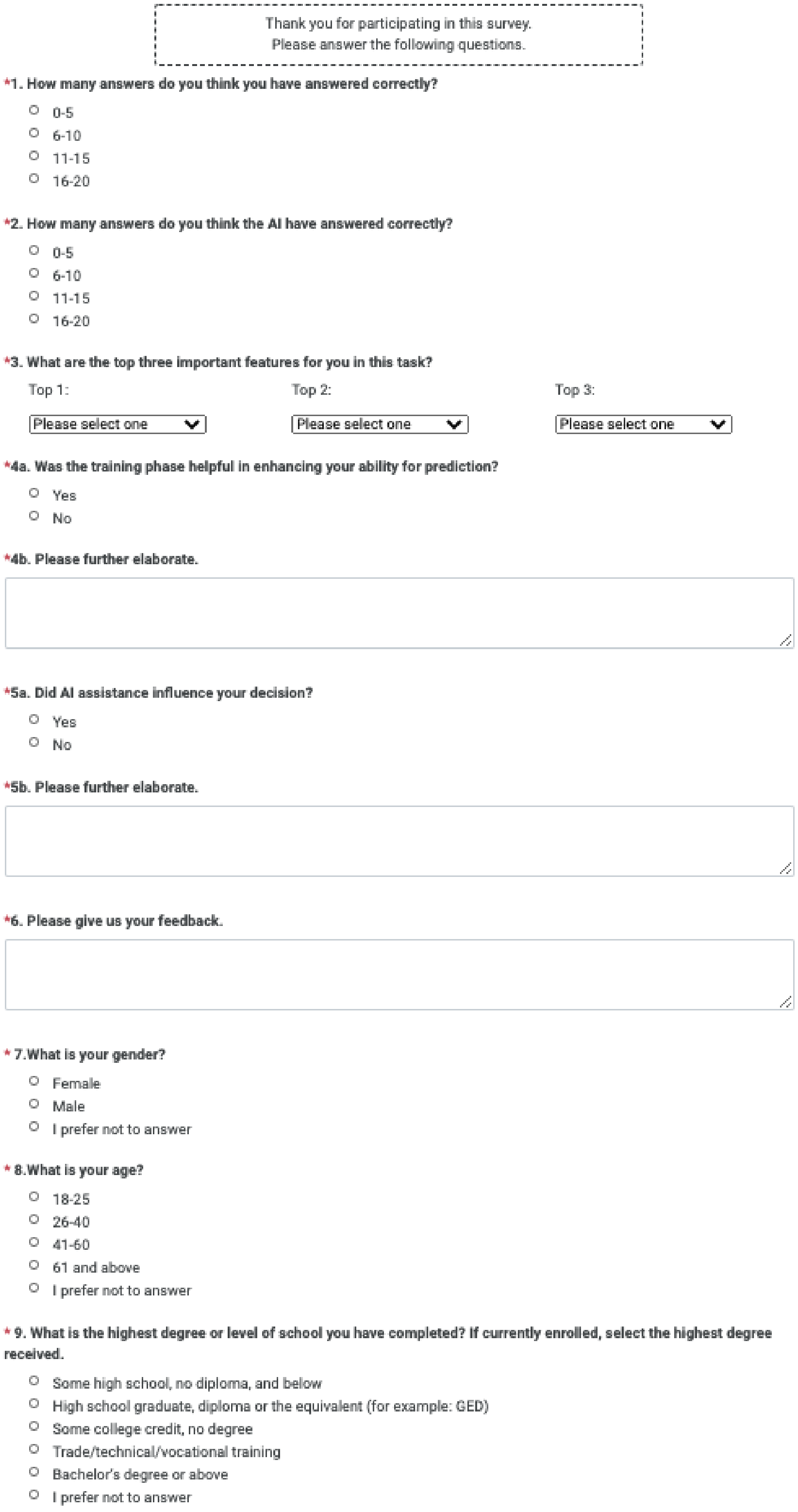
(b) Median time taken by users in COMPAS.

(c) Median time taken by users in BIOS.

Figure 6.12: Median of time taken by MTurk users in each explanation type.

