Journal of Risk Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjrr20

Combining uncertainty information with AI recommendations supports calibration with domain knowledge

Harishankar Vasudevanallur Subramanian, Casey Canfield, Daniel B. Shank & Matthew Kinnison

To cite this article: Harishankar Vasudevanallur Subramanian, Casey Canfield, Daniel B. Shank & Matthew Kinnison (2023) Combining uncertainty information with AI recommendations supports calibration with domain knowledge, Journal of Risk Research, 26:10, 1137-1152, DOI: 10.1080/13669877.2023.2259406

To link to this article: https://doi.org/10.1080/13669877.2023.2259406

JOURNAL OF RISK RESEARCH
2023, VOL. 26, NO. 10, 1137–1152
https://doi.org/10.1080/13669877.2023.2259406

Check for updates

Combining uncertainty information with AI recommendations supports calibration with domain knowledge

Harishankar Vasudevanallur Subramanian[a], Casey Canfield[a], Daniel B. Shank[b] and Matthew Kinnison[a]

[a]Engineering Management and Systems Engineering, Missouri University of Science and Technology, Rolla, MO, USA; [b]Psychological Science, Missouri University of Science and Technology, Rolla, MO, USA

ABSTRACT

The use of Artificial Intelligence (AI) decision support is increasing in high-stakes contexts, such as healthcare, defense, and finance. Uncertainty information may help users better leverage AI predictions, especially when combined with their domain knowledge. We conducted a human-subject experiment with an online sample to examine the effects of presenting uncertainty information with AI recommendations. The experimental stimuli and task, which included identifying plant and animal images, are from an existing image recognition deep learning model, a popular approach to AI. The uncertainty information was predicted probabilities for whether each label was the true label. This information was presented numerically and visually. In the study, we tested the effect of AI recommendations in a within-subject comparison and uncertainty information in a between-subject comparison. The results suggest that AI recommendations increased both participants' accuracy and confidence. Further, providing uncertainty information significantly increased accuracy but not confidence, suggesting that it may be effective for reducing overconfidence. In this task, participants tended to have higher domain knowledge for animals than plants based on a self-reported measure of domain knowledge. Participants with more domain knowledge were appropriately less confident when uncertainty information was provided. This suggests that people use AI and uncertainty information differently, such as an expert versus second opinion, depending on their level of domain knowledge. These results suggest that if presented appropriately, uncertainty information can potentially decrease overconfidence that is induced by using AI recommendations.

ARTICLE HISTORY
Received 18 January 2023
Accepted 4 September 2023

KEYWORDS
Overconfidence; artificial intelligence; uncertainty; human-AI teams; risk communication

Introduction

Increasingly, artificial intelligence (AI) recommender systems are being integrated into high-risk contexts ranging from healthcare to the military to the finance system. However, numerous high-profile failures have highlighted the limitations of these systems (Dastin 2018; Sohn 2023). Experimental evidence suggests that human-AI teams tend to perform better than either alone (Bansal, Wu, and Zhou 2021; Rosenberg et al. 2018), especially for domain experts (Bien et al.

CONTACT Casey Canfield canfieldci@mst.edu Engineering Management and Systems Engineering, MO University of Science and Technology, Rolla, MO 65409, USA
Supplemental data for this article is available online at https://doi.org/10.1080/13669877.2023.2259406

2018; Lakhani and Sundaram 2017; Patel et al. 2019; Bai et al. 2020). However, human-AI teams with novices tend to be less accurate than AI alone (Green and Chen 2019; Grgić-Hlača, Engel, and Gummadi 2019; Lin et al. 2020).

For successful human-AI teams, it may be helpful for users to understand the uncertainty associated with specific AI recommendations, particularly in risky contexts where they may struggle to understand the reliability of the AI (Elder et al. 2022). In general, users tend to weigh AI recommendations like an expert's when making decisions (Ashktorab et al. 2020; Wang, Molina, and Sundar 2020), sometimes even giving it more weight than a human expert (Logg, Minson, and Moore 2019). Although users are inclined to rely on AI recommendations, they may adjust their strategy if they have domain expertise or know that the task is difficult to predict in general. In addition, there is a risk of AI recommendations inducing overconfidence. Users may have poor calibration and believe they and/or the AI system are better at a task than they are (Moore and Healy 2008). However, little research has explored how AI recommendations influence confidence and therefore induce overconfidence.

