You Complete Me: Human-AI Teams and Complementary Expertise

Qiaoning Zhang
University of Michigan-Ann Arbor
Ann Arbor, Michigan, United States
qiaoning@umich.edu

Matthew L. Lee
Toyota Research Institute
Los Altos, California, United States
matt.lee@tri.global

Scott Carter
Toyota Research Institute
Los Altos, California, United States
scott.carter@tri.global

ABSTRACT

People consider recommendations from AI systems in diverse domains ranging from recognizing tumors in medical images to deciding which shoes look cute with an outfit. Implicit in the decision process is the perceived expertise of the AI system. In this paper, we investigate how people trust and rely on an AI assistant that performs with different levels of expertise relative to the person, ranging from completely overlapping expertise to perfectly complementary expertise. Through a series of controlled online lab studies where participants identified objects with the help of an AI assistant, we demonstrate that participants were able to perceive when the assistant was an expert or non-expert within the same task and calibrate their reliance on the AI to improve team performance. We also demonstrate that communicating expertise through the linguistic properties of the explanation text was effective, where embracing language increased reliance and distancing language reduced reliance on AI.

CCS CONCEPTS

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

KEYWORDS

human-AI teams, explainable AI, trust, complementary expertise

ACM Reference Format:
Qiaoning Zhang, Matthew L. Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *CHI '22: ACM CHI, May 01–05, 2022, New Orleans, LA.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3491102.3517791

1 INTRODUCTION

People team up with AI-powered systems to accomplish critical tasks including identifying tumors in medical images and detecting obstacles in L2 automated driving, as well as more mundane tasks such as automated grammar checking and phishing attack detection. The dream of the ideal human-AI partnership relies on the premise that the expertise of the AI system complements the expertise of the human, which allows the partnership to accomplish even more than each actor alone. Bansal et al. [5] found that human-AI team performance was higher in tasks in which the correlation of errors between the human and AI was lower, in other words, when there was a higher degree of complementary expertise between the human and AI.

However, it is not always easy for a human user to know when an AI system may be an expert (having high accuracy) in some parts of a task but be an non-expert (having low accuracy) in others parts of the task. It is thus difficult for the human user to know when to trust and rely on the AI system's recommendation or override it. Humans can calibrate their trust by observing the performance of their AI partner over time and comparing it to their own performance. Yin et al. [50] found that higher trust ratings were associated with an AI system that is observed to perform better than the human partner. Though, in practice, AI systems often do not perform uniformly better or worse in a task relative to their human partners. For example, an AI radiology assistant may detect certain types of tumors more accurately than the radiologist but less accurately than the radiologist with other types of tumors. The human partner must dynamically learn and calibrate themselves to the nuances of AI system's specific expertise to know when to defer to or override the system's recommendation. But how well can people detect and learn when an AI assistant may be highly accurate in some parts of the task but not others? Furthermore, is it ideal to have perfect complementarity or does some overlap of expertise help to build trust in the system? Finally, how does the degree of complementary expertise affect human-AI team performance?

In this paper, we report on a series of controlled online lab studies that investigate these questions. In the first study, we designed a shape identification task in which human participants had to identify the category of the shape with the help of an AI assistant that had different degrees of complementary expertise to the human. Humans leveraged their innate expertise to recognize real shapes such as circles, squares, and triangles, but had no expertise in recognizing novel "fake" shapes we generated for this study. In different experimental conditions, we manipulated the level of complementary expertise of the AI assistant from completely overlapping with the human (i.e., being highly accurate for only the real shapes) to completely complementary to the human (i.e., being highly accurate for only the novel fake shapes). The results demonstrate that participants were able to perceive when the AI assistant was an expert or non-expert within the same task and relied on the AI more often as its complementary expertise increased, leading to greater task performance (number of shapes identified correctly). However, subjective measures of trust did not follow a similar trend across conditions of different complementary expertise, highlighting an

CHI '22, May 01–05, 2022, New Orleans, LA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517791

opportunity for the AI to better communicate its expertise to their human partner to build appropriate trust.

Providing natural language explanations that communicate confidence information [47] is a common method to moderate trust when conveying the expertise of an AI system. However, as Zhang et al. [52] found, explanations can undermine team performance if they overload the human user with too much extra information to consider. To address this concern, we leverage prior work showing that people naturally rely on subtle linguistic cues to infer the intentional stance and beliefs of the other person (or AI assistant) and to explain behavior [18]. Existing AI systems such as smart agents often employ particular linguistic cues such as the first person point-of-view pronoun "I" or cues that reveal the AI's belief state such as "Your thermostat turned off the heat because it *thinks* you are not home". Specifically, we investigate how these cues convey psychological distance (the cognitive separation between the self and other entities such as persons, events, or times [26]) to indicate perspective and allow the user to infer the AI's confidence or expertise. We investigate two methods for manipulating psychological distance that are already commonly used but understudied in textual explanations: belief markers and point-of-view (POV). A belief marker in an explanation is a word that indicates the strength of the explainer's belief in the explanation, and allows the listener or explainee to form impressions about the explainer and the explanation [29]. Belief markers lie in a spectrum from distancing to embracing. Belief markers with higher distance (distancing markers) include terms that imply lower certainty such as *think* and *believe*, whereas belief markers with lower distance (embracing markers) include words that imply higher certainty such as *realize* and *know* [29, 30]. Point-of-view (POV), on the other hand, refers to statements either in the more distancing 3rd-person perspective (using terms like "the system", "Siri", "it") or in the less distancing 1st-person perspective (using terms like "I", "We"). In a second study, we added natural language explanations to the experimental paradigm of our first study and conducted a systematic investigation of the effects of belief markers and point-of-view on trust, reliance, and performance. The results show that embracing and distancing language had a significant effect, with embracing and distancing language inducing higher agreement and lower agreement with the AI, respectively, providing a subtle mechanism to support appropriate reliance on the AI according to when its recommendation is correct or incorrect.

Our work makes two key contributions. In the area of AI-assisted decision making, our work demonstrates that people are able to detect and rely on an AI assistant that performs both expertly and non-expertly in different parts of the same task, a common case in practice but not yet systematically studied. In the area of Explainable AI, our work systematically investigates the distancing/embracing effects of belief markers and point-of-view markers that are commonly employed in AI explanations.

2 BACKGROUND

2.1 AI-Assisted Decision Making

Many AI systems that provide input for high-impact decision making are already integrated into a variety of work domains including child welfare risk assessment [9], recidivism prediction [10], predictive policing [39], and clinical decision support [46]. These examples follow the paradigm of AI-assisted Decision Making [4, 22] where 1) both the human and AI system are presented with a decision, 2) the AI system provides its recommendation, 3) the human makes the final decision that either agrees with or rejects the recommendation, and 4) the outcome of the final decision is revealed along with a reward if correct. Along with many other HCI studies [38, 51, 52], we follow a similar paradigm in our experiments.

AI systems rarely work perfectly in practice and often make errors in their recommendations, leading the human astray. In other cases, the AI system may actually be correct but the human teammate may incorrectly reject the recommendation. The success of the human-AI team depends on the human knowing when to rely or not rely on the AI. Bansal et al. [3] found that humans can form mental models of the AI system's error boundary by observing the system's performance over time in a novel object identification task. Updating the error boundary in the middle of the task could also degrade team performance because it becomes incompatible with the human's mental model. Gero et al. [17] similarly showed how players who had a more accurate mental model of the AI assistant performed better in a game. In our work, we investigate whether humans with a priori expertise can identify the error boundaries of an AI system that has expertise in some aspects of the task and rely on it appropriately.

2.2 Complementary Expertise in Human-AI Teams

One of the most compelling reasons for using human-AI teams is that humans and machines often have complementary strengths. Indeed, Bansal et al. [5] found that team performance was higher for tasks where there was only a small correlation of errors between human and AI. In particular, relying on the AI improved team performance for tasks in which the human's and AI's expertise complemented each other, in contrast to other tasks in which expertise was more overlapping. Feng & Boyd-Graber [15] describe examples of when humans and AI have complementary abilities in different output spaces such as in a trivia game where humans are good at chaining evidence and solving wordplay whereas an AI can leverage its ability to memorize every poem or book ever written to identify quotes faster than a human. We adopt a related working definition of complementary expertise to be the degree to which the AI system is more accurate than the human in areas of the task that the human partner tends to be less accurate. An analogy for an AI that has high complementary expertise in a trivia game would be if the AI was an expert in an obscure topic (e.g., numismatics) but had little knowledge in a common topic that the human partner had expertise (e.g., cliches).

Moreover, the expertise of human can affect the trust and use of AI systems. Nourani et al. [36] studied domain experts and how AI errors shown early in a sequence of recommendations can erode trust in the AI that can be difficult to rebuild, whereas novices with low domain expertise displayed automation bias and tended to trust the AI blindly. Even domain experts like clinicians can be susceptible to automation bias when the AI system proactively offers a suggestion before the clinicians make their own initial assessment

[25]. Schaffer et al. [40] found that overconfident participants who reported high task familiarity relative to their actual expertise said they trusted their AI partner more but in fact did not rely on it. Prior work [5, 28, 52] studied AI systems that have a single fixed level of expertise or performance relative to the human.

In our work, we explore the common case in practice where the human leverages their existing expertise along with a new AI system that performs correctly and incorrectly in different parts of the task. We explicitly manipulate the expertise of the AI system (i.e., how accurate its recommendations are) to have more or less overlap with the expertise with the human, to study how it effects team performance, reliance, and trust.

2.3 Trust and Reliance on AI Systems

Appropriate trust and reliance are central facilitators of team performance for both human-human and human-AI teams. Reliance can be defined as the continued relationship on the basis of one party's dependable habits toward the other, whereas trust is a special case of reliance where one party is relying specifically on the good will of the other party [2]. An integrative model by Mayer et al. [32] identifies the bases of team trust to include ability, integrity, and benevolence.

In the scope of our work using the AI-assisted Decision Making paradigm, we assume the benevolence (committed to the team's well-being) and integrity (behaves ethically) of the AI partner, and will focus on how learned trust is formed based on the human's perceived ability/expertise of the AI partner. With respect to ability, Lee & Moray [24] proposed that trust depends on human-machine joint performance, system errors, and the human's prior trust. For prior trust, Kocielnik et al. [19] and Cai et al. [7] showed that setting initial expectations of the particular types of errors (false positives vs false negatives) can help increase user satisfaction and acceptance of an AI system. Similarly, Yin et al. [50] found that trust is a function of the human's expectation of the system's accuracy but also the observed system accuracy. People were more likely to trust in a model if the model's observed accuracy is higher than their own, in other words, when the model has greater overall expertise than the self. Yu et al. [51] found that trust in a system can increase over time when the system performs at over a 70% accuracy threshold. In our work, we focus on how appropriate trust and reliance are built up when partnering with an AI at different levels of expertise within the same task relative to the human.

2.4 Explanation Language in AI Systems

Providing explanations can help people who do not possess complete in-depth knowledge of the system to understand, predict, and therefore trust AI systems [33]. Communicating an AI system's intentions, state, capacity, and upcoming actions, can help users form and reinforce an accurate mental model [21]. This representation of the AI system allows users to continuously cycle through the process of perceiving information, comprehending the provided explanation, anticipating future actions, and taking the appropriate precautions in unforeseen circumstances [35, 42]. Körber et al. [20] found that users felt more strongly that they had understood the system, the reason for the AI system's actions, and intervention request when they were provided explanations.

Besides helping people to form a correct mental model, explanations can also help clarify the responsibilities of the users and the AI system. By explicitly explaining the system actions and clarifying the task to users, explanations help to highlight the status of both parties as partners through cooperative perception and to assist users in understanding whether they (an internal determinant) or the AI system (an external determinant) are mainly responsible for the behavior of the system [43], and consequently to improve the interaction of a user with the AI system [35].

Explanations are closely associated with human attitudes towards and acceptance of AI systems. For example, Lai & Tan [23] found that providing an appropriate explanation for AI-assisted decision making increased people's trust in AI agents, acceptance of the recommendation, and team performance. In addition, evidence suggests that rational explanations are effective for users who report being unfamiliar with a task in terms of trust and performance [41]. Cai et al. [6] found that example-based explanations can lead to a better understanding of the system and trust (i.e., capability and benevolence). However, the positive effect of explanation has not been supported by the work of Zhang et al. [52], who found that even though providing explanations or confidence information can help humans better assess the abilities of the AI system and calibrate their trust to the system's error boundary, the additional information can lead to cognitive overload and reduce overall team performance. Thus, there is a need for more subtle, less cognitively demanding methods for AI systems to communicate their confidence and rationales.

Explanations using natural language text is common in intelligent virtual agents [48] and intelligent systems [11, 13, 45] because it leverages people's innate ability to understand language, which does not require additional learning or high cognitive effort. Prior literature found that linguistic structure plays an important role in the effectiveness of explanations. DeGraaf & Malle [12] posited that people expect explanations from AI systems to follow the linguistic framework used to explain human behaviors. For example, Harbers et al. [18] showed that people prefer explanations that reflect the intentional stance of the AI, including its beliefs, desires, and mental states influencing its decisions. One natural way for explanations to indicate perspective and communicate confidence [31] is to adjust the text's psychological distance, the cognitive separation between the self and other entities such as persons, events, or times [26]. Belief markers and point-of-view are two linguistic devices that can manipulate psychological distance [29, 33]. Belief markers that increase distance (distancing markers) include words that imply less certainty such as *think* and *believe*, whereas belief markers that decrease distance (embracing markers) include words that imply more certainty such as *realize* and *know* [29]. For example, when answering the question "Why is this shape a triangle?", the explanations "I *think* it has three sides" and "I *know* it has three sides" both use the same reasoning, but the use of different belief markers conveys different impressions: the use of "*think*" seems to convey the idea that the explainer may not be certain of the reasons. Research in linguistic semantics shows that different verbs can convey different degrees of speaker confidence in the truth of a statement, for example, verbs such as "*know*" presuppose the truth, so we should expect to hear them only from speakers who are certain about what they are saying. On the other hand, verbs such as "*believe*" convey

no such presupposition [27]. Point-of-view markers, on the other hand, refer to statements in the third-person perspective or the first-person perspective. In general, the use of first-person point-of-view (e.g., I/we) invites the explainee to infer that the reason is a belief of the explainer and to form an embracing impression. However, the third-person point-of-view (e.g., she/he/it) only implies some sort of agreement from the explainer, which leads to a more distancing impression [33] (e.g., "She thinks it has three sides"). In prior studies of health narratives, using the 1st person POV has been shown to be more effective for enhancing message involvement and producing persuasive outcomes than using the 3rd person POV [8]. However, even though most text-based explanations use these markers, past work on the impacts of the linguistic structure in AI explanations is limited. In fact, a survey paper by Lai et al. [22] highlighted a research opportunity to better define the design space of AI assistance elements and to identify what might help people better assist decision makers rather than driven by new modeling or technical capabilities. In our work, in Study 2, we systematically investigate how explanations with different levels of psychological distance (operationalized using point of view and belief markers) influence trust and reliance on an AI system and its interactions with expertise. But first in the next section, we will describe Study 1, which investigates how the degree of complementary expertise affects human-AI team performance, trust, and reliance.

3 STUDY 1: COMPLEMENTARY EXPERTISE

Our first study focused on collaborating with an AI with different degrees of complementary expertise relative to the human participant. We investigate the following research question:

- Research Question 1 (RQ1): How does the degree of complementary expertise affect team performance, reliance behavior, and trust in AI-assisted decision making?