Details for experiments on Mechanical Turk. We report the median time taken by the users to complete each task. The median time taken for ICPSR, COMPAS, and BIOS are 9'55", 9'16", and 8'59" respectively. In Fig. 6.12, we show the median time taken for each explanation type. We are reporting the median time taken due to a few outliers in the data collected where user is inactive for a long period of time during the study.

6.1.5 Survey Questions



Thank you for participating in this survey.
Please answer the following questions.

*1. How many answers do you think you have answered correctly?

○ 0-5
○ 6-10
○ 11-15
○ 16-20

*2. How many answers do you think the AI have answered correctly?

○ 0-5
○ 6-10
○ 11-15
○ 16-20

*3. What are the top three important features for you in this task?

Top 1:	Top 2:	Top 3:
Please select one	Please select one	Please select one

*4a. Was the training phase helpful in enhancing your ability for prediction?

○ Yes
○ No

*4b. Please further elaborate.

*5a. Did AI assistance influence your decision?

○ Yes
○ No

*5b. Please further elaborate.

*6. Please give us your feedback.

* 7.What is your gender?

○ Female
○ Male
○ I prefer not to answer

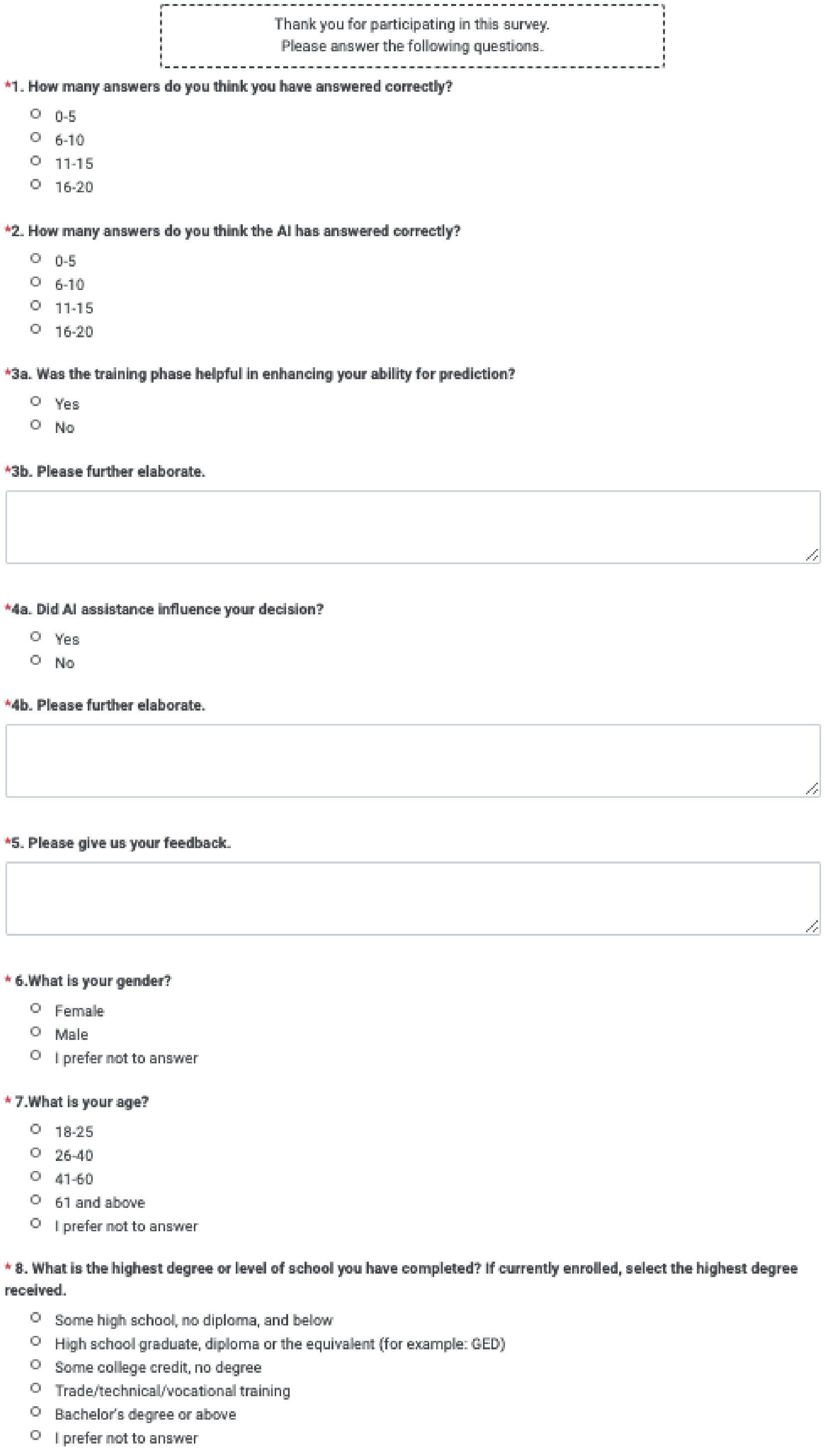
* 8.What is your age?

