Three Maxims for Developing Human-Centered AI for Decision Making

A dissertation

submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Daniel S. Weld, Chair

Jamie Morgenstern

Besmira Nushi

Program Authorized to Offer Degree:

Computer Science and Engineering

University of Washington

Abstract

Three Maxims for Developing Human-Centered AI for Decision Making

Gagan Bansal

Chair of the Supervisory Committee:

Professor Daniel S. Weld

Computer Science and Engineering

We focus on AI-advised decision making, where AI systems (*e.g.*, classifiers) are deployed to assist users to make better decisions (*e.g.*, in healthcare, finance, and criminal justice). While the dominant development practice deploys the most "accurate" autonomous AI to assist users, we argue that in order for AI to *augment* users, we should shift the focus of research to developing *human-centered* AI (HCAI). HCAI systems additional requirements atop those of autonomous AI. They are not just capable but also: trustworthy and dependable, they communicate and coordinate their reasoning with users, and complement users' expertise. We specifically develop and study three relevant maxims for developing HCAI systems: 1) help users understand *when to trust* AI recommendations, 2) *preserve user's mental model of AI's trustworthiness*, and 3) train AI to *optimize for team performance*.

Through experiments on various tasks that involve AI-assisted decision making, we show that a) contrary to expectations, current XAI methods may be insufficient for helping users understand when to rely on AI recommendations. b) It is easier for users to create a mental model of AI's trustworthiness when its error boundary (*i.e.*, regions where it errs) is simple and deterministic. c) The current practice of updates to AI systems (*e.g.*, to improve its accuracy) can result in models that violate user trust, *e.g.*, by introducing errors on examples on which the system was previously correct; however, we also show that its possible to create models that preserve trust by considering *compatibility* of updates during the training process. d) For a simple setting, we formally show that by accommodating the user's mental model in the AI's training process, we can train a model that results in higher a human-AI team performance than the team performance achieved with the most accurate AI. Finally, we discuss open problems and future work in

developing HCAI including enabling explanatory dialogs (as opposed to static, one-shot explanations) and enabling user control of AI behavior. Overall, the problems and results in this thesis show the richness and interdisciplinary nature of the challenge of developing human-centered AI.

Acknowledgements

The work presented in this thesis could not have been possible without the support of many people. I am deeply grateful to my advisor, Dan Weld, for mentorship and inspiration throughout graduate school. I learned from Dan the very fundamentals of the research process— choosing problems, critiquing literature, framing research questions, and designing experiments. Interacting with Dan helped me understand and deepen my love for science and creativity and, more importantly, helped me situate my reason to continue a research career in this understanding. Finally, I am thankful for his care and friendship during times of distress and impromptu classes on sailing, outdoors, and photography.

My internships at Microsoft Research and collaborations with Besmira Nushi, Ece Kamar, and Eric Horvitz introduced me to the rich problem space of human-AI interaction which ultimately became the subject of this thesis. In fact the work presented in the Chapters 3, 4, and 5 began during my 2018 and 2019 summer internships at MSR. Besa not just provided low- and high-level advice during my internships but also throughout graduate school. From Ece, I learned the value of clearly presenting my arguments and research. And, Eric was a consistent source of energy and confidence.

I am also thankful to the rest of my committee members, Jamie Morgenstern and Linda Boyle, for their feedback and questions that helped improve this thesis. The Chapter 2 of this thesis could not have been possible without the help from Sherry Wu, Joyce Zhou, and Raymond Fok who made substantially large contributions that ultimately helped this work turn into a CHI paper. Marco Ribeiro provided useful, critical feedback on almost all my conference submissions. Gonzalo Ramos and Alison Smith helped broaden my perspectives on human-centered AI.

The work presented in this thesis could not have been possible without the funding from many agencies including ONR grants N00014-18-1-2193 and N00014-21-1-2707, NSF RAPID grant 2040196, the Univer-

sity of Washington WRF/Cable Professorship, the Allen Institute for Artificial Intelligence (AI2), Microsoft Research, and The Future of Life Foundation.

I am thankful to the Allen School staff for creating such a welcoming environment for graduate students. Elise Dorough made my life easy on innumerable number of occasions. Sandy Kaplan provided helpful feedback on my paper drafts and helped me become a better writer.

My graduate school experience would not have been complete without the support of my peers at UW and MSR, many of whom are now some of my closest friends: Jonathan Bragg, Danielle Bragg, Chris Lin, Mandar Joshi, Srini Iyer, Sachin Mehta, Maaz Ahmad, Anupreet Porwal, Sherry Wu, Jim Chen, Xiao Lin, Vikash Kumar, Marco Ribeiro, Eunice Jun, Svet Kolev, Deepali Aneja, William Agnew, Joyce Zhou, Raymond Fok, Ben Lee, Ramya Ramakrishnan, Andi Peng, Olivia Baddeley, Tracy Tran, and friends from UW Salsa Club.

One of my favorite part of being a graduate was getting to interact with extremely smart students. And, I was fortunate to work with many excellent Masters, undergraduate and high-school advisees: Diana Iftimie, Lynsey Liu, Ziyao Huang, Joyce Zhou, Jake Sippy, Prithvi Tarale, and Cindy Su.

Finally, I am thankful to my family for providing unconditional love and support throughout graduate school and for encouraging me to choose my path. While my grandfather, Jaswant Rai Bansal, passed away this Fall and did not get to see me finish my thesis, I know he was very proud of my decision to pursue a career in research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays, AI systems are not just being used to automate tasks. Instead people work with an AI teammate, *e.g.*, because the team may help perform better than either the AI or human alone [Nagar and Malone, 2011; Patel *et al.*, 2019; Kamar *et al.*, 2012], or because legal requirements may prohibit complete automation [GDPR, 2020]. From a rational perspective, for such human-AI teams, just like for any team, improving the performance of the whole team is more important than improving the performance of an individual member (*e.g.*, just the AI). Yet, to date for the most part, the AI community has focused on maximizing the individual accuracy of machine-learned models. Implicitly assuming that these models will be used for automation. This raises an important question: Is the most accurate AI the best possible teammate for users?

We argue that the answer may be "No." For instance, considering human-human teams, *Is the best-ranked tennis player necessarily the best doubles teammate?* Clearly not — teamwork puts additional demands on players besides high individual performance. Reliable human partners are not just capable but also trustworthy and dependable, they communicate and coordinate their reasoning, and complement their teammate's skills. Similarly, we argue that the development of high-performing human-AI teams may require creating AI that exhibit additional abilities that are beyond high accuracy and *human-centered*, such as reliable and consistent behavior, ability to explain and justify their actions and recommendations, offering users control and feedback, optimizing for joint human-AI performance, and enabling *appropriate* trust and reliance. The latter, in fact, is the motivation behind much work in intelligible AI [Caruana *et al.*, 2015; Weld and Bansal, 2019 05] and post-hoc explainable AI [Ribeiro *et al.*, 2016].

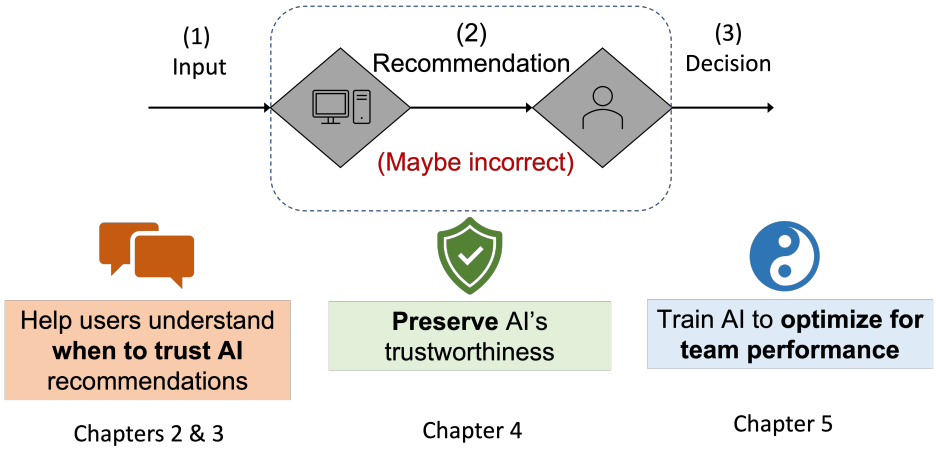


Figure 1.1: Three maxims for developing human-centered AI for AI-assisted decision making.

In this work, we consider human-AI teams where an AI assists a human decision maker by offering its recommendations but humans retains the agency to make the final decision (Figure 1.1), a context also known as *AI-assisted decision making* [Bansal *et al.*, 2019b]. We focus on AI-assisted decision making because its a formal model that represents the simplest kind of human-AI teamwork, and thus is a good place to build a precise understanding of human-AI interaction. We consider three human-centered maxims for such developing such AI:

M1. *Help users understand when to trust AI recommendations:* In Chapter 2, we analyze prior works that evaluate efficacy of AI explanations at improving human-AI team performance and show that they only observed improvements when the AI, alone, outperformed both the human and the best human-AI team [Bansal *et al.*, 2021b]. This raises an important research question: can explanations lead to *complementary performance*, i.e., with accuracy higher than both the human and the AI working alone? To answer this, we conduct new studies across multiple NLP tasks (classification and question answering) where human and AI have a comparable performance. While our experiments show that AI augmentation produced complementary improvements, accuracy was not increased by state-of-the-art explanations compared to simply displaying AI's confidence. Explanations increased the chance that users accept the AI's recommendation regardless of its correctness. In addition to the effect of explanations, in Chapter 3, we highlight two key properties of an AI's *error boundary* (*i.e.*, regions where it errs), *parsimony* and *stochasticity*, that affect humans' ability to create mental models of AI's

trustworthiness [Bansal *et al.*, 2019a]. Experiments show that we people find it easier to create such mental models when the system's error boundary is simple and deterministic.

M2. *Preserve user's mental model of AI's trustworthiness:* In Chapter 4 we show that updates that increase AI performance may actually hurt team performance, e.g., by creating behavior that is at odds with prior user experience and mental model of AI's trustworthiness [Bansal *et al.*, 2019b]. We introduce a practical re-training objective that can improve the compatibility of updates. Experiments across three datasets show that our approach can create updates that are more compatible, while maintaining high accuracy.

M3. *Train AI to optimize for human-AI team performance:* In Chapter 5 we suggest that directly modeling a human-AI team's collaborative process can help learn better AI teammates. Specifically, to optimize the team performance we maximize the team's expected utility, expressed in terms of quality of the final decision, cost of verifying, and accuracy of teammates [Bansal *et al.*, 2021b]. Our experiments with linear and non-linear models on real-world, high-stakes datasets show that the most accurate AI may not lead to highest team performance and show the benefit of modeling user's mental model during training through improvements in expected team utility across datasets, considering parameters such as human skill and the cost of mistakes. We discuss the shortcoming of current optimization approaches beyond well-studied loss functions such as log-loss, and encourage future work on AI optimization problems motivated by human-AI teaming.

In Chapter 6, we conclude with a discussion of lessons learnt and two ongoing and future works for developing HCAI. First, since current – static – explanations may be insufficient to help users decide when to trust AI recommendations, how can create better explanations? Since results from human psychology show that the process of explaining is a conversation between the explainer and explainee [Miller, 2018], we propose enabling interactive explanatory dialogs that allow users to drilldown into the system's reasoning and ask follow up questions. Second, since HCAI systems will remain imperfect, how can we allow users to not just understand and predict their mistakes but also to correct system recommendations and/or reasoning? One option is to allow users to give feedback in terms of the system's explanatory vocabulary. However, just like feature representations used by AI systems may be incomplete, their explanatory vocabulary may remain

incomplete. Thus, we specifically focus on the problem of allowing users to refine an AI's explanatory vocabulary and control its behavior.

Through all these maxims, this thesis shows that the challenge of developing human-centered AI and improving human-AI interaction is interdisciplinary in nature: Solving it requires marrying ideas from artificial intelligence, machine learning, and human-computer interaction to further our understanding of what kind of AI models we train and how they interface people.

Chapter 2

The Effect of Explanations on Team Performance

Many researchers motivate explainable AI with studies showing that human-AI team performance on decision-making tasks improves when the AI *explains* its recommendations. However, prior studies observed improvements from explanations only when the AI, alone, outperformed both the human and the best team. Can explanations help lead to *complementary performance*, where team accuracy is higher than either the human or the AI working solo? Such a performance level is only possible if the explanations help users better decide when to trust the model's recommendations (Maxim 1). We conduct mixed-method user studies on three datasets, where an AI with accuracy comparable to humans helps participants solve a task (explaining itself in some conditions). While we observed complementary improvements from AI augmentation, they were *not increased* by explanations. Rather, explanations increased the chance that humans will accept the AI's recommendation, regardless of its correctness (bolstering blind trust rather than appropriate trust). Our result poses new challenges for human-centered AI: Can we develop explanatory approaches that encourage appropriate trust in AI, and therefore help generate (or improve) complementary performance?

Figure 2.1: (Best viewed in color) Do AI explanations lead to complementary team performance? In a team setting, when given an input, the human uses (usually imperfect) recommendations from an AI model to make the final decision. We seek to understand if automatically generated explanations of the AI's recommendation improve team performance compared to baselines, such as simply providing the AI's recommendation, R, and confidence. (A) Most previous work concludes that explanations improve team performance (*i.e.*, $\Delta_A > 0$); however, it usually considers settings where AI systems are much more accurate than people and even the human-AI team. (B) Our study considers settings where human and AI performance is comparable to allow room for complementary improvement. We ask, "Do explanations help in this context, and how do they compare to simple confidence-based strategies?" (Is $\Delta_B > 0$?).

2.1 Introduction

Although the accuracy of Artificial Intelligence (AI) systems is rapidly improving, in many cases, it remains risky for an AI to operate autonomously, *e.g.*, in high-stakes domains or when legal and ethical matters prohibit full autonomy. A viable strategy for these scenarios is to form *Human-AI teams*, in which the AI system augments one or more humans by recommending its predictions, but people retain agency and have accountability on the final decisions. Examples include AI systems that predict likely hospital readmission to assist doctors with correlated care decisions [Bayati *et al.*, 2014 10; Caruana *et al.*, 2015; Wiens *et al.*, 2016] and AIs that estimate recidivism to help judges decide whether to grant bail to defendants [Angwin *et al.*, 2016; Hayashi and Wakabayashi, 2017]. In such scenarios, it is important that the human-AI team achieves *complementary performance* (*i.e.*, performs better than either alone): From a decision-theoretic perspective, a rational developer would only deploy a team if it adds utility to the decision-making process [Morgenstern and Von Neumann, 1953]. For example, significantly improving decision accuracy by closing deficiencies in automated reasoning with human effort, and vice versa [Horvitz and Paek, 2007; Tan *et al.*, 2018].

Many researchers have argued that such human-AI teams would be improved if the AI systems could *explain their reasoning*. In addition to increasing trust between humans and machines or improving the

speed of decision making, one hopes that an explanation should help the responsible human know when to trust the AI's suggestion and when to be skeptical, *e.g.*, when the explanation doesn't "make sense." Such *appropriate reliance* [Lee and See, 2004] is crucial for users to leverage AI assistance and improve task performance [Bilgic, 2005]. Indeed, at first glance, it appears that researchers have already confirmed the utility of explanations on tasks ranging from medical diagnosis [Cai *et al.*, 2019; Lundberg *et al.*, 2018 10 01], data annotation [Schmidt and Biessmann, 2019] to deception detection [Lai and Tan, 2019]. In each case, the papers show that, when the AI provides explanations, team accuracy reaches a level higher than human-alone.

However, a careful reading of these papers shows another commonality: in every situation, while explanations are shown to help raise team performance *closer* to that of the AI, one would achieve an even better result by stripping humans from the loop and letting the AI operate autonomously (Figure 2.1A & Table 2.1). Thus, the existing work suggests several important open questions for the AI and HCI community: Do explanations help achieve *complementary performance* by enabling humans to anticipate when the AI is potentially incorrect? Furthermore, do explanations provide significant value over simpler strategies such as displaying the AI's uncertainty? In the quest to build the best human-machine teams, such questions deserve critical attention.

To explore these questions, we conduct new experiments where we control the study design, ensuring that the AI's accuracy is *comparable* to the human's (Figure 2.1B). Specifically, we measure the human skill on our experiment tasks and then control AI accuracy by purposely selecting study samples where AI has comparable accuracy. This setting simulates situations where there is a strong incentive to deploy human-AI teams, *e.g.*, because there exists more potential for complementary performance (by correcting each other's mistakes), and where simple heuristics such as blindly following the AI are unlikely to achieve the highest performance.

We selected three common-sense tasks that can be tackled by crowd workers with little training: sentiment analysis of book and beer reviews and a set of *LSAT* questions that require logical reasoning. We conducted large-scale studies using a variety of explanation sources (AI versus expert-generated) and strategies (explaining just the predicted class, or explaining other classes as well). We observed complementary performance on every task, but — surprisingly — explanations did not appear to offer benefit compared to

simply displaying the AI's confidence. Notably, explanations increased reliance on recommendations even when the AI was incorrect. Our result echoes prior work on inappropriate trust on systems [Kaur *et al.*, 2020; Mitchell *et al.*, 2019], *i.e.*, explanations can lead humans to either follow incorrect AI suggestions or ignore the correct ones [Stumpf *et al.*, 2009]. However, using end-to-end studies, we go one step further to quantify the impact of such over-reliance on objective metrics of team performance.

As a first attempt to tackle the problem of blind reliance on AI, we introduce *Adaptive Explanation*. Our mechanism tries to reduce human trust when the AI has low confidence: it only explains the predicted class when the AI is confident, but also explains the alternative otherwise. While it failed to produce significant improvement in final team performance over other explanation types, there is suggestive evidence that the adaptive approach can push the agreement between AI predictions and human decisions towards the desired direction.

Through extensive qualitative analysis, we also summarize potential factors that should be considered in experimental settings for studying human-AI complementary performance. For example, the difference in expertise between human and AI affects whether (or how much) AI assistance will help achieve complementary performance, and the display of the explanation may affect the human's collaboration strategy. In summary:

2.1 We highlight an important limitation of previous work on explainable AI: While many studies show that explaining predictions of AI increases team performance (Table 2.1), they all consider cases where the AI system is significantly more accurate than both the human partner and the human-AI team. In response, we argue that AI explanations for decision-making should aim for complementary performance, where the human-AI team outperforms both solo human and AI.

2.2 To study complementary performance, we develop a new experimental setup and use it in studies with 1626 users on three tasks[1] to evaluate a variety of explanation sources and strategies. We observe complementary performance in every human-AI teaming condition.

2.3 However, surprisingly, we do not observe any significant increase in team performance by communicating explanations, compared to simply showing the AI's confidence. Explanations often increased accuracy when the AI system was correct but, worryingly, *decreased* it when the AI erred, resulting in

[1]All the task examples and the collected experiment data are available at `https://github.com/uw-hai/Complementary-Performance`.

Domain	Task	Performance				
		Metric	Human alone	AI alone	Team	Complementary?
Classification	Deceptive review [Lai and Tan, 2019]	Accuracy ↑	51.1%	**87.0%**	74.6%	✗
	Deceptive review [Lai *et al.*, 2020]	Accuracy ↑	54.6%	**86.3%**	74.0%	✗
	Income category [Zhang *et al.*, 2020]	Accuracy ↑	65%	**75%**	73%	✗
	Loan defaults [Green and Chen, 2019]	Norm. Brier ↑	0	**1**	0.682	✗
	Hypoxemia risk [Lundberg *et al.*, 2018 10 01]	AUC ↑	0.66	**0.81**	0.78	✗
	Nutrition prediction [Buçinca *et al.*, 2020]	Accuracy ↑	0.46	**0.75**	0.74	✗
QA	Quiz bowl [Feng and Boyd-Graber, 2019]	*"AI outperforms top trivia players."*				✗
Regression	House price [Poursabzi-Sangdeh *et al.*, 2019]	Avg. Absolute Error ↓	$331k	**$200k**	$232k	✗

Table 2.1: Recent studies that evaluate the effect of automatically generated explanations on human-AI team performance. While explanations did improve team accuracy, the performance was not complementary — acting autonomously, the AI would have performed even better. For papers with multiple domains or experiments, we took one sample with the most comparable human and AI performance. ↑ (or ↓) indicates whether the metric should be maximized (or minimized).

a minimal net change — even for our adaptive explanations. Through qualitative analysis, we discuss potential causes for failure of explanations, behavioral differences among tasks, and suggest directions for developing more effective AI explanations.

2.2 Background and Related Work

Explanations can be useful in many scenarios where a human and AI interact: transparently communicating model predictions [Ribeiro *et al.*, 2016; Feng and Boyd-Graber, 2019; Koh and Liang, 2017; Kaur *et al.*, 2020; Bilgic, 2005], teaching humans tasks like translation [Glassman *et al.*, 2015] or content moderation [Jhaver *et al.*, 2019], augmenting human analysis procedure [Jhaver *et al.*, 2019] or creativity [Clark *et al.*, 2018], legal imperatives [Miller, 2018; Weld and Bansal, 2019 05], etc. Various studies have evaluated the effect of explanations from different dimensions, including whether the explanation improves users' trust in the AI [Yu *et al.*, 2019; Kunkel *et al.*, 2019] or enables users to simulate the model predictions [Poursabzi-Sangdeh *et al.*, 2019; Chandrasekaran *et al.*, 2018 10], or assists developers to debug models [Bhatt *et al.*, 2020; Kaur *et al.*, 2020].

In this paper, we focus explicitly on *AI-assisted decision making* scenarios [Bansal *et al.*, 2019b; Wang *et al.*, 2019], where an AI assistant (*e.g.*, a classification model) makes recommendations to a human (*e.g.*, a judge), who is responsible for making final decisions (*e.g.*, whether or not to grant bail). In particular,

we assess performance in terms of the *accuracy* of the human-AI team. While other metrics can be used for evaluation (more discussed in Section 2.6.1), we directly evaluate end-to-end team accuracy for three reasons. First, deploying such a human-AI team is ideal if it achieves *complementary performance*, *i.e.*, if it outperforms both the AI and the human acting alone. Second, evaluating explanations using proxy tasks (such as whether humans can use it to guess the model's prediction) can lead to different, misleading conclusions for achieving best team performance than an end-to-end evaluation [Buçinca *et al.*, 2020]. Third, AI-assisted decision making is often listed as a major motivation for AI explanations. In recent years numerous papers have employed user studies to show that human accuracy increases if the AI system explains its reasoning for tasks as diverse as medical diagnosis, predicting loan defaults, and answering trivia questions. However, as summarized in Table 2.1, complementary performance was not observed in any of these studies – in each case, adding the human to the loop *decreased* performance compared to if AI had acted alone.

For example, in Lai *et al.* [Lai *et al.*, 2020; Lai and Tan, 2019], MTurk workers classified deceptive hotel reviews with predictions from SVM and BERT-based models, as well as explanations in the form of inline-highlights. However, models outperformed every team (see Table 1 and Figure 6 in [Lai *et al.*, 2020]). Zhang *et al.* [Zhang *et al.*, 2020] noticed the superior behavior of the models in Lai *et al.*'s work, and evaluated the accuracy and trust calibration where the gap between human and the AI performances was less severe. Still, on their task of income category prediction, their Gradient Boosted Trees model had 10% higher accuracy compared to their MTurk workers, which seemed borderline "comparable" at best. Furthermore, when run autonomously, their AI model performed just slightly better than the best team (see Section 4.2.2 and Figure 10 in [Zhang *et al.*, 2020]). A similar performance trend is observed on tasks other than classification. In Sangdeh *et al.* [Poursabzi-Sangdeh *et al.*, 2019], MTurk workers predicted house price using various regression models that generated explanations in terms of most salient features. Their models' predictions resulted in lowest error (See Figure 6 in [Poursabzi-Sangdeh *et al.*, 2019]). In Feng *et al.* [Feng and Boyd-Graber, 2019], experts and novices played Quiz Bowl with recommendation from Elastic Search system. The system explained its predictions by presenting training examples that were influential, and using inline-highlights to explain the connection between question and evidence. However, Feng *et al.* do not report the exact performance of the AI on their study sample, but mention its superiority in Section 3.1 in [Feng and Boyd-Graber, 2019] pointing out that it outperforms top trivia players. One possible exception is

Explain. Strategies	Explain. Sources	Tasks
Explain-Top-1 ●	AI ●	*Beer* ●
Explain-Top-2 ●	Expert	*Amzbook*
Adaptive		*LSAT*

Table 2.2: An overview of our tasks, explanation strategies and sources. We ran our pilot studies (Section 2.3.2) with conditions marked with ● . Based on the pilot results, we added adaptive explanations and expert explanations (Section 2.3.3). Along with two additional domains, these form the conditions for our final study conditions (Section 2.4.1).

Bligic & Mooney (2005) [Bilgic, 2005], who probably achieved complementary performance on their task of recommending books to users. However, they did not compare explanations against simple baselines, such as showing the book title or the system confidence (rating).

At least two potential causes account for the absence of complementary performance in these cases. First, task design may have hindered collaboration: previous researchers considered AI systems whose accuracy was substantially higher than the human's, leading to a small zone with potential for complementary performance (see Figure 2.1). For example, this may have made it more likely that human errors were a superset of the AI's, reducing the possibility of a human overseer spotting a machine mistake. Second, even when the task has the potential for complementary performance, it is unclear if the collaboration mechanisms under study supported it. Collaboration factors like incentives, the format of explanations, and whether AI's uncertainty was displayed may drive the human towards simple, less collaborative heuristics, such as "always trust the AI" or "never trust the AI."

2.3 Setup and Pilot Studies

To better understand the role of explanations in producing complementary performance, we enlarge the zone of potential complementarity by matching AI accuracy to that of an average human,[2] and investigate multiple explanation styles on several domains (Section 2.3.1). As Table 2.2 summarizes, we first designed and conducted pilots studies (Sections 2.3.2) and used them to inform our final study and hypotheses (Section 2.4).

[2]Of course, complementary performance may be possible even in situations when one of the team partners is significantly more accurate than the other. For example, a low-accuracy team member may be valuable if their errors are independent, because they may be able to spot mistakes made by the team majority. However, it is more difficult to observe complementary performance in such settings, so we first consider the case where humans and AI have similar accuracy. If explanations cannot provide value in such settings, it will be even more difficult to show complementary performance when teammates have disparate skills.

2.3.1 Choice of Tasks and Explanations

Since our interest is in *AI-assisted decision making*, we studied the effect of *local* explanations on team performance – that is, explaining each individual recommendation made by a model [Ribeiro *et al.*, 2016]. This contrasts with providing a global understanding of the full model all at once (*e.g.*, [Lakkaraju *et al.*, 2017]).

We conducted experiments on two types of tasks: text classification (sentiment analysis) and question answering. Text classification because it is a popular task in natural language processing (NLP) that has been used in several previous studies on human-AI teaming [Feng and Boyd-Graber, 2019; Lai *et al.*, 2020; Lipton, 2018 06; Nguyen, 2018 06; Zhang *et al.*, 2020; Hase and Bansal, 2020 07] and because it requires little domain expertise, and is thus amenable to crowdsourcing. Specifically, we selected two sentiment analysis datasets to improve the generalization of our results: beer reviews [McAuley *et al.*, 2012] and book reviews [He and McAuley, 2016]. More details about these datasets are in Section 2.4.2. While there exist various local explanation approaches for text classification, we rely on *local saliency explanations*, which explain a single prediction in terms of the importance of input features (*e.g.*, each word) towards the model's prediction (*e.g.*, positive or negative sentiment).

As commonly practiced in previous work [Lai and Tan, 2019; Lai *et al.*, 2020; Feng and Boyd-Graber, 2019], we display explanations with *inline-highlights*, *i.e.*, directly highlighting the explanation in the input text, so the user need not go back and forth between input and the explanation. While there exist other explanatory approaches, such as feature-importance visualization [Lundberg *et al.*, 2018 10 01; Green and Chen, 2019; Weerts *et al.*, 2019; Narayanan *et al.*, 2018] (more suitable for tabular data) or communicating influential training examples [Yang *et al.*, 2020; Koh and Liang, 2017] (more suitable for images), these techniques are not ideal for text because they add an additional cognitive cost to mapping the explanation to the respective text. Figure 2.2 shows one example beer review.