To answer this question, we formulated these hypotheses:

- ***H1:*** *Human-AI team performance is higher when the AI has higher degrees of complementary expertise.*
- ***H2:*** *People rely more on an AI system that has higher degrees of complementary expertise.*
- ***H3:*** *People have higher subjective trust in an AI system that has a higher degree of complementary expertise.*

To test these hypotheses, we designed a shape identification experiment in which participants relied on their innate knowledge of familiar shapes, but in order to identify unfamiliar shapes they also had to learn when to rely on an AI that had a certain combination of expertise in some familiar and unfamiliar shapes.

3.1 Task and User Interface Design

To test our hypotheses, we needed to design a task in which all participants would be total experts for some trials without any training but total non-experts in other trials, in order to mirror the type of real-life situation that requires working with a partner. Accordingly, we developed an object identification task in which participants were shown a series of images of 2D shapes and were asked to select to which of six shape categories it belonged. The six shape categories included three *Regular* shapes, Rectangle, Triangle, Circle, and three *Fake* shapes, Senectus, Pharetra, and Ultrices. The Regular shapes of Rectangle, Triangle, and Circle were already familiar to all participants, and each instance within a category varied randomly in fill color, length of sides, and corner angles (for the Triangle). In contrast to the Regular shapes that participants were experts at identifying, the Fake shapes were artificially created and named for this study, using Bezier curves as sides for a closed 2D shape with particular border and interior patterns (dots, dashes, or dots&dashes). Fake shapes categories are distinguished by the number of sides and whether the border pattern and interior fill patterns match or are different. Shape design criteria are shown in Appendix Table 3. Each example of a Fake shape varied randomly in category-irrelevant features including its fill color, length of sides, and angles of its curved sides (Appendix Figures 32, 33, and 34 show examples of the three Fake shape categories).

To manipulate degrees of complementary expertise in a between-subjects experiment, we paired each participant with a version of an AI partner, named ShapeBot, that had a specific combination (or level) of expertise in identifying shapes and recommending an answer to the participant. In the condition with the highest complementary expertise, ShapeBot was configured to correctly identify all Fake shapes, but performed poorly (seeming to guess randomly) on the Regular shapes that participants would already know how to identify on their own. In this case, the expertise of the AI and the human perfectly complemented each other. In the condition with the lowest complementary expertise, ShapeBot was configured to correctly identify only Regular shapes, and performed poorly (seeming to guess randomly) on the Fake shapes. In this case, the expertise of the AI and human were completely overlapping. We also tested all combinations of conditions in between those two extremes (see 3.2.1). At the beginning of the study, participants were introduced to ShapeBot and told that it will make recommendations that participants could choose to agree or disagree with in their final decision.

Participants experienced the task through a web interface shown in Figure 1. There were in total 42 trials (7 trials per object category), randomly ordered. For each task trial, participants followed these steps: 1) the object is shown and participants make their initial guess (Figure 1a); 2) the AI shows its recommendation (Figure 1b); and 3) the participant makes a final decision that either agrees with or rejects the AI's recommendation; 4) feedback is shown indicating whether the AI's recommendation and participant's final guess were correct (Figure 1c) or incorrect (Figure 1d) as well as the number of points earned (up to 2 points: 1 point for a correct first guess + 1 point for a correct final guess).

3.2 Study Design

We designed a between-subject experiment with complementary expertise level as the factor to investigate how it affects team performance, behavioral reliance, and subjective trust. The study design and task were reviewed and approved by a third party outside our research group according to our organization's standard protocol.

3.2.1 Independent Variable. The independent variable in this study, degree of complementary expertise, consisted of four levels. Participants naturally had perfect expertise in identifying the 3 types of Regular shapes (i.e., they can correctly identify 100% of examples), but had no expertise in the new unfamiliar Fake shapes (i.e., random

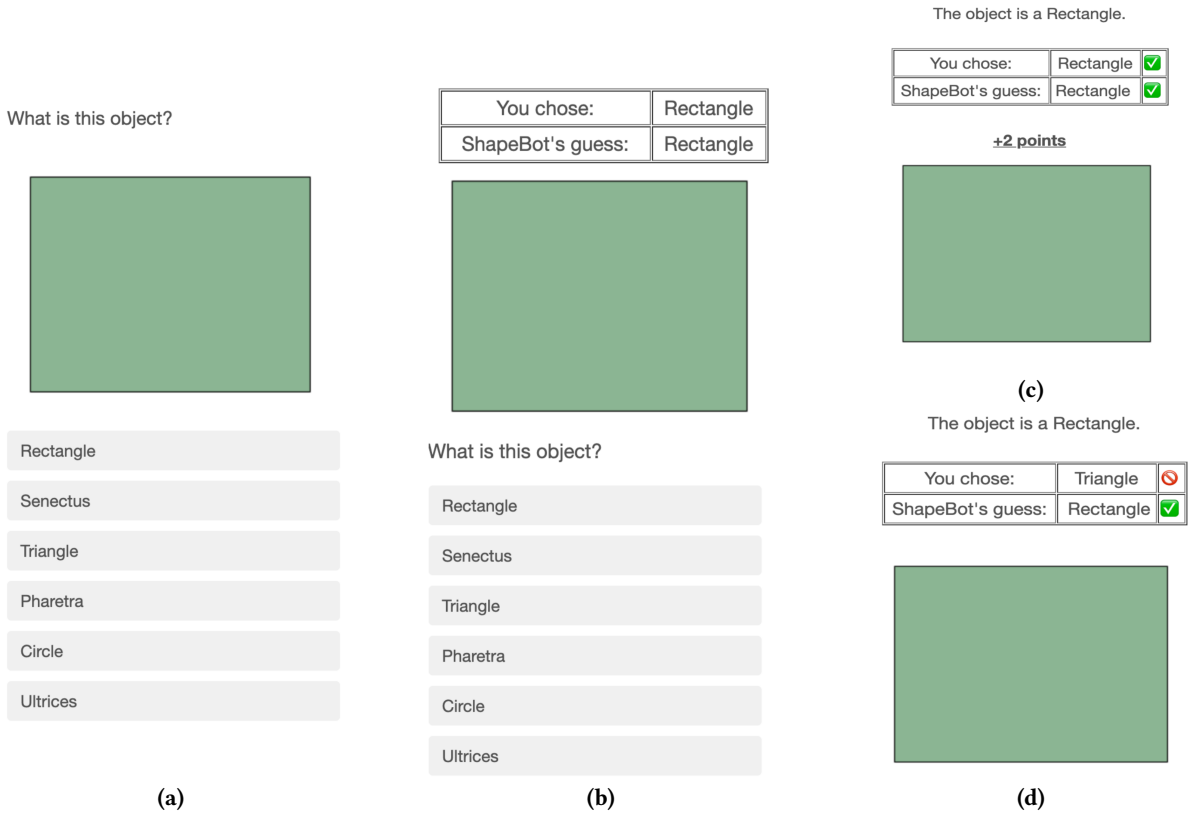


Figure 1: The task interface for study 1. Participants first guessed the object's category (a), were shown the AI's guess (b), and then made their final decision. Participants then saw a feedback screen indicating the correctness of their decision, the AI's guess, as well as bonus points awarded for the task, if applicable (c and d).

guessing). To make the expertise symmetric, we designed the AI assistant, ShapeBot, to have perfect expertise in an equal number (3) of shape categories as the participant. However, ShapeBot's expertise is distributed among the Regular and Fake shape categories differently according to the level of complementary expertise in each experimental condition. In the lowest level (0), there was a complete overlap in the expertise, with ShapeBot sharing the exact same expertise in the three Regular shape categories as the participant and identifying the three Fake shapes poorly (correctly identifying them at chance or roughly 1/3 of trials). In the highest level (3), ShapeBot perfectly complemented the participant and had expertise only in the three Fake shape categories, and identified the three Regular shapes poorly (correctly identifying them at chance or roughly 1/3 of trials). See Table 1 for details.

3.2.2 Dependent Variables. The dependent variables in this study include performance, reliance behavior, and trust.

- **Performance**: We used two performance metrics for this study: first guess performance, representing how often participants guessed the shape correctly before seeing the AI's recommendation, and final decision performance, representing how often participants guessed the shape correctly after seeing the AI's recommendation.

Table 1: Levels of Complementary Expertise

	ShapeBot correctly identifies	
Complementarity Level	Regular shapes[a]	Fake shapes [b]
0 (None)	3 of 3	0 of 3
1 (Small)	2 of 3	1 of 3
2 (Moderate)	1 of 3	2 of 3
3 (Complete)	0 of 3	3 of 3

[a]Easily recognized by humans
[b]Not easily recognized by humans

- **Reliance Behavior**: To measure participants' reliance on the AI, we used two behavioral indicators, agreement frequency and switch-to-agree frequency, defined below.
  (1) Agreement Frequency: how often the participant's final decision agreed with the AI system's recommendation.
  (2) Switch-to-agree Frequency: how often participants changed their first guess to match the AI system's recommendation as their final decision.
- **Trust**: Subjective trust was measured using the Scale of Trust in Automated Systems developed by Muir & Moray

[34]. The 7-item scale comprises six dimensions measured with a Likert scale: competence, predictability, dependability, responsibility, reliability, and faith. Responses were summed for a composite score of self-reported subjective trust.

3.3 Participants

We recruited a total of 178 participants from Amazon Mechanical Turk. To ensure data quality, we included only workers whose prior task approval rating of at least 95% and had a minimum of 1,000 approved tasks. After removing responses that indicated inattentive participants (those who made more than 2 mistakes when identifying Regular shapes), we were left with data from 160 participants.

For participating in the study, each respondent received a payment consisting of a base rate of $1.50 and a performance-based bonus payment, which depended on the total points earned (1 point =$0.02) for correctly identifying shapes. Participants received a median bonus of $1.20 and took a median of 11.12 minutes to complete the task, for a median hourly rate of $14.57.

3.4 Study Procedure

Upon accepting the Human Intelligence Task (HIT) on Amazon Mechanical Turk, participants were directed to our survey on Qualtrics and were briefed on the experiment's purpose and that their participation was voluntary. After signing a consent form, they read the task instructions and scoring scheme. To help participants understand the task procedure, they were given five training trials. In each trial, six objects (that were different from the objects in the main part of the task) were shown including three Regular shapes (Square, Pentagon, Ellipse) and three Fake shapes (Pretium, Sceleris, Trapezium). After finishing all training trials, participants were asked if they understood the task procedure and proceeded to the main task.

After completing the 42 randomly-ordered task trials, participants were required to rate their trust and reliance in the AI system, and to estimate the AI system's ability to identify objects. Further, participants completed a questionnaire that collected information on their age, gender, education, and race/ethnicity.

3.5 Results

To test our hypotheses, we used the Mixed Linear Model package in SPSS Statistics 26 to conduct one-way ANOVAs with the level of complementary expertise as the main factor on the dependent variables. The alpha level was set at 0.05 for statistical tests, and a Bonferroni alpha correction was used for post hoc comparisons. For the sake of brevity, we focus on statistically significant results, and selectively report statistically insignificant results to address our specific hypotheses.

3.5.1 Manipulation Check. To verify that participants were able to recognize the AI system's expertise in identifying different shapes, participants were asked to rate how well ShapeBot correctly identified objects from each shape type on a 7-item scale (Not well at all to Extremely Well). Results confirmed that the manipulation of complementary expertise degree affected the participants as intended. Specifically, participants' rating of ShapeBot's ability to identify the Regular shapes decreased as the level of complementary expertise increased ($F(3,159)=28.507$, $p<0.001$, $\eta^2=0.354$), and the ability in identifying the unfamiliar Fake shapes increased as the level of complementary expertise increased ($F(3,159)=35.763$, $p<0.001$, $\eta^2=0.407$).

3.5.2 Performance. Performance was measured using two indicators: first guess and final decision performances. For Regular shapes, not surprisingly there were no significant difference in these two indicators across the expertise conditions because participants relied solely on their innate ability to recognize regular shapes (see Appendix A.2.1 for details).

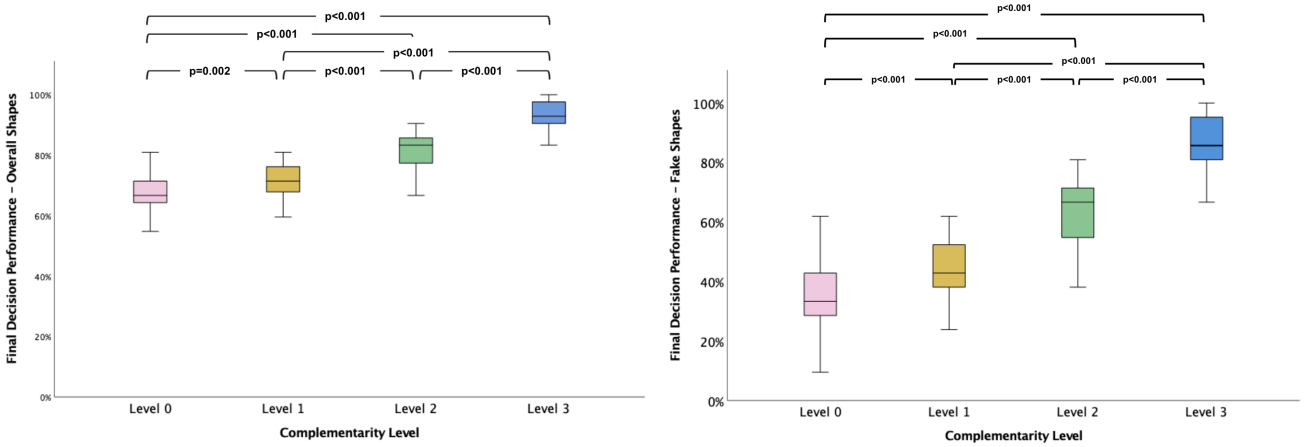
First Guess Performance: We calculated the percentage of trials in which the participant correctly guessed the correct shape *before* seeing the AI's recommendation. Participants are expected to perform well on this measure only if they innately can identify the shape (e.g., Regular shapes) or have learned the shape after seeing similar examples in previous trials. The results combining all shape categories showed that there were no significant differences ($F(3, 159)=0.281$, $p=0.839$, $\eta^2=0.005$) across different levels of complementary expertise. Likewise, the results for only the Fake shapes also did not reveal any significant effects of complementarity on the ability to guess correctly on the first guess ($F(3, 159)=0.339$, $p=0.797$, $\eta^2=0.006$). As expected, participants were universally poor at guessing the Fake shapes at approximately chance (1/3) level.

Final Decision Performance: We calculated the percentage of trials in which the participant correctly selected the correct shape *after* seeing the AI's recommendation, which represents the main outcome of the AI-assisted decision making. The results combining all shape categories showed there was a significant effect of the level of complementary expertise on the final decision performance ($F(3, 159)=145.732$, $p<0.001$, $\eta^2=0.737$), with post hoc comparisons with a Bonferroni adjustment revealing that all conditions were significantly higher than the others. Performance increased as the level of complementarity increased as shown in Figure 2a.

Considering only the Fake shapes, the results showed that the level of complementary expertise had a significant effect on the final decision of Fake shape categories ($F(3, 159)=150.537$, $p<0.001$, $\eta^2=0.743$), with post hoc comparisons revealing that all conditions significantly different than the others. As shown in Figure 2b, participants had more correct final decisions as the level of complementary expertise increased. As the AI's expertise in Fake shapes increasingly complemented the human's expertise in Regular shapes, the better the human-AI team performed in correctly identifying the shape of the object.