Providing uncertainty information with AI predictions

Uncertainty measures the lack of knowledge about an outcome. In the context of AI recommendations, uncertainty can be measured as the probability that a prediction matches the truth, which is also called the system confidence (Antifakos et al. 2005; Bhatt et al. 2021). Deep learning models (a common approach to AI) are nonparametric and there are no well-established methods for estimating confidence intervals, which are typically used to evaluate uncertainty for statistical models. However, researchers are continuing to develop methods, such as Bayesian approximation and ensemble modeling, to improve uncertainty quantification in deep learning (Abdar et al. 2021).

Regardless of the specific metric, uncertainty can be communicated *via* text, numbers, and visuals as appropriate depending on task and individual characteristics (Bhatt et al. 2021; van der Bles et al. 2019). For example, an outcome can be as described as 'very likely' or a '90% probability' or visualized in a graph (Lipkus and Hollands 1999). Combining text, numerical, and visual communications can increase accuracy and confidence (Gkatzia, Lemon, and Rieser 2016). Communicating uncertainty is distinct from AI explainability, which aims to increase transparency by describing the logic between the inputs and outputs (Gunning et al. 2019). Thus, explainability describes what an AI system knows while uncertainty describes what the AI system does not know (Tomsett et al. 2020).

For AI predictions, there is mixed evidence on the effectiveness of providing uncertainty information to improve decision-making. In some studies, providing uncertainty information increases accuracy (Bansal, Wu, and Zhou 2021; Fernandes et al. 2018; Gkatzia, Lemon, and Rieser 2016), often because users trust the AI more (Antifakos et al. 2005). However, studies have also found no effect of the use of uncertainty information. For example, Subramanian et al. (2020) found no effect of uncertainty information when visualized as a bar graph. In Antifakos et al. (2005), users wanted additional information when the AI's confidence level was below 50%, suggesting that 50% may be a key threshold for user trust.

Most research on AI recommendations aims to calibrate trust, rather than confidence (Buçinca et al. 2020; Wang and Yin 2021; Zhang, Liao, and Bellamy 2020). One of the most widely accepted definitions of trust is 'the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party' (Mayer, Davis, and Schoorman 1995). In the human-AI interaction context, Ferrario, Loi, and Viganò (2020) describe trust as the willingness of a user to rely on AI predictions for decision making, without wanting to review the AI's capabilities. While a user can trust an AI in general, they must also determine whether to trust each individual recommendation, which may not always be correct. Uncertainty information can improve trust calibration, but accuracy only improves if the user has sufficient domain knowledge to complement the AI (Zhang, Liao, and Bellamy 2020). In contrast, a confidence measure reflects the perception of the human-AI team as a whole and the likelihood that together they have identified the correct choice. Self-confidence can ultimately influence their reliance behavior if poorly performing users blame themselves rather than the AI (Chong et al. 2022). There is also a risk that communicating uncertainty may reduce trust, a phenomenon that is often observed when human experts express uncertainty about a prediction (van der Bles et al. 2019).

Effect of domain knowledge

Experts can generally perform a task much better than novices. For example, Snow et al. (2008) found that they needed four novices to label an item to the same accuracy as one expert in a natural language task. In general, as the task difficulty increases, the need for a human expert to be in the loop also increases (Maadi, Akbarzadeh Khorshidi, and Aickelin 2021). Logg, Minson, and Moore (2019) found that experts tended to under-rely on algorithmic versus human recommendations, hurting their accuracy. In contrast, novices with insufficient domain knowledge tended to over-rely on recommendation systems (Wang and Benbasat 2013), especially when uncertainty information was provided (Bussone, Stumpf, and O'Sullivan 2015).

Experts may be in a better position to navigate AI recommendations, but may need appropriate meta-information like uncertainty information to evaluate whether to trust and therefore rely on, specific recommendations (Feng and Boyd-Graber 2019). There is some evidence that users who are overconfident in their own abilities are also more likely to adopt algorithmic advice (Piehlmaier 2022; Zhang, Xu, and Palma 2019). Thus, users who rely on AI recommendations may have a higher propensity for overconfidence.

