Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance

Gagan Bansal*
Tongshuang Wu*
bansalg@cs.washington.edu
wtshuang@cs.washington.edu
University of Washington

Joyce Zhou†
Raymond Fok†
jyzhou15@cs.washington.edu
rayfok@cs.washington.edu
University of Washington

Besmira Nushi
besmira.nushi@microsoft.com
Microsoft Research

Ece Kamar
eckamar@microsoft.com
Microsoft Research

Marco Tulio Ribeiro
marcotcr@microsoft.com
Microsoft Research

Daniel S. Weld
weld@cs.washington.edu
University of Washington &
Allen Institute for Artificial Intelligence

ABSTRACT

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? 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. 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?

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Interactive systems and tools*; • **Computing methodologies** → Machine learning.

KEYWORDS

Explainable AI, Human-AI teams, Augmented intelligence

*Equal contribution.
†Made especially large contributions.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445717

ACM Reference Format:
Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3411764.3445717

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 [8, 13, 15, 78] and AIs that estimate recidivism to help judges decide whether to grant bail to defendants [2, 30]. 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 [73]. For example, significantly improving decision accuracy by closing deficiencies in automated reasoning with human effort, and vice versa [35, 70].

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* [46] is crucial for users to leverage AI assistance and improve task performance [10]. Indeed, at first glance, it appears that researchers have already confirmed the utility of explanations on tasks ranging from medical diagnosis [14, 53], data annotation [67] to deception detection [43]. In each case, the papers show that, when the AI provides explanations, team accuracy reaches a level higher than human-alone.

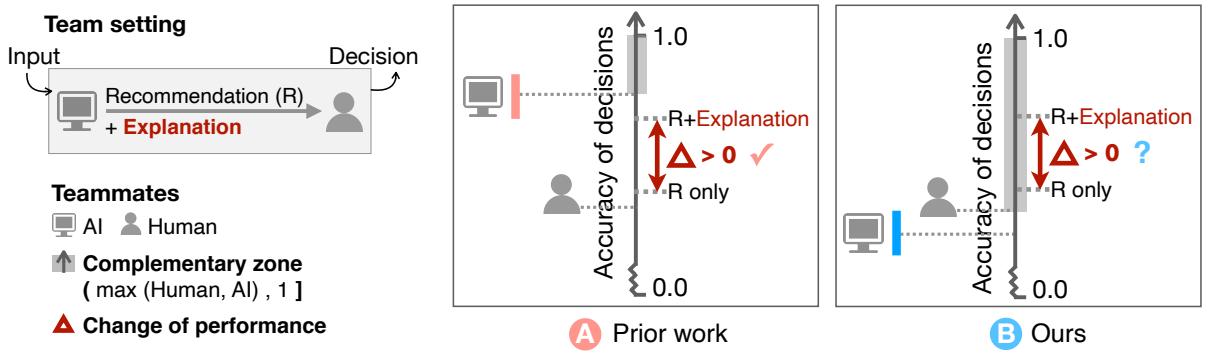

Figure 1: (Best viewed in color) Do AI explanations lead to complementary team performance? In a team setting, when given an input, the human uses (usually imperfect) recommendations from an AI model to make the final decision. We seek to understand if automatically generated explanations of the AI's 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 $\Delta_B > 0$?).

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 1A & Table 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 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 [38, 58], *i.e.*, explanations can lead humans to either follow incorrect AI suggestions or ignore the correct ones [13, 69]. 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:

- We highlight an important limitation of previous work on explainable AI: While many studies show that explaining predictions of AI increases team performance (Table 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.
- 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.

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

- 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 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 BACKGROUND AND RELATED WORK

Explanations can be useful in many scenarios where a human and AI interact: transparently communicating model predictions [10, 21, 38, 40, 66], teaching humans tasks like translation [3, 26] or content moderation [36], augmenting human analysis procedure [36] or creativity [17], legal imperatives [57, 77], etc. Various studies have evaluated the effect of explanations from different dimensions, including whether the explanation improves users' trust in the AI [41, 81] or enables users to simulate the model predictions [16, 65], or assists developers to debug models [9, 38].