Although their first guesses were not impacted by level of the AI's complementary expertise, human-AI team performance on the final decision (after seeing the recommendation of the AI) significantly increased as expertise complementary increased, particularly for cases in which the participant had low expertise (the Fake shape categories). **Therefore, H1, which hypothesized that increased complementary expertise degree leads to higher human-AI team performance, is supported.**

3.5.3 Reliance Behaviors. Reliance was measured using two behaviors: agreement frequency and switch-to-agree frequency. For trials showing Regular shapes, the results of these two measures revealed that people tended to rely on their own expertise to make



(a) All Shapes **(b) Fake shapes**

Figure 2: Final decision performance (in percent) across four levels of complementary expertise (Level 0: Completely overlapping expertise; Level 3: Completely complementary expertise) for correctly identifying: (a) All shapes; (b) Fake shapes. As the level of complementary expertise increased, participants were able to correctly identify the object more often.

the first and final guesses, as we had expected because they were already familiar with Circles, Triangles, and Rectangles (see Appendix A.2.2 for details). In the following sections, we focus on the results for all shapes and Fake shapes only, because participants need AI assistance for Fake shapes.

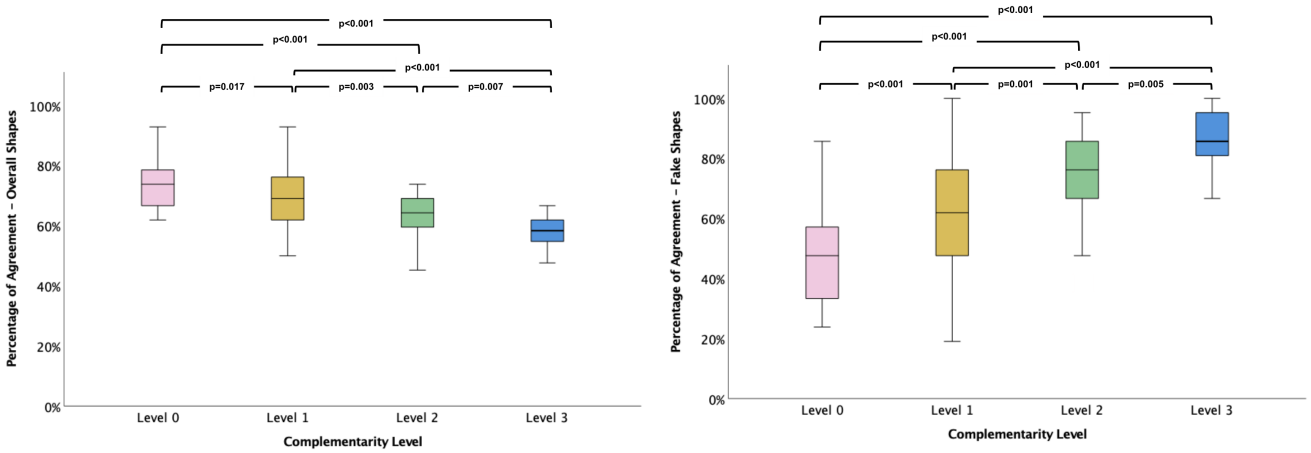
Agreement Frequency: The trials in which the participant's final guess was the same as the AI's recommendation were counted as agreement. First looking at all the trials (both Regular and Fake shapes), there was a significant effect of complementary expertise degree on agreement frequency ($F(3,159)=32.900$, $p<0.001$, $\eta^2=0.388$), as shown in Figure 3a. Post hoc comparisons with a Bonferroni adjustment showed that all complementarity levels were significantly different from the others. As the level of complementarity increased, participants tended to agree with the AI less often (Figure 3a).

For Fake shapes only, participants' agreement with the AI's recommendation also increased significantly as the level of complementary expertise increased ($F(3,159)=45.417$, $p<0.001$, $\eta^2=0.466$), as shown in Figure 3b. Post hoc comparisons showed that all complementarity levels were significantly different from the others. As the AI was better at Fake shapes (but made poorer recommendations for the Regular shapes), and participants tended to agree with and rely on the AI's recommendation for Fake shapes.

We also investigated how the degree of complementary expertise affected agreement with the AI over time across multiple trials of shape identification. Participants were shown seven examples of each of the three Fake shapes in random order. We used the trial order number (1 through 7) as a factor in our analysis to determine whether agreement changed across trials and focus on looking at its interaction with the level of complementary expertise. The results show that people's agreement over time differed significantly across different levels of complementary expertise ($F(18,3359)=2.958$, $p=0.007$, $\eta^2=0.005$). More specifically, in the early trials, the agreement frequency with the AI is similar across all levels of complementary expertise (because participants have not had a chance to learn the expertise of the AI). However, as participants encountered more trials and more recommendations by the AI, they agreed more frequently over time ($\beta=0.117$, $p=0.005$) for the highest complementary level (Level 3), shown as the top red line in Figure 4, while in contrast, they agreed less frequently over time ($\beta=-0.145$, $p<0.001$) for the lowest complementarity level (Level 0), as shown in the bottom blue line in Figure 4. The pattern provides evidence that when the AI demonstrates its expertise (or lack thereof) in areas of the task (in this case, Fake shapes) that the participant has little to no expertise, the participant is able to discern the AI's (non-)expertise and adjust their reliance on its recommendations over time, instead of just blindly agreeing with the AI.

Switch-to-agree Frequency: We investigated another measure of reliance on the AI: how often participants switched their answers to match the AI after seeing the AI's recommendation. This measure tracks the affirmative behavior to both agree and rely on the AI's recommendation that differed from the participant's initial guess. For all shapes combined, the results showed that complementary expertise degree had a significant effect on the switch-to-agree frequency ($F(3,159)=27.621$, $p<0.001$, $\eta^2=0.347$), as shown in Figure 5a. As the level of complementary expertise increased, the amount of switching to agree with the AI also increased. Post hoc analyses revealed that the lowest level of complementary (level 0) had significantly lower amount of switching than the other levels.

A similar pattern was found for identifying only the Fake shapes, where the switch-to-agree frequency was significantly influenced by the level of complementary expertise ($F(3,159)=29.101$, $p<0.001$, $\eta^2=0.359$), as shown in Figure 5b. When using the AI system with the lowest level of complementary expertise (level 0), participants switched less often than in all others conditions. As the AI better complemented the expertise of the participant with better recommendations in the Fake shapes, participants were able to notice this and switch their answers after seeing the AI's recommendation.



(a) All Shapes **(b) Fake shapes**

Figure 3: Agreement frequency (in percent) across four levels of complementary expertise (Level 0: Completely overlapping expertise; Level 3: Completely complementary expertise) when identifying: (a) All shapes; (b) Fake shapes. For All Shapes that include both Regular and Fake Shapes, participants disagreed more often with the AI's recommendation as the AI's expertise in Regular shapes decreased, as expected. For Fake Shapes only, agreement with the AI's recommendation increased as the AI's expertise in Fake shapes increased (more complementary to participant's expertise), also as expected.

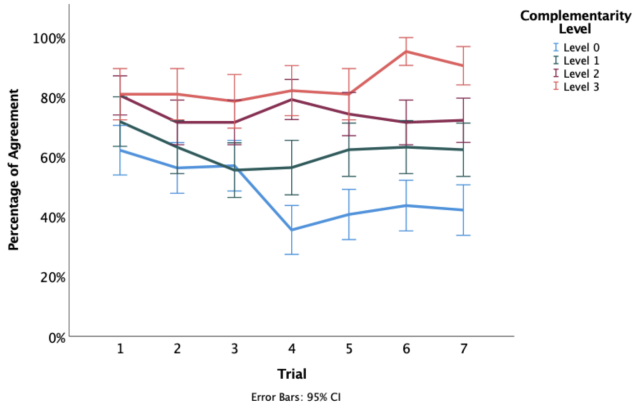
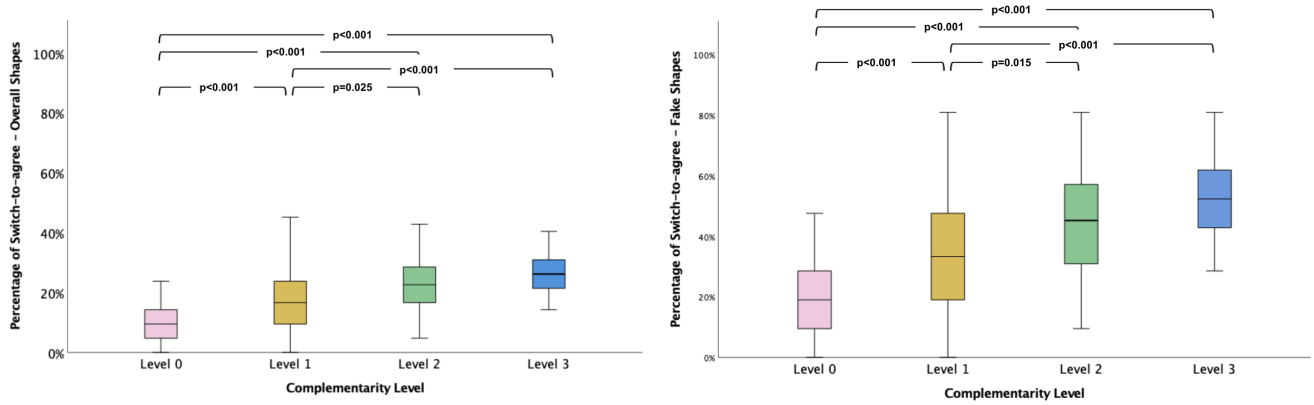


Figure 4: The change of agreement frequency over trials across four complementarity levels (Fake shapes). As participants encountered more trials and more recommendations by the AI, they agreed with the AI more frequently over time for highest complementary level (Level 3), as shown the top red line), while in contrast, they agreed less frequently over time for the lowest complementarity level (Level 0), as shown in the bottom blue line.

We also considered how often participants, after seeing the AI's recommendation, switched their initial answer to another choice that did *not* match the AI's recommendation. In fact in our data, participants rarely exhibited this behavior. Including these small number of cases into the analysis yielded nearly identical results (see Switch Frequency in Appendix A.2.3). Therefore, when participants switched their guess, it was to align their guess to the AI's recommendation (rather than choosing a third option that neither the AI nor the participant guessed initially).

Overall, participants behaved differently in identifying the Regular and Fake shapes. For Regular shapes, as expected, people relied solely on their own expertise to make decisions. However, for identifying unfamiliar Fake shapes, the results showed that participants relied on the AI's recommendation to make their final decision. Moreover, two behavior measures of reliance, agreement frequency and switch-to-agree frequency, both increased along with the increased level of complementary expertise. Participants were able to perceive over time when the AI is correct and can be relied upon, and align their final answer to match the AI appropriately. **Thus, H2, which stated that people rely on the AI system more with increased complementary expertise degree, is supported, particularly for cases in which the participant has low expertise (Fake shapes) and can benefit from AI assistance.**

3.5.4 Trust. We measured subjective trust in the AI system using a validated questionnaire developed by Muir & Moray [34]. The results showed that there was a significant effect of complementary expertise on trust ($F(3,159)=5.141$, $p=0.002$, $\eta^2=0.090$). See Figure 6. Post hoc comparisons revealed that participants in complementarity level 0 (mean=3.648) had lower subjective trust ratings than those assigned to level 3 (mean=4.689, $p<0.001$) condition. However, subjective trust did not monotonically increase with higher levels of complementary expertise, with no significant difference between two intermediate levels 1 (mean=4.231) and 2 (mean=4.089) in which the AI had partial complementary expertise with the participant. **Thus, H3, which hypothesized that trust in AI system increases along with the increased complementary expertise degree, was partially supported.**



(a) All Shapes **(b) Fake shapes**

Figure 5: Switch-to-agree frequency (in percent) across four levels of complementary expertise (Level 0: Completely overlapping expertise; Level 3: Completely complementary expertise) for identifying: (a) All shapes; (b) Fake shapes. After making their first guess and then seeing the AI's recommendation, participants switched their guess to agree with the AI's recommendation more often since the AI's expertise in Fake Shapes better complemented the participant's expertise in Regular shapes.

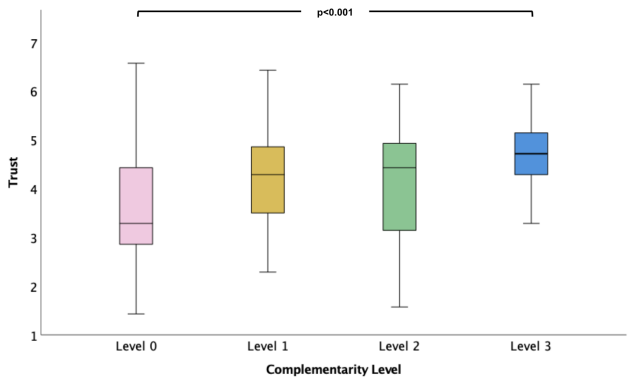


Figure 6: Subjective trust ratings (from the Scale of Trust in Automated Systems [34]) across four levels of complementary expertise (Level 0: Completely overlapping expertise; Level 3: Completely complementary expertise). Overall, there is a trend that subjective trust increases as the level of complementary expertise increased. Post hoc comparisons reveal that the only pairwise difference was between the lowest and highest levels of complementary expertise, with no statistical difference among the levels where the AI exhibited expertise in at least one shape category with which the participants needed help (i.e., a Fake shape).

Summary The results show that when the AI had greater complementary expertise (i.e., less overlap with the participant's expertise [Regular shapes] but more expertise with Fake shapes, with which participants were unfamiliar), participants were able to correctly identify more shapes and achieve higher performance. Further analysis showed that participants also switched their answers to match the AI's recommendation more often as the level of complementary expertise increased, indicating that participants were able to detect the AI's expertise in Fake shapes over time and rely on it more. However, ratings of subjective trust did not follow the same clear pattern of increase, with no significant difference between two intermediate levels (2 & 3) in which the AI had partial complementary expertise with the participant. In these conditions where there is the bot is good at some Regular shape categories and some Fake shape categories, it is more difficult for the participant to detect for which cases the AI has expertise or not. Thus, there is an opportunity for the AI to better communicate its expertise to their human partner. In our next study, we investigate how embracing or distancing language in explanations can help humans calibrate better to the expertise of their AI partner.

4 STUDY 2: EMBRACING AND DISTANCING LANGUAGE IN EXPLANATIONS

The results of study 1 showed that participants had difficulty calibrating their trust in the AI in cases of partial complementary expertise. Since in most real world settings expertise will not overlap or complement perfectly, it is important to understand how to close this gap.

Linguistic manipulation of psychological distance (such as the use of embracing or distancing language) is a key method to help communicate confidence and support an appropriate level of trust between parties [31]. In our second study, we leverage this insight to investigate the following research question:

- Research Question 2 (RQ2): How does embracing or distancing language in AI explanations affect trust, reliance behavior, and team performance in AI-assisted decision making across different levels of complementary expertise?

To answer this question, we formulated the following hypotheses:

- **H4a:** *Embracing language* in AI explanations will communicate the AI system's *high confidence* in its recommendation, leading to *higher* reliance on the AI's recommendation.
- **H4b:** *Distancing language* in AI explanations will communicate the AI system's *low confidence* in its recommendation, leading to *lower* reliance on the AI's recommendation.
- **H5:** Higher reliance on an AI system with *high* expertise (i.e., mostly makes correct recommendations) will result in *higher* team performance, whereas higher reliance on an AI system with *low* expertise (i.e., mostly makes incorrect recommendations) will result in *lower* team performance.
- **H6:** Providing an explanation that includes the features and logic used by the AI will help participants learn new expertise.
- **H7:** Embracing language in AI explanations will increase the human partner's subjective trust in the AI system.