We also experimented with Law School Admission Test (LSAT) questions[3] because it is more challenging. In this task, every question contains four options with a unique correct answer (Figure 2.3). Again, answering *LSAT* questions requires no specialized knowledge except common-sense reasoning skills, such as recognizing logical connections and conflicts between arguments [Yu *et al.*, 2020]. Because in this case

[3] https://en.wikipedia.org/wiki/LawSchoolAdmissionTest

it is unclear how inline-highlights could be used to communicate logical constructs (*e.g.*, contradiction may not be visible by highlighting the input alone), we turned to narrative explanations which justify a candidate answer in natural language. We explain these in more detail in Section 2.4.2.

2.3.2 Pilot Study on Sentiment Classification

To iterate on the hypotheses and the associated explanation conditions for our main study (detailed later in Section 2.4), we conducted a pilot study on one of our datasets (*Beer*). The between-subject pilot study asked crowdworkers to judge the sentiment of 50 beer reviews with assistance from a logistic regression classifier in three conditions, each condition with 50 workers. One condition *only* showed the model prediction and confidence; the other two also included the following common **explanation** strategies[4]:

1. *Explain-Top-1* explains just the predicted class by highlighting the most influential words for that class.
2. *Explain-Top-2* explains the top two predicted classes, and unlike Explain-Top-1, it also color codes and highlights words for the other sentiment class.

The two strategies closely align with the design in prior work [Wang *et al.*, 2016; Lin *et al.*, 2017; Lai and Tan, 2019], and have been shown to be beneficial (Table 2.1). Explain-Top-2 also corresponds to Wang *et al.*'s suggestion to mitigate heuristic biases by explaining "multiple outcomes" [Wang *et al.*, 2019].

Observations We summarize our findings from the pilot study:

1. Contrary to many prior works, we observed *no significant changes or improvements in aggregated team accuracy by displaying either type of explanations.*
2. That said, *explaining just the predicted class (Explain-Top-1) performed better than explaining both (Explain-Top-2).*
3. We also observed that *explanations increased reliance on recommendations even when they were incorrect*: explaining the predicted class slightly improved performance (compared to confidence only) when the recommendation was correct but decreased performance when it was incorrect.
4. This effect was less pronounced in Explain-Top-2, presumably because it *encouraged users to consider alternatives and hence deterred over-reliance.* In Figure 2.2, for example, if counter-argument

[4]the saliency scores were based on feature weights learned by the linear model [Green and Chen, 2019; Lai *et al.*, 2020]

(d) was not highlighted, participants could easily stop reading at the highlighted first sentence and overlook the negative ending.

5. Finally, *participants indicated that they wanted higher quality explanations.* Crowd-workers were confused when explanations did not seem to correlate with model behavior.

Because we made similar observations in our main study, we defer detailed discussions and implications of these observations to Section 2.5.1 and Figure 2.4.

2.3.3 Additional Explanation Strategies/Sources

Added Strategy: Adaptive Explanations. The pilot study indicated that Explain-Top-2 was more beneficial than Explain-Top-1 when the classifier made mistakes, but not otherwise. Relying on the commonly seen correlations between mistakes and low-confidence [Hendrycks and Gimpel, 2017], we developed a new dynamic strategy, *adaptive explanation*, that switches between Explain-Top-1 and Explain-Top-2 depending on the AI's confidence. This method explains the top-two classes only when the classifier confidence is below a task- and model-specific threshold (described later in Section 2.4.2), explaining only the top prediction otherwise. Intuitively, it was inspired by an efficient assistant that divulges more information (confessing doubts and arguing for alternatives) only when it is unsure about its recommendation. Adaptive explanations can also be viewed as changing explanation according to *context* [Abowd *et al.*, 1999]. While we limit the our context to the AI's confidence, in general, one could rely on more features of the human-AI team, such as the user, location, or time [Horvitz, 1999; Lim *et al.*, 2009].

Added Source: Expert-Generated Explanations. Users in our pilot study were confused when the explanations did not make intuitive sense, perhaps due to either the quality of the underlying linear model-based AI. While we test state-of-the-art models in the final study, we also added *expert-generated* explanations to serve as an upper bound on explanation quality. We describe their annotation process in Section 2.4.2.

2.4 Final Study

Based on our pilot studies, we formulated our final hypotheses and used them to inform our final conditions and their interface (Section 2.4.1). We then tested these hypotheses for several tasks and AI systems

(Section 2.4.2) through crowdsourcing studies (Section 2.4.3).

2.4.1 Hypotheses, Conditions, and Interface

We formulated the following hypotheses for sentiment analysis:

H1 *Among current explanation strategies, explaining the predicted class will perform better than explaining both classes.*

H2 *The better strategies, Explain-Top-1, will still perform similarly to simply showing confidence.*

H3 *Our proposed Adaptive explanations, which combines benefits of existing strategies, will improve performance.*

H4 *Adaptive explanations would perform even better if AI could generate higher quality explanations.*

Since generating AI explanations for *LSAT* was not feasible (Section 2.3.3), we slightly modified the hypothesis for *LSAT*: we omitted the hypothesis on explanation quality (**H4**) and tested the first three hypotheses using expert- rather than AI-generated explanations.

Conditions For both domains, we ran two baseline conditions: unassisted users (**Human**), as well as a simple AI assistance that shows the AI's recommendation and confidence but *no explanation* (**Team (Conf)**). We use this simple assistance because it serves as a stronger and broadly acknowledged baseline than the alternative, *i.e.*, displaying AI's recommendation *without* confidence. First, most ML models can generate confidence scores that, in practice, correlate with the model's true likelihood to err [Hendrycks and Gimpel, 2017]. Second, displaying uncertainty in predictions can help users can make more optimal decisions [Joslyn and LeClerc, 2013; Dong and Hayes, 2012; Fernandes *et al.*, 2018; Nadav-Greenberg and Joslyn, 2009; Gkatzia *et al.*, 2016 08]. Hence, we focus on evaluating whether the explanations provide *additional* value when shown alongside confidence scores. In the rest of the paper, we indicate the explanation conditions using the following template: **Team (Strategy, Source)**. For example, Team (Explain-Top-1, AI) indicates the condition that shows the AI's explanations for the top prediction.

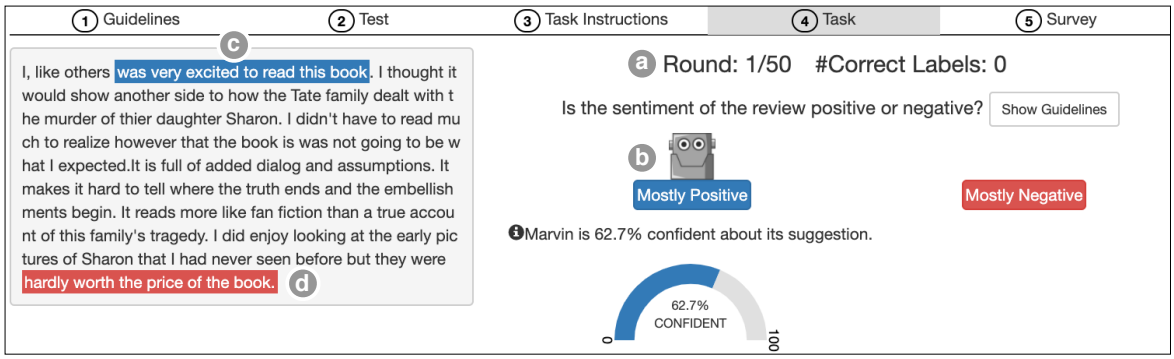


Figure 2.2: A screenshot of the **Team (Adaptive, Expert)** condition for the *Amzbook* reviews dataset. Participants read the review (left pane) and used the buttons (right pane) to decide if the review was mostly *positive* or *negative*. The right pane also shows progress and accuracy (a). To make a recommendation, the AI (called "Marvin") hovers above a button (b) and displays the confidence score under the button. In this case, the AI incorrectly recommended that this review was positive, with confidence 62.7%. As part of the explanation, the AI highlighted the most positive sentence (c) in the same color as the *positive* button. Because confidence was low, the AI also highlights the most negative sentence (d) to provide a counter-argument.

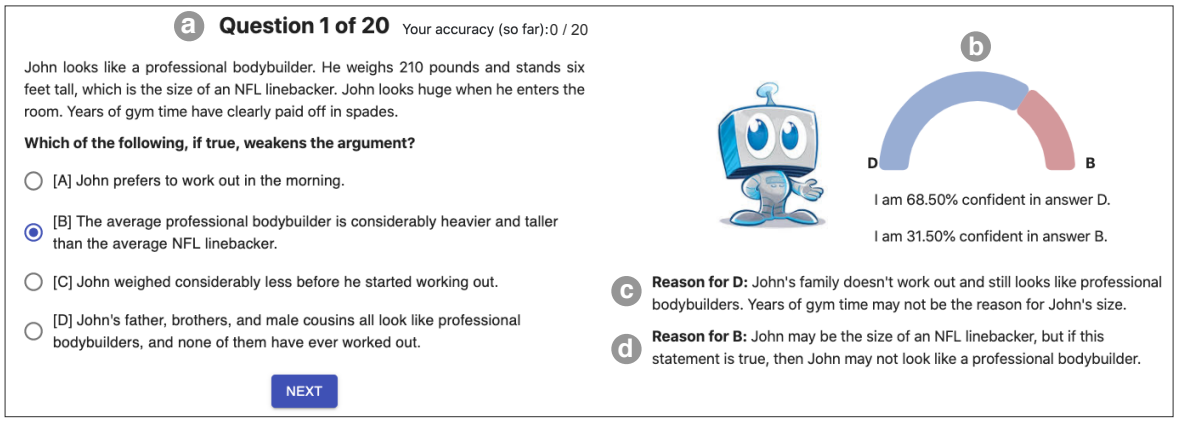


Figure 2.3: A screenshot of **Team (Adaptive, Expert)** for *LSAT*. Similar to Figure 2.2, the interface contained a progress indicator (a), AI recommendation (b), and explanations for the top-2 predictions (c and d). To discourage participants from blindly following the AI, all AI information is displayed on the right. In (b), the confidence score is scaled so those for top-2 classes sum to 100%.

Interface Figure 2.2 shows an example UI for sentiment classification for Team (Adaptive, AI). In all explanation conditions, explanations are displayed as inline highlights, with the background color aligned with the positive/negative label buttons. The highlight varies by condition, *e.g.*, Team (Adaptive, AI) has a similar display to Figure 2.2, except that the AI picks multiple short phrases, instead of a full sentence. In Team (Explain-Top-1, AI) the counter-argument (d) is always missing, and in Team (Conf) no explanations are highlighted. Figure 2.3 shows a screenshot of the user interface for *LSAT* in the Team (Adaptive, Expert)

condition.

2.4.2 AI Model, Study Samples and Explanations

Sentiment Classification

Training data. To prepare each dataset (*Beer* and *Amzbook*) for training classification models, we binarized the target labels, split the dataset into training and test sets (80/20 split), removed class imbalance from the train split by oversampling the minority class, and further split the training set to create a validation set.

AI Model. For each dataset, we fine-tuned a RoBERTa-based [Liu *et al.*, 2019] text classifier from AllenNLP[5] on the training dataset and performed hyper-parameter selection on the validation set.

Task examples. For each domain, we selected 50 examples from the test set to create our study sample. We first conducted additional pilot studies to establish the accuracy of unassisted users, which we observed were 87% for *Beer* and 85% for *Amzbook*[6] We then selected 50 unambiguous examples so that the AI's accuracy was 84% (*i.e.*, comparable to human accuracy), with equal false positive and false negative rates. The filtering was for keeping the task objective: If the ground-truth answer was unclear, one cannot compute or compare the accuracy of decisions.

Explanations. To generate saliency explanations, we used LIME, which is a popular post hoc method [Ribeiro *et al.*, 2016]. We chose this setup because the combination of RoBERTa and LIME was consistently ranked the highest among the various systems we tried in an explainer comparison study with human judges (details in Appendix). Despite offering accurate predictions, RoBERTa generated poorly calibrated confidence scores, a common issue with neural networks [Guo *et al.*, 2017], which we mitigated with *post hoc calibration* (isotonic regression [Barlow and Brunk, 1972]) on the validation set.

In particular, for Adaptive explanation, we used the classifier's median confidence as the threshold to have an equal number of 25 examples displayed as Explain-Top-1 and Explain-Top-2, respectively. The thresholds were 89.2% for *Beer* and 88.9% for *Amzbook*. We happened to explain 18 correctly predicted and 7 incorrectly predicted examples with Explain-Top-2 for both datasets (leaving 1 incorrect and 24 correct cases with Explain-Top-1). While one might learn a better threshold from the data, we leave that to future

[5] `https://demo.allennlp.org/sentiment-analysis`

[6] Again, each condition containing 50 crowd-workers. We estimated the human accuracy on all the three datasets with another 150 crowd-workers.

work. As for expert-generated explanations, one author created expert explanations by selecting one short, convincing text phrase span for each class (positive or negative).

LSAT

AI Model. We finetuned a RoBERTa model[7] on ReClor [Yu *et al.*, 2020], a logic-reasoning dataset that contains questions from standardized exams like the *LSAT* and GMAT.[8]

Task examples. We selected 20 examples from an *LSAT* prep book [Team, 2017]. We verified that our questions were not easily searchable online and were not included in the training dataset. We selected fewer *LSAT* questions than for sentiment analysis, because they are more time consuming to answer and could fatigue participants: *LSAT* questions took around a minute to answer, compared to around 17 seconds for *Beer* and *Amzbook*. The RoBERTa model achieved 65% accuracy on these examples, comparable to the 67% human accuracy that we observed in our pilot study.

Explanations. We found no automated method that could generate reasonable explanations (unsurprising, given that explanations rely on prior knowledge and complex reasoning); instead, we used expert explanations exclusively, which is again based on the prep book. The book contains explanations for the correct answer, which one author condensed to a maximum of two sentences. Since the book did not provide explanations for alternative choices, we created these by manually crafting a logical supporting argument for each choice that adhered to the tone and level of conciseness of the other explanations. Experts only generated explanations and did not determine the model predictions or its uncertainties.

2.4.3 Study Procedure

Sentiment Classification For the final study, participants went through the following steps: 1) A landing page first explained the payment scheme; the classification task was presented (here, predicting the sentiment of reviews); and they were shown dataset-specific examples. 2) To familiarize them with the task and verify their understanding, a screening phase required the participant to correctly label four of six reviews [Liu *et al.*, 2016 06]. Only participants who passed the gating test were allowed to proceed to the main task. 3) The main task randomly assigned participants to one of our study conditions (Section 2.3.3) and presented

[7] Based on the opensource implementation: `https://github.com/yuweihao/reclor`.

[8] `https://en.wikipedia.org/wiki/GraduateManagementAdmissionTest`

condition-specific instructions, including the meaning and positioning of AI's prediction, confidence, and explanations. Participants then labeled all 50 study samples one-by-one. For a given dataset, all conditions used the same ordering of examples. The participants received immediate feedback on their correctness after each round of the task. 4) A post-task survey was administered, asking whether they found the model assistance to be helpful, their rating of the usefulness of explanations in particular (if they were present), and their strategy for using model assistance.

We recruited participants from Amazon's Mechanical Turk, limiting the pool to subjects from within the United States with a prior task approval rating of at least 97% and a minimum of 1,000 approved tasks. To ensure data quality, we removed data from participants whose median labeling time was less than 2 seconds or those who assigned the same label to all examples. In total, we recruited 566 (*Beer*) and 552 (*Amzbook*) crowd workers, and in both datasets, 84% of participants passed the screening and post-filtering. Eventually, we collected data from around 100 participants (ranging from 93 to 101 due to filtering) per condition.

Study participants received a base pay of $0.50 for participating, a performance-based bonus for the main task, and a fixed bonus of $0.25 for completing the survey. Our performance-based bonus was a combination of linear and step functions on accuracy: we gave $0.05 for every correct decision in addition to an extra $0.50 if the total accuracy exceeded 90% or $1.00 if it exceeded 95%. The assigned additional bonuses were intended to motivate workers to strive for performance in the complementary zone and improve over the AI-only performance [Ho *et al.*, 2015]. Since we fixed the AI performance at 84%, humans could not obtain the bonus by blindly following the AI's recommendations. Participants spent 13 minutes on average on the experiment and received an average payment of $3.35 (equivalent to an hourly wage of $15.77).

Modifications for LSAT For the *LSAT* dataset, we used a very similar procedure but used two screening questions and required workers to answer both correctly. We used a stricter passing requirement to avoid low-quality workers who might cheat, which we observed more for this task in our pilots. We again used MTurk with the same filters as sentiment classification, and we post hoc removed data from participants whose median response time was less than three seconds. 508 crowd workers participated in our study, 35% of whom passed the screening and completed the main task, resulting in a total of 100 participants per condition.

Participants received a base pay of $0.50 for participating, a performance-based bonus of $0.30 for each

correct answer in the main task, and a fixed bonus of $0.25 for completing an exit survey. They received an additional bonus of $1.00, $2.00, and $3.00 for reaching an overall accuracy of 30%, 50%, and 85% to motivate workers to answer more questions correctly and perform their best. The average completion time for the LSAT task was 16 minutes, with an average payment of $6.30 (equals an hourly wage of $23.34).

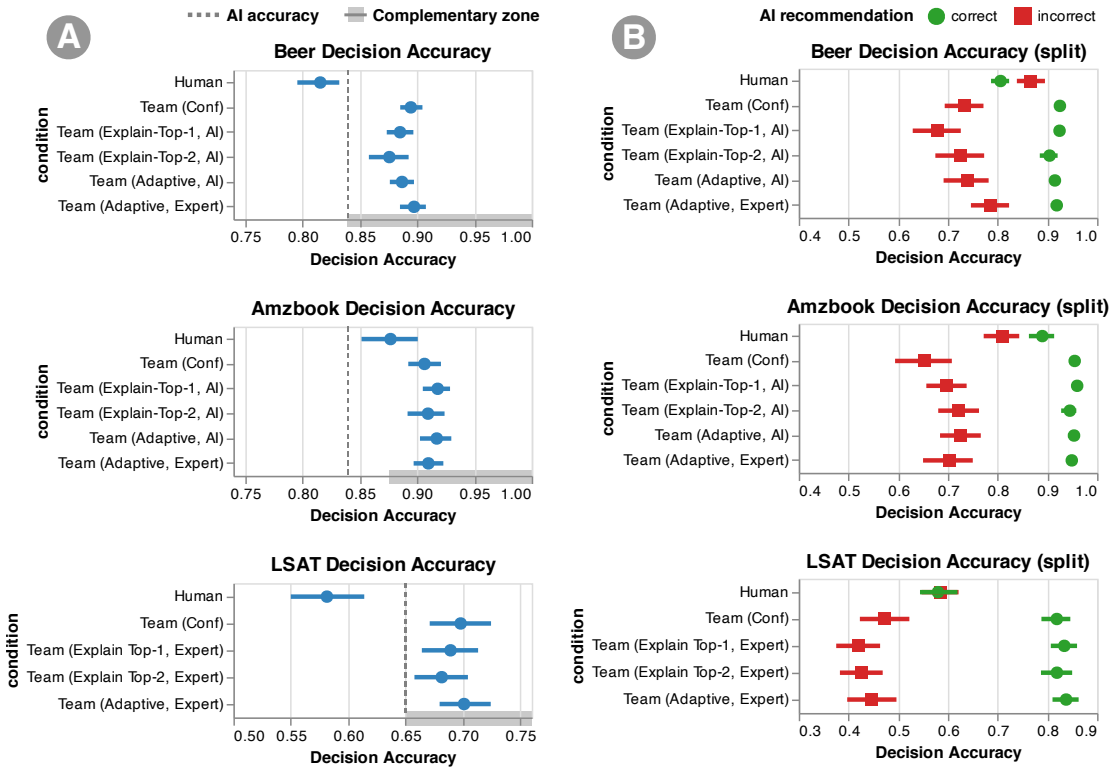


Figure 2.4: Team performance (with average accuracy and 95% confidence interval) achieved by different explanation conditions and baselines for three datasets, with around 100 participants per condition. (A) Across every dataset, all team conditions achieved complementary performance. However, we did not observe significant improvements from using explanations over simply showing confidence scores. (B) Splitting the analysis based on the correctness of AI accuracy, we saw that for *Beer* and *LSAT*, Explain-Top-1 explanations worsened performance when the AI was incorrect, the impact of Explain-Top-1 and Explain-Top-2 explanations were correlated with the correctness of the AI's recommendation, and Adaptive explanations seemed to have the potential to improve Explain-Top-1 when the AI was incorrect, and to retain the higher performance of Explain-Top-1 when the AI was correct.

2.5 Results

2.5.1 Effect of Explanation on Team performance

Figure 2.4A shows the team performance (*i.e.*, accuracy of final decision) for each domain and condition. We tested the significance of our results using Student's T-tests with Bonferroni correction.

The baseline team condition, Team (Conf), achieved complementary performance across tasks. For *Beer*, providing AI recommendations and confidence to users increased their performance to ($\mu = 0.89 \pm \sigma = 0.05$), surpassing both AI ($0.84$) and unassisted human accuracy (0.82 ± 0.09). Similarly, Team (Conf) achieved complementary performance for *Amzbook* and *LSAT*, with relative gains of 2.2% and 20.1% over unassisted workers (Figure 2.4A).

We did not observe a significant difference between Explain-Top-1 and Explain-Top-2, or that **H1** was not supported. For example, in Figure 2.4A of *Beer*, explaining the top prediction performed marginally better than explaining the top-two predictions, but the difference was not significant (z=0.85, p=.40). The same was true for *Amzbook* (z=0.81, p=.42) and *LSAT* (z=0.42, p=.68).

We did not observe significant improvements over the confidence baseline by displaying explanations. For example, for *Beer*, Team (Conf) and Team (Explain-Top-1, AI) achieved similar performance, with the accuracy being 0.89 ± 0.05 vs. 0.88 ± 0.06 respectively; the difference was insignificant (z=-1.18, p=.24). We observed the same pattern for *Amzbook* (z=1.23, p=.22) and *LSAT* (z=0.427, p=.64). As a result, we could not reject our hypothesis **H2** that Explain-Top-1 performs similar to simply showing confidence . This result motivates the need to develop new AI systems and explanation methods that provide true value to team performance by supplementing the model's confidence, perhaps working in tandem with confidence scores.

Though designed to alleviate the limitations of Explain-Top-1 and Explain-Top-2 in our experiments, **we did not observe improvements from using Adaptive explanations.** For example, we did not observe any significant differences between Team (Adaptive, AI) and Team (Conf) for *Beer* (z=-1.02, p=.31) or *Amzbook* (z=1.08, p=.28). We did not observe significant differences between Team (Adaptive, Expert) and Team (Conf) for *LSAT* (z=0.16, p=.87). More surprisingly, switching the source of Adaptive explanation to expert-generated did not significantly improve sentiment analysis results. For example, in Figure 2.4A, the

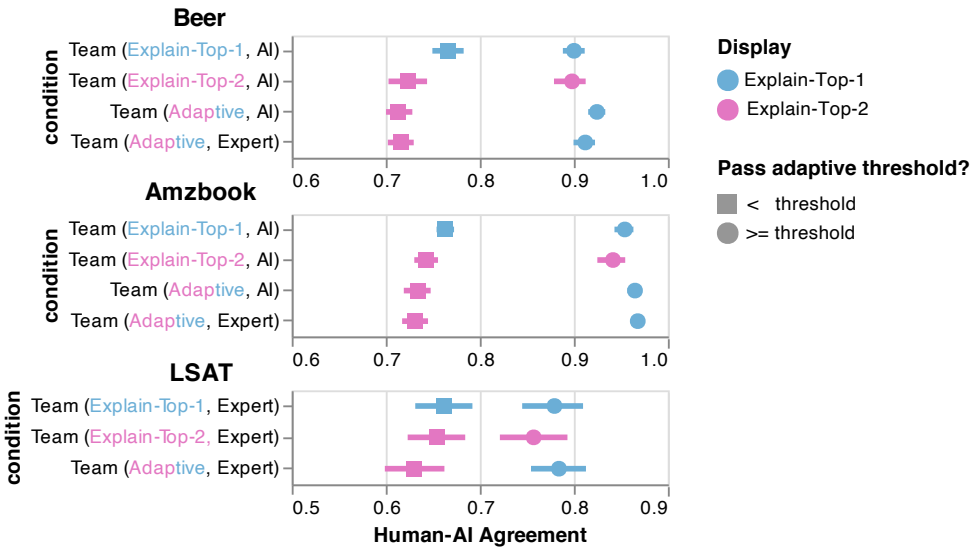


Figure 2.5: Relative agreement rates between humans and AI (*i.e.*, does the final human decision match the AI's suggestion?) for various conditions, with examples split by whether AI's confidence exceeded the threshold used for Adaptive explanations. Across the three datasets, Adaptive explanations successfully reduced the human's tendency to blindly trust the AI (*i.e.*, decreased agreement) when it was uncertain and more likely to be incorrect. For example, comparing Team (Explain-Top-1, AI) and Team (Adaptive, AI) on low confidence examples that did not pass the threshold (rectangles), participants in Explain-Top-2 (pink rectangles) were less likely to agree with the AI compared to those who saw Explain-Top-1 (blue rectangles).

differences in performance between Team (Adaptive, Expert) and Team (Adaptive, AI) were insignificant: *Beer* (z=1.31, p=.19) and *Amzbook* (z=-0.78, p=.43). As such, we could not reject the null hypotheses for either **H3** or **H4**.

While Adaptive explanation did not significantly improve team performance across domains, further analysis may point a way forward by combining the strengths of Explain-Top-1 and Explain-Top-2. Split the team performance by whether the AI made a mistake (Figure 2.4B), we observe that explaining the top prediction lead to better accuracy when the AI recommendation was correct but worse when the AI was incorrect, as in our pilot study. This is consistent with Psychology literature [Koehler, 1991], which has shown that human explanations cause listeners to agree even when the explanation is wrong, and recent studies that showed explanations can mislead data scientists into overtrusting ML models for deployment [Kaur *et al.*, 2020]. While these results were obtained by measuring user's subjective ratings of trust, to the best of our knowledge, our studies are the first to show this phenomenon for explanation and end-to-end decision making with large-scale studies. As expected, in *Beer*, Adaptive explanations improved performance over

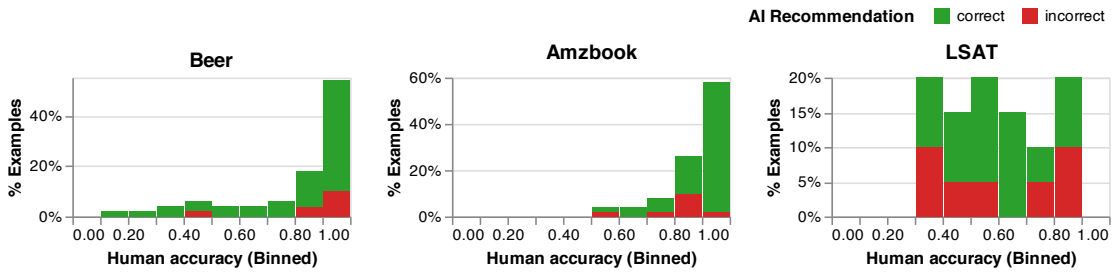


Figure 2.6: The distribution of study examples as a function of average human accuracy. For each domain, examples on the right were easy for most humans working alone. Both *Beer* and *LSAT* show a distribution that shows potential for complementary team performance: humans can correct easy questions mistaken by the AI (red bars towards the right), and, conversely, the AI may add value on examples where humans frequently err (green bars towards the left). In contrast, *Amzbook* showed less potential for this kind of human-AI synergy, with less "easy for human" questions (bars towards the left).

Explain-Top-1 when the AI was incorrect and improved performance over Explain-Top-2 when the AI was correct, although the effect was smaller on other datasets.