In this paper, we focus explicitly on *AI-assisted decision making* scenarios [6, 74], 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 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 [12]. 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 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.* [42, 43], 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 [42]). Zhang *et al.* [83] 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 [83]). A similar performance trend is observed on tasks other than classification. In Sangdeh *et al.* [65], 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 [65]). In Feng *et al.* [21], 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 [21] pointing out that it outperforms top trivia players. One possible exception is Bligic & Mooney (2005) [10], 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 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."

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 3.1). As Table 2 summarizes, we first designed and conducted pilots studies (Sections 3.2) and used them to inform our final study and hypotheses (Section 4).

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 [66]. This contrasts with providing a global understanding of the full model all at once (*e.g.*, [45]).

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

[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.

Domain	Task	Performance				
		Metric	Human alone	AI alone	Team	Complementary?
Classification	Deceptive review [43]	Accuracy ↑	51.1%	**87.0%**	74.6%	✗
	Deceptive review [42]	Accuracy ↑	54.6%	**86.3%**	74.0%	✗
	Income category [83]	Accuracy ↑	65%	**75%**	73%	✗
	Loan defaults [27]	Norm. Brier ↑	0	**1**	0.682	✗
	Hypoxemia risk [53]	AUC ↑	0.66	**0.81**	0.78	✗
	Nutrition prediction [12]	Accuracy ↑	0.46	**0.75**	0.74	✗
QA	Quiz bowl [21]	*"AI outperforms top trivia players."*				✗
Regression	House price [65]	Avg. Absolute Error ↓	$331k	**$200k**	$232k	✗

Table 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).

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

Table 2: An overview of our tasks, explanation strategies and sources. We ran our pilot studies (Section 3.2) with conditions marked with ● . Based on the pilot results, we added adaptive explanations and expert explanations (Section 3.3). Along with two additional domains, these form the conditions for our final study conditions (Section 4.1).

teaming [21, 29, 42, 50, 62, 83] 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 [56] and book reviews [31]. More details about these datasets are in Section 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 [21, 42, 43], 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 [11, 27, 53, 61, 76] (more suitable for tabular data) or communicating influential training examples [3, 40, 80] (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 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 3). Again, answering *LSAT* questions requires no specialized knowledge except common-sense reasoning skills, such as recognizing logical connections and conflicts between arguments [82]. Because in this case 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 4.2.

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 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 [24, 43, 49, 75], and have been shown to be beneficial (Table 1). Explain-Top-2 also corresponds to Wang *et al.*'s suggestion to mitigate heuristic biases by explaining "multiple outcomes" [74].

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*

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

[4] the saliency scores were based on feature weights learned by the linear model [27, 42]

hence deterred over-reliance. In Figure 2, for example, if counter-argument (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 5.1 and Figure 4.

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 [32], 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 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* [1]. 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 [34, 48].

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 4.2.

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 4.1). We then tested these hypotheses for several tasks and AI systems (Section 4.2) through crowdsourcing studies (Section 4.3).

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 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 [32]. Second, displaying uncertainty in predictions can help users can make more optimal decisions [19, 22, 25, 37, 60]. 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.

Interface. Figure 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, 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 3 shows a screenshot of the user interface for *LSAT* in the Team (Adaptive, Expert) condition.

4.2 AI Model, Study Samples and Explanations

4.2.1 *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 [52] 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 [66]. 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

[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.