In other words, we hypothesized that using embracing language in the AI's explanation will lead to higher reliance and distancing language will lead to lower reliance on the AI's recommendation. An open question we aimed to answer was: does embracing/distancing language increase/decrease reliance, regardless of whether the AI's recommendation was ultimately correct or incorrect? If the AI's recommendation is in fact correct, then higher reliance will result in better team performance. However, in real life, the AI's recommendation may or may not actually be correct, and under-reliance on a correct recommendation or over-reliance on an incorrect recommendation will lead to poorer team performance. Next we describe how we enhanced the design of our first study to address these hypotheses.

4.1 Task and User Interface Design

We used the same object identification task as in Study 1 with the same three Regular shapes types (Rectangle, Circle, Triangle) and the same three Fake shape types (Senectus, Pharetra, Ultrices). The task steps remained the same (Figure 1) with the only change being the addition of a dialog box (Figure 7) with an image of ShapeBot and a written explanation shown with the ShapeBot's recommendation.

The explanations were designed to follow a standard template:

[*Point-of-view (POV)] [Belief marker*] *this is a* [*AI recommendation*] *because* [*Point-of-view (POV)] [Belief marker*] *it has* [*reasons*].

For example, "I think this is a Triangle because I believe it has three sides," which uses a 1st person POV ("I") and distancing belief marker ("think...believe")

When ShapeBot makes a wrong recommendation, it is a classification error–not a perception error, so even though the recommended shape is ultimately incorrect, the reason always accurately describes the object shown. For example, if shown a Triangle but ShapeBot incorrectly classifies it as a Circle, it would say "I think this is a Circle because I believe it has three sides."

4.2 Study Design

We designed a factorial between-subject experiment with complementarity level (4 levels), point-of-view (POV) (2 levels), and belief marker (2 levels) as the factors to investigate how it affects team performance, behavioral reliance, and subjective trust. A control condition with no explanation was also included. Each participant was assigned to a version of ShapeBot that had a particular combination of expertise across Regular and Fake shapes and explained its recommendation using a particular combination of POV and belief markers, except in the control condition that had no explanation text at all. The study design and task were reviewed and approved by a third party outside our research group according to our organization's standard protocol.

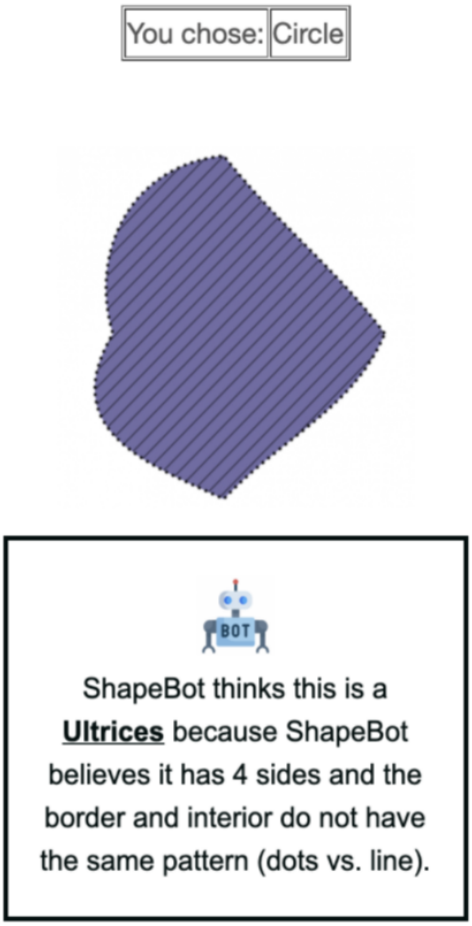


Figure 7: The task interface shown to participants, including the AI's recommendation and explanation, in Study 2.

4.2.1 Independent Variables. The manipulation of complementarity levels in study 2 was the same as the study 1, which has four levels as shown in Table 1. Two other independent variables, point-of-view (POV) and belief marker, were used to manipulate language distance in this study. Point-of-view had two levels: the 1st-person POV ("I") and the 3rd-person POV ("ShapeBot"). Belief marker also had two levels: embracing markers ("knows...realizes") and distancing markers ("thinks...believes"). Combining the POV and belief marker for different distancing levels resulted in four explanation

conditions plus one no-explanation control condition. See Table 2 for details.

4.2.2 Dependent Variables & Analysis. The dependent variables and their measurements in this Study 2 are the same as Study 1, including performance (the percent of shapes correctly identified), reliance behavior (agreement percentage and switch-to-agree percentage), and subjective trust (as measured by the Trust in Automated Systems scale developed by Muir & Moray [34]).

In Study 1, we observed that participants mostly ignored ShapeBot's recommendations for the Regular shapes, so in this study, we focus our analysis only on the trials with Fake shapes because these trials are where participants may actually consider ShapeBot's recommendations.

Another important difference from the previous analysis is that we are interested in understanding how embracing or distancing language (as operationalized with POV and belief markers) increases or decreases reliance on the AI's recommendation, both for cases when the AI is an "Expert" (i.e., mostly makes correct recommendations) and cases when the AI is a "Non-Expert" (i.e., mostly makes incorrect recommendations). Recall that ShapeBot is assigned to be an expert between zero to three Fake shapes categories, in different conditions of complementary expertise. Thus we analyze the effect of explanations on two subsets of data:

- **Expert AI cases**: the trials that present shapes that ShapeBot identifies correctly 100% of the time.
- **Non-Expert AI cases**: the trials that present shapes that ShapeBot identifies poorly being correct approximately at the level of random guessing.

According to Hypothesis H5, embracing language leading to higher reliance in Expert AI trials should result in higher team performance. However, embracing language leading to higher reliance in Non-Expert trials should result in lower team performance.

4.3 Participants

For Study 2, a total of 395 participants were recruited from Amazon Mechanical Turk and fully completed the study task, and 37 inattentive responses were excluded for making more than two mistakes when identifying the Regular shapes, resulting in a dataset with 358 participants. Participants were paid a base rate ($1.50) and a performance bonus payment based on the final points earned (1 point =$0.02). Participants received an average bonus of $1.26 and took a median of 14.06 minutes to complete the task, for a median hourly rate of $11.78. All participants provided informed consent prior to beginning the study.

4.4 Results

To test our hypotheses, we used the Mixed Linear Model package in SPSS Statistics 26, with an alpha level set at 0.05 and Bonferroni alpha correction for post-hoc comparisons. For the sake of brevity, we focus on statistically significant results, and selectively report statistically insignificant results to address our specific hypotheses. Also, we focus only on Fake shapes because participants could benefit from ShapeBot's recommendation for only Fake shapes based on the result of Study 1. To evaluate the independent effects of using POV (1st person or 3rd person) and belief markers (embracing or distancing) on each the dependent variables, we conduct a two-way ANOVA. To identify whether each linguistic factor had an effect significantly different than having no explanation at all, we conducted a one-way ANOVA with the three levels (two levels of the linguistic factor plus no explanation.)

4.4.1 Manipulation Check. Similar to Study 1, the manipulation of complementary expertise degree was effective—participants recognized the differences across conditions. Specifically, ratings of how well the AI system performed in identifying the Regular shapes decreased ($F(3, 357)=56.556$, $p<0.001$, $\eta^2=0.324$), and in identifying the Fake shapes increased ($F(3, 357)=98.048$, $p<0.001$, $\eta^2=0.454$) as the level of complementary expertise increased.

In terms of the impact of the level of complementary expertise, the results corroborate the finding from Experiment 1 that people tend to have higher final guess performance ($F(3, 357)=306.537$, $p<0.001$, $\eta^2=0.722$), reliance (Agreement Frequency: $F(3, 357)=59.178$, $p<0.001$, $\eta^2=0.334$; Switch-to-agree Frequency: $F(3, 357)=46.404$, $p<0.001$, $\eta^2=0.282$), and subjective trust ($F(3, 357)=5.377$, $p=0.001$, $\eta^2=0.044$) when interacting with an increased complementary expertise degree AI system. In the following sections, we focus only on the effect of using embracing or distancing language in the explanation.

4.4.2 Performance. **First Guess Performance** Participants were given a chance to make a first guess at the shape before seeing ShapeBot's recommendation. Given that Fake shapes were unfamiliar to participants at the beginning of the task, how well they correctly identified the shape on their first guess (i.e., % correct) can indicate participants' ability to learn to identify these novel shapes based on previous cases shown and the explanations provided by ShapeBot.

(1) **Expert AI cases**: Considering the shapes that ShapeBot was assigned to identify correctly 100%, we first performed a two-way ANOVA to measure the effects of the point-of-view (POV) and belief markers on first guess performance. Both POV ($F(1, 2953)=5.724$, $p=0.017$, $\eta^2=0.002$) and belief marker ($F(1, 2953)=30.745$, $p<0.001$, $\eta^2=0.010$) had significant effects on first guess performance. Post hoc analysis with a Bonferroni adjustment revealed that participants made better first guesses with an explanation using the 1st person POV (mean=43.132%) than the 3rd person POV (mean=41.642%) (Figure 8). An explanation using the embracing belief marker (mean=44.113%) led to a higher first guess performance than with the distancing belief marker (mean=40.661%) (Figure 9). No evidence was found to support the interaction between POV and markers on first guess performance.
To identify whether each POV and belief marker had an effect significantly different than having no explanation at all, we conducted a one-way ANOVA with the three levels (two levels of the linguistic factor plus a third level of no explanation). For POV, a one-way ANOVA with three levels (1st person, 3rd person, no explanation) showed significant differences among the three levels ($F(2, 3723)=66.914$, $p<0.001$, $\eta^2=0.035$) (Figure 10). Post hoc analysis with a Bonferroni adjustment revealed that providing explanations with either the 1st person POV (mean=42.857%) or the 3rd person POV

Table 2: Examples of combining POV and Belief Marker in the language used in explanations.

		Point of View	
		1st Person POV	3rd Person POV
Belief Marker	Embracing Marker	*I* **know** this is a [AI recommendation] because *I* **realize** it has [reasons].	*ShapeBot* **knows** this is a [AI recommendation] because *ShapeBot* **realizes** it has [reasons].
	Distancing Marker	*I* **think** this is a [AI recommendation] because *I* **believe** it has [reasons].	*ShapeBot* **thinks** this is a [AI recommendation] because *ShapeBot* **believes** it has [reasons].

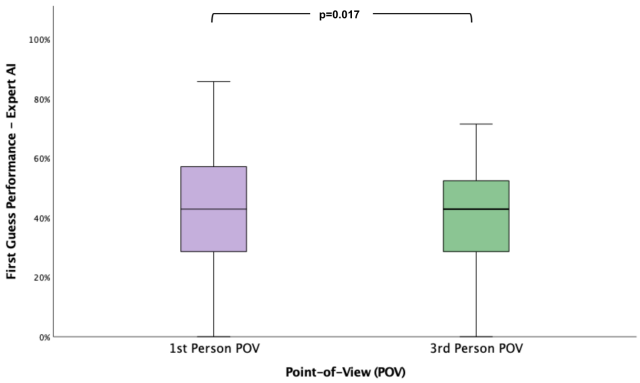


Figure 8: First guess performance (in percent) between two point-of-view levels (Expert AI cases). The participants makes a first guess before seeing the AI's recommendation and thus represents how well they have learned to identify the shape on their own based on previous cases shown. Participants learned the novel Fake shapes better with 1st person POV than with 3rd person POV.

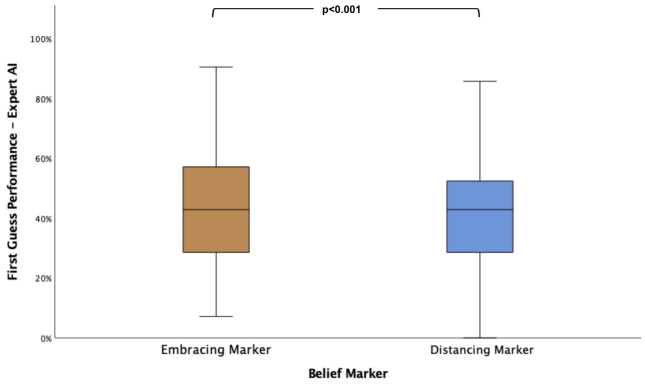


Figure 9: First guess performance (in percent) between two belief marker levels (Expert AI cases). The participants makes a first guess before seeing the AI's recommendation and thus represents how well they have learned to identify the shape on their own based on previous cases shown. Participants learned the novel Fake shapes better with embracing marker than with distancing marker.

(mean=41.661%) led to significantly higher first guess performance than having no explanation (mean=34.805%). For belief markers, a one-way ANOVA with three levels (embracing, distancing, no explanation) showed significant differences among the three levels ($F(2, 3723)=81.106$, $p<0.001$, $\eta^2=0.042$) (Figure 11). Post hoc analysis revealed all levels were significantly different, with the highest first guess performance using the embracing marker (mean=44.006%), followed by the distancing marker (mean=40.679%), and finally by no explanation (mean=34.805%).

(2) **Non-Expert AI cases**: Considering the trials in which the AI was assigned to identify the object poorly (being correct only 2 out of 7 times), a two-way ANOVA showed a statistically significant effect of belief marker ($F(1, 2967)=11.232$, $p<0.001$, $\eta^2=0.004$) on first guess performance. Post hoc analysis with a Bonferroni adjustment revealed that participants were more likely to correctly identify the shape on their first guess with explanations using the distancing belief marker (mean=39.661%) than the embracing belief marker (mean=37.683%) (Figure 12). We also found a significant interaction effect ($F(1, 2967)=4.930$, $p=0.026$, $\eta^2=0.002$) between belief markers and POV on the first guess performance (Figure 13). We note that the most embracing combination (embracing belief marker and 1st person POV) resulted in the worst performance among the explanation conditions.
For POV, a one-way ANOVA with three levels (1st person, 3rd person, no explanation) showed significant differences ($F(2, 3814)=10.428$, $p<0.001$, $\eta^2=0.005$). Post hoc analysis with a Bonferroni adjustment revealed that providing explanations with either the 1st person POV (mean=38.707%) or the 3rd person POV (mean=38.919%) led to significantly higher first guess performance than when no explanation was provided (mean=36.009%) (Figure 14). In addition, for belief marker, a one-way ANOVA with three levels (embracing, distancing, no explanation) showed significant differences among levels ($F(2, 3814)=15.797$, $p<0.001$, $\eta^2=0.008$). Post hoc analysis revealed all levels were significantly different, where participants had the highest first guess performance with an explanation using the distancing marker (mean=39.689%), followed by the embracing marker (mean=37.776%), and finally with no explanation having the lowest first guess performance (mean=36.009%) (Figure 15).

As expected, providing AI explanations that expose the relevant features of the shapes useful for categorizing them helped participants pay attention to these features and learn to identify the shapes themselves before seeing ShapeBot's recommendation

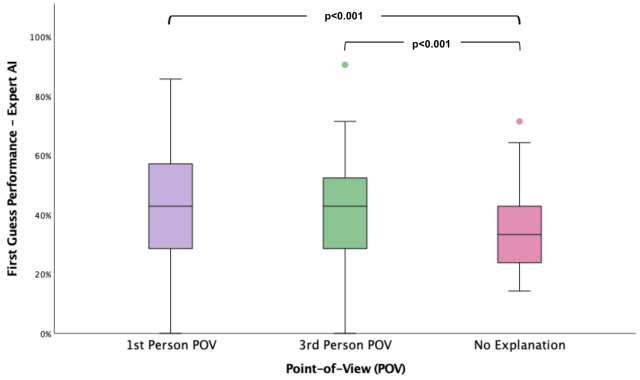


Figure 10: First guess performance (in percent) across point-of-view levels and no explanation (Expert AI cases). Participants learned the novel Fake shapes significantly better with explanations using either the 1st person POV or 3rd person POV than with No Explanation.