Aim of study

More research is needed to understand how communicating uncertainty with AI recommendations influences human-AI team performance (Gkatzia, Lemon, and Rieser 2016). To address this gap, this study evaluates the effect of providing AI predictions with uncertainty information on accuracy and confidence for an image recognition task. In addition, this relationship is examined for users at varying levels of domain knowledge.

Methods

Participants

We recruited 201 participants from Prolific, an online participant recruitment platform (Peer et al. 2017). Eligible participants were over 18 years old and spoke English. Overall, 49% were male and 48% had completed a 4-year college degree. The average age was 33, ranging from 18 to 64 years old. All participants provided informed consent and were compensated $5. This study was approved by the University of Missouri Institutional Review Board (#2021926).

Design

Participants performed an image recognition task in a mixed-subject design with and without AI recommendations (within-subjects). In addition, participants were randomly assigned to receive or not receive uncertainty information (between-subjects). Due to the high number of planned tests, we use $\alpha < 0.01$ for interpretation to reduce false positives. The data, R code, and experimental materials are available on Open Science Framework at https://osf.io/bjgu9/.

Stimuli

The stimuli were images of plants and animals from the ImageNet database, commonly used for training and testing AI models. The AI recommendations were derived from an existing deep learning model's supplementary materials, including 88 images with five label predictions and the associated predicted probabilities displayed as bars (Krizhevsky, Sutskever, and Hinton 2017). We restricted the stimuli to plants and animals to facilitate measuring domain knowledge. Of the 42 images of plants and animals, we selected 32 images where the image label was a focus of the image. Overall, there were 19 images of plants and 13 of animals. The accuracy of the AI's first recommendation (referred to as AI accuracy) was 62% for animals and 58% for plants.

For each image, there were six potential labels. In addition to the five labels provided by Krizhevsky, Sutskever, and Hinton (2017) we either added the correct label (in cases where it was not included in the provided labels) or another similar but incorrect label. The AI recommendations were ordered such that the AI's best recommendation was listed first, although this was not always correct. Overall, 19 of the 32 images included the correct label as the first recommendation, 5 as the second recommendation, 2 as the third recommendation, 1 as the fourth recommendation, 0 as the fifth recommendation, and 5 as the sixth recommendation (i.e. the AI did not provide the correct label). For the sixth recommendation, the predicted probability was always less than 1%. Since the sixth recommendation did not come from the AI, this represented a prediction with extremely low confidence. This was necessary to measure accuracy because the AI's recommendations did not include the correct label for 5 of the 32 images.

For the participants assigned to receive uncertainty information, each AI recommendation included a numerical measure of the probability the label was correct and a color-coded bar to enhance salience. The bars were color-coded green (100% – 76% confident), yellow (75% – 51%), orange (50% – 26%), or red (25% – 0%). An example of the AI recommendation with and without uncertainty information is shown in Figure 1. Overall, 11 of the images had a probability for the correct label between 100% and 76%, 4 between 75% and 51%, 3 between 50% and 26%, and 14 images between 25% and 0%.

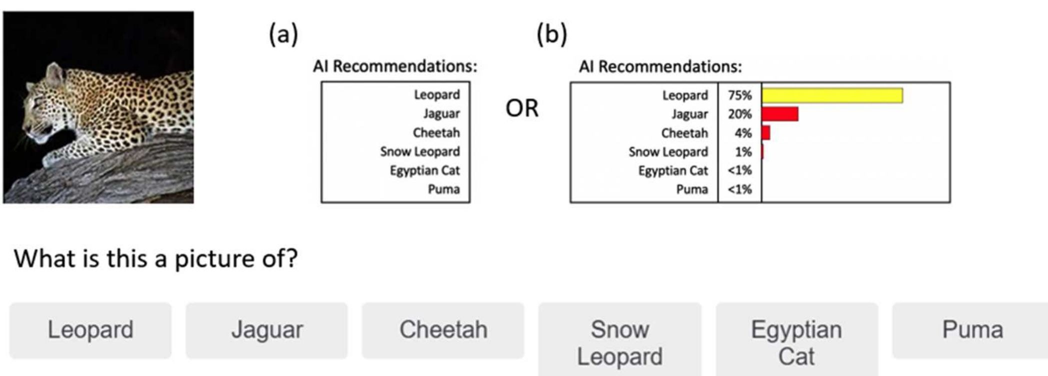

Figure 1. Example stimulus for (a) AI recommendations alone and (b) AI recommendations with uncertainty information.