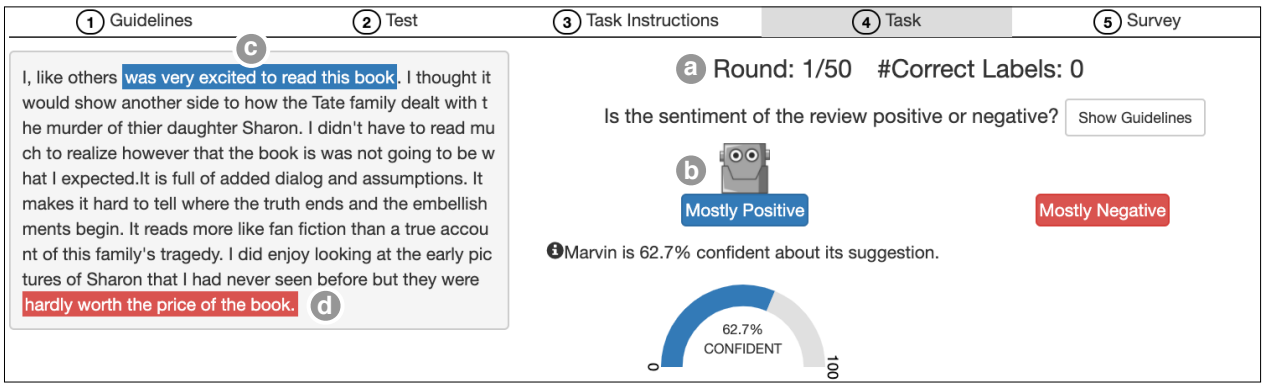

Figure 2: A screenshot of the Team (Adaptive, Expert) condition for the *Amzbook* reviews dataset. Participants read the review (left pane) and used the buttons (right pane) to decide if the review was mostly *positive* or *negative*. The right pane also shows progress and accuracy (a). To make a recommendation, 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.

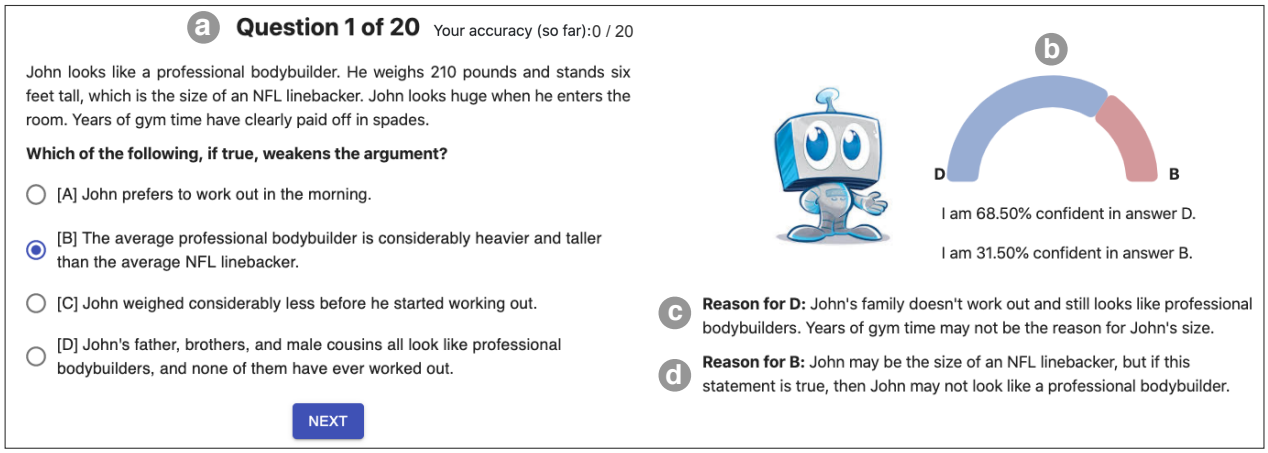

Figure 3: A screenshot of Team (Adaptive, Expert) for *LSAT*. Similar to Figure 2, the interface contained a progress indicator (a), AI recommendation (b), and explanations for the top-2 predictions (c and d). To discourage participants from blindly following the AI, all AI information is displayed on the right. In (b), the confidence score is scaled so those for top-2 classes sum to 100%.

[28], which we mitigated with *post hoc calibration* (isotonic regression [7]) 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 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).

4.2.2 *LSAT.*

AI Model. We finetuned a RoBERTa model[7] on ReClor [82], 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 [71]. 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.

[7] Based on the opensource implementation: https://github.com/yuweihao/reclor.
[8] https://en.wikipedia.org/wiki/GraduateManagementAdmissionTest

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.

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 [51]. 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 3.3) and presented 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 [33]. 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).