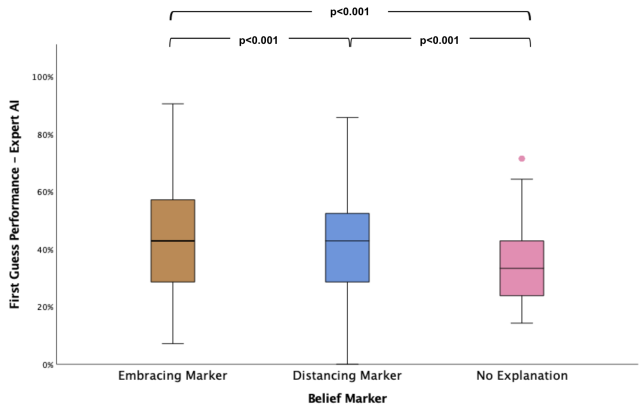


Figure 11: First guess performance (in percent) across belief marker levels and no explanation (Expert AI cases). Using the embracing belief marker in explanation text resulted in the highest learning of novel shapes by participants, followed by the distancing belief marker, and then by no explanation with the lowest.

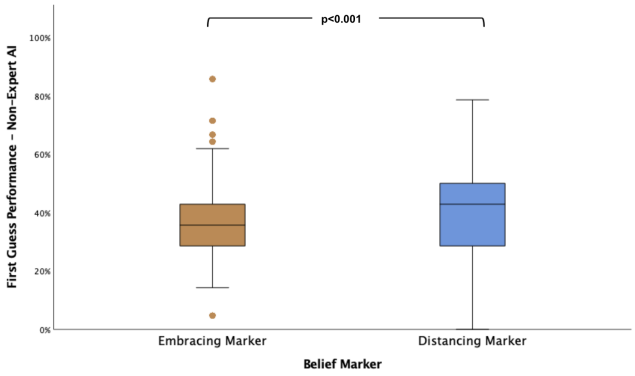


Figure 12: First guess performance (in percent) between two belief marker levels (Non-Expert AI cases). Participants learned the novel Fake shapes better with the distancing marker than the embracing marker.

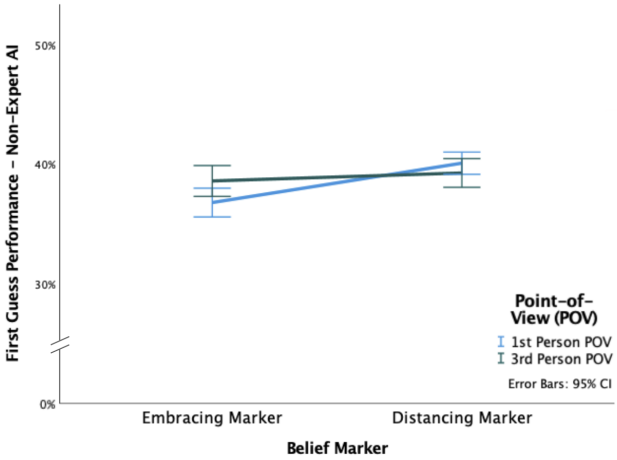


Figure 13: The two-way interaction between POV and belief marker on first guess performance (Non-Expert AI cases). The explanation style using the most embracing combination (1st person POV + embracing belief marker) had the lowest first guess performance, that is, the least amount of learning. Embracing a non-expert AI resulted in learning incorrect categorizations.

in each trial. Furthermore, the results show that embracing language (using 1st person POV, "I" and embracing belief markers, "know...realize") with an Expert AI helped participants learn the shapes better than distancing language. For a Non-Expert AI that displayed incorrect recommendations, we saw the expected higher learning when using the distancing language (using distancing belief markers, "think...believe") because it helped participants to have lower confidence in the incorrect recommendations. **Thus, H6 which hypothesized that providing an explanation that includes the features and logic used by the AI will help participants learn new expertise is confirmed.**

Final Decision Performance After making their first guess, participants were shown the AI's recommendation along with an explanation and then made their final decision about the type of the shape. How often their final decision correctly identifies the shape is a measure of human-AI team performance.

(1) **Expert AI cases**: When considering the shapes that the AI was assigned to identify correctly 100%, we first performed a two-way ANOVA to measure the effects of the point-of-view (POV) and belief markers on final decision performance. Belief marker ($F(1, 2953)=27.102$, $p<0.001$, $\eta^2=0.009$) and the interaction between POV and belief marker ($F(1, 2953)=12.869$, $p<0.001$, $\eta^2=0.004$) had significant effects on final decision performance. Post hoc analysis with a Bonferroni adjustment revealed that participants correctly identified the shapes more often with explanations using the embracing belief

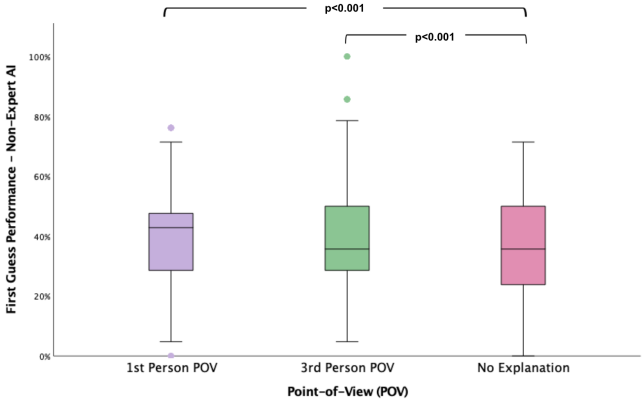


Figure 14: First guess performance (in percent) across point-of-view levels and no explanation (Non-Expert AI cases). Participants learned the novel Fake shapes significantly better either with 1st person POV or 3rd person POV than with No Explanation.

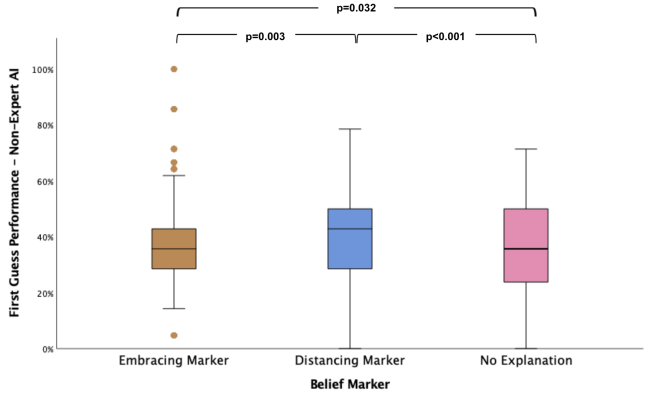


Figure 15: First guess performance (in percent) across belief marker levels and no explanation (Non-Expert AI cases). Using the distancing belief marker in explanation text resulted in the highest learning of novel shapes by participants, followed by the embracing belief marker, and then by no explanation with the lowest.

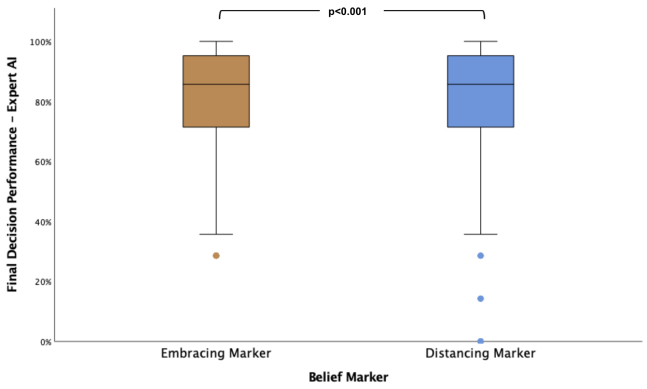


Figure 16: Final decision performance (in percent) between belief marker (Expert AI cases). Providing explanation with embracing marker led to significantly higher number of shapes correctly than distancing belief markers.

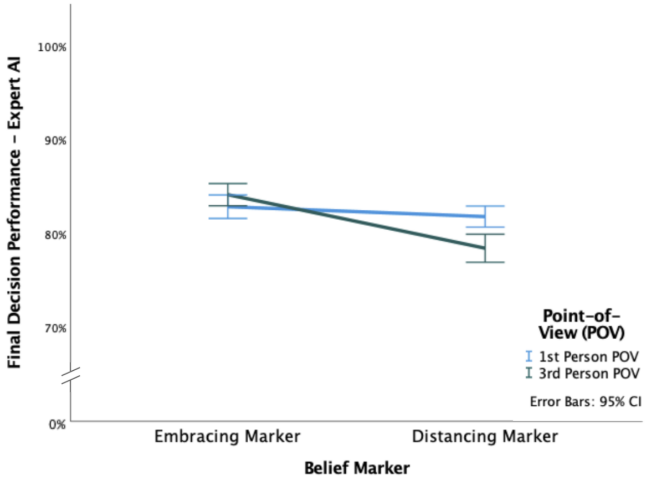


Figure 17: The two-way interaction between POV and belief marker on final decision performance (Expert AI cases). The explanation style using the most distancing combination (3rd person POV + distancing belief marker) had the lowest final decision performance.

marker (mean=83.499%) than using distancing belief marker (mean=80.122%) (Figure 16). We also observed in the two-way interaction (Figure 17) that the most distancing of explanation conditions (combining the distancing belief marker and 3rd person POV) had the lowest percentage of correctly identified shapes. Both factors of belief marker and POV seem to indicate that using distancing language in explanations for an AI that identifies shapes well leads to lower team performance.

To determine whether the effects of POV and belief markers on final decision performance were significantly different than having no explanation at all, we conducted a one-way ANOVA with three levels for each factor. For POV, a one-way ANOVA with three levels (1st person, 3rd person, no explanation) showed significant differences ($F(2, 3723)=10.519$, $p<0.001$, $\eta^2=0.006$). Post hoc analysis with a Bonferroni adjustment revealed that participants had significantly higher final decision performance with explanations using either the 1st person POV (mean=82.264%) or the 3rd person POV (mean=81.329%) than with no explanation (mean=78.571%) (Figure 18). For belief marker, a one-way ANOVA with three levels (embracing, distancing, no explanation) showed significant differences ($F(2, 3723)= 22.205$, $p<0.001$, $\eta^2=0.012$). Post hoc analysis revealed that participants had significantly higher final decision performance with explanations using embracing belief markers (mean=83.561%) than with

either distancing markers (mean= 80.205%) or no explanation (mean=78.571%) (Figure 19).

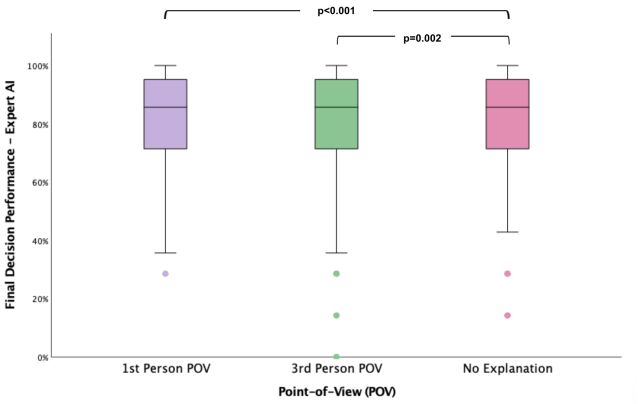


Figure 18: Final decision performance (in percent) across points-of-view levels and no explanation (Expert AI cases). Providing any explanation led to higher learning of the shapes than no explanation at all.

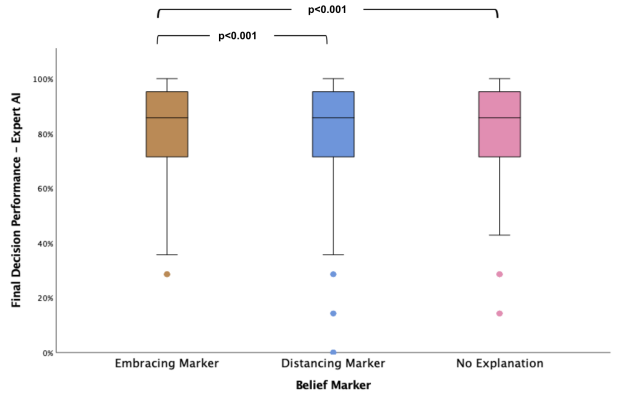


Figure 19: Final decision performance (in percent) across belief markers level and no explanation (Expert AI cases). Providing any explanation with embracing belief marker led to higher final decision performance than distancing belief markers and no explanation at all.

(2) **Non-Expert AI cases**: Considering the trials in which the AI was assigned to identify the object poorly (being correct only 2 out of 7 times), a two-way ANOVA with POV and belief markers did not show any significant effects on final decision performance. Neither POV ($F(1, 2967)=3.480$, $p=0.062$, $\eta^2=0.001$), belief marker ($F(1, 2967)=0.066$, $p=0.798$, $\eta^2<0.001$), nor the two-way interaction between these two factors ($F(1, 2967)=3.221$, $p=0.073$, $\eta^2=0.001$) had any significant impact on the final decision performance. For the (non-significant) effect of POV, final decision performance was highest for the 3rd person POV (mean=38.318%), followed by the 1st person (mean=37.346%). Even though the more distancing 3rd person POV had higher performance than the embracing

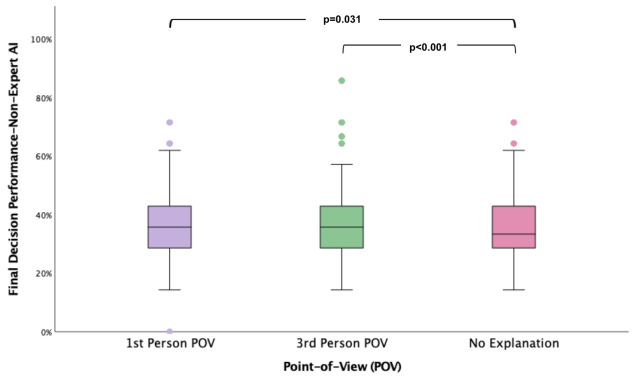


Figure 20: Final decision performance (in percent) across point-of-view levels and no explanation (Non-Expert AI cases). Participants had the better final decision performance with either the 1st person POV or the 3rd person POV than having No Explanation.

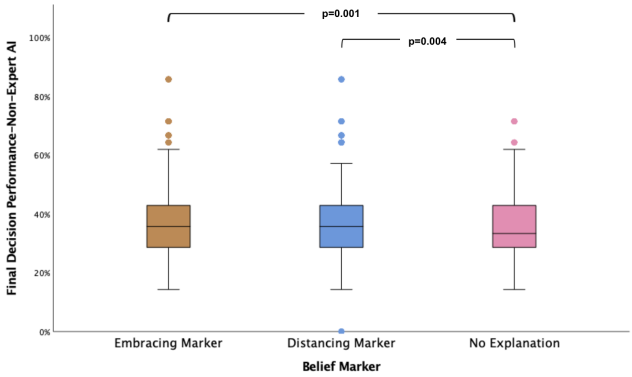


Figure 21: Final decision performance (in percent) across belief marker levels and no explanation (Non-Expert AI cases). Using either the distancing or embracing belief marker in explanation text resulted in higher final decision performance than having no explanation.