While Figure 2.4B shows team performance, the promising effects of Adaptive explanations are clearer if we study the agreement between AI predictions and human decisions (Figure 2.5). Adaptive explanations seem to encourage participants to consider the AI more when it is confident and solve the task themselves otherwise. Unfortunately, as our experiments show, the effect of using Adaptive did not seem sufficient to increase the final team accuracy, possibly for two reasons: (1) in high confidence regions (circles in Figure 2.5), not only did workers have to agree more, but they also had to identify cases where the model failed with very high confidence (unknown unknowns [Lakkaraju *et al.*, 2017]). Identifying unknown unknowns could have been a difficult and time-consuming task for workers, and they may have needed other types of support that we did not provide. (2) In low confidence regions (rectangles), not only did workers have to disagree more, but they also had to be able to solve the task correctly when they disagreed. Explain-Top-2 explanations might have enabled them to suspect the model more, but it is unclear if they helped participants make the right decisions. This indicates that more sophisticated strategies are needed to support humans in both situations. We discuss some potential strategies in Section 2.6.3.

Differences in expertise between human and AI affects whether (or how much) AI assistance will help achieve complementary performance. To understand how differences in expertise between the human and AI impact team performance, we computed the average accuracy of unassisted users on study examples and overlaid the AI's expertise (whether the recommendation was correct) in Figure 2.6. The figure helps

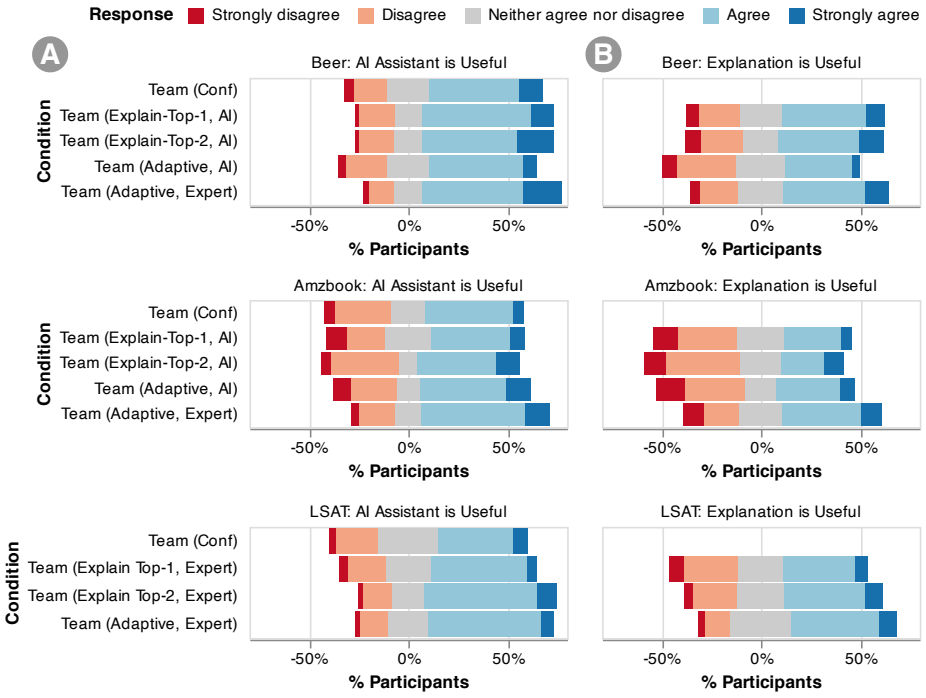


Figure 2.7: Analysis of participant responses to two statements: (A) "AI's assistance (*e.g.*, the information it displayed) helped me solve the task", and (B) "AI's explanations in particular helped me solve the task." Across datasets, a majority of participants found AI assistant to be useful, and they rated all the conditions similarly, with a slight preference towards Team (Adaptive, Expert). In contrast to AI's overall usefulness, fewer participants rated explanations as useful, particularly Explain-Top-2 explanations. Participants also had a clearer preference for higher-quality (expert) Adaptive explanations.

explain why users benefited more from AI recommendations for both *Beer* and *LSAT* datasets. There was a significant fraction of examples that the AI predicted correctly but humans struggled with (green bars to the left), while the same was not true for *Amzbook* (where AI recommendations did not help as much). Further, when the AI was incorrect, explaining predictions on *Amzbook* via Explain-Top-1 improved the performance by 5% over showing confidence (Figure 2.4B), but it decreased the performance for *Beer* and *LSAT*. One possible explanation is that most AI mistakes were predicted correctly by most humans on *Amzbook* (red bars were mostly towards the right). After observing clear model mistakes, participants may have learned to rely on them less, despite the convincing-effect of explanation. Participants' self-reported collaboration approaches supported our guess – *Amzbook* participants reportedly ignored the AI's assistance the most (Section 2.5.3). That said, other confounding effects such as the nature of the task (*e.g.*, binary classification vs. choosing between multiple options) should also be studied.

2.5.2 Survey Responses on Likert Scale Questions

Two of the questions in our post-task survey requested categorical ratings of AI and explanation usefulness.[9]

AI usefulness: While participants generally rated AI assistance useful (Figure 2.7A), the improvements in ratings between most explanations and simply showing confidence were marginal. The difference was more clear for high-quality adaptive explanations; for *Beer*, 70% of the participants rated AI assistance useful with Team (Adaptive, Expert) in contrast to 57% with Team (Conf). We observed a similar pattern on *Amzbook* (65% vs. 49%) and *LSAT* (63% vs. 45%), though on *LSAT*, Team (Explain-Top-2, Expert) received slightly higher ratings than Team (Adaptive, Expert) (66% vs. 63%).

Explanation usefulness: Figure 2.7B shows that participants' ratings for the usefulness of explanations were lower than the overall usefulness of AI's assistance (in A). Again, expert-generated Adaptive explanations received higher ratings than AI-generated ones for *Beer* (53% vs. 38%) vs. *Amzbook* (50% vs. 40%). This could indicate that showing higher quality explanations improves users' perceived helpfulness of the system. However, it is worth noting that this increased preference did not translate to an improvement in team performance, which is consistent with observations made by Buçinca *et al.* [Buçinca *et al.*, 2020] that show that people may prefer one explanation but make better decisions with another.

[9]Since we did not pre-register hypotheses for the subjective ratings and only analyzed them post-hoc, we do not perform/claim statistical significant analysis on these metrics.

Codes	Definitions and Examples	#Participants
Overall Collaboration Approach (codes are mutually exclusive)		
`Mostly Follow AI`	The participant mostly followed the AI. *"I went with Marvin most times."*	23 (6%)
`AI as Prior Guide`	Used AI as a starting reference point. *"I looked at his prediction and then I read the passage."*	190 (47%)
`AI as Post Check`	Double-checked after they made their own decisions. *"I ignored it until I made my decision and then verified what it said."*	102 (25%)
`Mostly Ignore AI`	Mostly made their own decisions without the AI. *"I didn't. I figured out the paragraph for myself."*	90 (22%)
The Usage of Explanation (codes can overlap)		
`Used Expl.`	Explicitly acknowledged they used the explanation. *"I skimmed his highlighted words."*	138 (42%)
`Speed Read`	Used explanations to quickly skim through the example. *"I looked at Marvin's review initially then speed read the review. "*	29 (9%)
`Validate AI`	Used the explanation to validate AI's reasoning. *"Marvin focuses on the wrong points at times. This made me cautious when taking Marvin's advice."*	17 (5%)
The Usage of Confidence (codes can overlap)		
`Used Conf.`	Explicitly acknowledged they used the confidence. *"I mostly relied on Marvin's confident levels to guide me."*	90 (22%)
`Conf. Threshold`	Was more likely to accept AI above the threshold. *"If Marvin was above 85% confidence, I took his word for it."*	24 (6%)
Others (codes can overlap)		
`Fall Back to AI`	Followed the AI's label if they failed to decide. *"I used it if I was unsure of my own decision."*	54 (13%)
`Updated Strategy`	Changed their strategy as they interacted more. *"I decided myself after seeing that sometimes Marvin failed me. "*	12 (2%)

Table 2.3: The codebook for participants' descriptions of how they used the AI, with the number of self-reports.

2.5.3 Qualitative Analysis on Collaboration

To better understand how users collaborated with the AI in different tasks, we coded their response to the prompt: "Describe how you used the information Marvin (the AI) provided." Two annotators independently read a subset of the responses to identify emergent codes and, using a discussion period, created a codebook (Table 2.3). Using this codebook, for each team condition and dataset, they coded a sample of 30 random worker responses: 28 were unique and 2 overlapped between annotators, allowing us to compute inter-annotator agreement. Our final analysis used 409 unique responses after removing 11 responses deemed to be of poor quality (Table 2.3). We scored the inter-annotator agreement with both the Cohen's κ and the raw overlap between the coding. We achieved reasonably high agreements, with an average $\mu(\kappa) = 0.71, \sigma(\kappa) = 0.18$ (the average agreement was $93\% \pm 6.5\%$). We noticed the following, which echo the performance differences observed across datasets:

Most participants used the AI's recommendation as a prior or to double-check their answers. For all datasets, more than 70% of the participants mentioned they would partially take AI's recommendation into consideration rather than blindly following AI or fully ignoring it (Figure 2.8). Participants used the AI as a prior guide more than as a post-check for sentiment analysis, but not for *LSAT*, which aligns with our interface design: for *LSAT*, AI recommendations were on a separate pane, encouraging users to solve the task on their own before consulting the AI.

Participants ignored the AI more on domains where AI expertise did not supplement their expertise. Figure 2.8 shows that while only 11% of *LSAT* participants claimed that they mostly ignored the AI, the ratio doubled (*Beer*, 23%) or even tripled (*Amzbook*, 30%) for sentiment analysis. As discussed in Figure 2.6, this may be due to correlation differences between human and AI errors for different datasets: *Amzbook* participants were less likely to see cases where AI was more correct than they were, and therefore they may have learned to rely less on it. For example, one participant in *Amzbook* mentioned, *"I had initially tried to take Marvin's advice into account for a few rounds, and stopped after I got 2 incorrect answers. After that I read all of the reviews carefully and followed my own discretion."*

In contrast, a *Beer* participant relied more on the AI once realizing it could be correct: *"At first I tried reading the passages and making my own judgments, but then I got several items wrong. After that, I just switched to going with Marvin's recommendation every time."*

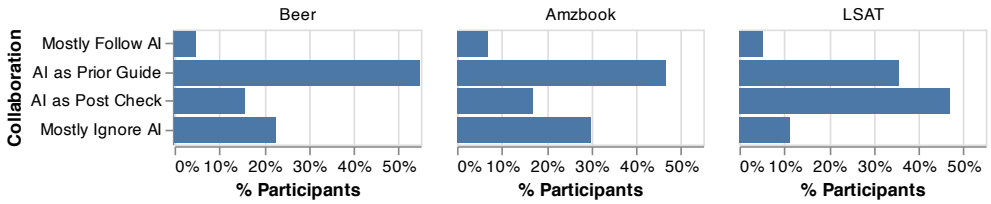


Figure 2.8: Instead of ignoring or strictly following the AI, participants reported taking the AI information into consideration most of the time. They most frequently used AI as a prior guide in sentiment analysis, but used it as post-check in *LSAT*. They were also more likely to ignore the AI in sentiment analysis than in *LSAT*.

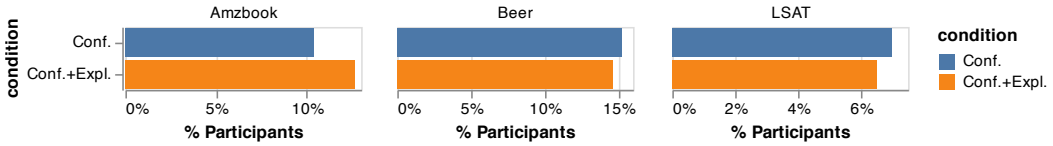


Figure 2.9: Comparing the occurrence of `Used Conf.` in just the confidence condition and in those with explanations, we saw a similar proportion of users that explicitly acknowledged using confidence, regardless of whether they saw an explanation.

In addition to the user's collaboration behavior, these differences between domains may have affected our quantitative observations of team performance. For example, a small difference between human and AI expertise (distribution of errors) means that the improvement in performance when the AI is correct would be less substantial. In fact, in Figure 2.4B, if we compare the team performance when the AI is correct, the difference between team conditions and the human baseline is least substantial for *Amzbook*.

Some participants developed mental models of the AI's confidence score to determine *when* to trust the AI. Among participants who mentioned they used confidence scores (90 in total), 27% reported using an explicit confidence threshold, below which they were likely to distrust the AI. The threshold mostly varied between 80 to 100 (83 ± 8 for *Beer*, 89 ± 7 for *Amzbook*, and 90 ± 0 for *LSAT*) but could go as low as 65, indicating that users built different mental models about when they considered AI to be "trustworthy." While this observation empirically shows that end-users develop mental model of trust in AI-assisted decision making [Bansal *et al.*, 2019a], it more importantly shows how the AI's confidence is a simple, yet salient feature via which users create a mental model of the AI's global behavior [Gero *et al.*, 2020]. Note that across all three domains, the same proportion of participants self-reported using AI's confidence scores regardless of whether they saw explanations (Figure 2.9).

Furthermore, **some participants consigned the task to AI when they were themselves uncertain.** For

example, 13% participants mentioned that they would go with the AI's decision if they were on the fence by themselves: *"There were some that I could go either way on, and I went with what Marvin suggested."* These user behaviors are similar to observations in psychology literature on *Truth-Default Theory* [Levine, 2014], which shows that people exhibit *truth-default* behavior: by default, people are biased to assume that the speaker is being truthful, especially when *triggers* that raise suspicion are absent. Furthermore, our participants' distrust in low-confidence recommendations is also consistent with examples of triggers that cause people to abandon the truth-default behavior.

Explanations can help participants validate the AI's decisions, and the inline-highlight format helped participants speed up their decision making. Among the participants who explicitly mentioned using explanations, 27% in *Beer* and 32% in *Amzbook* reported that they used them to read the review text faster. Since *LSAT* explanations required reading additional text, we did not expect *LSAT* users to find this benefit. Interestingly, for *Beer* and *Amzbook*, while a small percentage of users (17%) reported using the explanations to validate the AI's decisions (see Figure 2.3), only 2% did so in *LSAT*. This could be because *LSAT* is a harder task than sentiment analysis, and verifying AI's explanations is costlier. Other participants mostly mentioned that they would supplement their own reasoning with the AI's: *"I read the Marvin rationale and weighed it against my intuition and understanding."*

2.6 Discussion & Future Directions

Though conducted in a limited scope, our findings should help guide future work on explanations and other mechanisms for improving decision making with human-AI teams.

2.6.1 Limitations

As mentioned in Section 2.2, AI explanations have other motivations not addressed by this paper. Our work, as well as the papers listed in Table 2.1, evaluated team performance along one dimension: accuracy of decisions. We did not explore the benefits on other metrics (*e.g.*increasing speed as reported by some users in Section 2.5.2), but in general, one may wish to achieve complementary performance on a multi-dimensional metric. In fact, research shows that large collaborative communities like Wikipedia require AI systems that balance multiple aspects, *e.g.*, reducing human effort, improving trust and positive engagement [Smith *et*

al., 2020]. We encourage future research to extend the definition of complementarity, and to evaluate the impact of explanations on those dimensions accordingly.

Further, we restricted ourselves to tasks amenable to crowdsourcing (text classification and question answering), so our results may not generalize to high-stakes domains with expert users such as medical diagnosis. We also note that the effectiveness of explanations may depend on user expertise, a factor that we did not explore. Investigating this in our framework would either require recruiting lay and expert users for the same task [Feng and Boyd-Graber, 2019] or utilizing a within-subject experimental design to measure user expertise.

Finally, we only explored two possible ways to present explanations (highlighting keywords and natural language arguments). While these methods are widely adopted [Wang *et al.*, 2016; Lin *et al.*, 2017; Lai and Tan, 2019], alternative approaches may provide more benefit to team performance.

2.6.2 Explaining AI for Appropriate Reliance

One concerning observation was that explanations increased blind trust rather than appropriate reliance on AI. This is problematic especially in domains where humans are required in the loop for moral or legal reasons (*e.g.*, medical diagnosis) and suppose the presence of explanations simply soothes the experts (*e.g.*, doctors), making them more compliant so they blindly (or become more likely to) agree with the computer. Encouraging human-AI interactions like these seems deeply unsatisfactory and ethically fraught. Importantly, while prior works also observed instances of inappropriate reliance on AI [Wang *et al.*, 2019; Croskerry, 2009; Kaur *et al.*, 2020; Mitchell *et al.*, 2019], our studies quantified its effect on team performance. Since the nature of the proxy tasks can significantly change the human behavior, they can lead to potential misleading conclusions [Buçinca *et al.*, 2020]. The emphasis of the complementary team performance in end-to-end tasks can *objectively* evaluate the extent of such issues or about the effectiveness of a solution.

Our Adaptive Explanation aims to encourage the human to think more carefully when the system had a low confidence. While the relative agreement rates showed that the Explain-Top-2 explanation might cue the humans to suspect the model's veracity (Figure 2.5), the method was not sufficient to significantly increase the final team accuracy (Figure 2.4). This is perhaps because users still have to identify high-confidence

mistakes (unknown-unknowns) and solve the task when the AI is uncertain (Section 2.5.1). A followup question is, then, what kind of interactions would help humans perform correctly when the AI is incorrect?

Explanations should be informative, instead of just convincing. Our current expert explanations did not help any more than the AI explanations, which may indicate that having the ML produce the *maximally convincing* explanation — a common objective shared in the design of many AI explanation algorithms — might be a poor choice for complementary performance [Buçinca *et al.*, 2020]. A more ideal goal is explanations that accurately *inform* the user – such that the user can quickly gauge through the explanation when the AI's reasoning is correct and when it should raise suspicion. A successful example of this was seen with Generalized Additive Models (GAMs) for healthcare, where its global explanations helped medical experts suspect that the model had learned incorrect, spurious correlations (*e.g.*a history of asthma reduces the risk of dying from pneumonia [Caruana *et al.*, 2015]). We hope future research can produce explanations that better enable the human to effectively catch AI's mistakes, rather than finding plausible justifications when it erred.

High complementary performance may require adapting beyond confidence. Since approaches based on confidence scores make it difficult to spot unknown-unknowns, instead it may be worthwhile to design explanation strategies that adapt based on the *frequency* of agreement between the human and AI. For example, instead of explaining why it believes an answer to be true, the AI might play a devil's advocate role, explaining its doubts — even when it agrees with the human. The doubts can even be expressed in an interactive fashion (as a back and forth conversation) than a set of static justifications, so to avoid cognitive overload. For example, even if the system agrees with the user, the system can present a high-level summary of evidence for top-K alternatives and let the user drill down, *i.e.*, ask the system for more detailed evidence for the subset of alternatives that they now think are worth investigating.

2.6.3 Rethinking AI's Role in Human-AI Teams

Comparable accuracy does not guarantee complementary partners. Rather, in an ideal team, the human and AI would have minimally overlapping mistakes so that there is a greater opportunity to correct each other's mistakes. In one of our experiment domains (*Amzbook*), AI errors correlated much more strongly

with humans' than in others, and thus we saw relatively smaller gains in performance from AI assistance (Figure 2.6). As recent work has suggested [Wilder *et al.*, 2020 07; Madras *et al.*, 2018; Mozannar and Sontag, 2020; Bansal *et al.*, 2021a], it may be useful to directly optimize for complementary behavior by accounting for the human behavior during training, who may have access to a different set of features [Varshney *et al.*, 2018].

Furthermore, the human and AI could maximize their talents in different dimensions. For example, for grading exams, AI could use its computation power to quickly gather statistics and highlight commonly missed corner cases, whereas the human teacher could focus on ranking the intelligence of the student proposed algorithms [Glassman *et al.*, 2015]. Similarly, to maximize human performance at Quiz Bowl, Feng and Graber [Feng and Boyd-Graber, 2019] designed interaction so that the AI memorized and quickly retrieved documents relevant to a question, a talent which humans lacked because of cognitive limitations; however, they left the task of combining found evidence and logical reasoning to human partners. Future research should explore other ways to increase synergy.

The timing of AI recommendations is important. Besides the types of explanations, it is also important to carefully design *when* the AI provides its viewpoint. All of our methods used a workflow that showed the AI's prediction (and its explanation) to the human, before they attempted to solve the problem on their own. However, by presenting an answer and accompanying justification upfront, and perhaps overlaid right onto the instance, our design makes it almost impossible for the human to reason independently, ignoring the AI's opinion while considering the task. This approach risks invoking the anchor effect, studied in Psychology [Englich *et al.*, 2006] and introduced to the AI explanation field by Wang et al. [Wang *et al.*, 2019] — people rely heavily on the first information that is presented by others when making decisions. This effect was reflected in an increase in the use of the "`AI as Prior Guide`" collaboration approach in the sentiment analysis domain, compared to *LSAT* (Figure 2.8).

Alternate approaches that present AI recommendations in an asynchronous fashion might increase independence and improve accuracy. For example, pairing humans with slower AIs (that wait or take more time to make recommendation) may provide humans with a better chance to reflect on their own decisions [Park *et al.*, 2019]. Methods that embody recommendations from management science for avoiding group-think [Macleod, 2011] might also be effective, *e.g.*, showing the AI's prediction after the human's

initial answer or only having the AI present an explanation if it disagreed with the human's choice. We note that these approaches correspond to the Update and Feedback methods of Green & Chen [Green and Chen, 2019], which *were* effective, albeit not in the complementary zone. Another approach is to limit the AI's capabilities. For example, one might design the AI to summarize the best evidence for all possible options, without giving hard predictions, by training *evidence agents* [Perez *et al.*, 2019 11]. However, by delaying display of the AI's recommendation until after the human has solved the task independently or restricting to only per class evidences, one may preclude improvement to the *speed* of problem solving, which often correlates to the cost of performing the task.

As a result, there is a strong tension between the competing objectives of speed, accuracy, and independence; We encourage the field to design and conduct experiments and explore different architectures for balancing these factors.

2.7 Conclusions

Previous work has shown that the accuracy of a human-AI team can be improved when the AI explains its suggestions, but these results are only obtained in situations where the AI, operating independently, is better than either the human or the best human-AI team. We ask if AI explanations help achieve *complementary* team performance, *i.e.*whether the team is more accurate than either the AI or human acting independently. We conducted large-scale experiments with more than 1,500 participants. Importantly, we selected our study questions to ensure that our AI systems had accuracy comparable to humans and increased the opportunity for seeing complementary performance. While all human-AI teams showed complementarity, none of the explanation conditions produced an accuracy significantly higher than the simple baseline of showing the AI's confidence — in contrast to prior work. Explanations increased team performance when the system was correct, but they decreased the accuracy on examples when the system was wrong, making the net improvement minimal.

By highlighting critical challenges, we hope this paper will serve as a "Call to action" for the HCI and AI communities: and AI communities. In future work, characterize when human-AI collaboration can be beneficial (*i.e.*, when both parties complement each other), developing explanation approaches and coordination strategies that result in a complementary team performance that exceeds what can be produced by

simply showing AI's confidence, and communicate explanations to increase understanding rather than just to persuade. At the highest level, we hope researchers can develop new interaction methods that increase complementary performance beyond having an AI telegraph its confidence.

Chapter 3

Role of Mental Models in Human-AI Teams

Decisions made by human-AI teams (*e.g.*, AI-advised humans) are increasingly common in high-stakes domains such as healthcare, criminal justice, and finance. Achieving high *team* performance depends on more than just the accuracy of the AI system: Since the human and the AI may have different expertise, the highest team performance is often reached when they both know how and when to complement one another. We focus on a factor that is crucial to supporting such complementary: the human's mental model of the AI capabilities, specifically the AI system's *error boundary* (*i.e.* knowing "When does the AI err?"). Awareness of this lets the human decide when to accept or override the AI's recommendation. We highlight two key properties of an AI's error boundary, *parsimony* and *stochasticity*, and a property of the task, *dimensionality*. We show experimentally how these properties affect humans' mental models of AI capabilities and the resulting team performance. We connect our evaluations to related work and propose goals, beyond accuracy, that merit consideration during model selection and optimization to improve overall human-AI team performance.

3.1 Introduction

In Chapter 2, we concluded that there is very sparse evidence that existing AI explanations help improve performance into the complementary performance zone. This lack of appropriate reliance indicates that there exists a mismatch between the actual error boundary of AI (*i.e.*, regions where it errs) and its mental model that users create (*e.g.*, through explanations).

In settings where the human is tasked with deciding when and how to make use of the AI system's recommendation, extracting benefits from the collaboration requires the human to build insights (i.e., a mental model) about multiple aspects of the capabilities of AI systems. A fundamental attribute is recognition of whether the AI system will succeed or fail for a particular input or set of inputs (Maxim 1). For situations where the human uses the AI's output to make decisions, this mental model of the AI's error boundary—which describes the regions where it is correct versus incorrect—enables the human to predict when the AI will err and decide when to override the automated inference.

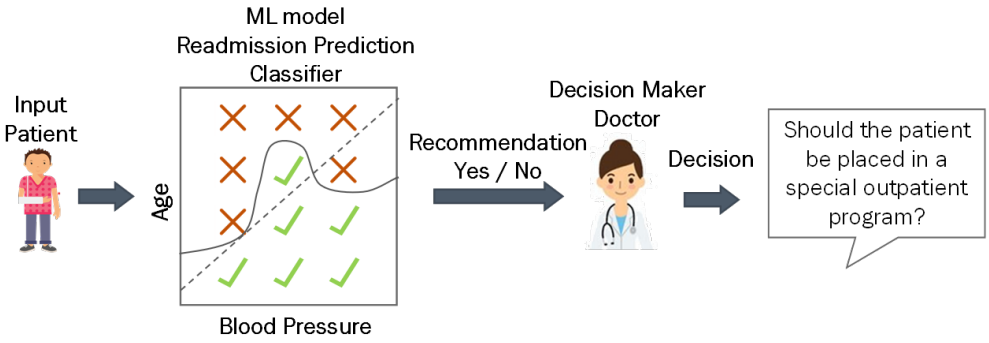


Figure 3.1: AI-advised human decision making for readmission prediction: The doctor makes final decisions using the classifier's recommendations. Check marks denote cases where the AI system renders a correct prediction, and crosses denote instances where the AI inference is erroneous. The solid line represents the AI error boundary, while the dashed line shows a potential human mental model of the error boundary.

We focus here on *AI-advised human decision making*, a simple but widespread form of human-AI team, for example, in domains like medical diagnosis, candidate screening for hiring, and loan approvals. Figure 3.1 illustrates an example of AI-advised human decision making for healthcare [Wiens *et al.*, 2016; Caruana *et al.*, 2015]. A doctor, using advice from a binary classifier, must decide whether to place a patient in a special (but costly) outpatient program. For a given input, the AI first recommends an action and the human then decides whether to trust or override it to make a final decision. This kind of human-AI collaboration is one formulation of teaming, where there is a binary trust model where the human either trusts or distrusts (and discards) the output of the AI system's influences. We consider the binary trust model instead of situations where the output of the AI system can have varying influence on human decision makers. Team performance in AI-advised human decision making depends on how well the human understands the AI's error boundary. A mismatch between the human's mental model and the true error boundary can lead to

sub-optimal decisions, such as: (1) the human may trust the AI when it makes an erroneous recommendation, (2) the human may not trust the AI when it makes a correct recommendation. These decision can lower productivity and/or accuracy.

We define properties of an AI's error boundary that affect human's ability to form an accurate mental model, such as *parsimony* and *non-stochasticity*. Intuitively, an error boundary is parsimonious if it is simple to represent. For example, an error boundary that can be described via a minimal number of features or conjunctive expressions on those features is considered to be parsimonious. A non-stochastic error boundary can be modeled with a small set of features that reliably and cleanly distinguishes successes from errors without uncertainty. Another factor that relates to a humans' ability to create a mental model of the error boundary is the task dimensionality, which we characterize by the number of features defining each instance.