5 RESULTS

5.1 Effect of Explanation on Team performance

Figure 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 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 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

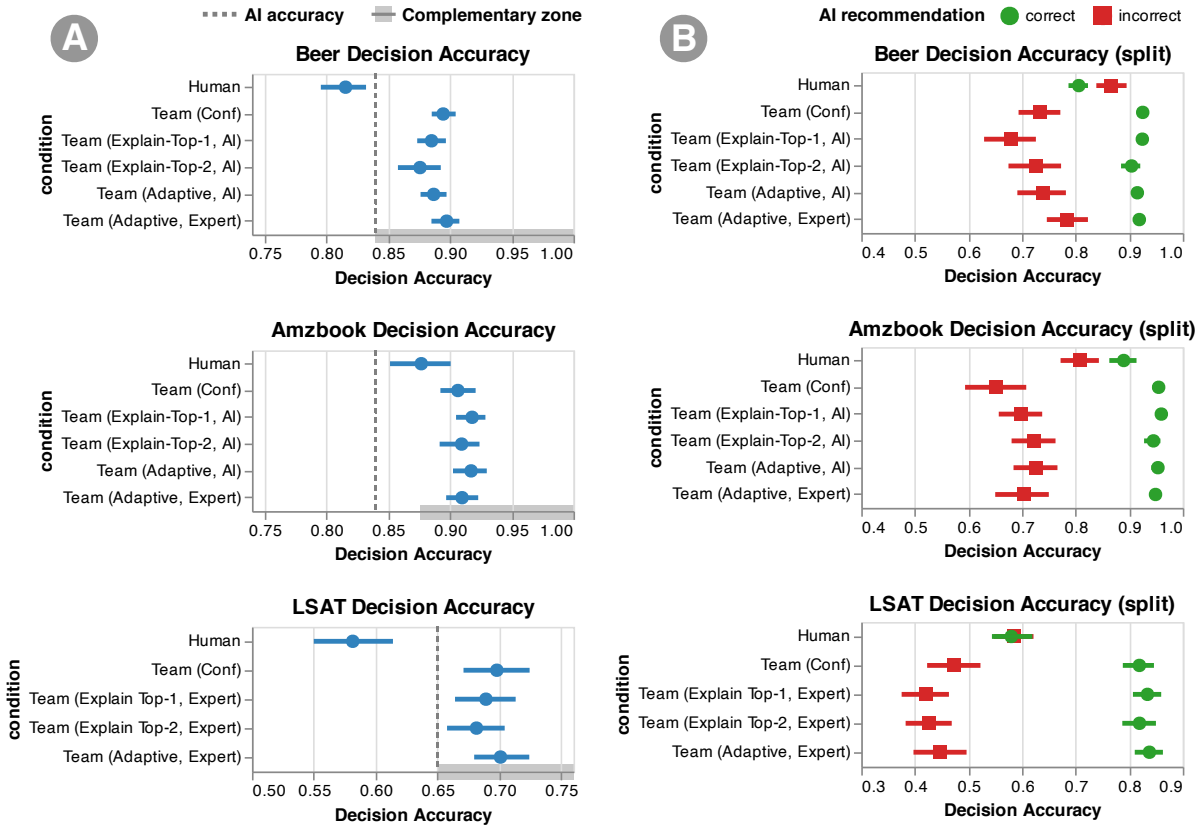

Figure 4: Team performance (with average accuracy and 95% confidence interval) achieved by different explanation conditions and baselines for three datasets, with around 100 participants per condition. (A) Across every dataset, all team conditions achieved complementary performance. However, we did not 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.

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 4A, the 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 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 [39], 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 [38]. 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 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 4B shows team performance, the promising effects of Adaptive explanations are clearer if we study the agreement between AI predictions and human decisions (Figure 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 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 [44]). Identifying unknown unknowns could have been a difficult and time-consuming task for workers, and they