Measures

Participants identified each image *via* a multiple-choice question, 'What is this a picture of?' Participants chose a label from the AI recommendations presented in random order to measure pre-AI accuracy. In addition, participants rated their pre-AI confidence *via* 'How confident are you in your answer?' on a Likert scale where 1 = not confident at all (0–20%), 5 = extremely confident (80–99%), and 6 = absolutely confident (100%). These responses were then quantified as the minimum values for each level of the Likert scale.

Participants again answered, 'What is this a picture of?' to measure post-AI accuracy with choices provided in recommendation order. This design is summarized in Figure 2. For each image, participants rated post-AI confidence using the same confidence scale and post-AI usefulness, 'How useful was the AI in recognizing the image?' on a Likert scale where 1 = not useful at all and 6 = extremely useful. We also recorded the time per image for the pre-AI and post-AI evaluation. Overconfidence was calculated as the difference between participants' average accuracy and average confidence ratings.

Three attention checks were combined into a single score to evaluate data quality. First, after the initial instructions, a multiple-choice question asked, 'What was mentioned as the correct answer to the image provided in the instructions?' (Howler Monkey). Second, there was an embedded attention check in the image recognition task with the same image as the example in the instructions (Howler Monkey). Lastly, after the image recognition task, a multiple-choice question asked, 'How many AI recommendations did you get for each image?' (answer: 6).

Participants rated their domain knowledge *via* two questions, 'How well can you identify plants?' and 'How well can you identify animals?' on a Likert scale where 1 = not well at all and 6 = extremely well. Participants responded to an AI usefulness scale adopted from Venkatesh and Davis (2000) as well as an AI reliability scale adopted from Madsen and Gregor (2000) on a 7-point Likert scale where 1 = Strongly disagree, 4 = Neither agree nor disagree, and 7 = Strongly agree. In addition, participants rated overall perceived difficulty, 'How difficult was this task?' on a Likert scale where 1 = extremely easy and 5 = extremely difficult. Participants also rated perceived trustworthiness, 'How trustworthy was the AI?' on a Likert scale where 1 = very untrustworthy and 5 = very trustworthy. Lastly, we measured age, gender, and education level. Due to significant skew, a log transformation was used to normalize age.

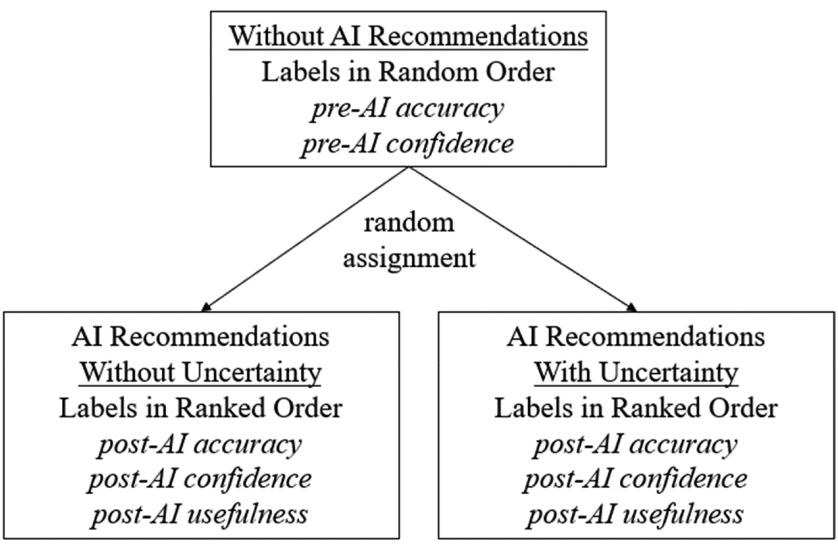

Figure 2. Summary of accuracy and confidence measures.

Procedure

Participants reviewed instructions about the task and answered the first attention check question. Then, they completed 33 randomly ordered image recognition tasks, which each included a pre-AI as well as post-AI assessment (see Figure 2). The 33 tasks included 32 experimental stimuli and 1 embedded attention check. Afterwards, participants completed the final attention check and reported their domain knowledge, AI usefulness, AI reliability, overall perceived difficulty, AI trustworthiness, and demographics.