○ 18-25
○ 26-40
○ 41-60
○ 61 and above
○ I prefer not to answer

* 9. What is the highest degree or level of school you have completed? If currently enrolled, select the highest degree received.

○ Some high school, no diploma, and below
○ High school graduate, diploma or the equivalent (for example: GED)
○ Some college credit, no degree
○ Trade/technical/vocational training
○ Bachelor's degree or above
○ I prefer not to answer

(a) Survey questions for ICPSR and COMPAS.



Thank you for participating in this survey.
Please answer the following questions.

*1. How many answers do you think you have answered correctly?

○ 0-5
○ 6-10
○ 11-15
○ 16-20

*2. How many answers do you think the AI has answered correctly?

○ 0-5
○ 6-10
○ 11-15
○ 16-20

*3a. Was the training phase helpful in enhancing your ability for prediction?

○ Yes
○ No

*3b. Please further elaborate.

*4a. Did AI assistance influence your decision?

○ Yes
○ No

*4b. Please further elaborate.

*5. Please give us your feedback.

* 6.What is your gender?

○ Female
○ Male
○ I prefer not to answer

* 7.What is your age?

○ 18-25
○ 26-40
○ 41-60
○ 61 and above
○ I prefer not to answer

* 8. What is the highest degree or level of school you have completed? If currently enrolled, select the highest degree received.

○ Some high school, no diploma, and below
○ High school graduate, diploma or the equivalent (for example: GED)
○ Some college credit, no degree
○ Trade/technical/vocational training
○ Bachelor's degree or above
○ I prefer not to answer

(b) Survey questions for BIOS.

Figure 6.13: Survey questions.

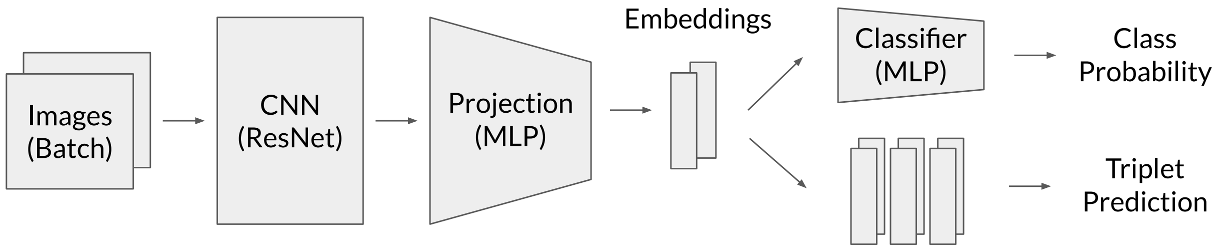


Figure 6.14: Architecture of the human-compatible representations model.