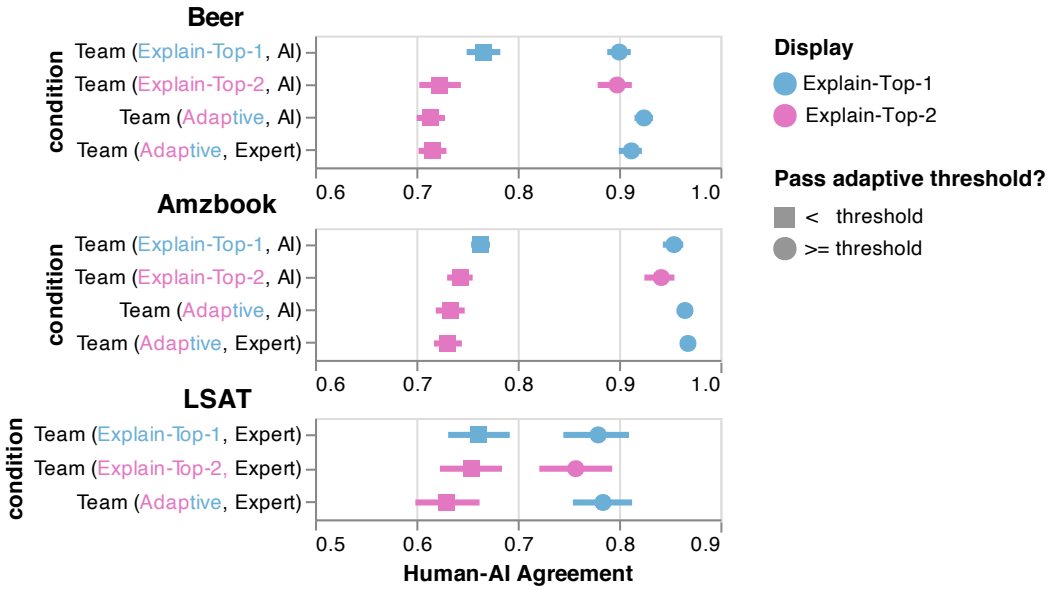

Figure 5: Relative agreement rates between humans and AI (*i.e.*, does the final human decision match the AI's suggestion?) for various conditions, with examples split by whether AI's confidence exceeded the threshold used for Adaptive explanations. Across the three datasets, Adaptive explanations successfully 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).

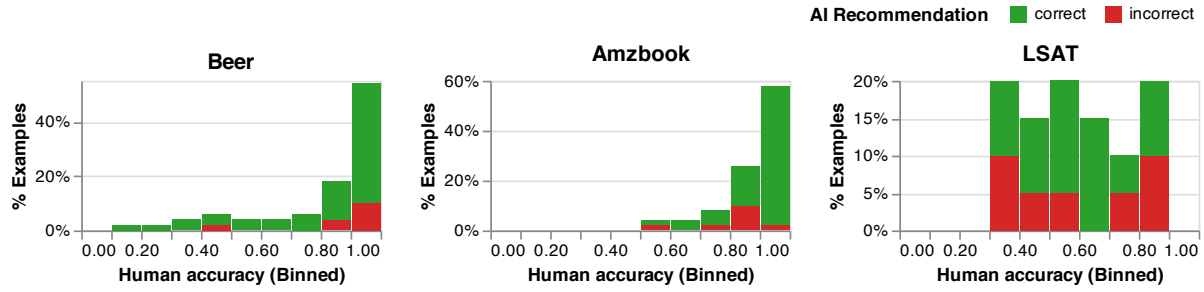

Figure 6: The distribution of study examples as a function of average human accuracy. For each domain, examples on the right were easy for most humans working alone. Both *Beer* and *LSAT* show a distribution that shows potential for complementary team performance: humans can correct easy questions mistaken 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).

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 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 6. The figure helps 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 4B), but it decreased the performance for *Beer* and *LSAT*. One