Results

Overall, 74% of the participants passed all three attention checks. More than 90% of the participants responded correctly to two of the three attention checks. Only 78% of participants correctly identified the number of AI recommendations for each image, which was asked near the end of the experiment. No participants were removed for performance on the attention checks, but the attention score was included as a control in the analysis. Results are reported in the following sections for the effect of AI recommendations and uncertainty information in general as well as in the context of varying domain knowledge. The key outcomes of interest include accuracy and confidence.

Effect of AI recommendations

In a pre-post comparison, providing rank-ordered AI recommendations increased accuracy and confidence. As reported in Table 1, post-AI accuracy was higher than pre-AI accuracy. However, the average performance in the post-AI condition did not exceed the accuracy of simply selecting the AI's first recommendation. The pre-and post-AI accuracy was positively correlated for

Table 1. Summary of measures by conditions.

	Within-subjects		Between-subjects	
	Pre-AI M (SD)	Post-AI M (SD)	No Uncertainty M (SD)	Uncertainty M (SD)
Participants	201		99	102
	All images (32 total, AI accuracy = 59%)			
Accuracy	36% (8%)	54% (8%)	52% (9%)	57% (5%)
Confidence	43% (15%)	57% (16%)	56% (16%)	58% (15%)
Overconfidence	7% (16%)	3% (17%)	5% (18%)	1% (15%)
Time/image(s)	16 (19)	10 (6)	10 (6)	10 (5)
AI usefulness[a]		3.6 (0.8)	3.5 (0.8)	3.6 (0.7)
Task difficulty[b]		3.2 (1.0)	3.2 (1.1)	3.1 (1.0)
AI trustworthiness[b]		3.5 (1.0)	3.5 (1.0)	3.5 (1.0)
	Plant images (19 total, AI accuracy = 58%)			
Accuracy	27% (9%)	50% (10%)	46% (12%)	53% (7%)
Confidence	39% (14%)	53% (17%)	52% (18%)	55% (16%)
Overconfidence	1% (21%)	2% (17%)	3% (18%)	1% (16%)
Time/image(s)	14 (9)	11 (7)	11 (7)	11 (7)
AI usefulness[a]		3.5 (0.8)	3.5 (0.8)	3.6 (0.7)
Plant knowledge[a]		2.0 (1.0)	2.0 (1.2)	1.9 (0.8)
	Animal images (13 total, AI accuracy = 62%)			
Accuracy	48% (12%)	61% (9%)	59% (10%)	63% (7%)
Confidence	49% (17%)	63% (15%)	62% (16%)	64% (15%)
Overconfidence	1% (21%)	2% (17%)	3% (18%)	1% (16%)
Time/Image (s)	18 (44)	10 (5)	10 (6)	10 (4)
AI usefulness[a]		3.7 (0.8)	3.6 (0.9)	3.7 (0.7)
Animal knowledge[a]		3.4 (1.2)	3.5 (1.2)	3.3 (1.1)

[a]6-point Likert scale.
[b]5-point Likert scale.

Table 2. Linear mixed regression models for effect of AI recommendations on accuracy with (1) all, (2) plant, and (3) animal stimuli.

	Model 1 All B (SE)	Model 2 Plants B (SE)	Model 3 Animals B (SE)
Intercept	0.41 (0.06)***	0.37 (0.07)***	0.35 (0.08)***
AI	0.19 (0.01)***	0.25 (0.02)***	0.17 (0.02)***
Domain knowledge		0.01 (0.01)*	0.02 (0.01)**
Knowledge*AI		−0.01 (0.01)	−0.01 (0.01)
Average time/image(s)	0.00 (0.00)	0.00 (0.00)	−0.00 (0.00)
Attention score	−0.00 (0.01)	0.00 (0.01)	0.00 (0.01)
Task difficulty	0.00 (0.00)	0.00 (0.01)	−0.00 (0.01)
AI trustworthiness	−0.00 (0.00)	0.00 (0.01)	−0.00 (0.01)
log(age)	−0.02 (0.02)	−0.05 (0.02)*	0.03 (0.02)
Male	−0.01 (0.01)	−0.01 (0.01)	−0.02 (0.01)
College	0.01 (0.01)	0.03 (0.01)*	−0.01 0(.01)
Number of images	32	19	13
N	398	398	398

*$p < .05$, **$p < .01$, and ***$p < .001$.

each participant, $r(200) = 0.39$, $p < .001$. That this correlation is not higher suggests that participants benefited from AI support, regardless of their initial performance. As expected, participants spent significantly more time on the pre-AI image task than post-AI task, even though there was an additional question for the post-AI evaluation, $t(200) = 4.19$, $p < .001$. There was no relationship between performance and how participants perceived the usefulness or reliability of the AI in general or the overall difficulty of the task.