1st person POV, as would be expected when casting doubt on an non-expert AI, it was not statistically higher.
However, an analysis using the one-way ANOVA with three levels for each factor showed significant differences among levels for both POV ($F(2, 3814)=8.421$, $p<0.001$, $\eta^2=0.004$) and belief marker ($F(2, 3814)=6.923$, $p<0.001$, $\eta^2=0.004$). Post hoc analysis with a Bonferroni adjustment revealed that having explanations using either the 1st person (mean=37.997%) or 3rd person POV (mean=38.318%) resulted in more shapes correctly identified than having no explanation (mean=35.891%) (Figure 20). Likewise, participants performed better when they had explanations using embracing (mean=37.997%) and distancing belief markers (mean=37.764%) than no explanation at all. (Figure 21).

We found significant effects on task performance (number of shapes correctly identified) for embracing and distancing language

in explanations for cases when the AI performed expertly. In particular, using the embracing belief marker ("knows...believes") led to a higher percentage of shapes identified than using the distancing belief marker. The combination of the distancing belief marker and the 3rd person POV led to the lowest percentage of shapes correctly identified. However, we did not observe the similar effects for the cases when the AI performed poorly. To further explain why embracing and distancing language resulted in higher team performance, we next consider whether embracing language led participants to rely on the AI's recommendation more often.

4.4.3 Reliance Behavior. To measure how much participants relied on the AI's recommendation, we consider how often participants agreed with the AI's recommendation as well as how often they switched their answer to match the AI.

Agreement Frequency: Agreement frequency measures how often participants' final decisions agreed with ShapeBot's recommendations.

(1) **Expert AI cases**: Considering the trials in which the AI was assigned to identify shapes 100% correctly, a two-way ANOVA found that belief marker ($F(1, 2953)=27.102$, $p<0.001$, $\eta^2=0.009$) and interaction between POV and belief marker ($F(1, 2953)=12.869$, $p<0.001$, $\eta^2=0.004$) had significant effects on how often participants agreed with the AI's recommendation. We note here that these factors followed the same pattern of significance as they had on final guess performance. Specifically, post hoc analysis with a Bonferroni adjustment showed that participants agreed with the AI's recommendation more frequently with explanations using embracing markers (mean=83.499%) than with distancing belief markers (mean=80.122%). The interaction between the POV and belief markers also indicated that the lowest agreement frequency was found in the most distancing of explanation condition (3rd person POV and distancing belief markers.)
For POV, a one-way ANOVA with three levels (1st person, 3rd person, no explanation) showed significant differences in agreement frequency ($F(2, 3723)=10.519$, $p<0.001$, $\eta^2=0.006$). Post hoc analysis with a Bonferroni adjustment revealed that participants had significantly higher frequency of agreement with the AI that used explanations with either the 1st person POV (mean=82.264%) or 3rd person POV (mean=81.329%) than with no explanation (mean=78.571%).
For belief marker, a one-way ANOVA with three levels (embracing, distancing, no explanation) showed significant differences in agreement frequency ($F(2, 3723)=22.205$, $p<0.001$, $\eta^2=0.012$). Post hoc analysis revealed that participants had significantly higher agreement with the AI that used explanations with the embracing belief marker (mean=83.561%) than with either distancing marker (mean=80.205%) or no explanation (mean=78.571%).

(2) **Non-Expert AI cases**: Considering the trials in which the AI was assigned to perform poorly, a two-way ANOVA showed that POV ($F(1, 2967)=27.736$, $p<0.001$, $\eta^2=0.009$) and belief marker ($F(1, 2967)=5.017$, $p=0.025$, $\eta^2=0.002$) had the significant effects on agreement frequency, while no evidence was found to support a significance of interaction effect between these two factors ($F(1, 2967)=0.044$, $p=0.843$, $\eta^2<0.001$). Specifically, post hoc analysis with a Bonferroni adjustment revealed participants agreed with the AI's recommendation significantly more frequently with explanations using the 1st person POV (mean=52.757%) than with the 3rd person POV (mean=49.599%) (Figure 22). In addition, for belief marker, post hoc analysis revealed that using embracing markers (mean=51.850%) led to significantly higher agreement with the AI's recommendation compared to the distancing marker (mean=50.507%) (Figure 23).
For POV, a one-way ANOVA with three levels (1st person, 3rd person, no explanation) showed significant differences in agreement frequency ($F(2, 3814)=16.397$, $p<0.001$, $\eta^2=0.009$). Post hoc analysis reveal that participants agreed with the AI's recommendation significantly less frequently in explanations using the 3rd person POV (mean=49.599%) compared with either using the 1st person POV (mean=52.653%) or no explanation at all (mean=53.129%) (Figure 24). For belief markers, a one-way ANOVA with three levels (embracing, distancing, no explanation) showed significant differences ($F(2, 3814)=5.994$, $p=0.003$, $\eta^2=0.003$). Post hoc analysis revealed that participants agreed with the AI's recommendation significantly less frequently in explanations using the distancing belief marker (mean=50.621%) than having no explanation at all (mean=53.129%) (Figure 25).

For Agreement Frequency, we found a consistent pattern that using the embracing belief marker resulted in higher agreement with the AI's recommendation, whereas the using the distancing belief marker resulted in lower agreement with the AI's recommendation. **Thus, H4a that hypothesized embracing language in AI explanations will communicate the AI system's high confidence in its recommendation, leading to higher reliance on the AI's recommendation is confirmed.**

We also found a consistent pattern that using the 3rd person POV tended to result in lower agreement with the AIs recommendation. It should be noted that we observe these effects in both the cases where the AI's recommendations are correct and incorrect. **Thus, H4a that hypothesized distancing language in AI explanations will communicate the AI system's low confidence in its recommendation, leading to lower reliance on the AI's recommendation is confirmed.**

Using embracing or distancing language can induce higher or lower reliance on the AI's recommendation, regardless of the AI's expertise in making correct recommendations. In cases where the AI's recommendations are incorrect, using the embracing belief marker could induce automation bias and conversely using the more distancing 3rd person POV could reduce automation bias. To dive more deeply into what contributed to higher agreement, we next consider when participants switched their initial guess to match the AI's recommendation.

Switch-to-agree Frequency Switch-to-agree frequency measures how often participants switched their answers after seeing ShapeBot's recommendation to match the recommendation. The higher the switch-to-agree frequency, the higher reliance the participant would have on the AI.

(1) **Expert AI cases**: Switch-to-agree frequency was only significantly impacted by the interaction between POV and

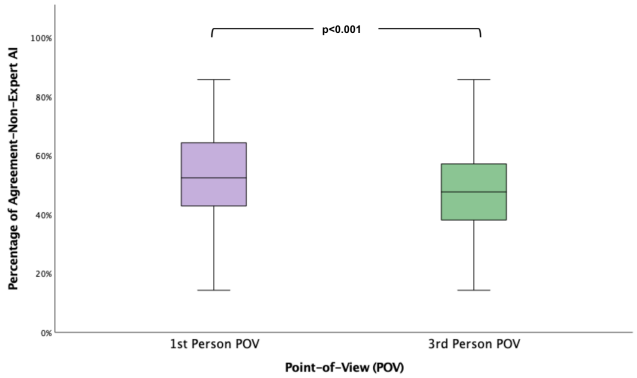


Figure 22: Agreement Frequency (in percent) between two point-of-view levels (Non-Expert AI cases). People agreed with the AI's recommendation significantly less frequently when presented with an explanation using the 3rd person POV than with 1st person POV.

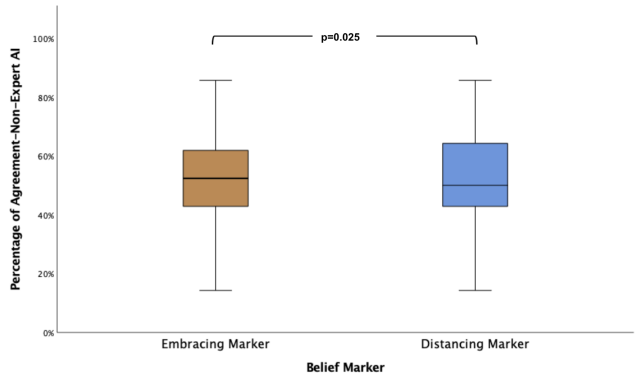


Figure 23: Agreement Frequency (in percent) between two belief marker levels (Non-Expert AI cases). Using a distancing marker led to a significantly lower agreement with the AI's recommendation compared to having embracing markers.

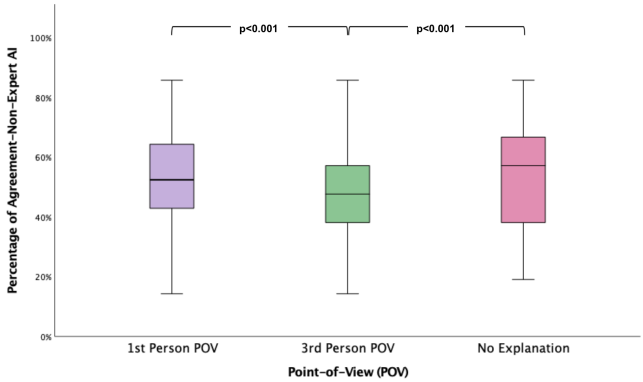


Figure 24: Agreement Frequency (in percent) across point-of-view levels and no explanation (Non-Expert AI cases). Participants agreed with the AI's recommendation less frequently when explanations used the 3rd person POV than with either the 1st person POV or no explanation.

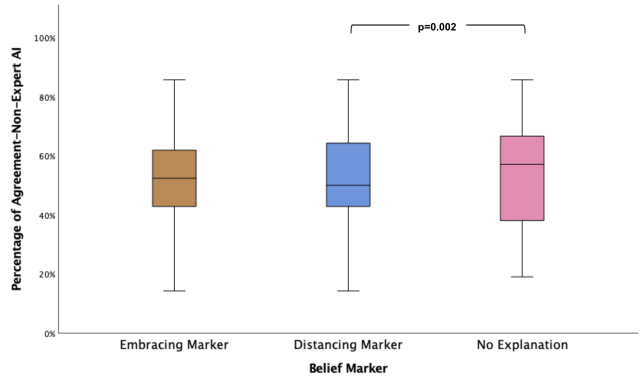


Figure 25: Agreement Frequency (in percent) across belief marker levels and no explanation (Non-Expert AI cases). The no explanation condition had a significantly higher agreement frequency than the condition with explanations using the distancing marker.

belief markers (F(1,2953)=14.749, p<0.001, η^2=0.005). We find a crossover effect (Figure 26) where as expected we see less switching-to-agree with distancing belief markers than with embracing belief markers, but only when used in combination with the 3rd person POV, and the opposite effect when using 1st person POV.

For POV, a one-way ANOVA with three levels (1st person, 3rd person, and no explanation) showed significant differences in switch-to-agree frequency (F(2, 3723)=11.847, p<0.001, η^2=0.006). Post hoc analysis with a Bonferroni adjustment revealed that having no explanation (mean=44.156%) had significantly higher switch-to-agree frequency compared to the conditions where explanations were presented with either 1st person POV (mean=40.097%) or 3rd person POV (mean=40.000%) (Figure 27). Similarly, for belief markers, a one-way ANOVA with three levels (embracing, distancing, and no explanation) showed significant differences in switch-to-agree frequency (F(2, 3723)=11.894, p<0.001, η^2=0.006). Post hoc analysis revealed the same pattern where the no explanation condition had a significantly higher switch-to-agree frequency than the conditions with embracing markers (mean=39.914%) or distancing markers (mean=40.167%) (Figure 28).

(2) **Non-Expert AI cases**: A two-way ANOVA showed the switch-to-agree frequency is significantly impacted by POV (F(1, 2967)=42.412, p<0.001, η^2=0.014) and the interaction between POV and belief markers (F(1, 2967)=5.459, p=0.020, η^2=0.002). Post hoc analysis with a Bonferroni adjustment revealed that having an explanation using the 3rd person POV (mean=22.630%) had a significantly lower switch-to-agree frequency than using the 1st person POV (mean=26.701%) (Figure 29). There was no evidence to indicate a significant effect of belief marker on switch-to-agree frequency. However,

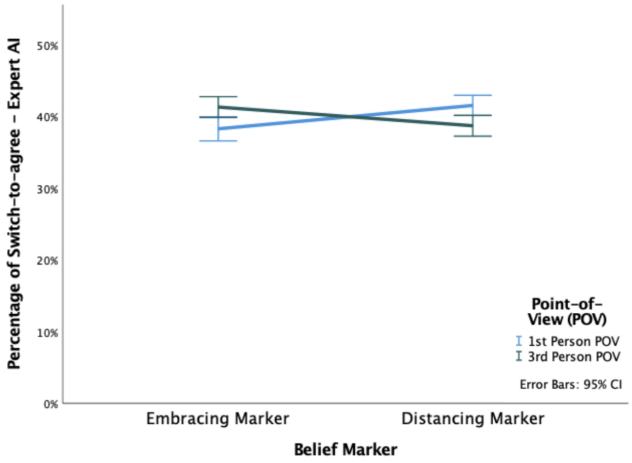


Figure 26: The two-way interaction between POV and belief marker on switch-to-agreement frequency (Expert AI cases). The most distancing combination of 3rd person POV and distancing belief marker resulted in the least amount of switching-to-agree.

examining the interaction effect between POV and belief marker (Figure 30), we see the lowest amount of switching is associated with the most distancing explanation combination (3rd person POV and distancing belief marker).

For POV, a one-way ANOVA with three levels (1st person, 3rd person, no explanation) found significant differences in switch-to-agree frequency ($F(2, 3184)=22.164$, $p<0.001$, $\eta^2=0.011$). Post hoc analysis with a Bonferroni adjustment revealed that using the 1st person POV (mean=26.803%) led to significantly higher switching-to-agree than having no explanation (mean=24.675%), whereas using the 3rd person POV (mean=22.630%) led to significantly lower switching-to-agree than having no explanation at all (Figure 31). For belief markers, a one-way ANOVA with three levels (embracing, distancing, no explanation) showed no significant differences when using distancing markers (mean=24.720%), embracing markers (mean=24.669%), and no explanation at all (mean=24.675%).

In the cases where the AI had expertise to give correct recommendations, we observed that having an explanation actually led to less switching than having no explanation at all. At first this may seem unexpected, but this can be explained by considering the participant's base rate of correctly guessing the shape correctly on the first guess before seeing the AI's recommendation. After seeing previous cases with an AI that provided (correct) explanations that referenced relevant shape features, participants were able to learn to pay attention to the relevant shape features and correctly identify them on their own more often than when not shown any explanations (see 4.4.2). Thus, participants did not need to switch their answer as often after seeing the AI's recommendation (which matched their initial guess anyway).

In contrast, for the cases where the AI did not provide good recommendations, participants were not able to learn the unfamiliar

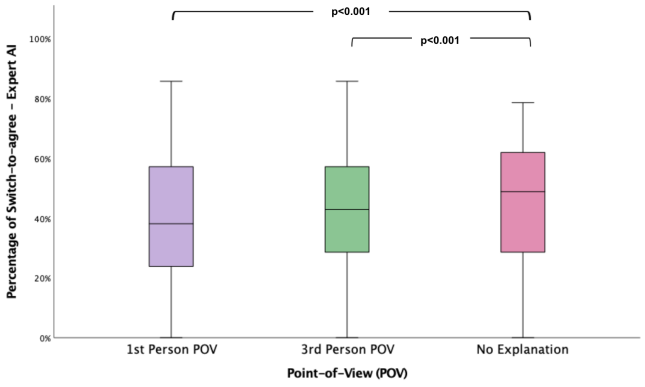


Figure 27: Switch-to-agree frequency (in percent) across point-of-view levels and no explanation (Expert AI cases). Providing an explanation either using the 1st person POV or 3rd person POV led to significantly less switching to agree with the AI than having no explanation at all.