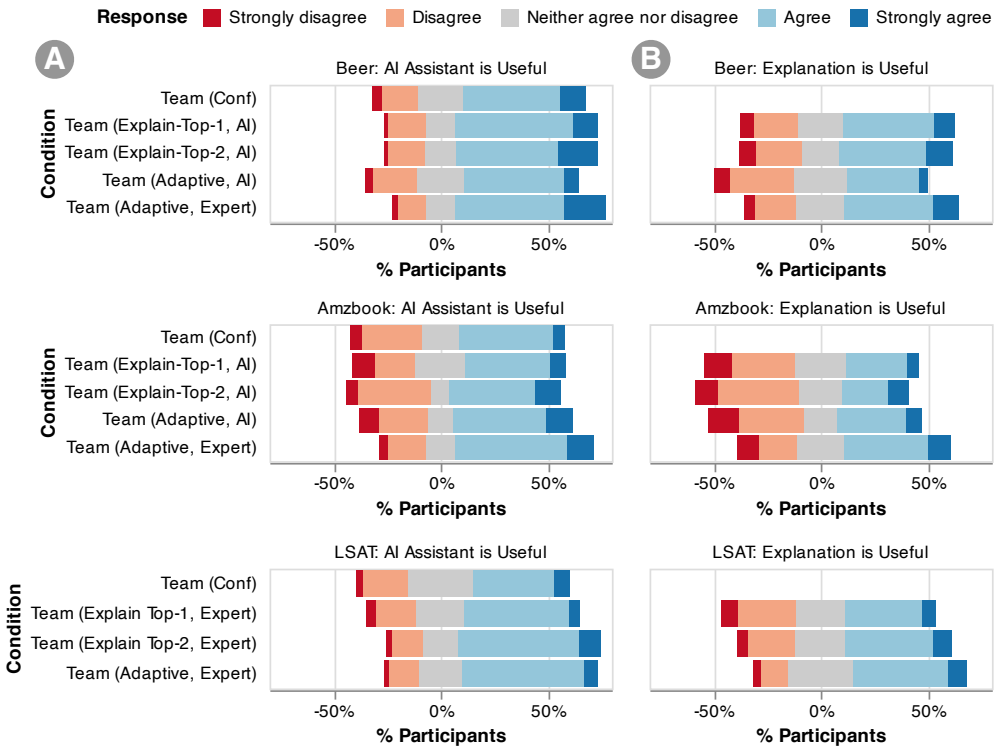

Figure 7: Analysis of participant responses to two statements: (A) "AI's assistance (*e.g.*, the information it displayed) helped me solve the task", and (B) "AI's explanations in particular helped me solve the task." Across datasets, a majority of participants found AI assistant to be useful, and they 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.

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 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.

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 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%).

[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.

Explanation usefulness: Figure 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.* [12] that show that people may prefer one explanation but make better decisions with another.

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 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 3). We scored the inter-annotator agreement with both the Cohen's κ

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

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% ± 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 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 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 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."*

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 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

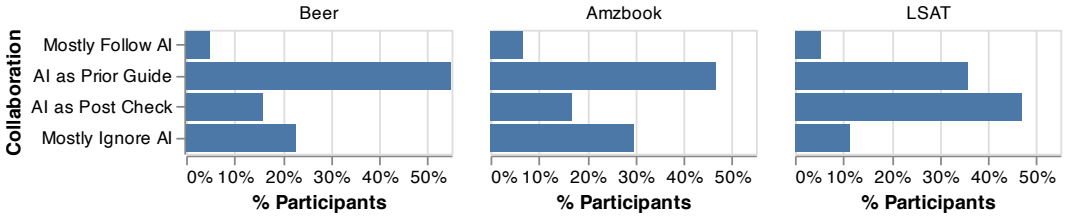

Figure 8: Instead of ignoring or strictly following the AI, participants reported taking the AI information into consideration most of the time. They most frequently used AI as a prior guide in sentiment analysis, but used it as post-check in *LSAT*. They were also more likely to ignore the AI in sentiment analysis than in *LSAT*.

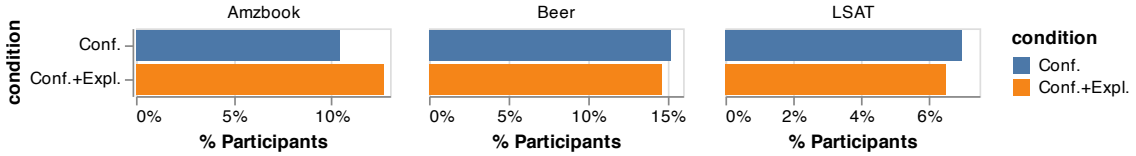

Figure 9: Comparing the occurrence of `Used Conf.` in just the confidence condition and in those with explanations, we saw a similar proportion of users that explicitly acknowledged using confidence, regardless of whether they saw an explanation.

be "trustworthy." While this observation empirically shows that end-users develop mental model of trust in AI-assisted decision making [5], 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 [23]. 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 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* [47], 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 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."*

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.