In addition, confidence was higher after participants received AI recommendations. As reported in Table 1, average post-AI confidence was higher than pre-AI confidence. Pre- and post-AI confidence was highly correlated, $r(200) = 0.72$, $p < .001$, suggesting that although AI recommendations increased confidence, participants anchored on their initial response.

In a paired *t*-test, participants were more overconfident before AI recommendations than after AI recommendations, $t(200) = 4.90$, $p < .001$. This suggests that AI recommendations improved participants' calibration without inducing worse overconfidence. More confident participants (post-AI) also tended to perceive the AI as more useful, $r(200) = 0.61$, $p < .001$.

In Table 2 Model 1, ranked AI recommendations significantly increased accuracy. Similarly, in Table 3 Model 1, ranked AI recommendations significantly increased confidence. None of the covariates were significant, suggesting that the AI alone was contributing to the improved accuracy and confidence. Models 2 and 3 are discussed in detail below.

Effect of uncertainty information

Providing uncertainty information with the rank-ordered AI recommendations increased accuracy but not confidence. As reported in Table 1, accuracy was higher when uncertainty information was provided, although it did not exceed the accuracy of the AI on average. In the uncertainty information, the AI reported a probability over 50% for the correct label for 15 out of the 32 images (i.e. less than half the time). This suggests that providing uncertainty information provided modest improvements above AI recommendations alone, likely because the uncertainty information was a relatively weak signal.

In Table 4 Model 1, uncertainty information increased accuracy. This suggests that participants could leverage additional information from the uncertainty information to improve their accuracy. In Table 4 Model 1, participants who perceived the AI as more useful had higher accuracy. Participants who perceived the AI as more useful may have been more likely to rely on the AI. In Table 5 Model 1, uncertainty information did not significantly increase confidence. Participants who perceived the AI as more useful had higher confidence, likely because participants who

Table 3. Linear mixed regression models for effect of AI recommendations on confidence with (1) all, (2) plant, and (3) animal stimuli.

	Model 1 All B (SE)	Model 2 Plants B (SE)	Model 3 Animals B (SE)
Intercept	0.61 (0.12)***	0.50 (0.12)***	0.37 (0.12)**
AI	0.15 (0.01)***	0.16 (0.02)***	0.27 (0.03)***
Domain knowledge		0.05 (0.01)***	0.07 (0.01)***
Knowledge*AI		−0.01 (0.01)	−0.04 (0.01)***
Average time/image(s)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Attention score	−0.05 (0.02)*	−0.04 (0.02)	−0.02 (0.02)
Task difficulty	−0.02 (0.01)*	−0.01 (0.01)	−0.02 (0.01)*
AI trustworthiness	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
log(age)	−0.04 (0.03)	−0.06 (0.03)	−0.03 (0.03)
Male	0.01 (0.02)	0.01 (0.02)	0.03 (0.02)
College	0.02 (0.02)	0.03 (0.02)	0.02 (0.02)
Number of images	32	19	13
N	398	398	398

*$p < .05$, **$p < .01$, and ***$p < .001$.

Table 4. Linear regression models for effect of uncertainty information on accuracy with (1) all, (2) plant, and (3) animal stimuli.