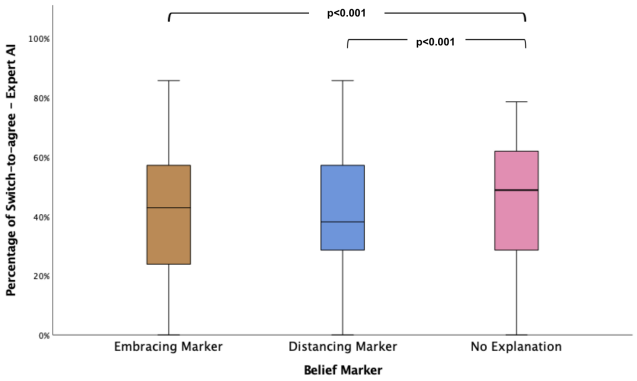


Figure 28: Switch-to-agree frequency (in percent) across levels of belief markers and no explanation (Expert AI cases). Providing an explanation either using an embracing marker or distancing marker led to significantly less switching to agree with the AI than having no explanation at all.

Fake shapes as well and had to rely more on the AI's recommendation to make their final decision. For these cases, we see a very similar pattern to that of Agreement Frequency, where using the more embracing 1st person POV in explanations led to more switching than having no explanation at all, and using the more distancing 3rd person POV led to less switching than having no explanation at all. We further observe some incremental effect of the distancing belief marker in combination with 3rd person POV leading to the lowest switching. These patterns add further evidence that hypotheses H4a and H4b that embracing and distancing language in explanations can impact the participant's reliance on the AI's recommendations.

4.4.4 Trust. Similar to Study 1, there was a significant effect of complementary expertise on subject trust ratings ($F(3, 357)=5.967$,

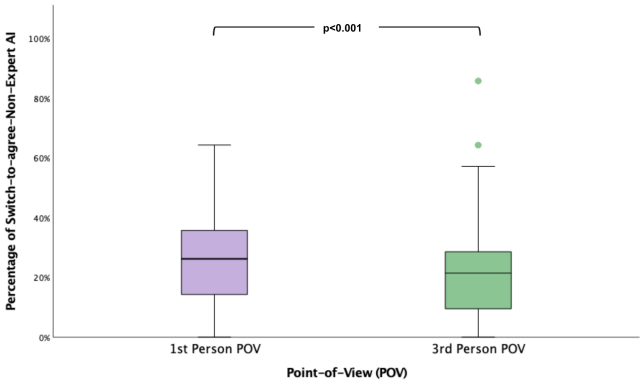


Figure 29: Box plots of switch-to-agreement frequency (in percent) between POV (Non-Expert AI cases). Providing an explanation using 1st person POV led to significantly more switching to agree with the AI than using 3rd person POV.

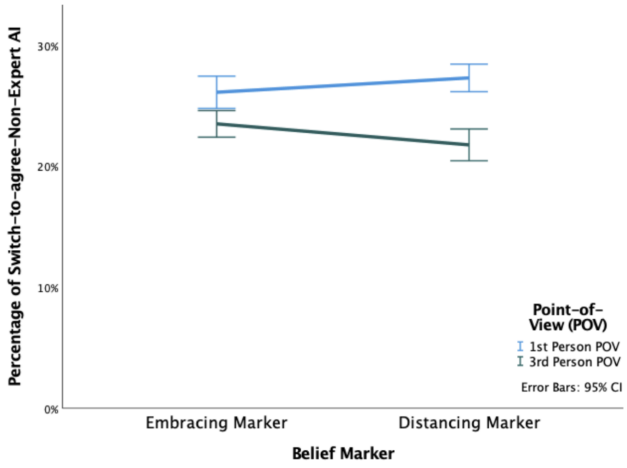


Figure 30: The two-way interaction between POV and belief marker on switch-to-agreement frequency (Non-Expert AI cases). The most distancing combination of 3rd person POV and distancing belief marker resulted in the least amount of switching-to-agree.

p<0.001, η^2=0.050). However, post hoc comparisons with a Bonferroni adjustment showed that only the highest level of complementary expertise (level 3) was significantly higher than levels 0 and 1. Similar to Study 1, we observe a general trend that higher trust is associated with the highest level of complementary expertise. There were no significant main effects of POV (F(1, 357)=0.382, p=0.537, η^2=0.001) and belief marker (F(1, 357)=0.013, p=0.910, η^2<0.001), two-way (p>0.050), nor three-way interactions (F(3, 357)=0.366, p=0.777, η^2=0.003) on trust in the AI system.

Including explanations along with the AI's recommendations, regardless of whether they used embracing or distancing language, did not affect the subjective trust that participants had in the AI system. **Thus, H7 which hypothesized that using embracing language in AI explanations will increase the human partner's subjective trust in the AI system is rejected.**

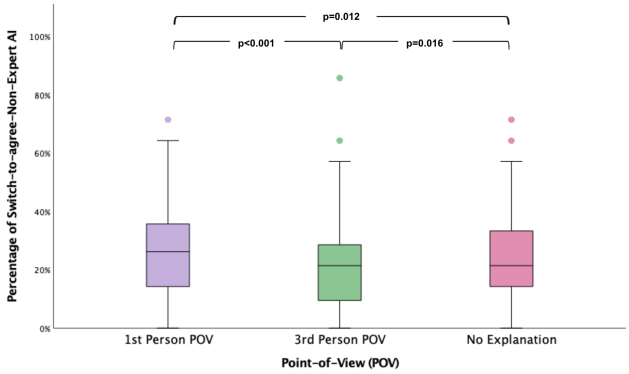


Figure 31: Box plots of switch-to-agreement frequency (in percent) across point-of-view levels and no explanation (Non-Expert AI cases). Providing an explanation using a distancing marker led to lowest switch-to-agree frequency, followed by no explanation at all, and explanation using embracing markers led to the highest switch-to-agree frequency.

Summary We found evidence to confirm H4a and H4b that explanations that use embracing language (i.e., 1st person POV and embracing belief markers "know...realize") resulted in higher reliance (agreement and switch-to-agree frequencies) on the AI's recommendations than distancing language (i.e., 3rd person POV and distancing belief markers "think...believe"), both in cases when the AI behaved like an expert to consistently make correct recommendations, as well as, in cases when the AI behaved like a non-expert to make mostly incorrect recommendations. Consequently, higher reliance on the expert AI resulted in participants being able to identify the shapes correctly more often yielding a better final decision performance. Higher reliance on a non-expert AI did result in slightly lower performance but it was not statistically significant, and thus H5 is partially supported.

With respect to performance metrics, we found that having any kind of explanation (regardless of distancing language) was effective at helping participants learn to identify new unfamiliar Fake shapes in the first guess before seeing the AI's recommendation, confirming H6. The explanations revealed the relevant features of the shapes and the logic used by the AI, allowing the participant to build their own classification model for the shapes.

We found no significant main effects of using embracing or distancing language (either POV or belief markers) on subjective trust in Study 2 (H7), though we replicated the same increase in trust as complementary expertise (and usefulness of the AI) increased that we found in Study 1. Subjective trust may thus depend more strongly on complementary expertise than the type of explanation, or our measure of subjective trust [34] was not sensitive enough to detect differences. Overall, we have demonstrated the effects of embracing and distancing language on human-AI team performance and reliance, and next we discuss the larger implications of our findings.

5 DISCUSSION

5.1 Complementary Expertise Supports Trust and Reliance

The ideal partnership between humans and AIs has been based on the premise of their complementary expertise. The hope is that human weaknesses can be compensated by the AI's strengths and vice versa. Results from prior work [5] suggest that when an AI's expertise better complements a human's expertise, performance and reliance on the AI tend to be higher. The results of our first study buttresses and extends these findings. Specifically, we found that when there was a complete overlap in expertise (where both the participant and the AI were good at identifying only all the Regular shapes, and both poor in Fake shapes), participants trusted and relied on the AI the least. Furthermore, when there was perfect complementarity (where the participant was good at identifying only in Regular shapes, and the AI was good at identifying all Fake shapes), participants trusted and relied on the AI the most. It is interesting to note that the increase in trust and reliance in the AI is not due to a higher absolute level of expertise (as was found in prior work [5, 50]), because in every condition of our study, the AI always was an expert in exactly three shape categories. Further, in the perfectly complementary case, participants still trusted and relied on the AI, despite the AI making obviously glaring mistakes when attempting to identify the Regular shapes (which humans might presume to be the easy shapes to identify). This indicates that subjective trust in an AI does not depend merely on absolute factors as such competence, predictability, and reliability (as measured by the Trust in Automated Systems scale [34] in our study), but that these factors can be influenced by the perceived usefulness of the AI for a task, which in turn is affected by complementary expertise.

5.2 Trust Calibration with Partially Complementary Expertise

Even though behavioral measures of reliance (agreeing, and switching to the recommendation of the AI) consistently increased with the degree of complementary expertise in our first study, ratings of subjective trust did not follow the same increasing pattern. In fact, there was no statistically significant difference in subjective trust between the partially complementary expertise levels 1 and 2 (Figure 6), when the AI was an expert in only one and only two Fake shape categories, respectively. In these conditions, the AI was not universally good or bad at the Fake shapes (as they were in levels 0 and 3), and thus participants needed to figure out which Fake shape categories the AI could be relied on to get right. Prior work [3] has found that error boundaries can be more difficult for people to detect and form mental models of when they are less parsimonious (i.e., include more dimensions/factors) and more stochastic. Therefore, it was likely more difficult for participants to form a mental model of the error boundary of AIs that had partially complementary expertise, which results in perceptions of the AI being less consistent or reliable in practice and affecting the participant's trust. Relative expertise can be an important factor when attempting to calibrate trust.

People calibrate their trust in an AI assistant based on the error boundary they perceive. Our results indicate that not all errors are equally useful for trust calibration. In fact, errors made by the AI in cases where the human is already a confident expert (e.g., identifying Regular shapes) can be safely ignored in the calibration process (or relegated to simply confirming the human's decision) because the human does not need the recommendation from the AI to perform well in those cases. Rather, attentional effort should be focused on the cases where the human has low expertise and requires help from the AI to perform well. For example, interactive AI systems can be designed to selectively reveal confidence information, explanations, or other examples only when dealing with cases where the human partner has low expertise. AI systems can assess the expertise of their human partner using a variety of methods, including common sense (what the average person is expected to know), years of experience (a resident doctor vs attending doctor), directly asking the human about their expertise (though this requires self-awareness and metacognition), or even inferring the human's expertise with known test cases at the beginning of the interaction.

5.3 Using Embracing and Distancing Language in Explanations

Prior work [11, 13, 45, 48] has described AI and robotic systems that used natural language in explanations and rationales for recommendations to increase transparency and interpretability for the user. There is also a thread of research [14] focused on automatically generating easily understandable natural language rationales of AI actions. In these examples, explanation statements almost always used belief markers and point-of-view markers, yet prior work has not systematically investigated their effects on trust, reliance, and performance for human-AI teams. In other words, past work leaves us with the question: Should designers of intelligent systems refer to the AI in the third person ("ShapeBot") or have it refer to itself in the first person ("I"), and should explanations use embracing words ("knows") or more distancing words ("thinks")? And in what situations should these be used?

The results of our second study demonstrate that the embracing or distancing effect of these linguistic features do, in fact, have an impact on whether humans rely on the AI's recommendation, across different levels of expertise. In particular, explanations that used embracing language such as embracing belief markers ("know...realize") resulted in significantly higher reliance on the AI's recommendation than distancing belief markers ("think...believe") or no explanation at all. In turn, higher reliance led to a higher number of AI-advised final decisions being correct, in the cases when the AI's recommendation was correct. In the cases when AI's recommendation was incorrect, we observed how the more distancing 3rd person POV led to lower agreement with the recommendation, as expected, but this lower agreement did not lead to significantly a higher final decision performance. We note that disagreement with an incorrect recommendation does not necessarily mean the participant will select the correct answer when there are more than two options. Thus, embracing and distancing effects may not be symmetric when there are multiple options available.

Comparing the effects of belief markers and POV, the effect sizes for belief markers were more often higher than POV, though both factors appear to contribute to the embracing or distancing

effect. Across our analyses, using the embracing belief marker was associated with higher reliance behavior (higher agreement frequency) than the distancing belief marker, leading to higher team performance (percentage of shapes identified correctly and learning novel shapes) when the AI was correct and leading to lower team performance when the AI was incorrect. For POV, we observe that the distancing effect of using the 3rd person POV was more consistent than the embracing effect of using the 1st person POV, with the 3rd person POV leading to lower reliance (agreement and switch-to-agree frequencies). Interestingly, the combination of POV and belief markers do seem to indicate some additive effects. For example, among the four different combinations, we found the combination with the highest psychological distance (3rd person POV + distancing belief marker) to result in the lowest reliance (agreement and switch-to-agree frequencies) and the lowest final decision performance in the cases with correct AI recommendations.

Based on our results, the simple use of the embracing belief marker terms such as "know" or "realize" in AI explanations can help communicate higher confidence than the use of the distancing belief marker such as "thinks" or "believes," potentially leading to higher reliance on the AI's recommendation. Combining POV with belief markers can also compound the embracing/distancing effects on reliance. Of course, inducing higher reliance is not always beneficial, especially when the AI makes incorrect recommendations, which will result in poor performance and automation bias. However, embracing language that induces a subtle form of automation bias can potentially be leveraged to increase the acceptability and reliance on an AI system, particularly in situations where the AI system knows it can outperform the human. Likewise, distancing language can be used in specific local cases when the AI has lower confidence in its recommendation, or when it knows its human partner likely has greater expertise.

One potential drawback of many types of explanations or confidence metrics is that they can introduce additional information that the human must attend to during the task, and thus can increase cognitive load [1] in already cognitively intense decision making tasks. Using distancing or embracing language, on the other hand, can be a subtle, non-numeric modality of conveying confidence information to the user in a way that does not overload existing numeric channels of cognition. Even though we have not explicitly compared belief markers and point-of-view markers with numeric confidence scores commonly used in prior work (e.g.,[52]), our work has demonstrated that they might have similar effects on (over-/under-)reliance on AI recommendations, without requiring any basic level of numeracy or initial training in how to interpret confidence scores.

5.4 Limitations and Future Work

We carefully designed the first study such that the AI system had expertise that included knowledge that almost any person would have (Regular shapes) as well as knowledge that no one could have (Fake shapes). While this was necessary to isolate external variables that could have polluted our study, it creates a natural trade-off with external validity. A next step for our work would be to understand the impact of different levels of complementary expertise in field settings and with an AI trained on real data. In many cases, the expertise of the AI can be tuned to different points on the ROC curve to better complement or overlap with the expertise of the human partner. Another approach to study the role of complementary expertise in the field can be to fix the expertise of the AI but include a range of expertise in the human partners from novices to experienced experts. Additionally, our methodological approach, a survey-based study, leaves out a deeper understanding of the mental models people construct while making their decisions. To address this concern, new survey-based studies or use more qualitative approaches can flesh out people's thought processes (as in [17]). Also, the first study focused on human-AI dyads exclusively. We could manipulate the expertise of humans [16] and explore more complex teams that involve multiple AIs and humans, or subgroups of cooperative human-AI dyads.