We investigate the effect of these properties by conducting controlled user studies using CAJA, which is an open-source and configurable platform that implements an abstract version of AI-advised human decision making [Bansal *et al.*, 2019b]. Our results demonstrate that parsimony and non-stochasticity of error boundaries improve people's ability to create a mental model. Moreover, the experiments characterize traits of how people create and update mental models over time, highlighting the need for potential guidance in this process. Given the importance of mental models for the ultimate goal of *team* performance, this work advocates for increased attention to properties necessary for effective human-centered AI. We make the following contributions:

3.1 We highlight an under-explored but significant research challenge at the intersection of AI and human computation research—the role of humans' mental models in team performance in AI-advised human decision making.

3.2 We identify two attributes of AI systems, parsimony and non-stochasticity of error boundaries, that may help humans learn better mental models of AI competence and therefore improve team performance.

3.3 Using an open-source, game-centric platform, we show that humans' mental models of AI competence are a critical component of achieving high team performance, provide insights into how humans build mental models in different settings, and demonstrate the desirability of parsimonious and non-

stochastic error boundaries.

3.4 We integrate these results with those of previous work to create a new set of guidelines to help developers maximize the team performance of human-centered AI systems that provide advice to people.

In Section 3.2 we formally define various concepts: AI-advised human decision making, error boundaries of AI, and mental models of error boundaries. In Section 3.3, we formulate desirable properties of error boundaries. In Section 3.4 we study their effect on mental models. We conclude with a discussion of recommendations for developing more human-centered ML.

3.2 Background

3.2.1 AI-advised human decision making

Following Bansal *et al.* (2019), we focus on a simple form of human-AI teamwork that is common in many real-world settings, such as a 30-day readmission classifier supporting a doctor [Bayati *et al.*, 2014 10] or a recidivism predictor supporting judges in courts [Angwin *et al.*, 2016]. We refer to situations where an AI system provides a *recommendation* but the human makes the final *decision* as *AI-advised human decision making* (Figure 3.1). The team solves a sequence of tasks, repeating the following cycle for each time, t.

S1: The environment provides an input, x^t.

S2: The AI (possibly mistaken) suggests an action, $h(x^t)$.

S3: Based on this input, the human makes a decision, u^t.

S4: The environment returns a reward, r^t, which is a function of the user's action, the (hidden) best action, and other costs of the human's decision (e.g., time taken).

The reward feedback in S4 lets the human learn when to trust the AI's recommendation. The cumulative reward R over T cycles is the team's performance. Throughout this paper, we will assume that the AI system is a machine learning (ML) classifier that maps an input $x \in$ X to an action y from the set of actions Y.

3.2.2 Error boundaries of ML models

The error boundary of model h is a function f that describes for each input x whether model output $h(x)$ is the correct action for that input: $f : (x, h(x)) \rightarrow \{T, F\}$. In other words, the *error boundary* defines the instances for which the model is correct. Note that this is not to be confused with the model's decision boundary, which outputs model predictions. The success of teamwork hinges on the human's recognizing whether to trust the AI model, making error boundaries a critical component of AI-advised human decision making. In fact, appropriate trust in automation is a topic that has received early attention [Lee and See, 2004] as determinant factor for designing systems that require people to manage and intervene during imperfect automation.

3.2.3 Human mental models of error boundaries

People create mental models for any system they interact with [Norman, 1988], including AI agents [Kulesza *et al.*, 2012]. In AI-advised human decision making, a simple definition for such a model would be $m : x' \rightarrow \{T, F\}$, indicating which inputs the human trusts the AI to solve correctly. Here, x' indicates the features that are available to the human. A more complex model might compute a probability and include additional arguments, such as the AI's output and its confidence. Further, there may exist a *representation mismatch*– the human may create a mental model in terms of features that are not identical to the ones used by the ML model. In fact, in real-world deployments, different team members may have access to different features. For example, a doctor may know information about a patient that is missing from electronic health records (*e.g.*, patient's compliance with taking medications), while an AI system may have access to the most recent results and trends in physiological state. However, mental models can be challenging to develop. Even when working within the same feature space, they may not be perfect because users develop them through a limited number of interactions, and humans have memory and computation limitations. To illustrate, the solid line in Figure 3.1 represents the AI error boundary, while the dashed line shows a possible human mental model of the error boundary.

3.3 Characterizing AI Error Boundaries

We now define properties that may influence peoples' ability to create a mental model of an AI's error boundary. The first two are the properties of the error boundary itself, while the third is a property of the task.

3.3.1 Parsimony

The parsimony of an error boundary f is inversely related to it representational complexity. For example, in Figure 3.1 parsimony corresponds to the geometric complexity of the error boundary (solid line). For AI error boundaries formulated in mathematical logic using disjunctive normal form, complexity depends on the number of conjuncts and literals in f. For example, a hypothetical model may yield incorrect recommendations for older patients with high blood pressure or younger patients with low blood pressure. In this case, the error boundary f would be expressed as $\{(age = \texttt{old} \wedge bloodPressure = \texttt{high}) \vee (age = \texttt{young} \wedge bloodPressure = \texttt{low})\}$, which has two conjunctions with two literals each. This error boundary is more complex and less parsimonious than one that instead uses only one conjunction and two literals.

In reality, an error boundary f may belong to any arbitrary function class. In this work, we choose to express f as a disjunction of conjunctions, where literals are pairs of features and values, so that in our controlled experiments we can vary the complexity of the error boundary and measure how it affects the accuracy of humans modeling the true error boundary. Any other choice of f would make it harder to do such a comparison and would make additional assumptions about the human representation.

3.3.2 Stochasticity

An error boundary f is non-stochastic if it separates all mistakes from correct predictions. For example, suppose that for the application in Figure 3.1, the error boundary $f_1 : \{age = \texttt{young} \wedge \textit{blood pressure} = \texttt{low}\}$ is non-stochastic; this means that the readmission classifier always errs for young patients with low blood pressure and is always correct for other patients. In contrast, consider another boundary, f_2, that separates only 90% of the inputs that satisfy f_1. That is, the model will now be correct for 10% of the young patients with low blood pressure, making f_2 a more stochastic error boundary than f_1.

In practice, an error boundary of a given model might be stochastic for three different reasons: generalization, representation mismatch between the AI and human, and inherent stochasticity in the outcome being predicted. Generalization may avoid overfitting by sacrificing instances close to the decision boundary for the sake of using a less complex, and hence more parsimonious, model (*e.g.* a polynomial of a lower degree). However, this may lead to a more stochastic error boundaries. Representation mismatch, for example, may result in a case where many instances that differ for the model appear equal to the human, who cannot understand why the model occasionally fails or succeeds. Finally, the learning model itself might also not be able to completely model the real-world phenomenon due to missing features or imperfect understanding of feature interactions.

In addition to the properties of the error boundary, the dimensionality of the task itself may affect the human's discoverability of the error boundary.

3.3.3 Task dimensionality

We quantify task dimensionality using the number of features defining each instance. With larger numbers of defining features, the search space of all possible error boundaries increases, which may affect how humans create mental models about error boundaries. In practice, using a larger number of features may improve AI accuracy but adversely affect mental models and thus team performance.

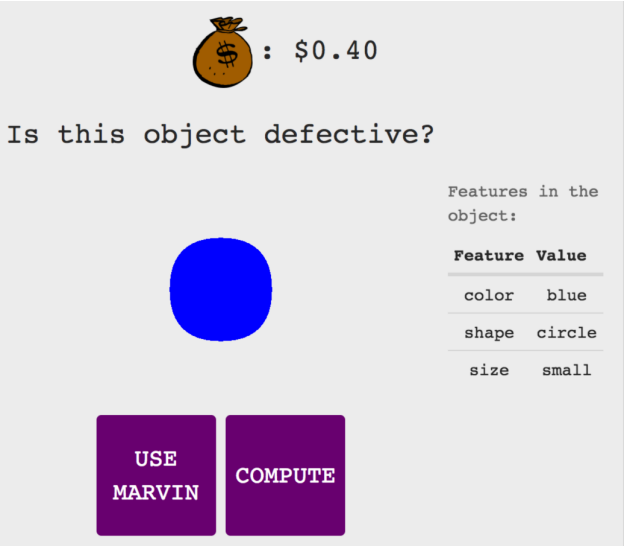


Figure 3.2: For each object, a subject can either choose to use Marvin's recommendation or perform the task independently.

	Marvin Correct	Marvin Wrong
Accept	$0.04	-$0.16
Compute	0	0

Table 3.1: Payoff matrix for the studies. As in high-stakes decisions, workers get 4 cents if they accept Marvin when it is correct, and lose 16 cents if they accept Marvin when wrong.

3.4 Experiments

3.4.1 Setup

We now present user studies we performed to build insights about the factors that may affect peoples' abilities to create a mental model of the AI. The studies were conducted using CAJA, an open-source, game-like platform that mimics AI-advised human decision making [Bansal *et al.*, 2019b]. CAJA is set up in an assembly line scenario, where the task of human subjects is to decide whether or not the objects going over the pipeline are defective (Figure 3.2). To decide on these labels, for each instance, subjects take a recommendation from an AI system called Marvin and, based on their mental model of Marvin, decide whether they should accept the AI recommendation or override it by clicking the compute button. After submitting a choice, the human receives feedback and monetary reward based on her final decision. Table 3.1 shows the payoff scheme used across these experiments, which aims to simulate high-stake decision making (*i.e.*, the penalty for an incorrect action is much higher than the reward for a correct one). In this game, subjects are not supposed to learn how to solve the task. In fact, the decision boundary is generated randomly and the only way for a participant to earn a high score is by learning the error boundary and relying on the Compute button to acquire the right prediction if Marvin is not to be trusted. This abstracts away human expertise in problem solving so the focus remains on the ability to learn the error boundary.

The CAJA platform enables control over many parameters relevant to AI-advised human decision making: task dimensionality, AI performance, length of interaction, parsimony and stochasticity of the error boundary, cost of mistakes, etc. In the human studies, we systematically vary these parameters and measure team performance to study the factors that affect humans' ability in developing mental models. Note that game parameters also distinguish between features that the machine reasons about and those that the human has access to. More specifically, the platform currently allows configurations where machine-visible features are a superset of human-visible ones, which is also the type of configuration we use in the next experiments.

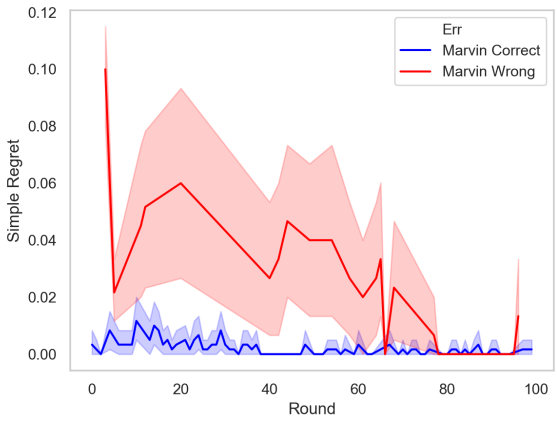


Figure 3.3: With more rounds of interaction, users perform closer to the optimal policy– the regret decreases and converges to zero for most users. Blue indicates the rounds when the AI system (Marvin) is correct and red indicates rounds when the AI makes an error. As mistakes are more costly (Table 3.1), in the beginning and when Marvin makes a mistake the difference between the optimal reward and reward earned by average worker is higher because the users have an incorrect mental model and fail to override the AI.

All studies were conducted on Amazon Mechanical Turk. For every condition we hired 25 workers and on average workers were paid an hourly wage of $20. To remove spam, we removed observations from workers whose performance was in the bottom quartile. We explain results in a question and answer format.

3.4.2 Results

Q1: *Do people create mental models of the error boundary? How do mental models evolve with interaction?*

We visualize action logs collected by CAJA to understand the evolution of mental models with more rounds of interaction. Figure 3.3 shows the average simple regret (*i.e.*, difference between the optimal and observed reward) of workers' actions over time (*i.e.*, rounds). The optimal policy is an oracle with access to Marvin's true error boundary that can thus always correctly choose when to trust Marvin. We observe that the simple regret decreases with more interactions, indicating that, on average, workers gradually learn the correct mental model and perform closer to the optimal policy.

Figure 3.4 shows the evolution of the mental model for one particular worker when Marvin's true error boundary is non-stochastic, uses one conjunction and two literals, and task dimensionality is three, *i.e.*, three features describe the problem space visible to the human. In the beginning, the worker makes more mistakes (more red crosses) because the mental model thus far is only partially correct. Eventually, the worker learns

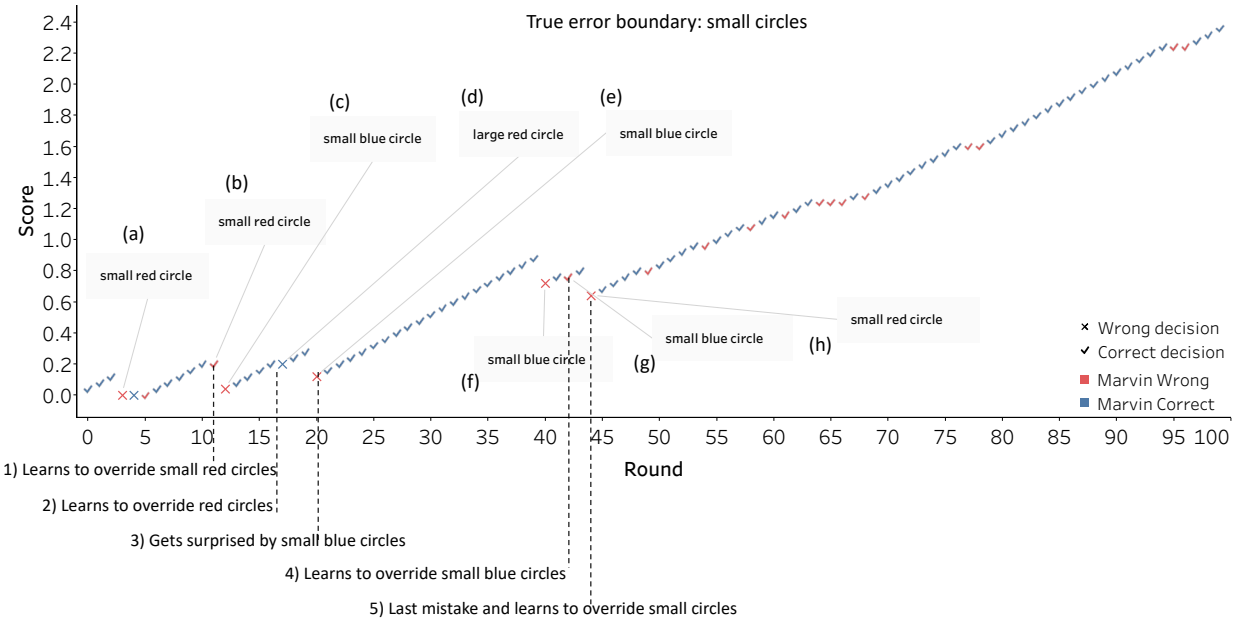


Figure 3.4: A visualization of a worker's behavior that shows how their mental model is refined with continuing interaction. Here, the score indicates the cumulative reward, and Marvin makes mistakes whenever the object is a small circle. Red markers indicate such rounds. Cross markers indicate if the worker's final decision was wrong. Hence, red crosses indicate a false accept (e.g., (a), (c), and (e)) and result in large negative reward. On the other hand, blue checks indicate a successful accept and result in a positive reward. Blue crosses indicate a false override and red checks indicate a true override. The figure contains a lot more crosses before round 45 than after. This indicates that the worker makes most of the wrong decisions in the first half of the interaction but eventually learns to act optimally. Annotations 1-5 describe the different stages of the worker's mental model. For example, by (1) the worker learns to override small red circles presumably because she learned from a previous wrong decision (a). However, since this mental model is only partially correct, in subsequent rounds (c, e, f) the worker makes wrong decisions for small blue circles. This causes surprise and confusion at first, but she eventually learns to override small blue circles by (4). But then in subsequent rounds she makes a wrong decision for a small red circle (5). After this mistake, the worker finally ties together lessons from all of her previous mistakes, figures out that small circles are problematic irrespective of the color, and acts optimally thereafter.

the correct model and successfully compensate for the AI (more red checks). Note that a mental model may be partially correct for two reasons: it is either over-generalized or over-fitted. An *over-generalized mental model* includes more cases in the error boundary than it should, for example, when the true error boundary is small circles, and the over-generalized mental model includes all small shapes. In contrast, an *over-fitted mental model* misses cases that the true error boundary contains. For example, point (c) of Figure 3.4 shows where the worker over-fit to small red circles when in fact errors occur for small circles. Clearly, a combination of both impartialities can also occur if the human tries to generalize too early on the incorrect feature

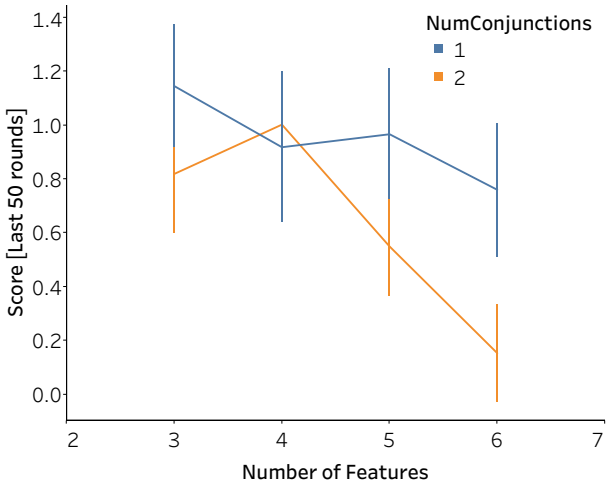
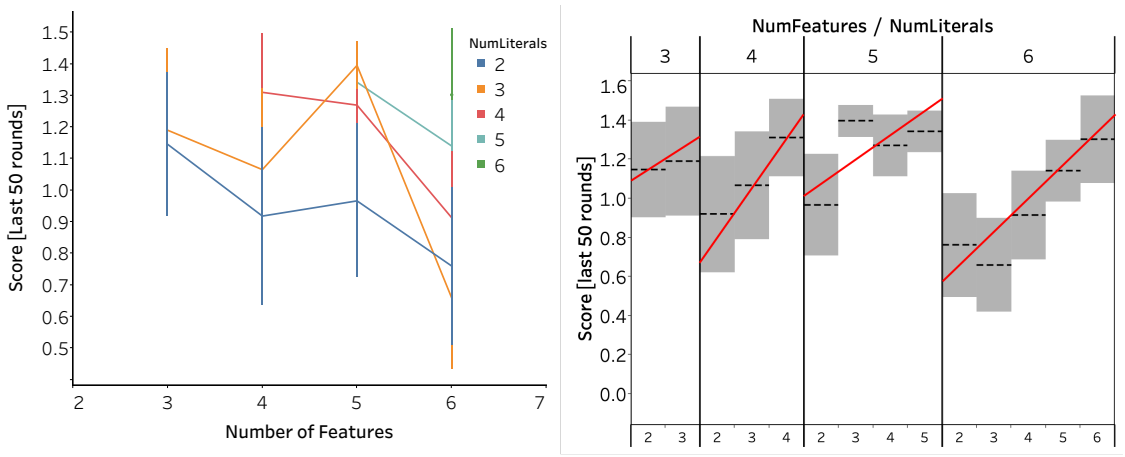


Figure 3.5: Team performance decreases as the number of conjuncts in the error boundary is increased. Number of literals were fixed to 2.

literal.

Q2: *Do more parsimonious error boundaries facilitate mental model creation?*

To answer this question, we compare team performance of many conditions that vary parsimony by changing the number of conjunctions and literals. We additionally vary the number of features to study the effect of parsimony for different task dimensionality. Figure 3.5 shows the overall team performance (cumulative score) for two boundaries of different complexity: a single conjunction with two literals (e.g., red and square), and two conjunctions with two literals each (e.g., red and square or small and circle). Different features may have different salience; therefore, for the same formula, we randomly assign different workers isomorphic error boundaries. For example, the error boundary (red and square) is isomorphic with the error boundary described by (blue and circle). Since error boundary complexity increases with the number of conjunctions, we observe that a more parsimonious error boundary (*i.e.*, a single conjunction) results in a higher team performance. Thus, our results demonstrate the value for learning ML models with parsimonious error boundaries, for example, by minimizing the number of conjunctions. In Figure 3.6, we observe that team performance generally decreases as the number of human-visible features increases, which is consistent with previous findings on human reasoning about features [Poursabzi-Sangdeh *et al.*, 2019].



(a) **(b)**

Figure 3.6: a) Team performance decreases as the task dimensionality increases (i.e., number of features). b) Re-visualization of a) that shows that, for a given number of features, team performance increases with the number of literals in the error boundary, because the errors become more specific. The solid red lines show this trend. Number of conjuncts was fixed to 1.

Q3: *Do less stochastic error boundaries lead to better mental models?*

In the previous experiments, Marvin's error boundary was non-stochastic (*i.e.*, Marvin made a mistake if and only if the object satisfied the formula). In practice, error boundaries may be fuzzier and not as clean. To understand the effect of stochasticity, we vary two parameters: $P(err|\neg f)$, the conditional probability of error if the object does not satisfy the formula, and $P(err|f)$, the conditional probability of error if the object satisfies the formula. In the non-stochastic experiments, we use a $P((err|\neg f) = 0, P(err|f) = 1))$ configuration. The other conditions that we experiment with are $(0, 0.85)$, $(0, 0.7)$, and $(0.15, 0.85)$. Of these, only the last condition is two-sided, meaning that errors can occur on both sides of the boundary although with less probability when the formula is not satisfied. All other conditions are one-sided.

Figure 3.7 shows that for the one-sided error case, the percentage of workers who make the correct decision (vertical axis) increases over time. In contrast, for two-sided error boundaries, workers find it challenging to learn the true error boundary. In addition, we observe that even for one-sided error case, increased stochasticity makes it difficult for participants to trust Marvin and learn a correct mental model. For example, the $(0, 0.7)$ condition has clearly more rounds where Marvin was correct (indicated by circles) and the percentage of people who trusted Marvin is less than 50%.

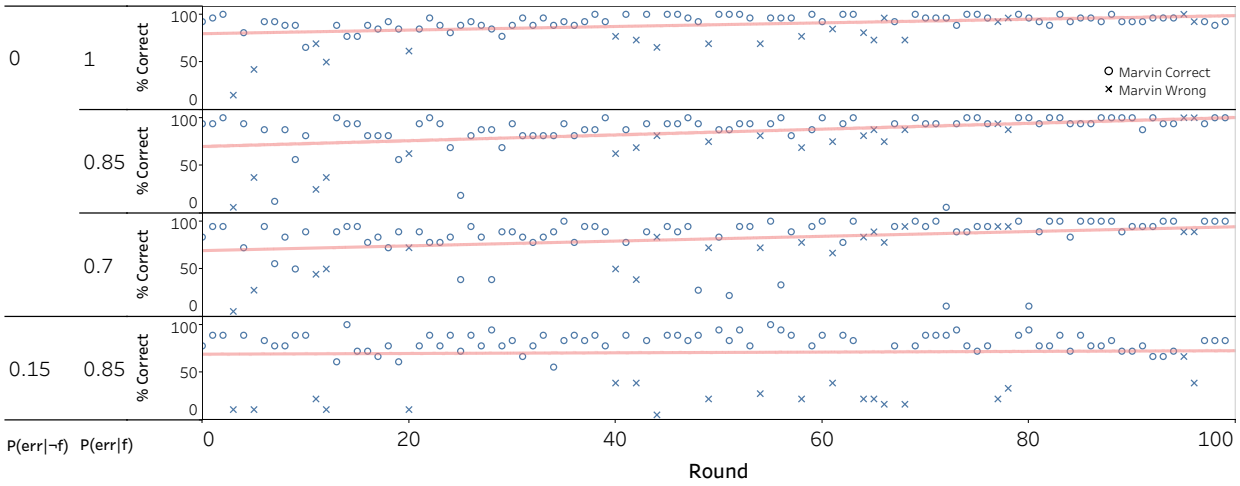


Figure 3.7: For one-sided error boundaries (the top three rows), the percentage of workers who choose the optimal action improves with time and reaches 100% – the positive slope of the best fit line shows this increasing trend. For the two-sided stochastic boundary (bottom row), the improvement is minimal and stays close to 50% – the slope of the best fit line is close to 0.

3.5 Related Work

Mental models for collaboration. Early work explored the importance of mental models for achieving high performance in group work [Grosz and Kraus, 1999; Mohammed *et al.*, 2010], human-system collaboration [Rouse *et al.*, 1992], and interface design [Carroll and Olson, 1988]. More recently, the impact of mental models has been revisited for better understanding human-in-the-loop systems [Chakraborti and Kambhampati, 2018] and for grounding human-AI collaboration within traditional HCI work [Kaur *et al.*, 2019]. Our work builds upon these foundations and studies the problem for AI-advised human decision making. While there exist many forms of mental modeling (*i.e.*, How does the system work?) and they are relevant for collaboration, this work focuses particularly on mental models about system performance (*i.e.*, When does the system err?), which are learned upon context and past experiences.

Backward compatibility. The closest relevant work to this study that also operates on mental models about error boundaries is presented in [Bansal *et al.*, 2019b] and focuses on the usefulness of such models during AI updates highlighting the importance of remaining backward compatible while deploying a new model. Backward compatibility is measured through comparing the errors of the previous and the updated version of the model and quantifying the percentage of all input instances that were correct in the previous version

that remain correct in the updated one. The work showed that error boundaries that are not backward compatible with previous versions of the model breaks mental models human have created in the process of collaboration, and showed that updates to a more accurate model that is not backward compatible can hurt team performance.

In traditional software design, backward compatibility is a well-studied software property [Bosch, 2009; Spring, 2005 11], used to denote software that remains compatible with a larger legacy ecosystem even after an update. In the field of AI/ML, a related notion to backward compatibility is catastrophic forgetting [Kirkpatrick *et al.*, 2017; Goodfellow *et al.*, 2013; McCloskey and Cohen, 1989], which is an anomalous behavior of neural network models that occurs when they are sequentially trained on more instances and forget to solve earlier instances over time. While forgetting in sequential learning is an important problem, backward compatibility is applicable to a larger set of update scenarios that do not necessarily require more data (*e.g.* different architecture or the same architecture but with different parameters).

Interpretability for decision-making. As learning models are being deployed to assist humans in taking high-stake decisions, the explainability of machine predictions is crucial for facilitating human understanding. Ongoing and prior research has contributed to improving the interpretability of such predictions either by building more interpretable models [Caruana *et al.*, 2015; Rudin, 2018; Lage *et al.*, 2018] or by imitating complex models via simpler but more explainable ones [Lakkaraju *et al.*, 2016; Tan *et al.*, 2018]. However, while explanations help with understanding, it is not yet clear under which conditions they improve collaboration and human productivity [Doshi-Velez and Kim, 2017; Poursabzi-Sangdeh *et al.*, 2019; Feng and Boyd-Graber, 2019]. For example, some explanations may describe how the system works but they do clearly disclose when it will fail and needs human intervention. In other cases, inspecting an explanation might take just as much time as solving the task from scratch (*i.e.*, high cognitive load). Both challenges motivate the need for predictable and easy-to-learn error boundaries, properties of which we study in our experimental evaluation. A promising direction is generating explanations of error boundaries themselves as a tool for users to quickly learn and remember various failure conditions. Recent work [Nushi *et al.*, 2018], uses decision trees to predict and visualize error boundaries for the purpose of debugging ML models but more work is needed on deploying and evaluating such tools for decision-making.