	Model 1 All B (SE)	Model 2 Plants B (SE)	Model 3 Animals B (SE)
Intercept	0.40 (0.07)***	0.36 (0.09)***	0.39 (0.09)***
Uncertainty information	0.06 (0.01)***	0.08 (0.03)**	0.04 (0.04)
Domain knowledge		0.01 (0.01)	0.01 (0.01)
Knowledge*uncertainty information		0.00 (0.01)	0.00 (0.01)
Average time/image(s)	−0.00 (0.00)	−0.00 (0.00)	0.00 (0.00)
Attention score	0.01 (0.01)	0.02 (0.02)	0.00 (0.02)
Average AI usefulness	0.03 (0.01)***	0.02 (0.01)*	0.03 (0.01)***
Task difficulty	0.01 (0.01)*	0.02 (0.01)**	0.01 (0.01)
AI trustworthiness	−0.01 (0.01)	−0.00 (0.01)	−0.01 (0.01)
log(age)	−0.00 (0.02)	−0.02 0(.02)	0.02 (0.02)
Male	−0.01 (0.01)	−0.01 (0.01)	−0.02 (0.01)
College	0.01 (0.01)	0.03 (0.01)*	−0.02 (0.01)
N	199	199	199
Adjusted R2	0.19	0.19	0.09
F	6.13***	5.23***	2.69**

*$p < .05$, **$p < .01$, and ***$p < .001$.

perceived the AI as more useful were more likely to rely on it. None of the other covariates were significant.

Effect of domain knowledge

In this context, AI recommendations were valuable regardless of domain knowledge. In general, participants had higher domain knowledge for animals than plants, $t(395) = 13$, $p < .001$. Participants with higher animal domain knowledge also tended to report higher plant domain knowledge, $r(200) = 0.40$, $p < .001$.

Participants generally performed better for the animal stimuli. As reported in Table 1, participants were systematically more accurate and confident when identifying animals. Similarly, the AI was also better at identifying animals than plants. As reported in Table 1, for plant images, post-AI accuracy was higher than pre-AI accuracy, but did not exceed the performance of the AI. For animal images, the average performance in the post-AI condition approaches the

Table 5. Linear regression models for effect of uncertainty information on confidence with (1) all, (2) plant, and (3) animal stimuli.

	Model 1 All B (SE)	Model 2 Plants B (SE)	Model 3 Animals B (SE)
Intercept	0.40 (0.12)***	0.34 (0.12)**	0.32 (0.12)**
Uncertainty information	0.00 (0.02)	0.10 (0.04)*	0.16 (0.05)**
Domain knowledge		0.04 (0.01)***	0.06 (0.01)***
Knowledge*uncertainty information		−0.05 (0.02)*	−0.04 (0.01)**
Average time/image(s)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Attention score	−0.03 (0.02)	−0.03 (0.02)	−0.01 (0.02)
Average AI usefulness	0.13 (0.01)***	0.12 (0.01)***	0.10 (0.01)***
Task difficulty	−0.02 (0.01)*	−0.01 (0.01)	−0.02 (0.01)**
AI trustworthiness	−0.02 (0.01)	−0.01 (0.01)	−0.01 (0.01)
log(age)	−0.05 (0.03)	−0.07 (0.03)*	−0.06 (0.03)
Male	0.01 (0.02)	0.02 (0.02)	0.02 (0.02)
College	0.02 (0.02)	0.02 (0.02)	0.04 (0.02)*
N	199	199	199
Adjusted R^2	0.38	0.39	0.41
F	14.29***	12.62***	13.26***

*$p < .05$, **$p < .01$, and ***$p < .001$.