6.2 Appendix for Chapter 3

6.2.1 Implementation Detail

The architecture of our model is presented in Fig. 6.14. We first encode image inputs using a Convolutional Neural Network (CNN), and then project the output into an high-dimension representation space with a projection head made of multi-layer perceptron (MLP). In our experiments we use one non-linear layer to project the output of the CNN into our representation space. For classifcation task we add an MLP classifier head. We also use one non-linear layer with softmax activation. For triplet prediction, we re-index the representations with the current triplet batch and calculate prediction or loss. We use the PyTorch framework [Paszke et al., 2019] and the PyTorch Lightning framework [Falcon et al., 2019] for implementation. Hyperparameters will be reported in §3.12 for models in the synthetic experiments and in §3.13 and §3.14 for models in the human experiments.

6.2.2 Computation Resources

We use a computing cluster at our institution. We train our models on nodes with different GPUs including Nvidia GeForce RTX2080Ti, Nvidia GeForce RTX3090, Nvidia Quadro RTX 8000, and Nvidia A40. All models are trained on one allocated node with one GPU access.

REFERENCES

Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics*, pages 11–18, 2007.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*, pages 1–16, 2021.

Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, and Bo Von Schéele. A case-based decision support system for individual stress diagnosis using fuzzy similarity matching. *Computational Intelligence*, 25(3):180–195, 2009.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. URL `https://bit.ly/2QsOf4P`.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365 (6456):885–890, 2019.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464, 2020.

Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 258–262. ACM, 2019a.

Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 4. ACM, 2019b.

Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019c. doi:10.1145/3359206. URL `https://doi.org/10.1145/3359206`.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 95–106, 2020.

Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *CoRR*, abs/1806.10574, 2018a. URL `http://arxiv.org/abs/1806.10574`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Yuxin Chen, Oisin Mac Aodha, Shihan Su, Pietro Perona, and Yisong Yue. Near-optimal machine teaching via explanatory teaching sets. In *International Conference on Artificial Intelligence and Statistics*, pages 1970–1978. PMLR, 2018b.

Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 559. ACM, 2019.

Chun-Wei Chiang and Ming Yin. You'd better stop! understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*, pages 120–129, 2021.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, 2019.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

Ewart J de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. A design methodology for trust cue calibration in cognitive agents. In *International conference on virtual, augmented and mixed reality*, pages 251–262. Springer, 2014.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

William Falcon et al. Pytorch lightning, 2019.

Shi Feng and Jordan Boyd-Graber. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 229–239, 2019.

Emma Frid, Ceslo Gomes, and Zeyu Jin. Music creation by example. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference*, pages 1–8, 2016.

Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. *arXiv preprint arXiv:2001.09219*, 2020.

Nikhil Ghosh, Yuxin Chen, and Yisong Yue. Landmark ordinal embedding. In *Advances in Neural Information Processing Systems*, pages 11502–11511, 2019.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

Ian Goodfellow, Y Bengio, and A Courville. Machine learning basics. In *Deep learning*, volume 1, pages 98–164. MIT press, 2016.

Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99. ACM, 2019a.

Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):50, 2019b.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

Yugo Hayashi and Kosuke Wakabayashi. Can ai become reliable source to support human decision making in a court scene? In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 Companion, pages 195–198, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3022198.3026338. URL `https://doi.org/10.1145/3022198.3026338`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of ICCV*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 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*, pages 1–13, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. ISSN 0092-8674. doi:https://doi.org/10.1016/j.cell.2018.02.010. URL `https://www.sciencedirect.com/science/article/pii/S0092867418301545`.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of NIPS*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2018.

Matthäus Kleindessner and Ulrike Luxburg. Uniqueness of ordinal embedding. In *Conference on Learning Theory*, pages 40–67, 2014.

Matthäus Kleindessner and Ulrike von Luxburg. Kernel functions based on triplet comparisons. In *Advances in Neural Information Processing Systems*, pages 6807–6817, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Janet L Kolodneer. Improving human decision making through case-based decision aiding. *AI magazine*, 12(2):52–52, 1991.

Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697. ACM, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Vivian Lai and Chenhao Tan. 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*, pages 29–38, 2019.

Vivian Lai, Han Liu, and Chenhao Tan. "why is 'chicago' deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 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*, pages 1675–1684, 2016.

Walter S Lasecki, Christopher D Miller, Iftekhar Naim, Raja Kushalnagar, Adam Sadilek, Daniel Gildea, and Jeffrey P Bigham. Scribe: deep integration of human and machine intelligence to caption speech in real time. *Communications of the ACM*, 60(9):93–100, 2017.