Modeling and communicating uncertainty in ML. Confidence calibration has been a topic of exten-

sive research, especially for embracing and communicating uncertainty in models that are inherently non-probabilistic. Foundational work in this space has proposed techniques for calibrating output scores for support vector machines [Platt, 1999], decision trees and Naïve Bayes models [Zadrozny and Elkan, 2001], and deep neural networks [Gal and Ghahramani, 2016; Gal and Ghahramani, 2016; Guo *et al.*, 2017]. Later work has proposed data collection algorithms for addressing overconfident predictions [Lakkaraju *et al.*, 2017; Bansal and Weld, 2018]. While confidence estimation and reporting is informative for decision-making, research in human-centered machine learning [Gillies *et al.*, 2016] and HCI shows that people have difficulties with correctly interpreting probabilistic statements [Handmer and Proudley, 2007] and even system accuracy itself [Yin *et al.*, 2019]. Research in the intersection of AI and HCI has found that interaction improves when setting expectations right about what the system can do and how well it performs [Kocielnik *et al.*, 2019; Amershi *et al.*, 2019]. This paper takes a step forward by proposing properties of ML models that can assist with setting the right expectations and evaluating them through controlled user studies. Moreover, we envision this line of work on making error boundaries predictable as complementary but also valuable for designing better confidence models as the defined properties. For example, parsimonious error boundaries are easier to generalize also from a statistical learning point of view, which would help with calibration.

3.6 Recommendations for Human-Centered AI

When developing ML models current practices solely target AI accuracy, even in contexts where models support human decision making. Our experiments reveal that error boundaries and task complexity can influence the success of teamwork. The user studies presented suggest the following considerations when developing ML models to be used in AI-advised human decision making:

3a Build AI systems with parsimonious error boundaries.

3b Minimize the stochasticity of system errors.

3c Reduce task dimensionality when possible either by eliminating features that are irrelevant for both machine and human reasoning or most importantly by analyzing the trade-off between the marginal gain of machine performance per added feature and the marginal loss of the accuracy of human mental models per added feature.

3d Based on results from Bansal *et al.* (2019), during model updates, deploy models whose error boundaries are *b*ackward compatible, *i.e.* by regularizing in order to minimize the introduction of new errors on instances where the user has learned to trust the system.

Given the importance of these properties on overall team performance, and potentially of other properties to be discovered in future work, it is essential to make such properties a part of considerations during model selection. For example, if a practitioner is presented with two different models, h_1 and h_2, of similar accuracy (e.g., such a situation could arise as a result of a grid search for hyper-parameter selection), and the error boundary f_1 is more stochastic than f_2, clearly h_2 would be the better choice. In a more complex situation, where h_2's accuracy is slightly inferior to h_1's, the practitioner must carefully estimate the potential loss in team accuracy attributed to human mistakes (*i.e.*, trusting the model when it is incorrect) due to stochasticity and compare this loss to the difference in accuracy between the two candidate models. Often, a negligible compromise in ML accuracy, can lead to for a higher gain in accuracy of overall teamwork. The same analysis could be employed to appraise the optimal tradeoffs when one human-aware property may be at odds with another (*e.g.*, making an error boundary more parsimonious might also make it more stochastic).

Model selection decisions also depend on the type of tools made available to users for learning and remembering error boundaries. For example, if users can access a scratchpad that records and summarizes the observed error boundary in real time, then they might be able to afford a slightly more complex error boundary.

Human-aware model selection should also be supported by making the presented properties part of the optimization problem is formulation while training either by including human-aware considerations in loss functions or by posing additional optimization constraints. The former technique has been used to combine backward compatibility in the loss function [Bansal *et al.*, 2019b] and to combine tree-based regularization to learn a more interpretable model [Wu *et al.*, 2018]; the latter has found application in domains like fair classification and healtcare [Dwork *et al.*, 2012; Zafar *et al.*, 2017; Ustun and Rudin, 2017]. More effort is needed to algorithmically ensure error boundary parsimony and non-stochasticity and combine such efforts for generating actionable confidence scores. This would reshape learning techniques to optimize for both the human in the loop or any other part of the ecosystem that requires reliable trust contracts to cooperate with the AI.

Finally, as human-AI collaboration becomes more pervasive, we foresee further opportunities to study human-AI team behavior in the open world, and for richer and more general forms of human-AI teams, for example, cases where AI recommendation directly updates human's belief in the final decision in contrast to our simplified notion of accept or override. Other opportunities include making interaction more natural by building computational models about what users have learned and by simplifying mental model creation using explanatory tools.

3.7 Conclusion

We studied the role of human mental models on the human-AI team performance for AI-advised human decision making for situations where people either rely upon or reject AI inferences. Our results revealed important properties that describe the error boundaries of inferences that can influence how well people can collaborate with an AI system and how efficiently they can override the AI when it fails. We find that systems with exactly the same accuracy can lead to different team performance depending upon the parsimony, non-stochasticity, and dimensionality of error boundaries. Future research opportunities include developing methods for integrating these considerations into algorithmic optimization techniques. While AI accuracy has been traditionally considered a convenient proxy for predicting human-AI team performance, our findings motivate investing effort to understand how to develop AI systems to support teamwork, in particular, in making properties of error boundaries more understandable and learnable when selecting an AI model for deployment.

Chapter 4

Updates in Human-AI Teams and Performance/Compatibility Tradeoff

AI systems are being deployed to support human decision making in high-stakes domains such as healthcare and criminal justice. In many cases, the human and AI form a team, in which the human makes decisions after reviewing the AI's inferences. A successful partnership requires that the human develops insights into the performance of the AI system, including its failures. We study the influence of *updates* to an AI system in this setting. While updates can increase the AI's predictive performance, they may also lead to behavioral changes that are at odds with the user's prior experiences and confidence in the AI's inferences (*i.e.*, violate the Maxim 2 which argues to preserve system's trustworthiness). We show that updates that increase AI performance may actually hurt *team* performance. We introduce the notion of the *compatibility* of an AI update with prior user experience and present methods for studying the role of compatibility in human-AI teams. Empirical results on three high-stakes classification tasks show that current machine learning algorithms do not produce compatible updates. We propose a re-training objective to improve the compatibility of an update by penalizing new errors. The objective offers full leverage of the performance/compatibility tradeoff across different datasets, enabling more compatible yet accurate updates.

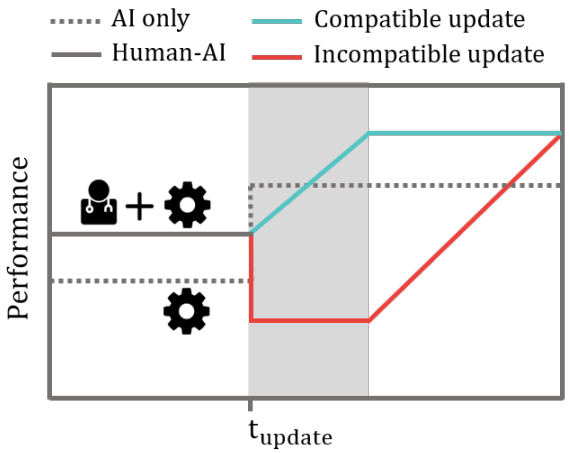


Figure 4.1: Schematized view of human-AI teams in the presence of AI updates. Human-AI teams perform better than either alone, but when the AI is *updated* its behavior may violate human expectations. Even if updates increase the AI's *individual* performance, they may reduce *team* performance by making mistakes in regions where humans have learned to trust the AI.

4.1 Introduction

A promising opportunity in AI is developing systems that can partner with people to accomplish tasks in ways that exceed the capabilities of either individually [Wang *et al.*, 2016; Kamar, 2016; Gaur *et al.*, 2016]. We see many motivating examples: a doctor using a medical expert system [Wang *et al.*, 2016], a judge advised by a recidivism predictor, or a driver supervising a semi-autonomous vehicle. Indeed, economists expect human-AI teams to handle many such tasks [Gownder *et al.*, 2017 06]. Despite rising interest, there is much to learn about creating effective human-AI teams and what capabilities AI systems should employ to be competent partners.

We study human-AI teams in decision-making settings where a user takes action recommendations from an AI partner for solving a complex task. The user considers the recommendation and, based on previous experience with the system, decides to accept the suggested action or take a different action. We call this type of interaction *AI-advised human decision making*. While there exist other important forms of human-AI collaboration (including human-advised AI decision making and more general collaborative decision making involving a mix of initiatives and emergent team behaviors), we focus on a specific interplay where the goal is to create AI systems that recommend actions to assist humans with decisions in high-stakes domains [Angwin *et al.*, 2016; Bayati *et al.*, 2014 10]. The motivation for AI-advised human decision

making comes from the fact that humans and machines have complementary strengths and abilities. While both human experts and machine-learned models are not perfect on a task like medical diagnosis or classifying objects in images, researchers have shown that their ideal combination could significantly improve performance [Wang *et al.*, 2016; Kamar *et al.*, 2012]. AI systems offer added benefits by speeding up decision making when humans can identify tasks where the AI can be trusted and no more human effort is needed [Lasecki *et al.*, 2012b; Lasecki *et al.*, 2012a].

It might be expected that improvements in the performance of AI systems lead to stronger team performance, but, as with human groups, individual ability is only one of many factors that affect team effectiveness [DeChurch and Mesmer-Magnus, 2010; Grosz, 1996]. Even in a simple collaboration scenario, in which an AI system assists a human decision maker with predictions, the success of the team hinges on the human correctly deciding when to follow the recommendation of the AI system and when to override. Unless the particular domain and the interaction allows the human to validate the correctness of the machine recommendation efficiently and effectively, extracting benefits from collaboration with the AI system depends on the human developing insights (i.e., a mental model) of when to trust the AI system with its recommendations. If the human mistakenly trusts the AI system in regions where it is likely to err, catastrophic failures may occur. Human-AI teams become especially susceptible to such failures because of discrepancies introduced by system updates that do not account for human expectations. The following example and Figure 4.1 illustrate this situation.

Example (PATIENT READMISSION)**.** *A doctor uses an AI system that is 95% accurate at predicting whether a patient will be readmitted following their discharge to make decisions about enlisting the patient in a supportive post-discharge program. The special program is costly but promises to reduce the likelihood of readmission. After a year of interacting with the AI, the doctor develops a clear mental model that suggests she can trust the AI-advised actions on elderly patients. In the meantime, the AI's developer trains and deploys a new 98% accurate classifier, which errs on elderly patients. While the AI has improved by 3%, the doctor is unaware of the new errors, and as a result of this outdated mental model, takes the wrong actions for some elderly patients.*

This example is motivated by real-world AI applications for reducing patient readmissions and other costly outcomes in healthcare [Bayati *et al.*, 2014 10; Wiens *et al.*, 2016; Caruana *et al.*, 2015], and motivates

the need for reducing the cost of disruption caused by updates that violate users' mental models. The problem with updates extends to numerous AI-advised human decision-making settings; similar challenges have been observed during over-the-air updates in the Tesla autopilot [O'Cane, 2018], and analogous issues arise in a variety of other settings when AI services being consumed by third-party applications, are updated.

Despite these problems, developers have almost exclusively optimized for AI performance. Retraining techniques largely ignore important details about human-AI teaming, and the mental model that humans develop from interacting with the system. The goal of this work is to make the human factor a first-class consideration of AI updates. We make the following contributions:

4.1 We define the notion of compatibility of an AI update with the user's mental model created from past experience. We then propose a practical adjustment to current ML (re)training algorithms — an additional differentiable term to the logarithmic loss — that improves compatibility during updates, and allows developers to explore the performance/compatibility tradeoff.

4.2 We introduce an open-source experimental platform[1] for studying how people model the error boundary of an AI teammate in the presence of updates for a an AI-advised decision-making task. The platform exposes important design factors (*e.g.*, task complexity, reward, update type) to the experimenter.

4.3 Using the platform, we perform user studies showing that, humans develop mental models of AI systems across different conditions, and that more accurate mental models improve team performance. More importantly, we show that updating an AI to increase accuracy, at the expense of compatibility, may *degrade* team performance. Moreover, experiments on three high-stakes classification tasks (recidivism prediction, in-hospital mortality prediction, and credit-risk assessment) demonstrate that: (i) current ML models are not inherently compatible, but (ii) flexible performance/compatibility tradeoffs can be effectively achieved via a reformulated training objective.

[1] Available at `https://github.com/gagb/caja`

4.2 AI-Advised Human Decision Making

As in Chapter 3, we focus on a simple, but common, model of human-AI teamwork that abstracts many real-world settings, *e.g.*, a 30-day readmission classifier supporting a doctor [Bayati *et al.*, 2014 10], a recidivism predictor supporting judges in courts [Angwin *et al.*, 2016]. In this setting, which we call *AI-advised human decision making*, an AI system provides a *recommendation*, but the human makes the final *decision*.

4.2.1 Trust as a Human's Mental Model of the AI

Cognitive psychology research shows that when people interact with any complex system, they create a mental model, which facilitates their use of the system [Norman, 1994]. Just as for other automated systems, humans create a mental model of AI agents [Kulesza *et al.*, 2012]. In AI-advised human decision making, valid mental models of the reliability of the AI output improve collaboration by helping the user to know when to trust the AI's recommendation. A perfect mental model of the AI system's reliability could be harnessed to achieve the highest team performance. A simple definition for such a model would be $m : x \rightarrow \{T, F\}$, indicating which inputs the human trusted the AI to solve correctly. A more complex model might compute a probability and include additional arguments, such as the AI's output, $h(x)$. In reality, mental models are not perfect [Norman, 1994]: users develop them through limited interaction with the system, and people have cognitive limitations. Furthermore, different team members may have access to different information about the situation. For example, a doctor may know things about a patient that are missing from electronic health records (*e.g.*, an estimate of the patient's compliance with taking medications), while an AI system may have access to the most recent results and trends in physiological state that are not tracked by physicians. In summary, users learn and evolve a model of an AI system's competence over the course of many interactions. In the experimental section, we show that these models can greatly improve team performance. Next, we study the problem of updating an AI system within the context of AI-assisted human decision making, and introduce the notion of compatibility.

4.3 Compatibility of Updates to Classifiers

Developers regularly update AI systems by training new models with additional or higher-quality training data, or by switching to an improved learning algorithm. Such updates presumably improve the AI's performance on a validation set, but the patient readmission example highlights how this is not always sufficient: updates can arbitrarily change the AI's error boundary, introduce new errors which violate user expectations and decrease team performance.

In software engineering, an update is *backward compatible* if the updated system can support legacy software. By analogy, we define that an update to an AI component is *locally compatible* with a user's mental model if it does not introduce new errors and the user, even after the update, can safely trust the AI's recommendations.

Definition (LOCALLY-COMPATIBLE UPDATE)**.** *Let $m(x)$ denote a mental model that dictates the user's trust of the AI on input x. Let $A(x, u)$ denote whether u is the appropriate action for input x. An update, h_2, to a learned model, h_1, is locally compatible with m iff*

$$\forall x, [m(x) \wedge A(x, h_1(x))] \Rightarrow A(x, h_2(x))$$

In other words, an update is compatible only if, for every input where the user trusts the AI and h_1 recommends the correct action, the updated model, h_2, also recommends the correct action. In the rest of this paper, we focus on situations where a classifier's predictions are actions. For instance, in the patient readmission example, if a classifier predicts that the patient will be readmitted in the next 30 days, the suggested action from the classifier would be to include the patient in a special post-discharge program.

4.3.1 Globally Compatible Updates

When developers are building an AI system that is used by many individuals, it may be too difficult to track individual mental models or to deploy different updated models to different users. In this situation, an alternative to creating locally compatible updates, is a *globally compatible update*. To make this notion precise, we observe that a developer who is updating a classifier with new training data goes through the following steps:

1. Collect initial training data D_1.
2. Train a model h_1 on D_1 and deploy h_1.
3. Collect additional data to create D_2, where $D_1 \subset D_2$.
4. Train h_2 on D_2.
5. If the performance of h_2 is higher than h_1, deploy h_2.

Similar steps can be formulated for a model update where the training data does not change ($D_2 = D_1$) but h_2 belongs to a different model class.

Definition (GLOBALLY-COMPATIBLE UPDATE)**.** *An updated model, h_2, is globally compatible with h_1, iff*

$$\forall x, A(x, h(x)) \Rightarrow A(x, h_2(x))$$

Note that a globally compatible update is locally compatible for *any* mental model. While global compatibility is a nice ideal, satisfying it for all instances is difficult in practice. More realistically, we seek to minimize the number of errors made by h_2's that were not made by h_1, since that will hopefully minimize confusion among users. To make this precise, we introduce the notion of a *compatibility score*.

Definition (COMPATIBILITY SCORE)**.** *The compatibility score $\mathcal{C}$ of an update h_2 to h_1 is given by the fraction of examples on which h_1 recommends the correct action, h_2 also recommends the correct action.*

$$\mathcal{C}(h_1, h_2) = \frac{\sum_x A(x, h_1(x)) \cdot A(x, h_2(x))}{\sum_x A(x, h_1(x))} \tag{4.1}$$

If h_2 introduces no new errors, $\mathcal{C}(h_1, h_2)$ will be 1. Conversely, if all the errors are new, the score will be 0.

4.3.2 Dissonance and Loss

To train classifiers, ML developers optimize for the predictive performance of h_2 by specifying, and minimizing, a classification loss L that penalizes low performance. The equation below shows the negative

logarithmic loss (also known as log loss or cross-entropy loss) for binary classification – a commonly used training objective in ML.

$$L(x, y, h_2) = y \cdot \log p(h_2(x)) + (1 - y) \cdot \log(1 - p(h_2(x)))$$

Here, the probability $p(h(x))$ denotes the confidence of the classifier that recommendation $h(x)$ is true, while y is the true label for x (*i.e.*, $A(x, y) = \textit{True}$). The negative log loss, like many other loss functions in machine learning, depends only on the true label and the confidence in prediction – it ignores the previous versions of the classifier and, hence, has no preference for compatibility. As a result, retraining using different data can lead to very different hypotheses, introduce new errors, and decrease the compatibility score. To alleviate this problem, we define a new loss function L_c expressed as the sum of classification loss and *dissonance*.

Definition (DISSONANCE)**.** *The dissonance $\mathcal{D}$ of h_2 to h_1 is a function $\mathcal{D} : x, y, h_1, h_2 \rightarrow \mathcal{R}$ that penalizes a low compatibility score. Furthermore, $\mathcal{D}$ is differentiable.*

$$\mathcal{D}(x, y, h_1, h_2) = \mathbb{1}(h_1(x) = y) \cdot L(x, y, h_2) \tag{4.2}$$

Recall that $\mathcal{C}(h_1, h_2)$ is high when both h_1 and h_2 are correct (Eqn 4.1). Dissonance expresses the opposite notion: measuring if h_1 is correct ($\mathbb{1}$ denotes an indicator function) and penalizing by the degree to which h_2 is incorrect. Equation 4.3 defines the new loss.

$$L_c = L + \lambda_c \cdot \mathcal{D} \tag{4.3}$$

Here, λ_c encodes the relative weight of dissonance, controlling the additional loss to be assigned to all new errors. We refer to this version as *new-error dissonance*. Just as with classification loss, there are other ways to realize dissonance. We explored two alternatives, which we refer to as *imitation* and *strict imitation* dissonance. Eqn 4.4 describes the imitation dissonance which measures the log loss between the prediction probabilities of h_1 and h_2:

$$\mathcal{D}'(x, y, h_1, h_2) = L(x, h_1, h_2) \tag{4.4}$$

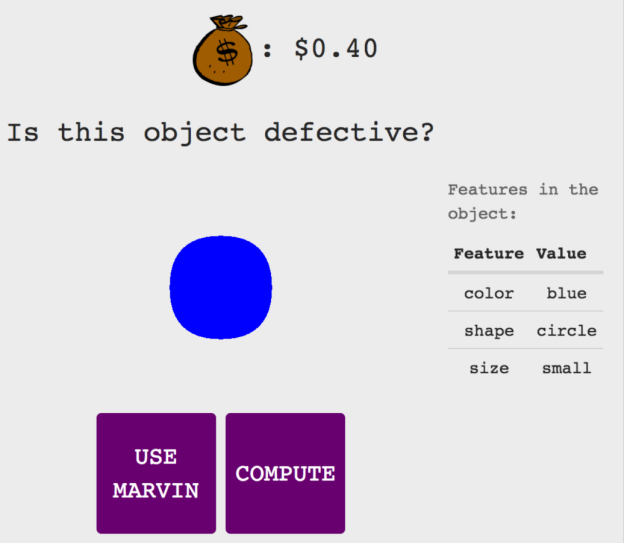


Figure 4.2: Screenshot of the CAJA platform for studying human-AI teams.

Eqn 4.4 is used in model distillation [Ba and Caruana, 2014; Hinton *et al.*, 2015], where the aim is to train a shallower, less expensive model by imitating the probabilities of larger, accurate model. Unfortunately, $\mathcal{D}'$ has the effect of nudging h_2 to mimic h_1's mistakes as well as its successes. Eqn 4.5 describes the strict imitation dissonance, which follows a similar intuition but it only adds the log loss between h_1 and h_2 when h_1 is correct.

$$\mathcal{D}''(x, y, h_1, h_2) = \mathbb{1}(h_1(x) = y) \cdot L(x, h_1, h_2) \tag{4.5}$$

Compared to dissonance, $\mathcal{D}$, $\mathcal{D}''$ still puts a larger emphasis on matching h_1's predictions (*vs.* the true labels, y), which we worried would hurt accuracy. Our experiments (*e.g.*, Figure 4.5) confirm this intuition and show the effect of varying λ_c on the performance/compatibility tradeoff.

4.4 Platform for Studying Human-AI Teams

How might we study the impact of AI accuracy, updates, compatibility, and mental models on the performance of AI-advised human decision making teams? Ideally, we would conduct user studies in real-world settings, varying parameters like the length of interaction, task and AI complexity, reward function, and the AI's behavior. All human-subjects research is challenging, but our setting poses special perplexities. Testing in real settings reduces or removes our ability to directly control the performance of the AI. Furthermore,

it may largely measure experts' differing experience in the domain, rather than their interactions with the AI. The importance of mental models for team success varies among domains and the interaction designed between the AI system and humans. When humans do not have an easy way to validate machine correctness, extracting value out of AI assistance depends on the ability of the human developing a mental model of the AI system.

To control for human expertise and the centrality of mental modeling, we developed the CAJA platform, which supports parameterized user studies in an assembly line domain that abstracts away the specifics of problem solving and focuses on understanding the effect of mental modeling on team success. CAJA is designed such that *no* human is a task expert (nor can they become one). In fact, the true label of decision problems in the platform is randomly generated so that people cannot learn how to solve the task. However, humans can learn when their AI assistant, Marvin, succeeds and when Marvin errs. Alongside, the human has access to a perfect problem-solving mechanism, which she can use (at extra cost) when she does not trust Marvin.

Specifically, CAJA is a web-based game, whose goal is to make classification decisions for a fixed number of box-like objects. For each object, the team follows the steps S1-S4 to decide whether the object is "defective" or not. In S1 a new object appears (*e.g.*, blue square), in S2 the AI recommends a label (*e.g.*, not-defective), in S3 the player chooses an action (*e.g.*, accept or reject the AI recommendation), and in S4 the UI returns a reward and increments the game score. The objects are composed of many features, but only a subset of them are made *human-visible*. For example, visual properties like shape, color, and size are visible, but the contents are not. In contrast, the AI has access to all the features but may make errors. At the beginning of the game, users have no mental model of the AI's error boundary. However, to achieve high scores, they must learn a model using feedback from step S4. Figure 4.2 shows a screenshot of the game at step S3.

CAJA allows study designers to vary parameters, such as the number of objects, number human-visible features, reward function, AI accuracy, and complexity of perfect mental model (number of clauses and literals in the error boundary and stochasticity of errors). Further, it enables one to study the effects of updates to AI by allowing changes to these parameters at any time step. In the next section, we use CAJA to answer various research questions.

	Accept	Compute
AI right	$0.04	0
AI wrong	-$0.16	0

Table 4.1: Reward matrix for the user studies. To mimic high-stakes domains, penalty for mistakes is set to high.

4.5 Experiments

We present experiments and results in two parts. First, using our platform, we conduct user studies to understand the impact of mental models and updates on team performance. Second, we simulate updates for three real-world, high-stakes domains and show how the retraining objective enables an explorable tradeoff between compatibility and performance that is not available in the original models.

User Studies. In user studies, we hired MTurk workers and directed them to the CAJA platform.[2] We informed them of the purpose of the study and provided a set of simple instructions to familiarize them with the task and the user interface: form a team with an AI, named Marvin, and label a set of 100 objects as "defective" or "not defective". Following AI-advised human decision making, to label an object, a worker can either accept Marvin's recommendation, which is initially correct 80% of the time, or use the "compute" option, which is a surrogate for the human doing the task herself perfectly but incurring an opportunity cost. Table 4.1 summarizes the reward function used in our studies. The matrix is designed in a way that it imitates a high-stakes scenario, i.e., the monetary penalty for a wrong decision is much higher than the reward for a correct decision. We found this design choice to be a good incentive for workers to learn and update their mental model on Marvin. Note that the expected value of a pure strategy (*e.g.*, always "Compute" or always "Accept," without considering the likelihood of Marvin's correctness) is zero. The only way to get a higher score is by learning when to trust Marvin. While the subjects are told Marvin's accuracy and the payoff matrix, they can only learn Marvin's error boundary gradually by playing the game. These design choices allow us to study the impact of mental models while controlling for human problem solving expertise — every player is able to solve problems perfectly, at a fixed cost, using "compute."

Q1: *Do better mental models of AI lead to higher team performance?*

[2]Workers were paid on average $20/hr, over the minimum wage in line with ethical guidelines for requesters [Dynamo,].

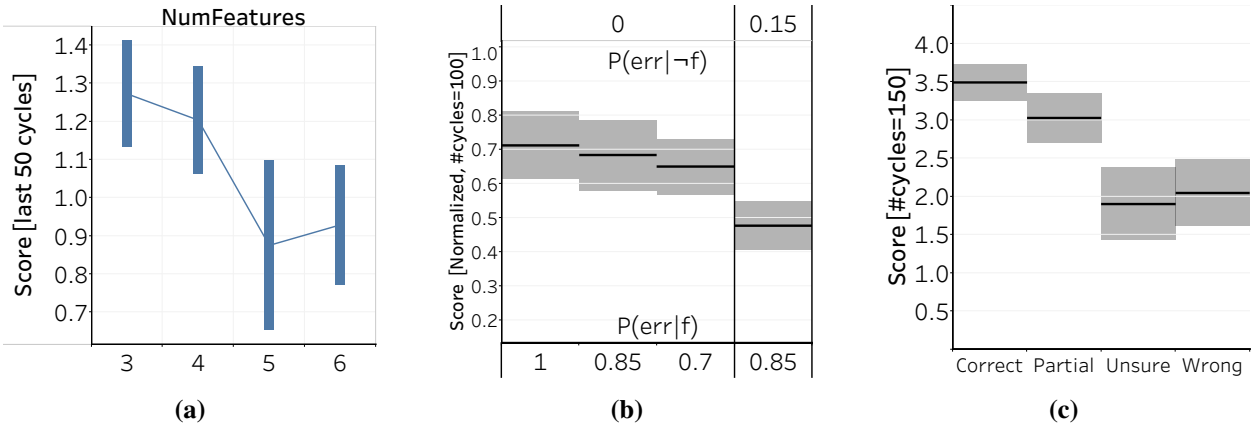


Figure 4.3: (a) Team performance decreases as we increase the number of human-visible features. (b) Team performance decreases with the stochasticity of errors. The decrease is much higher for two-sided errors. (c) Better mental models result in higher team performance. Wrong and Unsure mental models have the lowest performance.

To answer this, we conducted human studies that measured team performance across different conditions of complexity of task and error boundary. For each condition, we hired 25 MTurk workers, and filtered spammers by deleting data from workers in the bottom quartile. We varied the task complexity by varying the number of human-visible features. The complexity of error boundary f, expressed as a logical formula, is varied by changing the number of conjuncts and literals in f. For example, we tried one conjuncts containing two literals, one conjunct with three literals, and two conjuncts with two literals. Since many features can be used as literals, we chose them randomly to create different but isomorphic error boundaries. For example, worker A gets $f_A = (blue \cap square)$ and worker B gets $f_B = (red \cap circle)$. Figure 4.3a shows that, for one conjunct and two literals, team performance decreases with the number of features. We observed a similar behavior for other error boundaries (results omitted for space), and for the rest of these studies we set the number of conjuncts to one. Figure 4.3a shows that, as the number of features increases, team performance decreases because it becomes harder to create a mental model.