In the second study, we focused on helping humans adapt their level of reliance and trust in the AI system. Even though we found clear impacts on reliance behaviors, our measure of subjective trust may not have been sensitive enough to capture differences in subjective trust. Moreover, our approach was to manipulate the linguistic characteristics of the explanation that the bot offered. The choice to focus on language was driven both by past literature on psychological distance as well as the fact that it is relatively unexplored in past work in human-AI collaborative systems. However, there are many other ways that the bot's explanation could be aligned to its expertise. In addition to graphical representations of confidence (e.g., [5]) future studies could manipulate the bot's appearance; the bot could be made interactive and mimic human gestures, facial expressions, or other mannerisms that indicate confidence [37, 48, 49]. Future work could investigate the use of multimedia; the bot could be made to speak and the aural qualities of its voice could be manipulated to suggest higher or lower confidence [44].

6 CONCLUSION

AI systems are becoming ubiquitous. As a consequence, most of our everyday tasks will involve working with an AI system in one way or another. But how we manage this new type of collaboration is an open question. When we work with other people, we intuit that different people have different abilities for different tasks, and we expect to be able to complete tasks in our fields of expertise while delegating other tasks. Our work shows that when an AI agent has expertise that is complementary to their human partner, people are able to complete certain tasks on their own and rely on the advice the AI for other tasks, in much the same way they might with a human partner. People were able to detect and rely on the AI's expertise as it complemented the human's expertise more. However, people did not always trust the AI more, even as their expertise better complemented each other and their performance increased. When the AI had expertise in some areas but not in others that the human had low expertise, it was more difficult for the human to calibrate to the AI which affected subjective trust in the AI. This is perhaps not surprising—when two people collaborate they send each other signals via any number of explicit and implicit mechanisms that are not available to an AI agent. There are opportunities for AI systems to better assess the expertise of their human partners to focus their explanations and other mechanisms to increase transparency and reliance where they are needed most.

Our results shows that using embracing or distancing language can subtly affect people's reliance on the AI. Our work is also one example that for human-AI teams, the devil is in the details, and that the methods that the AI uses to communicate its expertise must be carefully designed. This suggests that there remains extensive opportunities for digital designers of all stripes to bring to bear their own expertise on the development of communication patterns in user-facing AI systems.

ACKNOWLEDGMENTS

Our special thanks to Nayeli Bravo for helping with data collection, Emily Sarah Sumner and Alexandre Filipowicz for helping with the videos, and Rumen Iliev for advice on this work.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[2] Annette Baier. 1986. Trust and Antitrust. *Ethics* 96, 2 (Jan. 1986), 231–260.

[3] 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. *HCOMP* 7, 1 (Oct. 2019), 2–11.

[4] 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. *AAAI* 33, 01 (July 2019), 2429–2437.

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 81)*. Association for Computing Machinery, New York, NY, USA, 1–16.

[6] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. 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, 258–262.

[7] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7.

[8] Meng Chen, Robert A Bell, and Laramie D Taylor. 2017. Persuasive effects of point of view, protagonist competence, and similarity in a health narrative about type 2 diabetes. *Journal of health communication* 22, 8 (2017), 702–712.

[9] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A Friedler and Christo Wilson (Eds.). PMLR, 134–148.

[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806.

[11] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable AI for Robot Failures: Generating Explanations that Improve User Assistance in Fault Recovery. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 351–360.

[12] Maartje M A de Graaf and Bertram F Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *2017 AAAI Fall Symposium Series*.

[13] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–19.

[14] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. 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, 263–274.

[15] 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.

[16] 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.

[17] 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*. Association for Computing Machinery, New York, NY, USA, 1–12.

[18] Maaike Harbers, Karel van den Bosch, and John-Jules Ch Meyer. 2009. A Study into Preferred Explanations of Virtual Agent Behavior. In *Intelligent Virtual Agents*. Springer Berlin Heidelberg, 132–145.

[19] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[20] Moritz Körber, Eva Baseler, and Klaus Bengler. 2018. Introduction matters: Manipulating trust in automation and reliance in automated driving. *Appl. Ergon.* 66 (Jan. 2018), 18–31.

[21] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. ieeexplore.ieee.org, 3–10.

[22] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. (Dec. 2021). arXiv:2112.11471 [cs.AI]

[23] 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.

[24] J Lee and N Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (Oct. 1992), 1243–1270.

[25] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 72)*. Association for Computing Machinery, New York, NY, USA, 1–13.

[26] Nira Liberman, Yaacov Trope, Elena Stephan, and Others. 2007. Psychological distance. *Social psychology: Handbook of basic principles* 2 (2007), 353–383.

[27] Alexandra Lorson, Chris Cummins, and Hannah Rohde. 2021. Strategic use of (un) certainty expressions. *Frontiers in Communication* 6 (2021), 28.

[28] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 78)*. Association for Computing Machinery, New York, NY, USA, 1–16.

[29] Bertram F Malle. 2006. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT Press.

[30] Bertram F Malle. 2006. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. MIT press.

[31] Bertram F Malle, Joshua Knobe, Matthew J O'Laughlin, Gale E Pearce, and Sarah E Nelson. 2000. Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *J. Pers. Soc. Psychol.* 79, 3 (Sept. 2000), 309–326.

[32] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 3 (July 1995), 709–734.

[33] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (Feb. 2019), 1–38.

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

[35] Frederik Naujoks, Yannick Forster, Katharina Wiedemann, and Alexandra Neukum. 2017. A Human-Machine Interface for Cooperative Highly Automated Driving. In *Advances in Human Aspects of Transportation*. Springer International Publishing, 585–595.

[36] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *HCOMP* 8, 1 (Oct. 2020), 112–121.

[37] Catherine Pelachaud. 2009. Studies on gesture expressivity for a virtual agent. *Speech Commun.* 51, 7 (July 2009), 630–639.

[38] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21, Article 237)*. Association for Computing Machinery, New York, NY, USA, 1–52.

[39] Ajay Sandhu and Peter Fussey. 2021. The 'uberization of policing'? How police negotiate and operationalise predictive policing technology. *Policing Soc.* 31, 1 (Jan. 2021), 66–81.

[40] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. 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, 240–251.

[41] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. 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, 240–251.

[42] Younho Seong and Ann M Bisantz. 2008. The impact of cognitive feedback on judgment performance and trust with decision aids. *Int. J. Ind. Ergon.* 38, 7 (July 2008), 608–625.

[43] N A Stanton and M S Young. 1998. Vehicle automation and driving performance. *Ergonomics* 41, 7 (July 1998), 1014–1028.

[44] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[45] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–17.

[46] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[47] 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*. Association for Computing Machinery, New York, NY, USA, 1–15.

[48] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 7–9.

[49] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. "Let me explain!": exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces* 15, 2 (June 2021), 87–98.

[50] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19, Paper 279)*. Association for Computing Machinery, New York, NY, USA, 1–12.

[51] 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.

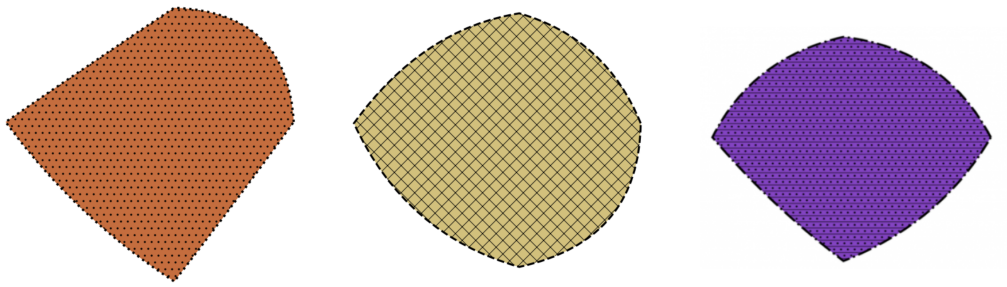
[52] Yunfeng Zhang, Q Vera Liao, and Rachel K E Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305.

A APPENDIX

A.1 Study 1 Design

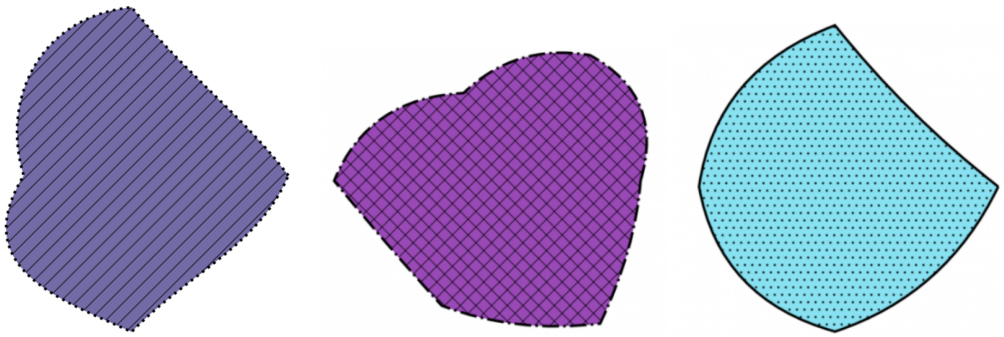
Table 3: Fake Shapes Design

Fake Shape Name	Number of Sides	Border and Interior Patterns
Senectus	4	Same dash pattern (both dots, both dashes, or both dot&dashes)
Pharetra	4	Different dash patterns
Ultrices	5	Same dash pattern (both dots, both dashes, or both dot&dashes)



(a) Senectus Example 1 **(b) Senectus Example 2** **(c) Senectus Example 3**

Figure 32: Three examples of the fake shape–Senectus, which have *4* sides and have border and interior patterns that *match* (both dots, both dashes, both dots&dashes).

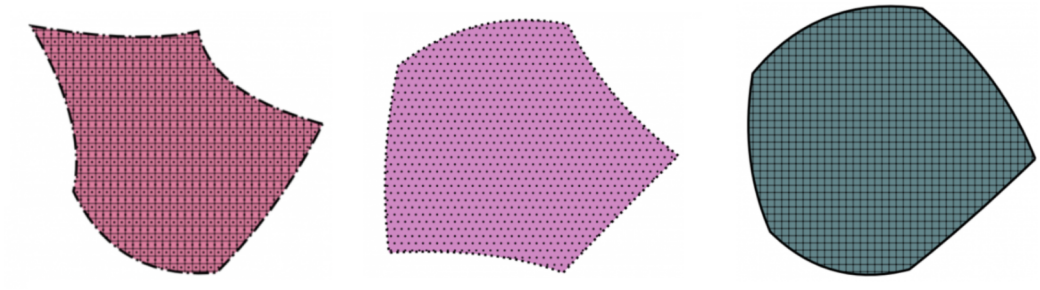


(a) Pharetra Example 1 **(b) Pharetra Example 2** **(c) Pharetra Example 3**

Figure 33: Three examples of the fake shape–Pharetra, which have *4* sides and have border and interior patterns that *do not match*.

Table 4: Total Number of Participants by Complementarity Level in Study 1.

Complementarity Expertise Condition	Number of Participants
Level 0	45
Level 1	39
Level 2	48
Level 3	28
Total	160



(a) Ultrices Example 1 **(b) Ultrices Example 2** **(c) Ultrices Example 2**

Figure 34: Three examples of the fake shape–Ultrices, which have *5* sides and have border and interior patterns that *match* (both dots, both dashes, both dots&dashes).

A.2 Study 1 Results

A.2.1 Performance-Regular shapes.

- First Guess Performance: For regular shapes, there was no significant effects of complementarity on first guess performance (F(3, 159)= 0.811, p=0.490, η^2=0.015). Participants were able to correctly guess Regular shapes on their first try without any trouble.
- Final Decision Performance: We found no significant difference in the final decision performance for Regular shapes across the conditions (F(3, 159)=2.381, p=0.072, η^2=0.044). Similar to the first guess performance in the Regular shape categories, participants' final guesses were nearly all correct.

A.2.2 Reliance Behaviors-Regular shapes.

- Agreement Frequency: For Regular shapes, participants' agreement with the AI's recommendation decreased significantly as the level of complementary expertise increased (F(3,159)= 16172.338, p<0.001, η^2=0.997).Post-hoc comparisons showed that participants would naturally agree with the recommendation less often as the AI made poorer recommendations for the Regular shapes that they were already familiar with.
- Switch-to-agree Frequency: For Regular shapes, there were no significant differences in switch-to-agree frequency among levels of complementarity (F(3,159)=0.337, p=0.799, η^2=0.006). As expected, participants were already experts at identifying Regular shapes in their first guess and rarely switched after seeing the AI's recommendation.

A.2.3 Switch Frequency-Fake Shapes.

- Switch Frequency: Switching occurred in trials when participant changed their guess after seeing the AI recommendation. For all shapes, the results show that there was a significant effect of complementary expertise on switch frequency (F(3,159)=23.656, p<0.001, η^2=0.313). As the level of complementary expertise increased, the amount of switching also increased. Post-hoc analyses reveal that the lowest level of complementary (level 0), when the AI's expertise completely overlapped with participant's expertise in only Regular shapes, had significantly lower amount of switching than the other levels.

A.3 Study 2

Table 5: Total Number of Participants by Complementarity Level, POV, and Belief Marker Groups in Study 2.

		Complementarity Level								
		Level 0		Level 1		Level 2		Level 3		
		Point-of-view								Total
		1st person	3rd person	1st person	3rd person	1st person	3rd person	1st person	3rd person	
Belief Marker	Embracing Marker	13	17	16	20	16	16	14	19	131
	Distancing Marker	18	17	25	19	19	18	18	17	151
No Explanation		18		22		22		14		76
Total		83		102		91		82		358

Table 6: Mean and Standard Deviation of First Guess Performance (Non-Expert AI system - POV*Belief Marker).

	1st Person POV (1st)		3rd Person POV (3rd)		Difference
	Mean	SD	Mean	SD	
Embracing marker (embr)	36.782	0.646	38.585	0.583	1st vs. 3rd: p=0.038
Distancing marker (dist)	40.070	0.544	39.252	0.583	
Difference	embr vs. dist: p<0.001				

Table 7: Mean and Standard Deviation of Final Decision Performance (Expert AI system - POV*Belief Marker).

	1st Person POV (1st)		3rd Person POV (3rd)		Difference
	Mean	SD	Mean	SD	
Embracing marker (embr)	82.579	0.699	84.142	0.635	
Distancing marker (dist)	81.807	0.613	78.437	0.644	1st vs. 3rd: p<0.001
Difference			embr vs. dist: p<0.001		

Table 8: Mean and Standard Deviation of Switch-to-agree Frequency (Expert AI system - POV*Belief Marker).

	1st Person POV (1st)		3rd Person POV (3rd)		Difference
	Mean	SD	Mean	SD	
Embracing marker (embr)	38.254	0.823	41.284	0.748	1st vs. 3rd: p=0.006
Distancing marker (dist)	41.514	0.722	38.679	0.758	1st vs. 3rd: p=0.007
Difference	embr vs. dist: p=0.003		embr vs. dist: p=0.014		

Table 9: Mean and Standard Deviation of Switch-to-agree Frequency (Non-Expert AI system - POV*Belief Marker).

	1st Person POV (1st)		3rd Person POV (3rd)		Difference
	Mean	SD	Mean	SD	
Embracing marker (embr)	26.108	0.685	23.498	0.617	1st vs. 3rd: p=0.005
Distancing marker (dist)	27.294	0.576	21.762	0.617	1st vs. 3rd: p<0.001
Difference			embr vs. dist: p=0.047		