Gaobo Liang and Lixin Zheng. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer methods and programs in biomedicine*, 187: 104964, 2020.

Shu-hsien Liao. Case-based decision support system: Architecture for simulating military command and control. *European Journal of Operational Research*, 123(3):558–567, 2000.

Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, Jennifer Skeem, et al. The limits of human predictions of recidivism. *Science advances*, 6(7):eaaz0652, 2020.

Adam Liptak. Sent to prison by a software program's secret algorithms, 2017.

Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Han Liu, Vivian Lai, and Chenhao Tan. 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, 2021.

Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan. Learning human-compatible representations for case-based decision support. In *International Conference on Learning Representations*, 2023.

Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. Novice-ai music co-creation via ai-steering tools for deep generative models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

R Duncan Luce. Individual choice behavior. 1959.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

Maranda McBride and Shona Morgan. Trust calibration for automated decision aids. *Institute for Homeland Security Solutions*, pages 1–11, 2010.

Jon McCormack, Toby Gifford, Patrick Hutchings, Maria Teresa Llano Rodriguez, Matthew Yee-King, and Mark d'Inverno. In a silent way: Communication between ai and improvising musicians beyond sound. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1334. URL `https://www.aclweb.org/anthology/P19-1334`.

John M McGuirl and Nadine B Sarter. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, 48(4): 656–665, 2006.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788): 89–94, 2020.

Stephanie M Merritt, Deborah Lee, Jennifer L Unnerstall, and Kelli Huber. Are well-calibrated users effective users? associations between calibration of trust and performance on an automation-aided task. *Human Factors*, 57(1):34–47, 2015.

C Metz, C Metz, N Tiku, I Lapowsky, K Finley, C Thompson, E Griffith, and M Spector. In two moves, alphago and lee sedol redefined the future. wired, 2016.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

Bonnie M Muir. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539, 1987.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436, 2021.

Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 101–110. ACM, 2002.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 1907–1916, 2015. URL `https://bit.ly/2ObMA1J`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52, 2021.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L. Ball, Marc Mendelson, Gary Maartens, Daniël J. van Hoving, Rulan Griesel, Andrew Y. Ng, Tom H. Boyles, and Matthew P. Lungren. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *npj Digital Medicine*, 3(1):1–8, September 2020. ISSN 2398-6352. doi:10.1038/s41746-020-00322-2. URL `https://www.nature.com/articles/s41746-020-00322-2`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of KDD*, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *Proceedings of ICML*, 2014.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Supreme Court of the United States. Daubert v. merrell dow pharmaceuticals, inc., 1993. 509 U.S. 579.

Supreme Court of Wisconsin. State of Wisconsin, Plaintiff-Respondent, v. Eric L. Loomis, Defendant-Appellant, 2016. URL `https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690`.

Mohammad Reza Taesiri, Giang Nguyen, and Anh Nguyen. Visual correspondence-based explanations improve ai robustness and human-ai team accuracy. In *Advances in Neural Information Processing Systems*, 2022.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 673–680, 2011. URL `https://bit.ly/2xAshnJ`.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. 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*, pages 107–118, 2020.

Yoshikazu Terada and Ulrike Luxburg. Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855, 2014.

Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, August 2020. ISSN 1546-170X. doi:10.1038/s41591-020-0942-0. URL `https://www.nature.com/articles/s41591-020-0942-0`.

United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. State court processing statistics, 1990-2009: Felony defendants in large urban counties., 2014.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6, 2012. URL `https://bit.ly/2O2TF8h`.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. 2017.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv:1907.03324*, 2019.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1180–1192, 2017.

Tongshuang Wu, Daniel S Weld, and Jeffrey Heer. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(4):1–27, 2019.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3090–3100, 2023.

Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 279. ACM, 2019.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 2020.

REFERENCES

Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics*, pages 11–18, 2007.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias, 2016.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3, 2016.

Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*, pages 1–16, 2021.

Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

Shahina Begum, Mobyen Uddin Ahmed, Peter Funk, Ning Xiong, and Bo Von Schéele. A case-based decision support system for individual stress diagnosis using fuzzy similarity matching. *Computational Intelligence*, 25(3):180–195, 2009.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. URL `https://bit.ly/2QsOf4P`.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365 (6456):885–890, 2019.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 454–464, 2020.

Carrie J Cai, Jonas Jongejan, and Jess Holbrook. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 258–262. ACM, 2019a.

Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 4. ACM, 2019b.

Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019c. doi:10.1145/3359206. URL `https://doi.org/10.1145/3359206`.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 95–106, 2020.

Chaofan Chen, Oscar Li, Alina Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *CoRR*, abs/1806.10574, 2018a. URL `http://arxiv.org/abs/1806.10574`.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Yuxin Chen, Oisin Mac Aodha, Shihan Su, Pietro Perona, and Yisong Yue. Near-optimal machine teaching via explanatory teaching sets. In *International Conference on Artificial Intelligence and Statistics*, pages 1970–1978. PMLR, 2018b.

Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 559. ACM, 2019.

Chun-Wei Chiang and Ming Yin. You'd better stop! understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*, pages 120–129, 2021.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4069–4082, 2019.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In

Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 120–128, 2019.

Ewart J de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. A design methodology for trust cue calibration in cognitive agents. In *International conference on virtual, augmented and mixed reality*, pages 251–262. Springer, 2014.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

William Falcon et al. Pytorch lightning, 2019.

Shi Feng and Jordan Boyd-Graber. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 229–239, 2019.

Emma Frid, Ceslo Gomes, and Zeyu Jin. Music creation by example. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th Web for All Conference*, pages 1–8, 2016.

Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. *arXiv preprint arXiv:2001.09219*, 2020.

Nikhil Ghosh, Yuxin Chen, and Yisong Yue. Landmark ordinal embedding. In *Advances in Neural Information Processing Systems*, pages 11502–11511, 2019.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

Ian Goodfellow, Y Bengio, and A Courville. Machine learning basics. In *Deep learning*, volume 1, pages 98–164. MIT press, 2016.

Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99. ACM, 2019a.

Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):50, 2019b.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

Yugo Hayashi and Kosuke Wakabayashi. Can ai become reliable source to support human decision making in a court scene? In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 Companion, pages 195–198, New York, NY, USA, 2017. Association for Computing Machinery. doi:10.1145/3022198.3026338. URL `https://doi.org/10.1145/3022198.3026338`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of ICCV*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 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*, pages 1–13, 2019.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y.L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018. ISSN 0092-8674. doi:https://doi.org/10.1016/j.cell.2018.02.010. URL `https://www.sciencedirect.com/science/article/pii/S0092867418301545`.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of NIPS*, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2018.

Matthäus Kleindessner and Ulrike Luxburg. Uniqueness of ordinal embedding. In *Conference on Learning Theory*, pages 40–67, 2014.

Matthäus Kleindessner and Ulrike von Luxburg. Kernel functions based on triplet comparisons. In *Advances in Neural Information Processing Systems*, pages 6807–6817, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Pang Wei Koh, Shiori Sagawa, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Janet L Kolodneer. Improving human decision making through case-based decision aiding. *AI magazine*, 12(2):52–52, 1991.

Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697. ACM, 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Vivian Lai and Chenhao Tan. 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*, pages 29–38, 2019.

Vivian Lai, Han Liu, and Chenhao Tan. “why is ‘chicago’ deceptive?” towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 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*, pages 1675–1684, 2016.

Walter S Lasecki, Christopher D Miller, Iftekhar Naim, Raja Kushalnagar, Adam Sadilek, Daniel Gildea, and Jeffrey P Bigham. Scribe: deep integration of human and machine intelligence to caption speech in real time. *Communications of the ACM*, 60(9):93–100, 2017.

Gaobo Liang and Lixin Zheng. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer methods and programs in biomedicine*, 187: 104964, 2020.

Shu-hsien Liao. Case-based decision support system: Architecture for simulating military command and control. *European Journal of Operational Research*, 123(3):558–567, 2000.