Next, we conducted a study (Figure 4.3b) to understand the impact of *stochasticity* in the error boundary on team performance. Stochasticity is defined using two conditional probabilities: $P(err|f)$ and $P(err|\neg f)$. That is, the probability of error if f is satisfied, and if it is not satisfied. To vary stochasticity, we chose the following four pairs of probabilities: (0.7, 0), (0.85, 0), (1.0, 0), and (0.85, 0.15). For the first three pairs, the errors are "one-sided": since $P(err|\neg f)$ is 0, the classifier makes a mistake only if the formula is satisfied.

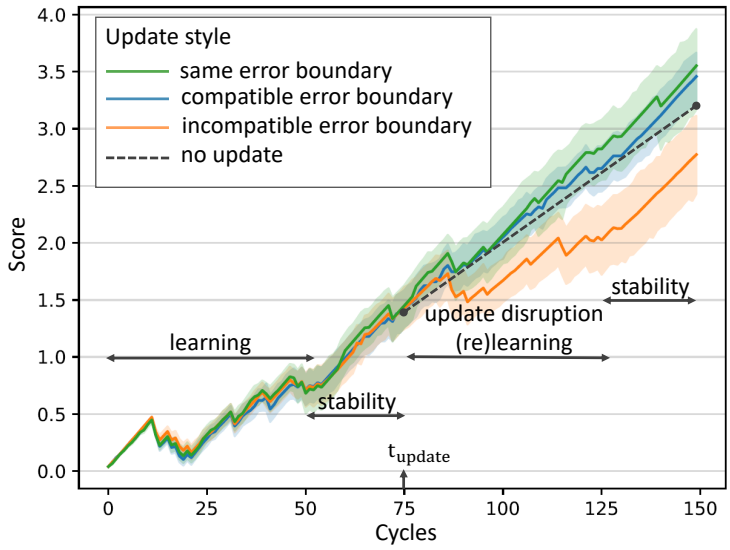


Figure 4.4: Team performance for different update settings. Compatible updates improve team performance, while incompatible updates hurt team performance despite improvements in AI accuracy.

In the last pair, errors are "two-sided": with a probability of 0.15, the classifier makes a mistake even if the formula is not satisfied. We fix the number of features to three and the number of literals to two. Figure 4.3b shows that as errors become more stochastic, it becomes harder to create a mental model, deteriorating team performance. The y-axis shows the score normalized by the score of the optimal policy because, as we vary stochasticity, the optimal policy's score changes.

Finally, in order to have a closer view of the quality of the workers' mental models, we ask them to self report when they thought Marvin was wrong. We manually labeled these reports as correct, partial, unsure, and wrong, without looking at their team performance. The label denotes how the worker's mental model compared to the true error boundary f. For example, correct denotes that the mental model and f were the same, wrong denotes no match, partial denotes an incomplete match, and unsure denotes that the worker was skeptical of their mental model. Figure 4.3c compares team performance on these groups. Workers with the correct mental model score the highest, followed by workers with a partially correct model. These observations confirm that better mental models contribute positively to team performance.

Q2: *Do more compatible updates lead to higher team performance than incompatible updates?*

To study the impact of updates, we set the number of cycles to 150, and at the 75th cycle, update the classi-

fier to a version that is 5% more accurate (80% $\rightarrow$ 85%). Then, we divide the participants into three groups: same error boundary, compatible error boundary, and incompatible error boundary. The same error boundary group receives an update improving accuracy, but the error boundary is unchanged. For the two other groups, the number of literals (features) in the error boundary changes from two to three. The update for the compatible error boundary group introduces no new errors; for example, if before the update the error boundary was $blue \cap square$, after the update it may change to $small \cap blue \cap square$. For the incompatible error boundary group, the error boundary introduces new errors violating compatibility. Figure 4.4 summarizes our results. We also show the performance of workers if no update was introduced (dashed line). It uses the no-update setting from experiments in Q1, and extrapolates from there assuming that the worker's mental model is already stable at the 75th cycle, meaning that the human-AI team has reached the maximum performance for the original setting and no further improvements are expected. The graph demonstrates two main findings on the importance of compatibility. First, a more accurate but incompatible classifier results in lower team performance than a less accurate but compatible classifier (no update). Second, compatible updates improve team performance. Moreover, the figure shows different stages during the interaction: the user learning the original error boundary, team stabilizes, update causes disruption, and performance stabilizes again. A central insight in the update stage is that the incompatible error boundary condition sacrifices the team score while workers have to relearn the new boundary. This insight shows that compatible updates not only improve team performance but they can also reduce the cost of retraining users after deploying system updates.

4.5.1 Experiments with High-Stakes Domains

Datasets. To investigate whether a tradeoff exists between performance and compatibility of an update, we simulate updates to classifiers for three domains: recidivism prediction (Will a convict commit another crime?)[Angwin *et al.*, 2016], in-hospital mortality prediction (Will a patient die in the hospital?) [Johnson *et al.*, 2016; Harutyunyan *et al.*, 2017], and credit risk assessment (Will a borrower fail to pay back?)[3]. We selected these high-stakes domains to highlight the potential cost of mistakes caused by incompatible updates in human-AI teams.

[3] https://community.fico.com/s/explainable-machine-learning-challenge

Classifier	Dataset	ROC h_1	ROC h_2	$\mathcal{C}(h_1, h_2)$
LR	Recidivism	0.68	0.72	0.72
	Credit Risk	0.72	0.77	0.66
	Mortality	0.68	0.77	**0.40**
MLP	Recidivism	0.59	0.73	0.53
	Credit Risk	0.70	0.80	0.63
	Mortality	0.71	0.84	0.76

Table 4.2: Although training on a superset of data increases classifier performance, compatability can be suprisingly low.

Q3: *Do current ML classifiers produce compatible updates?*

For this experiment, we first train a classifier h_1 on 200 examples and note its performance. Next, we train another classifier h_2 on 5000 examples and note its performance and compatibility score. We train both classifiers by minimizing the negative log loss. Table 4.2 shows the performance (area under ROC) and compatibility averaged over 500 runs for logistic regression (LR) and multi-layer perceptron (MLP) classifiers. We find that training h_2 by just minimizing log loss does not ensure compatibility. For example, for logistic regression and the in-hospital mortality prediction task, the compatibility score is as low as 40%. That is, 60% of the instances where h_1 was correct are now violated.

Q4: *Does there exist a tradeoff between the performance and the compatibility of an update to AI?*

For Q4 (and Q5), we learn the second classifier h_2 by minimizing L_c. As L_c depends also on the first classifier, we make its prediction available to the learner. We vary λ_c and summarize the resulting performance and compatibility scores across different datasets for the logistic regression and multi-layer perceptron classifiers in Figure 4.5 and for different definitions of dissonance (discussed in Q5). The figure shows that there exists a tradeoff between the performance of h_2 and its compatibility to h_1. This tradeoff is generally more flexible (flat) in the first half of the curves. This shows that, at the very least, one can choose to train via L_c and deploy a more compatible update without significant loss in accuracy. Although such updates are not fully compatible, they might still be relevant to be picked by the developer if the update is supported by efficient explanation techniques that can help users to better understand how the model has changed. In these cases, a more compatible update would also reduce the effort of user (re)training. In the second half,

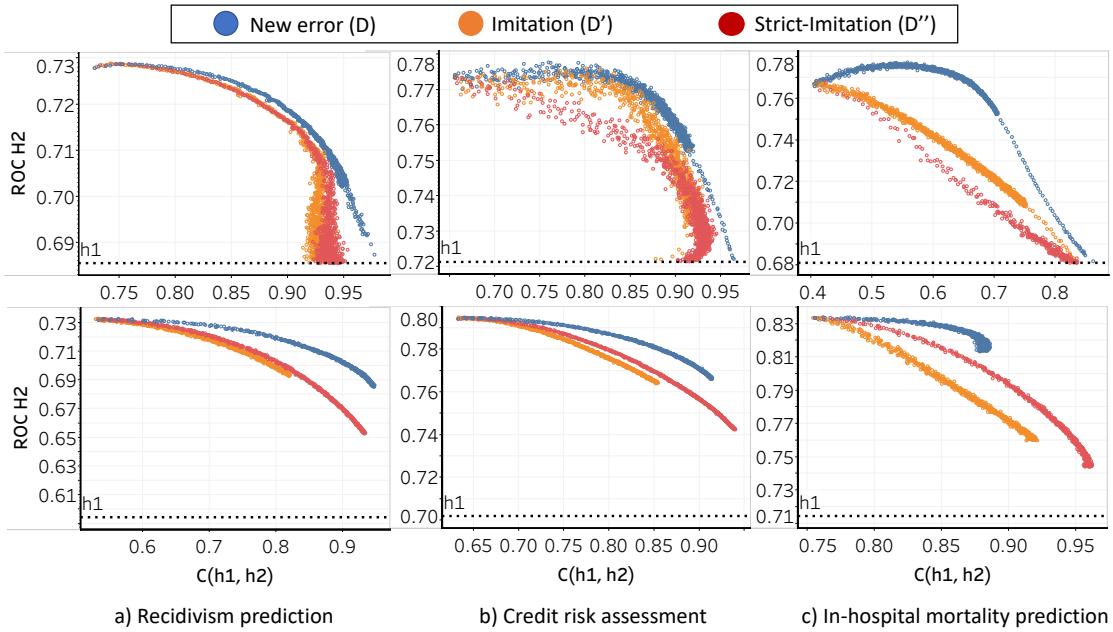


Figure 4.5: Performance vs. compatibility for a logistic regression and multi-layered perceptron classifiers. The reformulated training objective (L_c) offers an explorable performance/compatibility tradeoff, generally more forgiving during the first half of the curves. The training objective based on new-error dissonance performs the best, whereas the ones based on imitiation and strict-imitation dissonance perform worse since they imitate probabilities of a less accurate, and less calibrated model (h_1).

the tradeoff becomes more evident. High compatibility can sacrifice predictive performance. Look-up summaries similar to graphs shown in Figure 4.5 are an insightful tool for ML developers that can guide them select an accurate yet compatible model based on the specific domain requirements.

Q5: *What is the relative performance of the different dissonance functions?*

Figure 4.5 compares the performance of the new-error dissonance function ($\mathcal{D}$) with the imitation-based dissonances ($\mathcal{D}'$ and $\mathcal{D}''$). As anticipated, $\mathcal{D}$ performs best on all three domains. The definitions inspired by model distillation, $\mathcal{D}'$ and $\mathcal{D}''$, assume that h_1 is calibrated, and more accurate. Therefore, h_2 needs to remain faithful to only the correct regions of a less accurate model h_1. If these assumptions are violated, h_2 overfits to non-calibrated confidence scores of h_1, which hurts performance.

4.6 Discussion and Directions

The AI-assisted human decision-making problem assumes that there are instances for which the AI is more efficient (*e.g.*, higher accuracy, faster, or low resource usage), and the human can recognize when the AI is capable of doing so. Earlier, we discussed that one way for humans to recognize when to follow the AI's recommendations is by creating mental models. However, depending on the domain and the type of interaction design, the importance of mental modeling for team performance may vary. For example, if the human can quickly validate the correctness of the recommendation, or the human expertise improves over time to leave no room for machine contribution, then mental modeling may not be needed. Otherwise, the accuracy of the mental model limits team performance. Thus, compatibility of updates becomes an essential determinant of team performance, and developers should factor it in system design supported by guiding tools exploring the performance/compatibility tradeoff.

Varying the value of λ_c results in numerous models on the performance/compatibility spectrum. The decision to select the appropriate model depends on several factors, including the user ability to create a mental model, the cost of disruption, and whether there exist other alternative approaches for minimizing disruption caused by updates. For example, if the cost of disruption (both the cognitive cost and mistakes) is high, then we may use a high value for λ_c. A more formal approach would be to set λ_c algorithmically. For example, a λ_c could be selected to maximize expected utility expressed using a computational user model and future rewards.

A developer can use other complementary approaches to minimize disruption caused by low compatibility. One approach is to retrain the user, for example, by leveraging mechanisms from interpretable AI to explain the updated model to users or to explain differences between h_1 and h_2. However, this may not always be practical: (1) in practice, developers may push updates frequently, and since re-training requires user's additional time and effort, it may not be practical to subject experts to repeated re-training; (2) updates can arbitrarily change the decision boundary of a classifier, and as a result, require the user to re-learn a large number of changes; (3) re-training requires the developers to create an effective curriculum or generate a "change summary" based on the update. It is often impossible to compute such summaries in a human-interpretable way. For example, explaining the changes to a self-driving car may require the challenging task of mapping the feature representation used by the car (myriad of sensor data) to human-interpretable

concepts. Nevertheless, backward compatibility does not preclude retraining; these techniques are complementary to each other. In fact, more compatible updates can be an efficient mechanism to simplify the re-training process by minimizing the divergence between two models deployed consecutively. Yet another complementary approach is to share the AI's confidence in the prediction. Well-calibrated confidence scores can help a user to decide when or how much to trust the system. Unfortunately, confidence scores of ML classifiers are often not calibrated [Nguyen *et al.*, 2015] or a meaningful confidence definition may not exist due to the complexity of the task.

We formalized compatibility in terms of differences in model recommendations before and after an update, independent of mental models of users. An important future direction is to develop computational models of how people create and update mental models, and condition on the personalized experiences and cognitive capabilities of each user, drawing upon general findings about how people learn about phenomena via observation [Reber, 1989]. While this work distills trust as the essence of teamwork and presented results are applicable to a variety of use cases, promising extensions include developing blended studies in the real world that combine both factors of human problem solving and learned trust in AI.

4.7 Related Work

Prior seminal work explored the importance of mental models for achieving high performance in group work [Grosz and Kraus, 1999], human-system collaboration (Rouse *et al.* 1992), and interface design [Carroll and Olson, 1988]. Our work builds upon these foundations and studies the problem for AI-advised human decision making. Other work [Hoff and Bashir, 2015] highlights the connection between mental models and trust in systems. While many "layers" of trust exist, our work focuses on *learned trust*, which is built upon context and past experiences [Marsh and Dibben, 2003]. Previous work [Zhou *et al.*, 2017] investigated factors that affect user-system trust, e.g., model uncertainty and cognitive load. The platform proposed in this work enables human studies that can analyze the effect of such factors.

The field of software engineering also considers the problem of backward compatibility, seeking to design components that, after updates, remain compatible with a larger software ecosystem[Bosch, 2009; Spring, 2005 11; Tsantilis, 2009 10]. Machine learning research has explored related notions. *Stability* expresses the ability of a model to not significantly change its predictions given small changes in the train-

ing set [Bousquet and Elisseeff, 2001]. *Consistency*, which has application in ML fairness, is a property of smooth classifiers, which output similar predictions for similar data points [Zhou *et al.*, 2004]. *Catastrophic forgetting* is an anomalous behavior of neural network models that occurs when they are sequentially trained to perform multiple tasks and forget to solve earlier tasks over time [Kirkpatrick *et al.*, 2017]. While these concepts are fundamental for analyzing changing trends in continuously learned models, they do not consider human-AI *team* performance nor prior user experience. Related to our proposed retraining objective is the idea of *cost-sensitive* learning [Elkan, 2001], where different mistakes may cost differently; for example, false positives may be especially costly. However, in our case, the cost also depends on the behavior of the previous model h_1.

4.8 Conclusions

We studied how updates to an AI system can affect human-AI team performance and introduced methods and measures for characterizing and addressing the compatability of updates. We introduced CAJA, a platform for measuring the effect of AI performance and the effect of updates on team performance. Since humans have no experience with CAJA's abstract game, the platform controls for human problem-solving skill, distilling the essence of mental models and trust in one's AI teammate. Using CAJA, we presented experiments demonstrating how an update that makes an AI component more accurate can still lead to diminished human-AI team performance. We introduced a practical re-training objective that can improve the compatibility of updates. Experiments across three data sets show that our approach creates updates that are more compatible, while maintaining high accuracy. Therefore, at the very least, a developer can choose to deploy a more compatible model without sacrificing performance.

Chapter 5

Optimizing AI for Teamwork

AI practitioners typically strive to develop the most *accurate* systems, making an implicit assumption that the AI system will function autonomously. However, in practice, AI systems often are used to provide *advice* to people in domains ranging from criminal justice and finance to healthcare. In such *AI-advised decision making*, humans and machines form a *team*, where the human is responsible for making final decisions. But is the most accurate AI the best teammate? We argue "not necessarily" — predictable performance may be worth a slight sacrifice in AI accuracy. Instead, we argue that AI systems should be trained in a human-centered manner , directly optimized for *team performance* (Maxim 3). We study this proposal for a specific type of human-AI teaming, where the human overseer chooses to either *accept* the AI recommendation or *solve* the task themselves. To optimize the team performance for this setting we maximize the team's *expected utility*, expressed in terms of the quality of the final decision, cost of verifying, and individual accuracies of people and machines. Our experiments with linear and non-linear models on real-world, high-stakes datasets show that the most accuracy AI may not lead to highest team performance and show the benefit of modeling teamwork during training through improvements in expected team utility across datasets, considering parameters such as human skill and the cost of mistakes. We discuss the shortcoming of current optimization approaches beyond well-studied loss functions such as log-loss, and encourage future work on AI optimization problems motivated by human-AI collaboration.

Utility Matrix
(Cost of human effort $\lambda = 0.5$, Cost of mistake $\beta = 1$)

Meta-decision/Decision	Correct	Incorrect
Accept	1.0	-1.0
Solve	0.5	-1.5

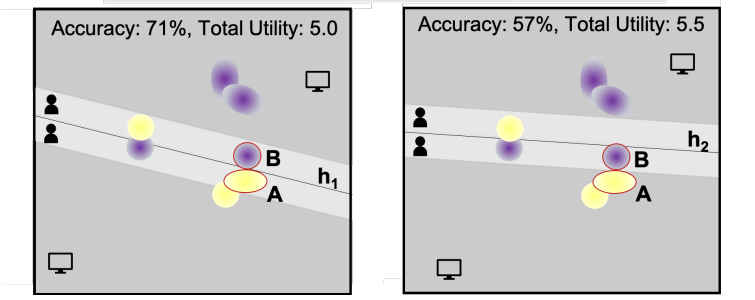


Figure 5.1: Consider a binary classification problem (purple vs. yellow). Assume each blob is uniformly distributed and of the same size. In a human-AI team, a more accurate classifier (h_1, left pane, learned using log-loss) may produce lower *team utility* than a less accurate model (h_2, right pane). Suppose the human can either quickly *accept* the AI's recommendation or *solve* the task themselves, incurring a cost λ in time or effort, to yield a more reliable result. The payoff matrix describes the utility of different outcomes. We explore the policy where humans accept recommendations when the AI is confident, but verify uncertain predictions (shown in the light grey region surrounding each hyperplane). While h_2 is less accurate than h_1 (because B is incorrectly classified), it results in a higher team utility: Since h_2 moved A outside the verify region, there are more *correctly classified* inputs on which the user can rely on the system.

5.1 Introduction

Many AI systems are developed for use in collaborative settings, where people work with an AI teammate. For example, numerous applications of AI have been designed as advisory tools, providing input to people who are tasked with making final decisions. Beyond the appropriateness of people making the final calls, the advisory role of AI systems may be obligatory; legal requirements may prohibit complete automation [GDPR, 2020]. Studies have demonstrated domains and tasks where human-AI teams may perform better than either the AI or human alone [Nagar and Malone, 2011; Patel *et al.*, 2019; Kamar *et al.*, 2012]. For human-AI teams, optimizing the performance of the whole team is more important than optimizing the performance of an individual member. Yet, to date, the AI community has focused on maximizing the individual accuracy of machine-learned models, assuming implicitly that this will optimize team performance. This raises an important question: Is the most accurate AI the best possible teammate for a human?

We argue that the most accurate model is not *necessarily* the best teammate. We show this formally, but the intuition is simple. Considering human-human teams, *Is the best-ranked tennis player necessarily*

the best doubles teammate? Clearly not—teamwork puts additional demands on participants that extend beyond individual performance on tasks, such as ability to complement and coordinate with one's partner. Similarly, creating high-performing human-AI teams may require training AI systems that exhibit additional *human-centered* properties, e.g., facilitating appropriate levels of trust and delegation. Implicitly, this is the motivation behind much work in intelligible AI, including efforts aimed at enhancing the understandability of complex AI inference [Horvitz *et al.*, 1986], interpretability of machine-learned models [Caruana *et al.*, 2015; Weld and Bansal, 2019 05], and performing post-hoc explanations of the output of models [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017]. We move beyond such general motivation and highlight the value of developing methods to model and optimize the collaborative process.

For example, consider the scenario when the system generates advice in which it is uncertain. In practice, users are likely to distrust such recommendations, and rightly so, because a low confidence is often correlated with erroneous predictions [Bansal *et al.*, 2021a; Hendrycks *et al.*, 2018]. In this work, we assume that, when systems have low confidence in their inferences, users will discard the recommendation and *solve* the task themselves, incurring a cost based in the required additional human effort. As a result, team performance depends on the AI accuracy only in the *accept region*, *i.e.*, the region where a user is actually likely to rely on AI. The singular objective of optimizing for AI accuracy (*e.g.*, using log-loss) may hurt team performance when the model has fixed inductive bias. Team performance will benefit from improving AI in the accept regions even if at the cost of performance over the complementary *solve regions* (Figure 5.1). While there exist other aspects of collaboration that can also be addressed via optimization techniques, such as model interpretability, supporting complementary skills [Wilder *et al.*, 2020 07], or enabling learning among partners, the problem we address in this paper is to account for team-based utility as a basis for collaboration. In sum:

5.1 We highlight an important direction in the field of human-centered AI: When paired with a human overseer, the most accurate ML model may not lead to the highest *team performance*. Specifically, we consider settings where, during training the system considers humans' mental model of the AI and how they make use of its recommendations. This setting complements recent advances in *learning to defer* where systems are trained when to refuse to share a recommendation to the overseer.

5.2 For a simple yet ubiquitous form of teamwork, we show that log-loss, the most popular loss function

for optimizing AI accuracy, can lead to suboptimal team performance and instead propose directly optimizing for human-AI team's utility. During training, the new objective considers and guides AI performance by considering various human and domain parameters, such as human accuracy, cost of human effort, and cost of mistakes.

5.3 We present experiments on real-world datasets and models that show improvements in expected team utility achieved by our method. We present qualitative analyses to understand how the re-trained model differs from the most accurate AI, and how the improvements in utility change as a function of domain parameters. We conclude with discussing optimization issues, loss-metric mismatch, and implications for optimizing team performance for more complex human-AI teams.

5.2 Problem Description

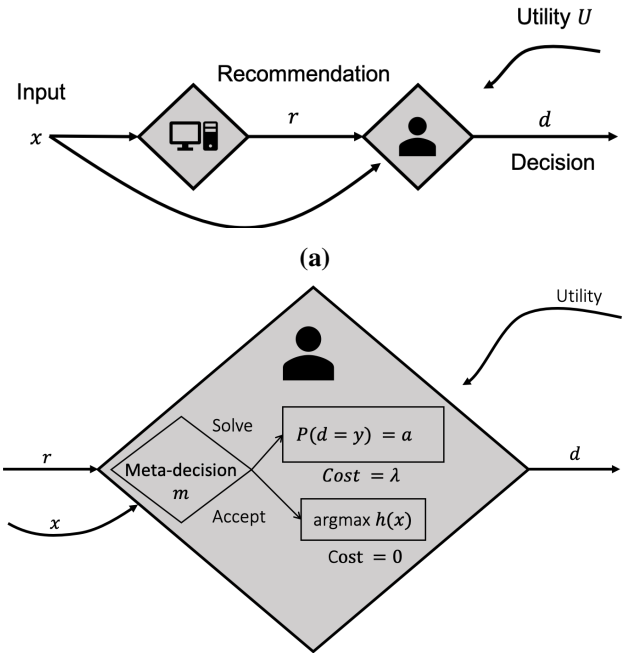


Figure 5.2: (a) AI-advised decision making. (b) To make a decision, the human either accepts or overrides a recommendation. The `Solve` meta-decision is costlier than `Accept`.

We focus on AI-advised decision making scenarios where a classifier h gives recommendations to a human decision maker to help them make decisions (Figure 5.2a). Suppose x is an n-dimensional feature

vector (i.e., $X \subset R^n$) and Y is a finite set of possible decisions. For example, for binary classification $Y = \{+, -\}$. If $h(x)$ denotes the classifier's output (*i.e.*, a probability distribution over the set of possible outcomes $\mathcal{Y}$) the recommendation r is a tuple consisting of the predicted label $\hat{y} = \arg\max h(x)$ and a confidence value $\max h(x)$, *i.e.*, $r \coloneqq (\hat{y}, \max h(x))$. Using this recommendation, the user computes a final decision d. The environment, in response, returns a utility which depends on the quality of the final decision and any cost incurred due to human effort. If the team classifies a sequence of instances, the objective is to maximize the cumulative utility. Before deriving a closed form equation for the objective, we describe the form of the human-AI collaboration we consider along with our assumptions. We study this simple setting as a step to exploring broader opportunities and challenges in team-centric optimization.

1. *User either accepts the recommendation or solves the task themselves:* The human computes the final decision by first making a meta-decision m: `Accept` or `Solve` (Figure 5.2b). In `Accept`, the user passes off the AI recommendation as the final decision. In contrast, in `Solve`, the user ignores the recommendation and computes the final decision themselves. Let m denote the function that maps an input instance and recommendation to a meta-decision in $\mathcal{M} = \{\texttt{Accept}, \texttt{Solve}\}$. Further, U denotes the utility function, which depends on the human meta-decision and final decision d (Figure 5.1). As a result, the optimal classifier h^* would maximize the team's expected utility:

$$h^* = \arg\max_h \mathbb{E}_{x,y}[U(m, d)] \tag{5.1}$$

2. *Mistakes are costly:* A correct decision results in unit reward. An incorrect decision results in a penalty $\beta \geq 1$.
3. *Solving the task is costly:* Since it takes time and effort for the human to perform the task themselves (*e.g.*, cognitive effort), we realistically assume that the `Solve` meta-decision costs more than `Accept`. Further, without loss of generality, we assume λ units of cost to `Solve` and zero cost to `Accept`. Note that even when the cost of `Accept` is non-zero and the reward for a correct decision is different than one, the utility function can still be transformed and simplified to the same form as in Table 5.1 and be optimized in the same way as we describe henceforth.

   Following the above specifications, we obtain the utility function in Figure 5.1. The values in the table originate from subtracting the cost of the action from the reward.

Meta-decision/Decision	Correct	Incorrect
`Accept [ A ]`	1	$-\beta$
`Solve [ S ]`	$1-\lambda$	$-\beta-\lambda$

Table 5.1: Utility as a function of meta-decision and decision.

4. *Human is uniformly accurate across decisions:* Let $a \in [0,1]$ denote the conditional probability that if the user solves the task, they will make the correct decision.

$$P(d = y | m = \texttt{S}) = a \tag{5.2}$$

5. *Human is rational:* The user chooses the meta-decision that results in highest expected utility. Further, the user trusts the classifier's confidence $h(x)[\hat{y}]$ as an accurate indicator of the recommendation's reliability, i.e., true conditional probability of prediction $\hat{y}$ being correct. As a result, the user will choose `Accept` if and only if the expected utility of `Accept` is greater than that of `Solve`.

$$\begin{aligned} \mathbb{E}[U(m = \texttt{A})] &\geq \mathbb{E}[U(m = \texttt{S})] \\ h(x)[\hat{y}] - (1 - h(x)[\hat{y}]) \cdot \beta &\geq a - (1-a) \cdot \beta - \lambda \\ h(x)[\hat{y}] &\geq a - \frac{\lambda}{1+\beta} \end{aligned}$$

Let $c(\beta, \lambda, a)$ denote the minimum value of confidence for which the user's meta-decision is `Accept`.

$$c(\beta, \lambda, a) = a - \frac{\lambda}{1+\beta} \tag{5.3}$$

This implies the human will follow the following threshold-based policy to make meta-decisions:

$$P(m = \texttt{A}) = \begin{cases} 1 & \text{if } h(x)[\hat{y}] \geq c(\beta, \lambda, a) \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

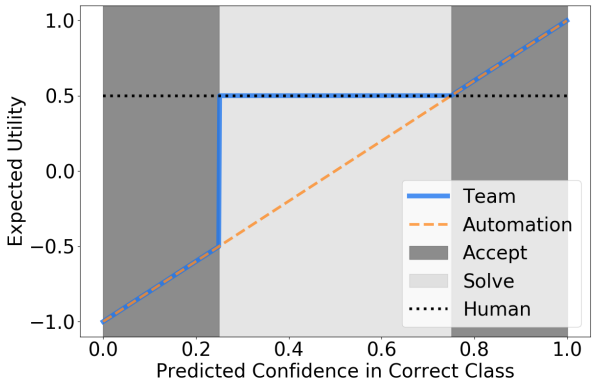


Figure 5.3: Visualization of expected utility when $\lambda = 0.5, \beta = 1$, and $a = 1$ (*i.e.*, the human is perfectly accurate but it costs them half a unit of utility to solve the task). In the `Accept` region, expected utility of the team is equal to expected utility of the automation, while in the `Solve` region it equals to the human utility. The negative team utility in the left-most region results from over-confident but incorrect recommendations to the human.