6.1 Limitations

As mentioned in Section 2, AI explanations have other motivations not addressed by this paper. Our work, as well as the papers listed in Table 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 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 [68]. 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 [21] 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 [24, 43, 49, 75], alternative approaches may provide more benefit to team performance.

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 [18, 38, 58, 74], 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 [12]. 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 5), the method was not sufficient to significantly increase the final team accuracy (Figure 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 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 [12]. 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 [15]). 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.

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 6). As recent work has suggested [4, 55, 59, 79], 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 [72].

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 [26]. Similarly, to maximize human performance at Quiz Bowl, Feng and Graber [21] 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 [20] and introduced to the AI explanation field by Wang et al. [74] — 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 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 [63]. Methods that embody recommendations from management science for avoiding group-think [54] 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 [27], 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* [64]. 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.

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.

ACKNOWLEDGMENTS

This material is based upon work supported by ONR grant N00014-18-1-2193, NSF RAPID grant 2040196, the University of Washington WRF/Cable Professorship, and the Allen Institute for Artificial Intelligence (AI2), and Microsoft Research. The authors thank Umang Bhatt, Jim Chen, Elena Glassman, Walter Lasecki, Qisheng Li, Eunice Jun, Sandy Kaplan, Younghoon Kim, Galen Weld, Amy Zhang, and anonymous reviewers for helpful discussions and comments.

REFERENCES

[1] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a Better Understanding of Context and Context-Awareness. In *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing* (Karlsruhe, Germany) *(HUC '99)*. Springer-Verlag, Berlin, Heidelberg, 304–307.

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

[3] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. 2018. Teaching Categories to Human Learners With Visual Explanations. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*. IEEE Computer Society, Salt Lake City, UT, USA, 3820–3828. https://doi.org/10.1109/CVPR.2018.00402

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2020. Optimizing AI for Teamwork. arXiv:2004.13102 [cs.AI]

[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 2–11. https://ojs.aaai.org/index.php/HCOMP/article/view/5285

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[7] Richard E Barlow and Hugh D Brunk. 1972. The isotonic regression problem and its dual. *J. Amer. Statist. Assoc.* 67, 337 (1972), 140–147.

[8] Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M. Mack, George Ruiz, Mark S. Smith, and Eric Horvitz. 2014. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. *PLOS ONE* 9, 10 (10 2014), 1–9. https://doi.org/10.1371/journal.pone.0109264

[9] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 648–657. https://doi.org/10.1145/3351095.3375624

[10] Mustafa Bilgic. 2005. Explaining Recommendations: Satisfaction vs. Promotion. , 13–18 pages.

[11] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[12] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[13] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, IEEE, Dallas, Texas, 160–169.

[14] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " 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 (2019), 1–24.

[15] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. 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* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[16] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human?. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 1036–1042. https://doi.org/10.18653/v1/D18-1128

[17] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

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

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

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

[21] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[22] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in*

Computing Systems (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173718

[23] 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. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376316

[24] Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting Recurrent and Attention-Based Neural Models: a Case Study on Natural Language Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4952–4957. https://doi.org/10.18653/v1/D18-1537

[25] Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. 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)*. Association for Computational Linguistics, Berlin, Germany, 264–268. https://doi.org/10.18653/v1/P16-2043

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

[27] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[28] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, Sydney, NSW, Australia, 1321–1330.