Zhiyuan “Jerry” Lin, Jongbin Jung, Sharad Goel, Jennifer Skeem, et al. The limits of human predictions of recidivism. *Science advances*, 6(7):eaaz0652, 2020.

Adam Liptak. Sent to prison by a software program’s secret algorithms, 2017.

Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Han Liu, Vivian Lai, and Chenhao Tan. 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, 2021.

Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan. Learning human-compatible representations for case-based decision support. In *International Conference on Learning Representations*, 2023.

Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J Cai. Novice-ai music co-creation via ai-steering tools for deep generative models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

R Duncan Luce. Individual choice behavior. 1959.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

Maranda McBride and Shona Morgan. Trust calibration for automated decision aids. *Institute for Homeland Security Solutions*, pages 1–11, 2010.

Jon McCormack, Toby Gifford, Patrick Hutchings, Maria Teresa Llano Rodriguez, Matthew Yee-King, and Mark d'Inverno. In a silent way: Communication between ai and improvising musicians beyond sound. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-1334. URL `https://www.aclweb.org/anthology/P19-1334`.

John M McGuirl and Nadine B Sarter. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, 48(4): 656–665, 2006.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg C Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788): 89–94, 2020.

Stephanie M Merritt, Deborah Lee, Jennifer L Unnerstall, and Kelli Huber. Are well-calibrated users effective users? associations between calibration of trust and performance on an automation-aided task. *Human Factors*, 57(1):34–47, 2015.

C Metz, C Metz, N Tiku, I Lapowsky, K Finley, C Thompson, E Griffith, and M Spector. In two moves, alphago and lee sedol redefined the future. wired, 2016.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

Bonnie M Muir. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies*, 27(5-6):527–539, 1987.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436, 2021.

Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 101–110. ACM, 2002.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *International Conference on Machine Learning*, pages 1907–1916, 2015. URL `https://bit.ly/2ObMA1J`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In

Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–52, 2021.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning.* The MIT Press, 2009.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L. Ball, Marc Mendelson, Gary Maartens, Daniël J. van Hoving, Rulan Griesel, Andrew Y. Ng, Tom H. Boyles, and Matthew P. Lungren. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *npj Digital Medicine*, 3(1):1–8, September 2020. ISSN 2398-6352. doi:10.1038/s41746-020-00322-2. URL `https://www.nature.com/articles/s41746-020-00322-2`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of KDD*, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Adish Singla, Ilija Bogunovic, Gábor Bartók, Amin Karbasi, and Andreas Krause. Near-optimally teaching the crowd to classify. In *Proceedings of ICML*, 2014.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Supreme Court of the United States. Daubert v. merrell dow pharmaceuticals, inc., 1993. 509 U.S. 579.

Supreme Court of Wisconsin. State of Wisconsin, Plaintiff-Respondent, v. Eric L. Loomis, Defendant-Appellant, 2016. URL `https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690`.

Mohammad Reza Taesiri, Giang Nguyen, and Anh Nguyen. Visual correspondence-based explanations improve ai robustness and human-ai team accuracy. In *Advances in Neural Information Processing Systems*, 2022.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 673–680, 2011. URL `https://bit.ly/2xAshnJ`.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. 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*, pages 107–118, 2020.

Yoshikazu Terada and Ulrike Luxburg. Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855, 2014.

Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, August 2020. ISSN 1546-170X. doi:10.1038/s41591-020-0942-0. URL `https://www.nature.com/articles/s41591-020-0942-0`.

United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics.

State court processing statistics, 1990-2009: Felony defendants in large urban counties., 2014.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6, 2012. URL `https://bit.ly/2O2TF8h`.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. 2017.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

Hilde JP Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv:1907.03324*, 2019.

James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1180–1192, 2017.

Tongshuang Wu, Daniel S Weld, and Jeffrey Heer. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(4):1–27, 2019.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang 'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian McAuley. Learning concise and descriptive attributes for visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3090–3100, 2023.

Qian Yang, Aaron Steinfeld, and John Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 279. ACM, 2019.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. *arXiv preprint arXiv:1801.05927*, 2018.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 2020.