5.2.1 Expected Team Utility

We now derive the equation for expected utility of recommendations for the teamwork that we described above. Let ψ denote the expected team utility on a given example.

$$\begin{aligned}
\psi(x, y) &= \mathbb{E}[U(m, d)] \\
&= P(m = \texttt{A}) \cdot \Big[P(d = y | m = \texttt{A}) \cdot 1 \\
&\quad + P(d \neq y | m = \texttt{A}) \cdot (-\beta) \Big] \\
&\quad + P(m = \texttt{S}) \cdot \Big[P(d = y | m = \texttt{S}) \cdot (1 - \lambda) \\
&\quad + P(d \neq y | m = \texttt{S}) \cdot (-\beta - \lambda) \Big]
\end{aligned}$$

Since upon `Accept`, the human returns the classifier's recommendation, the probability that the final decision is correct is the same as the classifier's predicted probability of the correct decision, *i.e.*, $P(d =$

$y|m = \texttt{A}) = h(x)[y]$. Substituting this and Equation 5.2 we obtain:

$$\begin{aligned}\psi(x,y) &= P(m=\texttt{A})\cdot\Big[(1+\beta)\cdot h(x)[y]-\beta\Big] \\ &\quad + P(m=\texttt{S})\cdot\Big[(1+\beta)\cdot a-\beta-\lambda\Big] \\ &= P(m=\texttt{A})\cdot\Big[(1+\beta)\cdot(h(x)[y]-a)+\lambda\Big] \\ &\quad + \underbrace{\Big[(1+\beta)\cdot a-\beta-\lambda\Big]}_{\text{constant}}\end{aligned}$$

Substituting human policy (Equation 5.4) we obtain:

$$\psi(x,y) = \begin{cases}(1+\beta)\cdot h(x)[y]-\beta & \text{if } h(x)[\hat{y}] \geq c(\beta,\lambda,a) \\ (1+\beta)\cdot a-\beta-\lambda & \text{otherwise}\end{cases} \tag{5.5}$$

Figure 5.3 visualizes the expected team utility of the classifier predictions as a function of confidence in the true label. We convert expected utility into a loss function by negating it, *i.e.*, $-\psi(x,y)$.

5.3 Experiments

Experiments in this section address the following questions:

RQ1 Can we train a classifier with higher utility than the most accurate classifier?

RQ2 How does the new model qualitatively differ from the most accurate model?

RQ3 How do the properties of the task affect improvements in utility (*e.g.*, human skill and cost of mistake)?

Datasets We experimented with two synthetic datasets and four real-world binary classification datasets: German credit lending dataset [Hofmann, 1994], FICO credit risk assessment [1], recidivism prediction [2], and MIMIC-3 mortality prediction [Harutyunyan *et al.*, 2017]. The real datasets are drawn from high-stakes domains where machine learning has already been deployed or has been discussed being employed to assist

[1] https://community.fico.com/s/explainable-machine-learning-challenge

[2] https://github.com/propublica/compas-analysis

Dataset	#Features	Size	Frac. Pos.
Scenario1	2	10000	0.43
Moons	2	10000	0.50
German	24	1000	0.30
Fico	39	9861	0.52
Recidivism	13	6172	0.46
MIMIC	714	21139	0.13

Table 5.2: Number of features and size of binary classification datasets used for experiments. The original Fico dataset contains 23 features but 39 after preprocessing categorical features into binary features.

human decision makers. On the synthetic datasets, Scenario1 dataset refers to a dataset we created by sampling 10000 points from the data distribution similar to Figure 5.1. Moons refers to the classic two moons non-linear classification problem.[3]

Model Training We experimented with two types of models: logistic regression and multi-layered perceptron (two hidden layers with 50 and 10 units). For each task (defined by a choice of task parameters, dataset, model, and loss) we optimized the loss using the Adam optimizer and also used standard, well-known training practices such as regularization, check-pointing the model best validation performance, and learning rate schedulers. We selected the best hyperparameters using five-fold cross validation, including values for the learning rate, batch size, patience, decay factor of the learning rate scheduler, and the L2 regularization weight. (Range of parameters detailed in the Appendix).

In initial experiments to optimize team utility, we observed that the classifier's loss (in this case, negative of expected utility) remained constant over the optimization process. This happened because, in practice, random initializations resulted in classifiers that were uncertain on most of the data distributions considered. By definition, the expected utility is flat and constant in regions of uncertainty (see Figure 5.3). Thus, the gradient was zero and uninformative over these ranges. To overcome this issue, we initialized the classifiers with the (already converged) most accurate classifier.

Metrics: Empirical and Expected Utility We evaluated our systems on two metrics of team utility: expected team utility (Equation 5.5) and empirical team utility, which draws discrete rewards from the pay-off described in Table 5.1. A key difference between expected and empirical utilities is that the former incentivizes systems that output a calibrated belief, *i.e.*, in the `Accept` region it assigns a score proportional to the system's confidence in the correct class (Figure 5.3). Empirical utility, in contrast, does not differentiate

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html`

		Logloss			Expected Utility Loss		
Classifier	**Dataset**	**Accuracy**	**Expected Util.**	**Emp. Util.**	**Δ Accuracy**	**Δ Expected Util.**	**Δ Emp. Util.**
Linear	Fico	0.729	0.487	0.575	-0.247	**0.013**	-0.075
	German	0.754	0.529	0.594	-0.015	0	-0.019
	MIMIC	0.881	0.694	0.8	-0.004	**0.066**	-0.035
	Moons	0.885	0.687	0.79	-0.02	**0.079**	-0.006
	recidivism	0.669	0.485	0.52	-0.17	**0.015**	-0.02
	Scenario1	0.858	0.524	0.593	-0.165	**0.102**	**0.061**
MLP	Fico	0.725	0.472	0.574	-0.244	**0.028**	-0.074
	German	0.752	0.53	0.618	-0.036	-0.027	-0.056
	MIMIC	0.881	0.719	0.799	-0.001	**0.049**	-0.029
	Moons	1	0.944	0.989	0	**0.049**	**0.006**
	Recidivism	0.674	0.467	0.521	-0.168	**0.033**	-0.021
	Scenario1	1	0.826	0.854	-0.1	**0.08**	**0.057**

Table 5.3: Comparison of accuracy, expected and empirical team utilities of classifiers optimized for log-loss (with a checkpoint on accuracy) and expected team utility (with a checkpoint on expected utility) using Adam for $\lambda = 0.5$, $a = 1.0$, $\beta = 1.0$. Observations averaged over 50 train/test splits. Δ indicates difference with respect to log-loss. Classifier trained to optimize expected team utility achieves higher expected utility at the cost of automation accuracy. However, we notice a mismatch between expected and empirical utilities– empirical utility *decreased* even though expected utility increased.

between a low- and a high-confidence recommendation in the `Accept` region as long as they are both correct (or both are incorrect).

Each metric offers different advantages. Maximizing empirical utility aligns well with existing non-probabilistic discrete metrics for evaluating ML classifiers (such as, accuracy, F1-score, and AUPRC), which exclusively focus on the discriminative power of models. In contrast, maximizing expected utility is critical for decision making under uncertainty, *i.e.*, when the outcome of decisions may be probabilistic and thus a rational agent should maximize for its decision's expected utility. In fact, the primary result of utility theory, the accepted, normative theory of action under uncertainty, is that ideal decisions are those that maximize expected utility [Morgenstern and Von Neumann, 1953]. Maximizing expected utility requires the use of calibrated probabilities, which is an aspect that is not reflected in empirical utility. Moreover, expected utility optimization is useful in cases when empirical evaluation of metrics is not feasible due to delayed reward in the real world or when the definition of empirical ground truth labels is soft and non-discrete.

Dataset	Expected Util LL	Emp. Util LL	ΔExpected Util (A)	Δ Emp. Util (B)	Δ^* Emp. Util (C)
Fico-2d	0.475	0.511	0.025	-0.011	-0.004
German-2d	0.514	0.6	0.076	-0.004	-0.016
MIMIC-2d	0.641	0.772	0.121	-0.009	0.005
Moons	0.767	0.813	0.016	-0.006	0.034
Recidivism-2d	0.478	0.518	0.022	-0.017	0.007
Scenario1	0.707	0.715	0.045	0.069	0.068

Table 5.4: Test performance of linear classifier that optimizes log-loss and team utility using brute-force optimization on two-dimensional domains. While we observe consistent improvements in the team's expected utility (column marked A) across domains, improvements in expected utility did not translate to improvements in *empirical utility* (values in column marked B are negative), indicating a mismatch between the expected and empirical metrics of team utilities. At the same time, exhaustive search shows existence of linear classifiers with higher empirical utility (column marked C). Values were averaged over five seeds. Observations in column C on Fico-2d and German-2d were negative on test set due to over-fitting.

5.3.1 Results

RQ1: Table 5.3 shows that the new classifier can improve expected team utility over log-loss. These improvements are often achieved by sacrificing the classifier's individual accuracy. For example, on Scenario1 the new linear classifier improved expected utility from 0.524 to 0.606 even though it was less accurate.

When we considered *empirical* utility, our method did not always result in improvements. For example, for the linear classifier, while on Scenario1, the empirical utility increased from 0.593 to 0.654, but on MIMIC it decreased from 0.8 to 0.765. Ideally, one would expect that an increase in expected team utility would be accompanied with proportional increase in empirical team utility. However, as Table 5.3 shows, this was often not the case.

While this mismatch between empirical and expected utilities seems counterintuitive, it is a well know problem; [Huang *et al.*, 2019] noticed a mismatch between various common ML evaluation metrics, such as log-loss, zero-one loss, and AUPRC. However, we still considered the possibility that, in practice, the mismatch perhaps resulted from stochastic optimization getting stuck in local minimas, and that a better optimization procedure would alleviate this mismatch. To pursue this conjecture, we developed two-dimensional versions of our dataset (by selecting two top most informative features) and trained linear classifiers using exhaustive search, which by definition cannot get stuck in local minimas. We again found a persistence of the mismatch between expected and empirical utilities (Table 5.4). In addition, we also noticed that there exist classifiers with higher empirical utility if the exhaustive search maximizes directly for empirical utility

(column C in Table 5.4), which further demonstrates the existence of the mismatch. [4]

These results provide evidence that the challenge with achieving comparable increases in empirical utility to those in expected utility is not only due to optimization issues (*e.g.*, local minimas and plateaus due to flatness of the expected utility curve in the `Solve` region). There exists a fundamental ML challenge of loss-metric mismatch, which was prominent in our setup. In the rest of the section, we present further analyses of improvements in the normative decision making metric of expected utility, which as described earlier, is useful in decision-making under uncertainty.

RQ2: While the metrics in Table 5.3 (change in accuracy and utility) provide a global understanding of the classifier behavior, here we attempt to understand *how* these improvements were achieved and whether the behavior of the new models is consistent with the original intuition. Figure 5.4 displays the difference in behavior (averaged over 50 seeds) between the classifiers produced by log-loss and the one that maximizes team utility on the Scenario1 and MIMIC dataset. Specifically, as shown in Figure 5.4, we visualize and compare the following behaviors of the two classifiers:

V1. Calibration using *reliability curves*, which compare system confidence and its true accuracy. A perfectly calibrated system, for example, will be 80% accurate on regions that is 80% confident. However, in practice, systems may be over- or under-confident.

V2. Distributions of confidence in predictions. For example, in Figure 5.4, the new classifier makes more high-confidence predictions than the most accurate classifier.

V3. Density of system accuracy as function of confidence in true label. Thus, the area under this curve indicates the system's total accuracy. Note that, for our setup, the area under the curve in the `Accept` region is more crucial.

V4. Density of expected utility as a function of confidence.

The classifier optimized for the team's expected utility results in dramatically different predictions than the classifier trained using log-loss: The new classifier sacrifices accuracy on the uncertain examples (`Solve` region) to make higher numbers of high-confidence predictions (`Accept` region). Most importantly, it also

[4] Note that directly optimizing for empirical utility is not effective via stochastic optimization.

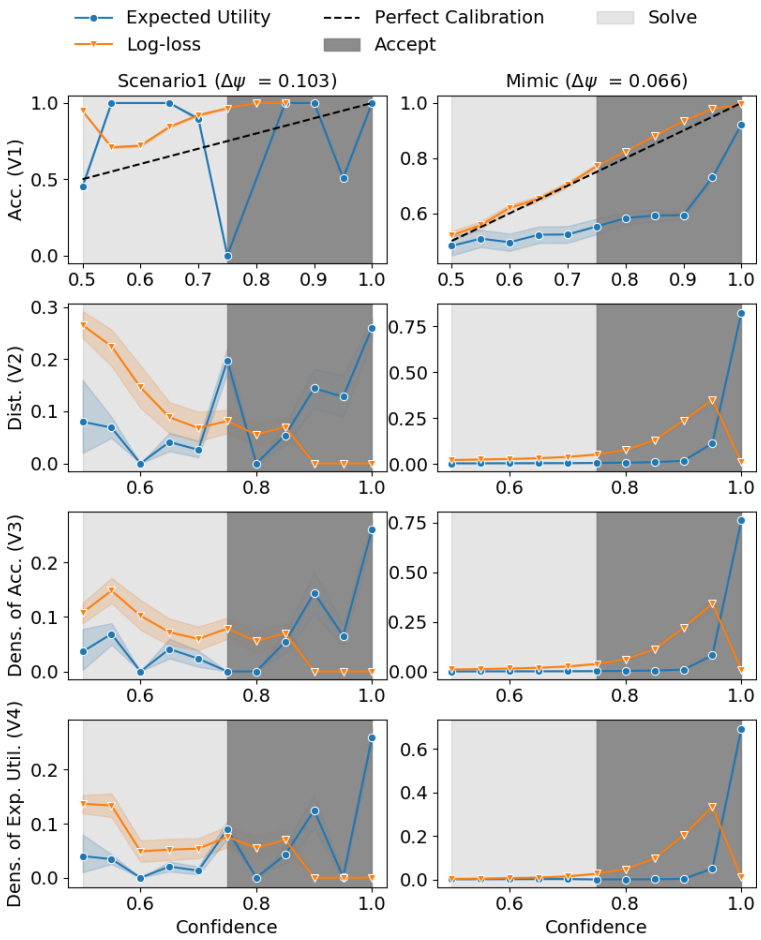


Figure 5.4: Behavior of linear classifiers that optimize log-loss and expected team utility on the Scenario1 and MIMIC datasets (observations averaged over 50 runs). The latter makes fewer predictions in the `Solve` region and also sacrifices accuracy in that region to increase it in `Accept`. We observed a behavior similar for the MLP model on all datasets (ommitted due to space constraints).

increases the density of system accuracy in the `Accept` region, which is where the system accuracy matters and contributes to team utility. Figure 5.4 illustrates the same behavior on the MIMIC in-hospital mortality prediction dataset.

An interesting exception was Fico, where the system learned to always be uncertain. This may make sense for the Fico domain because, as shown in Table 5.3, even though the most accurate linear classifier is 73% accurate on Fico, it achieves an expected team utility of 0.487. This is less than the expected utility achieved if humans solved the task alone. Hence, the more accurate classifier leads to lower expected team utility. We observed a similar behavior on recidivism prediction where the linear classifier led to team per-

dataset	a=0.8	a=0.9	a=1
Fico	0.257 (0.133)	0.337 (0.071)	0.487 (0.013)
German	0.397 (0.046)	0.444 (0.035)	0.529 (0)
MIMIC	0.625 (0.127)	0.644 (0.111)	0.694 (0.066)
Moons	0.582 (0.162)	0.616 (0.139)	0.687 (0.079)
Recidivism	0.155 (0.073)	0.292 (0)	0.485 (0.015)
Scenario1	0.224 (0.324)	0.364 (0.248)	0.524 (0.102)

Table 5.5: Expected utility of log-loss and improvements for linear classifiers (Δ Expected Util. shown in brackets) with varying human accuracy (a) and ($\lambda = 0.5$ and $\beta = 1.0$). Results averaged over 50 random seeds. Improvements in expected utility are higher when the human is less accurate.

dataset	β=1	β=3	β=5
Fico	0.487 (0.013)	0.474 (0.026)	0.481 (0.019)
German	0.529 (0)	0.427 (0.057)	0.367 (0.118)
MIMIC	0.694 (0.066)	0.58 (0.008)	0.543 (0)
Moons	0.687 (0.079)	0.637 (0.065)	0.594 (0.085)
Recidivism	0.485 (0.015)	0.495 (0.004)	0.498 (0.001)
Scenario1	0.524 (0.102)	0.501 (0.02)	0.5 (0)

Table 5.6: Expected uility of log-loss and improvements for linear classifiers (*i.e.*, Δ Expected Util., shown in brackets) with varying cost of mistakes (β) and ($\lambda = 0.5, a = 1.0$). Results averaged over 50 random seeds. On most datasets, gains diminish as the cost of mistakes increases.

formance lower than that associated with people making decisions unaided, even though the classifer had a 67.4% accuracy (Table 5.3). These cases illustrate timely concerns and questions of when and if an AI should be deployed to assist human decision-making, which we further discuss in the ethical statement.

RQ3: Since properties such as the accuracy of users and penalty of mistakes may be task-dependant (*e.g.*, an incorrect diagnosis may be costlier than incorrect loan approval), we varied human accuracy a and mistake penalty β to study the sensitivity in improvements in team utility to a wider range of these task parameters.

Table 5.5 shows improvements in expected utility as we vary human accuracy from 80% to 100% while keeping λ and β constant to 0.5 and 1, respectively. These three values of a result in three new values of optimal threshold $c(\beta, \lambda, a)$: 0.55, 0.65, and 0.75, thus gradually expanding the confidence region in which the user is likely to `Solve` because they themselves are more accurate. We notice higher improvements in expected utility from deploying a system when humans are less accurate, *e.g.*, Table 5.5 shows that, on Fico, improvement in expected utility is 0.133 when the human is 80% accurate whereas it is 0.013 when they are perfect. One explanation for this behavior is that when humans are less accurate there is greater value from system recommendations, which widens the `Accept` region and increases the scope where the AI can

provide value to the team.

Similarly, Table 5.6 shows the impact of varying cost of mistakes β on improvements. The three values of β increase the `Accept` threshold gradually from 0.75 to 0.91, and therefore shrink the size of the `Accept` region. Hence, we start observing smaller gains when the cost of mistake is high, *e.g.*, on the MIMIC dataset there are no gains, although the trend is also subject to the shape of expected utility and how easy it is to optimize it. In overall, the trend emphasizes once again that for extremely high-stake decisions, automation or AI recommendation may not always provide value.

5.4 Discussion and Future Work

Implications for complex human-AI teams While we investigated a simplified human-AI teamwork (as defined in Section 5.2), our setup allows extensions to more complex team and users. For example, one can relax our assumption that users are rational by modifying the human-policy in Equation 5.4, so that when the prediction confidence is greater than the threshold, the user chooses `Accept` with probability $p < 1$, instead of 1.0. Here, $1 - p$ denotes the probability of the user being irrational — assessed from historical data, if available. Similarly, in more complex situations users may make `Accept` and `Solve` decisions using a richer, more complex mental model instead of relying on just model confidence. Such scenarios are common in cases where the system confidence is an unreliable indicator of performance (*e.g.*, due to poor calibration), and, as a result, the user develops an understanding of system failures in terms of domain features. For example, Tesla drivers may learn to override the Autopilot considering such features as road, sun glare, and weather conditions. We can reduce the case where users have a complex mental model to the policy that we studied. Specifically, we can construct a new loss function in terms of human utility (in this case, constant) when the prediction belongs to the `Solve` region (as described by the user's mental model) and automation utility otherwise.

While the above extensions to our model are a start, even they may present challenges– If we cannot optimize empirical utility for our simplified case, it may be harder to optimize performance in the extensions as they contain more complex user behavior and the resultant loss surface is likely to be more complex, containing combinations of plateaus and local optima. In addition to these extensions, future work should also consider more general uses of AI recommendations in support of human decision making. For example,

we need to consider common uses that are not constrained to policies where a user either accepts an AI recommendation or relies completely on their own reasoning. It is natural to expect that users in human-AI teams will employ their own *evidential reasoning* to fuse AI inferences (and associated confidences if shared) with their own assessments. Furthermore, user's mental models may not be static; instead, they may change with time as users learn more about the AI. Mental models may also vary across users, as different people might have different propensity to accept machine recommendations.

Human-subject evaluations are an important next step to understand how factors such as biases, variations in user expertise, irrational behavior come to play in practice. Is our simple model of human behavior sufficient for our approach to yield gains in practice? We view our work as a fundamental first step showing the potential impact of a human-centered model and motivating additional work including real-world studies with human subjects. Over time, we hope to learn and incorporate rich (and individualized) models of human behavior into our framework and test them in real-world human-AI teams.

Empirical utility and auxiliary loss functions While optimizing for teamwork, we faced two fundamental optimization challenges. First, we observed an an inherent mismatch between empirical and expected utility as shown in the exhaustive experiments for two dimensional data, which hindered optimization on the empirical metric, which is often a central consideration in ML. Second, current optimization techniques were not always effective and in fact sometimes they did not change model behavior because the optimization approaches got stuck due to zero gradients and local minimas `Solve` region.

To support empirical utility maximization, in our initial analysis, we also experimented with an auxiliary loss function, as shown in Figure 5.5. However, in our experiments this loss function did not lead always lead to significant gains in empirical utility and when it did it only lead to marginal improvements. Based on these theoretical and practical challenges, we invite future work on machine learning optimization and human-AI collaboration to develop new optimization techniques that work well beyond log-loss, robustly over a more general set of loss of functions that can capture team utility.

Mental models and explainable AI To increase team performance we focused on adapting one AI property to user mental models– the AI should be more accurate on instances when the user is more likely to trust model recommendations. Similarly, future work should study whether other aspects of human-AI collabo-

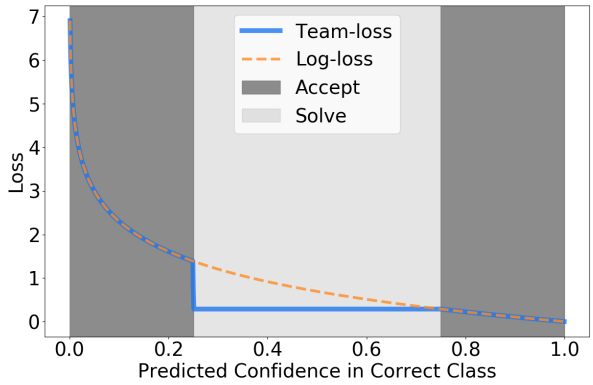


Figure 5.5: An example of an auxiliary loss function Team-loss defined as $= log(\psi(x, y) + K)$, which is equal to Log-loss in `Accept` region and constant otherwise. Here, K is a positive constant we added so that the logarithmic is valid.

ration can be improved by considering user mental models. For instance, mental models could help inform the nature of explanation given to the users. Such as, when users already trust the model, the system may be better off offering concise explanations. In contrast, when the users are likely to distrust the model, they system should be ready to offer detailed explanations/arguments supporting its prediction [Bansal *et al.*, 2021b]. Eventually, work on explainable AI aims to improves human-AI collaboration by providing a layer of communication between users and AI systems. Since explainability does not guarantee improvements in collaboration [Bansal *et al.*, 2021b], there is need to bring collaboration as an objective to every step of system development, starting from the training objective. We hope that this work paves the work for future directions towards uncovering how to develop AI systems for collaboration.

5.5 Related Work

Our approach is closely related to *maximum-margin classifiers*, such as an SVM optimized with the hinge loss [Burges, 1998], where a larger soft margin can be used to make high-confidence and accurate predictions. However, unlike our approach, it is not possible to directly plug the domain's payoff matrix (*e.g.*, in Table 5.1) into such a model. Furthermore, the SVM's output and margin do not have an immediate probabilistic interpretation, which is crucial for our problem setting. One possible (though computationally intensive) solution direction is to convert margin into probabilities, *e.g.*, using post-hoc calibration (*e.g.*,

Platt scaling [Platt, 1999]), and use cross-validation for selecting margin parameters to optimize team utility. While it is still an open question whether such an approach would be effective for SVM classifiers, in this work we focused our attention on gradient-based optimization.

Another related problem is *cost-sensitive learning*, where different mistakes incur different penalties; for example, false-negatives may be costlier than false-positives [Zadrozny *et al.*, 2003; Bach *et al.*, 2006]. A common solution here is up-weighting the inputs where the mistakes are costlier. Also relevant is work on *importance-based learning* where re-weighting helps learn from imbalanced data or speed-up training. However, in our setup, re-weighting the inputs makes less sense— the weights would depend on the classifier's output, which has not been trained yet. An iterative approach may be possible, but our initial analysis showed this approach is prone to oscillations. We leave exploring this avenue for future work.

A fundamental line of work that renders AI predictions more actionable (for humans) and better suitable for teaming is *confidence-calibration*, for example, using Bayesian models [Ghahramani, 2015; Beach, 1975; Gal and Ghahramani, 2016] or via *post-hoc* calibration [Platt, 1999; Zadrozny and Elkan, 2001; Guo *et al.*, 2017; Niculescu-Mizil and Caruana, 2005]. A key difference between these methods and our approach is that *team-loss re-trains* the model to improve on inputs on which users are more likely to rely on the AI predictions. The same contrast distinguishes our approach from *outlier detection* techniques [Hendrycks *et al.*, 2018; Lee *et al.*, 2017; Hodge and Austin, 2004].

Closely related is research on *learning to defer* [Madras *et al.*, 2018; Mozannar and Sontag, 2020] and *learning to complement* [Wilder *et al.*, 2020 07], where the classifier can abstain and defer/query the task to the user, while accounting for costs and benefits of intervention. While the "Solve" meta-decision in our framework corresponds to the defer action, our work differs from these works in two important ways. First, the defer action in prior work is system-initiated whereas in our case it is user-initiated and based on their mental model. Second, learning to defer does not preclude our methods, since users may create mental models even when the system does not defer and so the team may still benefit from training a model that accounts for user's mental model.

Other recent work that adjusts model behavior to accommodate collaboration includes *backward-compatibility for AI* [Bansal *et al.*, 2019b], where the model considers user interactions with a previous version of the system to preserve trust across updates. Recent user studies showed that when users develop mental mod-

els of AI system, properties besides accuracy are also desirable, such as *parsimonious* and *deterministic* error boundaries [Bansal *et al.*, 2019a]. Our approach is a first step towards implementing these desiderata within ML optimization itself. Other approaches regularize or constrain model optimization for other human-centered requirements such as local- or global-interpretability [Wu *et al.*, 2019 06] or fairness [Jung *et al.*, 2019; Zafar *et al.*, 2017].

5.6 Conclusions

We studied opportunity to train classifiers that optimize human-AI team performance. We showed the value of optimizing the expected utility of decision making of human-AI teams in contrast to traditional model optimization focusing solely on automation accuracy. Investigations and visualizations of classifier behavior before and after proposed optimization show that the methods can be harnessed to fundamentally change model behavior and improve the team utility. Changes in model behavior include (i) sacrificing model accuracy in low confidence regions for more accurate high-confidence predictions and (ii) increasing accuracy and number of high-confidence predictions. Such behaviors were observed in both synthetic and real-world datasets where AI is known to be employed as support for human decision makers, and across various domain parameters such as human accuracy and cost of mistake.