[29] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5540–5552. https://doi.org/10.18653/v1/2020.acl-main.491

[30] Yugo Hayashi and Kosuke Wakabayashi. 2017. 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* (Portland, Oregon, USA) *(CSCW '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 195–198. https://doi.org/10.1145/3022198.3026338

[31] Ruining He and Julian McAuley. 2016. 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* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 507–517. https://doi.org/10.1145/2872427.2883037

[32] Dan Hendrycks and Kevin Gimpel. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, Toulon, France, 1–12. https://openreview.net/forum?id=Hkg4TI9xl

[33] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web* (Florence, Italy) *(WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 419–429. https://doi.org/10.1145/2736277.2741102

[34] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

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

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

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

[38] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

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

[40] Pang Wei Koh and Percy Liang. 2017. Understanding Black-Box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, Sydney, NSW, Australia, 1885–1894.

[41] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[42] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376873

[43] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[44] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, San Francisco, California, USA, 2124–2132.

[45] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & Explorable Approximations of Black Box Models. arXiv:1707.01154 [cs.AI]

[46] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

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

[48] B. Lim, A. Dey, and D. Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, New York, NY, USA, 2119–2128. https://doi.org/10.1145/1518701.1519023

[49] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-Attentive Sentence Embedding. In *5th International Conference on Learning Representations, ICLR 2017, Conference Track Proceedings*. OpenReview.net, Toulon, France, 1–15. https://openreview.net/forum?id=BJC_jUqxe

[50] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (June 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[51] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H. Lin, Xiao Ling, and Daniel S. Weld. 2016. 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*. Association for Computational Linguistics, San Diego, California, 897–906. https://doi.org/10.18653/v1/N16-1104

[52] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL]

[53] 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. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2, 10 (01 Oct 2018), 749–760. https://doi.org/10.1038/s41551-018-0304-0

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

[55] David Madras, Toniann Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 6150–6160.

[56] Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning Attitudes and Attributes from Multi-Aspect Reviews. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM '12)*. IEEE Computer Society, USA, 1020–1025. https://doi.org/10.1109/ICDM.2012.110

[57] T. Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (February 2018), 1–38.

[58] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[59] Hussein Mozannar and David Sontag. 2020. Consistent Estimators for Learning to Defer to an Expert. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, Virtual, 7076–7087. http://proceedings.mlr.press/v119/mozannar20b.html

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

[61] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. arXiv:1802.00682 [cs.AI]

[62] Dong Nguyen. 2018. 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 1069–1078. https://doi.org/10.18653/v1/N18-1097

[63] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[64] Ethan Perez, Siddharth Karamcheti, Rob Fergus, Jason Weston, Douwe Kiela, and Kyunghyun Cho. 2019. 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)*. Association for Computational Linguistics, Hong Kong, China, 2402–2411. https://doi.org/10.18653/v1/D19-1244

[65] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Manipulating and Measuring Model Interpretability. arXiv:1802.07810 [cs.AI]

[66] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[67] Philipp Schmidt and Felix Biessmann. 2019. Quantifying Interpretability and Trust in Machine Learning Systems. arXiv:1901.08558 [cs.LG]

[68] C. Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376783

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

[70] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating Human + Machine Complementarity for Recidivism Predictions. arXiv:1808.09123 [cs.LG]

[71] LSAT Prep Books Team. 2017. *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, Beaumont, TX.

[72] Kush R. Varshney, Prashant Khanduri, Pranay Sharma, Shan Zhang, and Pramod K. Varshney. 2018. Why Interpretability in Machine Learning? An Answer Using Distributed Detection and Data Fusion Theory. arXiv:1806.09710 [stat.ML]

[73] J. von Neumann and O. Morgenstern. 1947. *Theory of games and economic behavior*. Princeton University Press, Princeton, New Jersey.

[74] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[75] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 606–615. https://doi.org/10.18653/v1/D16-1058

[76] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. arXiv:1907.03324 [cs.LG]

[77] Daniel S. Weld and Gagan Bansal. 2019. The Challenge of Crafting Intelligible Intelligence. *Commun. ACM* 62, 6 (May 2019), 70–79. https://doi.org/10.1145/3282486

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

[79] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to Complement Humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 1526–1533. https://doi.org/10.24963/ijcai.2020/212 Main track.

[80] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3377325.3377480

[81] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[82] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. In *8th International Conference on Learning Representations, ICLR 2020*. OpenReview.net, Addis Ababa, Ethiopia, 1–26. https://openreview.net/forum?id=HJgJtT4tvB

[83] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852