Chapter 6

Conclusions and Future Work

In this thesis, we focused on AI-assisted decision making and argued that for such tasks we should develop and deploy human-centered AI that is designed to augment user performance. Specifically, we discussed three maxims for developing such AI: first, we discussed how we need to help users determine when to trust the AI's recommendations and showed how despite massive interests in XAI there is actually sparse evidence that explanations help improve human AI teams. We argued that in order to be truly effective, these systems need to focus on achieving complementary performance and helping users achieve appropriate reliance on AI. Second, we showed how user ability to create a mental model of AI's trustworthiness is also impacted by the complexity and stochasticity of its error boundary. Specifically, we observed that its easier to create mental model of AI when its error boundary is simple and deterministic.

Third, in order to preserve user-system trust, we focused on AI updates and showed how even if updates increase the model's accuracy they can decrease human-AI team performance, by introducing behaviors that are at odds with user's expectations, *e.g.*, by erring on examples on which the AI was previously correct. But we also showed how we can mitigate this issue by training backwards compatible updates to gracefully update the performance of these models.

Finally, we looked at the problem of directly optimizing AI to improve human-AI team performance by accommodating user's mental model in the training process. And showed that even for a simple team, where the users mental model of AI trustworthiness was based on its confidence in its predictions, a lower accuracy model can achieve higher team performance that the most accuracy AI. However, optimizing the

team-centric loss functions results in new optimization challenges.

While AI-assisted decision making (with one human and one AI) and the maxims we described are a good starting point for developing a precise understanding of HCAI, they still represent a small part of the space of human-AI interaction. Thus, the next few years will be an exciting opportunity to expand the frontiers of research on human-centered AI: We will not just deepen our understanding of the existing properties for developing HCAI, but also identify new ones, *e.g.*, enabling explanatory dialogs with AI, enabling user control, etc. We will expand the breadth of this frontier and apply these principles to interactions with more complex ML model types (large language models, Github copilot), human-AI teams (*e.g.*, those involving multiple users or AI, mixed-initiative interactions), and domains (*e.g.*, planning, robotics, etc). Even with these expansions in the breadth and depth of the research frontier, like this thesis, we will ultimately need to guide all this research with the goal to augment people's abilities and ground solution direction in user needs and evaluation.

6.1 Future 1: Explanatory Dialogs

Most current research on XAI focuses on generating the best explanation for a given representation, *e.g.*, finding the most import features [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017] or prototypes for a model's prediction [Koh and Liang, 2017]. As shown in this thesis, empirically, there is sparse evidence that such current explanations help users decide when to trust AI recommendations. Hence, a key open problem is whether we can develop better explanations that promote appropriate reliance? In fact research from psychology tells us that explanations are a social process– they are a conversation between the explainer and the explainee [Hilton, 1990; Miller *et al.*, 2017; Weld and Bansal, 2019 05]. Hence, an important future direction is to enable *explanatory dialogs*, where the AI 1) supports multiple different ways to explain, 2) gives its users control to ask follow-up questions in order to drill down into the system's reasoning, and 3) conditions its responses based on what the users already likely know.

Figure 6.1 shows a schematic of an explanatory dialog for one specific task: helping users understand which concepts a deep neural network has learned. Specifically, the features learned by a node of a robust ResNet-50 trained on a bird classification dataset (CUB with 200 classes). While understanding the features learned by an neural network is an important problem in explainable AI, understanding and explaining such

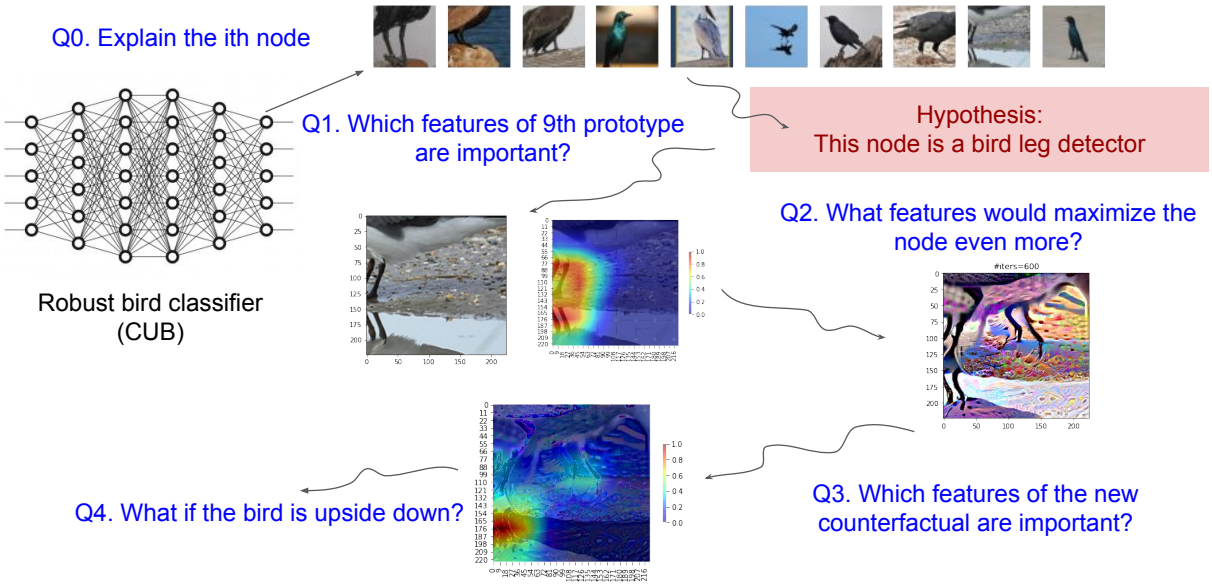


Figure 6.1: An example of an explanatory dialog for understanding a node of a robust NN for bird classification. 0) The user develops an initial hypothesis of the node's behavior by inspecting the training examples that activate it the most. 1) The user refines their understanding by asking the model to explain (using saliency maps) relevance of a prototype shown in previous iteration. 2) The user further verifies whether the node learned to detect bird legs by asking the UI for a counterfactual image that maximizes the node's activation. 3) The user again verifies to explain the relevance of the new prototype. 4) Since the model focused on the reflection of legs, the user asks another question to drilldown and verify whether the node can generally detect upside down legs, thus continuing the dialog. Our initial assessment reveals that model actually fails upon encounters upside-down birds.

robust neural networks is especially important because 1) in the presence of *adversarial examples* these models output more stable predictions than non-robust networks. And, 2) there exist anecdotal understanding in the research community that these models learn more human-interpretable features than their non-robust counterparts.

We plan to evaluate explanatory dialogs by asking the following questions and comparing them to baselines based on static explanations and model confidence: a) Does it help users understand AI behavior (*e.g.*, predict model outputs, detect errors, find error patterns)? For instance, an error patterns could be a slice of the feature space where the model's error rate (false-positive, -negative) is higher than average error rate on the entire test set. In the above figure, the slice was (approximately) described as "Model fails more when birds are hanging upside-down from branches." b) Does it help users improve the AI? For example, does training on more upside-down birds (*e.g.*, through data augmentation or labeling more samples) decrease the model's error rate? c) which explanations are essential for completeness of dialogs and which explanations are most useful (for the above tasks/metrics)? Are there frequent patterns in the questions (*e.g.*, frequently asked sequences) that users ask? Finally, since there may be a cost to dialog – longer interaction may takes more time and effort than static explanations, d) is there a tradeoff between the benefits from the insights and user time/effort?

6.2 Future 2: User Control of XAI and Explanatory Vocabulary Refinement

In the near future, perhaps, explanations may help people better identify incorrect model predictions and reasoning. But how can we allow the users to also give feedback to these models and correct their predictions and/or reasoning? For instance, explaining AI models requires defining an *explanatory vocabulary*. This is the human-understandable representation in terms of which the model explains its reasoning. For example, GAMs use semantic features of the task [Caruana *et al.*, 2015], concept-bottleneck models and TCAV use human-concepts [Kim *et al.*, 2017 11], and LIME (for image classification) used super-pixels [Ribeiro *et al.*, 2016]. In all these instances, the user could fix the system's reasoning by giving feedback in terms of this explanatory vocabulary. For example, by supplying their estimate of the importance of terms and updating the model [Lee *et al.*, 2020].

However, just like the features available to any ML model may be incomplete and different from the

features available to users, its explanatory vocabulary may also be incomplete— the explanatory vocabulary may not contain the terms or concepts that the user considers necessary for correctly reasoning about a given prediction. Thus an important sub-problem to enable user control of XAI models is to allow them to refine and extend a system's explanatory vocabulary.

Our ongoing work investigates the problem of vocabulary refinement in the context of scientific paper recommendation, where a black-box classifier (a linear model trained with SciBERT embeddings [Beltagy *et al.*, 2019]) recommends users (*e.g.*, researchers) papers relevant to a research feed/topic of their interest (*e.g.*, explainable AI or ML fairness). The system uses a post hoc explainer ([Ribeiro *et al.*, 2016]) to explain any prediction in terms of the salience of terms in its explanatory vocabulary that contains frequent unigrams and bigrams. In this case, the user can correct the system's reasoning, *e.g.*, by providing the correct importance of unigrams or bigrams for a prediction and then re-training the model using approaches such as LIMEADE [Lee *et al.*, 2020]. But what if the correct reasoning can't be expressed in terms of these unigrams or bigrams? For example, what if the correct reason a given paper was relevant to user was because it provided a theoretical perspective on explainable AI. The theoretical-ness of a text in this case may be dependent on presence or absence of mathematical symbols, theorems, and proofs. Thus the concept of a theory paper may not be present in the system's explanatory vocabulary and the user cannot make the model sensitive to this new concept. Our current solution direction borrows ideas from interactive machine learning [Amershi *et al.*, 2014] to first learn these concepts using a few shots and then develops algorithms to make the model sensitive to these concepts.

Using this setup, we plan to evaluate a) how much effort does it take to learn an abstract concept from user feedback? b) Can concepts be reused in multiple feeds? c) Does user feedback to refine the explanatory vocabulary improve the accuracy of the recommender?

6.3 Final Thoughts

The challenge of developing human-centered AI and improving human-AI interaction is interdisciplinary in nature: Solving it require marrying ideas from artificial intelligence, machine learning, and human-computer interaction to further our understanding of what kind of AI models we train and how they interface people. And we want to do all of this with the goal to responsibly augment human performance and experiences. The

three maxims presented in this thesis provide a good starting point towards this goal for AI-assisted decision making, but much there remains much to be done to address this challenge for more complex human-AI teams, domains, and AI models.

Bibliography

[Abowd *et al.*, 1999] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. Towards a better understanding of context and context-awareness. In *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing*, HUC '99, page 304–307, Berlin, Heidelberg, 1999. Springer-Verlag.

[Amershi *et al.*, 2014] S. Amershi, M. Cakmak, W. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

[Amershi *et al.*, 2019] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *CHI*, 2019.

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software across the country to predict future criminals and it's biased against blacks., 2016.

[Ba and Caruana, 2014] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *NIPS*, pages 2654–2662, 2014.

[Bach *et al.*, 2006] Francis R Bach, David Heckerman, and Eric Horvitz. Considering cost asymmetry in learning classifiers. *JMLR*, 2006.

[Bansal and Weld, 2018] Gagan Bansal and Daniel S. Weld. A coverage-based utility model for identifying unknown unknowns. In *AAAI*, 2018.

[Bansal *et al.*, 2019a] 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 *HCOMP*, 2019.

[Bansal *et al.*, 2019b] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *AAAI*, 2019.

[Bansal *et al.*, 2021a] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *AAAI*, 2021.

[Bansal *et al.*, 2021b] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of explanations on complementary team performance. In *CHI*, 2021.

[Barlow and Brunk, 1972] Richard E Barlow and Hugh D Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

[Bayati *et al.*, 2014 10] Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M. Mack, George Ruiz, Mark S. Smith, and Eric Horvitz. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLOS ONE*, 9(10):1–9, 2014-10.

[Beach, 1975] Barbara Heinrich Beach. Expert judgment about uncertainty: Bayesian decision making in realistic settings. *Organizational Behavior and Human Performance*, 14(1):10–59, 1975.

[Beltagy *et al.*, 2019] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *EMNLP*, 2019.

[Bhatt *et al.*, 2020] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 648–657, New York, NY, USA, 2020. Association for Computing Machinery.

[Bilgic, 2005] Mustafa Bilgic. Explaining recommendations: Satisfaction vs. promotion, 2005.

[Bosch, 2009] Jan Bosch. From software product lines to software ecosystems. In *SPLC*, pages 111–119, 2009.

[Bousquet and Elisseeff, 2001] Olivier Bousquet and André Elisseeff. Algorithmic stability and generalization performance. In *NIPS*, pages 196–202, 2001.

[Buçinca *et al.*, 2020] 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*, IUI '20, page 454–464, New York, NY, USA, 2020. Association for Computing Machinery.

[Burges, 1998] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[Cai *et al.*, 2019] 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. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[Carroll and Olson, 1988] John M Carroll and Judith Reitman Olson. Mental models in human-computer interaction. In *Handbook of human-computer interaction*, pages 45–65. Elsevier, 1988.

[Caruana *et al.*, 2015] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery.

[Chakraborti and Kambhampati, 2018] Tathagata Chakraborti and Subbarao Kambhampati. Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-ai collaboration. *arXiv preprint arXiv:1801.09854*, 2018.

[Chandrasekaran *et al.*, 2018 10] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make VQA models more predictable to a human? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1042. Association for Computational Linguistics, 2018-10.

[Clark *et al.*, 2018] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 329–340, New York, NY, USA, 2018. Association for Computing Machinery.

[Croskerry, 2009] Pat Croskerry. Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in health sciences education*, 14(1):27–35, 2009.

[DeChurch and Mesmer-Magnus, 2010] Leslie A DeChurch and Jessica R Mesmer-Magnus. The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology*, 95(1):32, 2010.

[Dong and Hayes, 2012] Xiao Dong and Caroline C Hayes. Uncertainty visualizations: Helping decision makers become more aware of uncertainty and its implications. *Journal of Cognitive Engineering and Decision Making*, 6(1):30–56, 2012.

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

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[Dynamo,] Dynamo. *Guidelines for Academic Requesters*.

[Elkan, 2001] Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, volume 17, pages 973–978, 2001.

[Englich *et al.*, 2006] Birte Englich, Thomas Mussweiler, and Fritz Strack. Playing dice with criminal sentences: The influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin*, 32(2):188–200, 2006.

[Feng and Boyd-Graber, 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, IUI '19, page 229–239, New York, NY, USA, 2019. Association for Computing Machinery.

[Fernandes *et al.*, 2018] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016.

[Gaur *et al.*, 2016] 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 (W4A)*, page 23. ACM, 2016.

[GDPR, 2020] GDPR. Art. 22 gdpr, automated individual decision-making, including profiling. `https://gdpr-info.eu/art-22-gdpr/`, 2020. [Online; accessed 14-January-2020].

[Gero *et al.*, 2020] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[Ghahramani, 2015] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452, 2015.

[Gillies *et al.*, 2016] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3558–3565. ACM, 2016.

[Gkatzia *et al.*, 2016 08] Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. Natural language generation enhances human decision-making with uncertain information. In *Proceedings of the 54th Annual Meeting*

of the Association for Computational Linguistics (Volume 2: Short Papers), pages 264–268. Association for Computational Linguistics, 2016-08.

[Glassman *et al.*, 2015] Elena L Glassman, Jeremy Scott, Rishabh Singh, Philip J Guo, and Robert C Miller. Overcode: Visualizing variation in student solutions to programming problems at scale. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(2):1–35, 2015.

[Goodfellow *et al.*, 2013] Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

[Gownder *et al.*, 2017 06] J. Gownder, C. Voce, L. Koetzle, C. LeClair, B. Purcell, S. Sridharan, C. Garberg, and D. Lynch. Techradar: Automation technologies, robotics, and AI in the workforce. Technical report, Forrester Research, 2017-06.

[Green and Chen, 2019] 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):1–24, 2019.

[Grosz and Kraus, 1999] Barbara J Grosz and Sarit Kraus. The evolution of sharedplans. In *Foundations of rational agency*. Springer, 1999.

[Grosz, 1996] Barbara J Grosz. Collaborative systems (AAAI-94 presidential address). *AI magazine*, 17(2):67, 1996.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org, 2017.

[Handmer and Proudley, 2007] John Handmer and Beth Proudley. Communicating uncertainty via probabilities: The case of weather forecasts. *Environmental Hazards*, 7(2):79–87, 2007.

[Harutyunyan *et al.*, 2017] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

[Hase and Bansal, 2020 07] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552. Association for Computational Linguistics, 2020-07.

[Hayashi and Wakabayashi, 2017] 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, page 195–198, New York, NY, USA, 2017. Association for Computing Machinery.

[He and McAuley, 2016] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 507–517, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee.

[Hendrycks and Gimpel, 2017] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017*, pages 1–12. OpenReview.net, 2017.

[Hendrycks *et al.*, 2018] Dan Hendrycks, Mantas Mazeika, and Thomas G Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[Hilton, 1990] D. Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65, 1990.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Ho *et al.*, 2015] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 419–429, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[Hodge and Austin, 2004] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[Hoff and Bashir, 2015] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434, 2015.

[Hofmann, 1994] Hans Hofmann. Statlog (German Credit Data) Data Set. `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`, 1994.

[Horvitz and Paek, 2007] Eric Horvitz and Tim Paek. Complementary computing: policies for transferring callers from dialog systems to human receptionists. *User Modeling and User-Adapted Interaction*, 17(1-2):159–182, 2007.

[Horvitz *et al.*, 1986] EJ Horvitz, DE Heckerman, BN Nathwani, and LM Fagan. The use of a heuristic problem-solving hierarchy to facilitate the explanation of hypothesis-directed reasoning. In *Proceedings of Medinfo, Washington, DC*, pages 27–31, 1986.

[Horvitz, 1999] Eric Horvitz. Principles of mixed-initiative user interfaces. In *CHI*, pages 159–166. ACM, 1999.

[Huang *et al.*, 2019] Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Angel Bautista, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. Addressing the loss-metric mismatch with adaptive loss alignment. *arXiv preprint arXiv:1905.05895*, 2019.

[Jhaver *et al.*, 2019] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Does transparency in moderation really matter? user behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27, 2019.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3, 2016.

[Joslyn and LeClerc, 2013] Susan Joslyn and Jared LeClerc. Decisions with uncertainty: The glass half full. *Current Directions in Psychological Science*, 22(4):308–315, 2013.

[Jung *et al.*, 2019] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*, 2019.

[Kamar *et al.*, 2012] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, pages 467–474, 2012.

[Kamar, 2016] Ece Kamar. Directions in hybrid intelligence: Complementing AI systems with human intelligence. In *IJCAI*, 2016.

[Kaur *et al.*, 2019] Harmanpreet Kaur, Alex Williams, and Walter S. Lasecki. Building shared mental models between humans and ai for effective collaboration. *ACM Conference on Human Factors in Computing Systems (CHI) Workshops*, 2019.

[Kaur *et al.*, 2020] 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*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[Kim *et al.*, 2017 11] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *ArXiv e-prints*, 2017-11.

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, page 201611835, 2017.

[Kocielnik *et al.*, 2019] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will you accept an imperfect ai?: Exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 411. ACM, 2019.

[Koehler, 1991] Derek J Koehler. Explanation, imagination, and confidence in judgment. *Psychological bulletin*, 110(3):499, 1991.

[Koh and Liang, 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*, ICML'17, page 1885–1894. JMLR.org, 2017.

[Kulesza *et al.*, 2012] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *CHI*, pages 1–10. ACM, 2012.

[Kunkel *et al.*, 2019] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[Lage *et al.*, 2018] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human-in-the-loop interpretability prior. *arXiv preprint arXiv:1805.11571*, 2018.

[Lai and Tan, 2019] 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*, FAT* '19, page 29–38, New York, NY, USA, 2019. Association for Computing Machinery.

[Lai *et al.*, 2020] 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*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[Lakkaraju *et al.*, 2016] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *KDD*, pages 1675–1684. ACM, 2016.

[Lakkaraju *et al.*, 2017] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models, 2017.

[Lasecki *et al.*, 2012a] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, pages 23–34. ACM, 2012.

[Lasecki *et al.*, 2012b] Walter S Lasecki, Jeffrey P Bigham, James F Allen, and George Ferguson. Real-time collaborative planning with the crowd. In *"HCOMP"*, 2012.

[Lee and See, 2004] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[Lee *et al.*, 2017] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

[Lee *et al.*, 2020] Benjamin Charles Germain Lee, Doug Downey, Kyle Lo, and Daniel S Weld. Limeade: A general framework for explanation-based human tuning of opaque machine learners. *arXiv preprint arXiv:2003.04315*, 2020.

[Levine, 2014] Timothy R Levine. Truth-default theory (tdt) a theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33(4):378–392, 2014.

[Lim *et al.*, 2009] B. Lim, A. Dey, and D. Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2119–2128, New York, NY, USA, 2009. ACM.

[Lin *et al.*, 2017] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*, pages 1–15. OpenReview.net, 2017.

[Lipton, 2018 06] Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018-06.

[Liu *et al.*, 2016 06] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H. Lin, Xiao Ling, and Daniel S. Weld. Effective crowd annotation for relation extraction. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 897–906. Association for Computational Linguistics, 2016-06.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NIPS*, 2017.

[Lundberg *et al.*, 2018 10 01] 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, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749–760, 2018-10-01.

[Macleod, 2011] Les Macleod. Avoiding "groupthink" a manager's challenge. *Nursing management*, 42(10):44–48, 2011.

[Madras *et al.*, 2018] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6150–6160, Red Hook, NY, USA, 2018. Curran Associates Inc.

[Marsh and Dibben, 2003] Stephen Marsh and Mark R Dibben. The role of trust in information science and technology. *Annual Review of Information Science and Technology*, 37(1):465–498, 2003.

[McAuley *et al.*, 2012] Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, page 1020–1025. IEEE Computer Society, 2012.

[McCloskey and Cohen, 1989] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[Miller *et al.*, 2017] T. Miller, P. Howe, and L. Sonenberg. Explainable ai: Beware of inmates running the asylum. In *IJCAI-XAI*, 2017.

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

[Mitchell *et al.*, 2019] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, 2019. Association for Computing Machinery.

[Mohammed *et al.*, 2010] Susan Mohammed, Lori Ferzandi, and Katherine Hamilton. Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management*, 36(4):876–910, 2010.

[Morgenstern and Von Neumann, 1953] Oskar Morgenstern and John Von Neumann. *Theory of games and economic behavior*. Princeton university press, 1953.

[Mozannar and Sontag, 2020] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR, 2020.

[Nadav-Greenberg and Joslyn, 2009] Limor Nadav-Greenberg and Susan L Joslyn. Uncertainty forecasts improve decision making among nonexperts. *Journal of Cognitive Engineering and Decision Making*, 3(3):209–227, 2009.

[Nagar and Malone, 2011] Yiftach Nagar and Thomas Malone. Making business predictions by combining human and machine intelligence in prediction markets. In *International Conference on Information Systems*. Association for Information Systems, 2011.

[Narayanan *et al.*, 2018] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation, 2018.

[Nguyen *et al.*, 2015] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015.

[Nguyen, 2018 06] Dong Nguyen. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078. Association for Computational Linguistics, 2018-06.

[Niculescu-Mizil and Caruana, 2005] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *ICML*, pages 625–632. ACM, 2005.

[Norman, 1988] Donald Norman. *The psychology of everyday things*. Basic Books, 1988.

[Norman, 1994] Donald A Norman. How might people interact with agents. *Communications of the ACM*, 37(7):68–71, 1994.

[Nushi *et al.*, 2018] Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. *HCOMP*, 2018.

[O'Cane, 2018] Sean O'Cane. Tesla can change so much with over-the-air updates that it's messing with some owners' heads. `www.theverge.com/2018/6/2/17413732/tesla-over-the-air-software-updates-brakes`, 2018.

[Park *et al.*, 2019] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15, 2019.

[Patel *et al.*, 2019] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine*, 2(1):1–10, 2019.

[Perez *et al.*, 2019 11] Ethan Perez, Siddharth Karamcheti, Rob Fergus, Jason Weston, Douwe Kiela, and Kyunghyun Cho. Finding generalizable evidence by learning to convince Q&A models. 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 2402–2411. Association for Computational Linguistics, 2019-11.

[Platt, 1999] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[Poursabzi-Sangdeh *et al.*, 2019] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability, 2019.

[Reber, 1989] Arthur S Reber. Implicit learning and tacit knowledge. *Journal of experimental psychology: General*, 118(3):219, 1989.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[Rouse *et al.*, 1992] William B Rouse, Janis A Cannon-Bowers, and Eduardo Salas. The role of mental models in team performance in complex systems. *IEEE Transactions on SMC*, 22(6):1296–1308, 1992.

[Rudin, 2018] Cynthia Rudin. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*, 2018.

[Schmidt and Biessmann, 2019] Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems, 2019.

[Smith *et al.*, 2020] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.

[Spring, 2005 11] Maximilian J Spring. Techniques for maintaining compatibility of a software core module and an interacting module, 2005-11. US Patent 6,971,093.

[Stumpf *et al.*, 2009] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662, 2009.

[Tan *et al.*, 2018] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. Investigating human + machine complementarity for recidivism predictions, 2018.

[Team, 2017] LSAT Prep Books Team. *LSAT prep book study guide: quick study & practice test questions for the Law School Admissions council's (LSAC) Law school admission test*. Mometrix Test Preparation, 2017.

[Tsantilis, 2009 10] Efstratios Tsantilis. Method and system to monitor software interface updates and assess backward compatibility, 2009-10. US Patent 7,600,219.

[Ustun and Rudin, 2017] Berk Ustun and Cynthia Rudin. Optimized risk scores. In *KDD*, 2017.

[Varshney *et al.*, 2018] Kush R. Varshney, Prashant Khanduri, Pranay Sharma, Shan Zhang, and Pramod K. Varshney. Why interpretability in machine learning? an answer using distributed detection and data fusion theory, 2018.

[Wang *et al.*, 2016] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

[Wang *et al.*, 2019] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA, 2019. Association for Computing Machinery.

[Weerts *et al.*, 2019] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. A human-grounded evaluation of shap for alert processing, 2019.

[Weld and Bansal, 2019 05] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6):70–79, 2019-05.

[Wiens *et al.*, 2016] Jenna Wiens, John Guttag, and Eric Horvitz. Patient risk stratification with time-varying parameters: a multitask learning approach. *JMLR*, 17(1):2797–2819, 2016.

[Wilder *et al.*, 2020 07] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1526–1533. International Joint Conferences on Artificial Intelligence Organization, 2020-07. Main track.

[Wu *et al.*, 2018] Mike Wu, Michael C Hughes, Sonali Parbhoo, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *AAAI*, 2018.

[Wu *et al.*, 2019 06] Tongshuang Wu, Daniel S. Weld, and Jeffrey Heer. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. *ACM Trans. Comput.-Hum. Interact.*, 26(4), 2019-06.

[Yang *et al.*, 2020] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 189–201, New York, NY, USA, 2020. Association for Computing Machinery.

[Yin *et al.*, 2019] 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*, pages 1–12, 2019.

[Yu *et al.*, 2019] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. Do i trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 460–468, New York, NY, USA, 2019. Association for Computing Machinery.

[Yu *et al.*, 2020] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020*, pages 1–26. OpenReview.net, 2020.

[Zadrozny and Elkan, 2001] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616. Citeseer, 2001.

[Zadrozny *et al.*, 2003] Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM*, volume 3, page 435, 2003.

[Zafar *et al.*, 2017] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

[Zhang *et al.*, 2020] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery.

[Zhou *et al.*, 2004] Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NIPS*, pages 321–328, 2004.

[Zhou *et al.*, 2017] Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP Conference on Human-Computer Interaction*, pages 23–39. Springer, 2017.