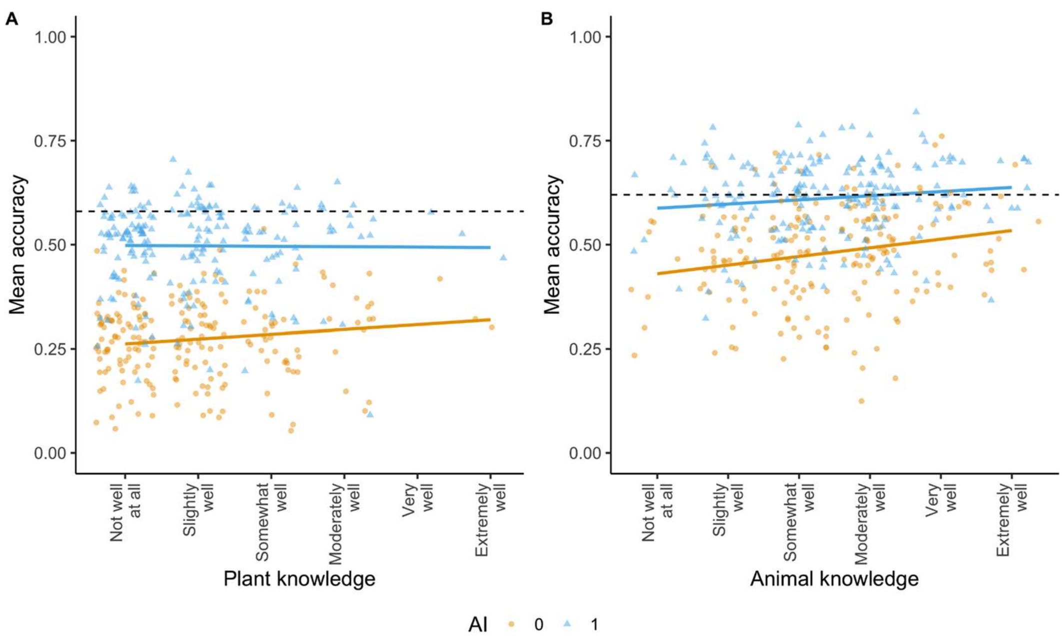

Figure 3. Comparison of pre (yellow circle) and post (blue triangle) accuracy separated by (a) plant and (b) animal domain knowledge. The dotted line represents the performance of the AI.

accuracy of the AI. This suggests that even when domain knowledge varies, AI recommendations help improve participants' performance.

Interaction of domain knowledge and AI recommendations

AI recommendations systematically increased accuracy and confidence across levels of domain knowledge. In Figure 3, there was no evidence of a relationship between accuracy and domain knowledge (i.e. no slope). For both plants and animals, AI recommendations shifted accuracy towards the AI accuracy and some participants were able to exceed it. In contrast, as shown

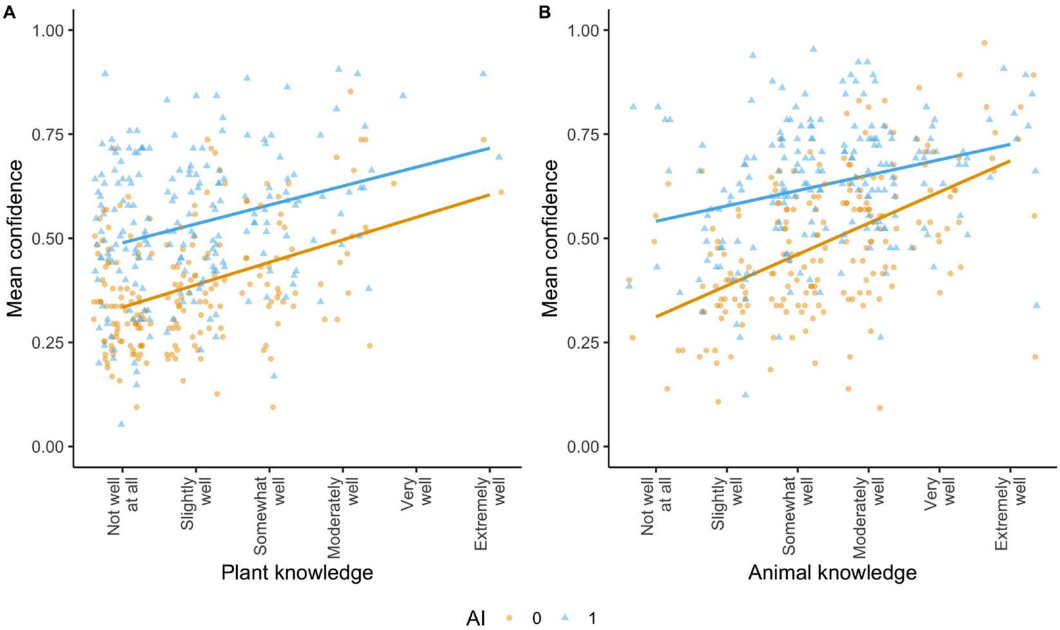

Figure 4. Comparison of pre (yellow circle) and post (blue triangle) confidence separated by (a) plant and (b) animal domain knowledge.

in Figure 4, participants with more domain knowledge tended to be more confident across both plants and animals. In addition, there is evidence of an interaction between domain knowledge and confidence for animal images. In Figure 4b, participants' confidence did not change after receiving AI recommendations with high animal domain knowledge. However, in Figure 3b, participants with high animal domain knowledge still showed improvements in accuracy from the AI recommendations. This suggests that AI recommendations did not contribute to overconfidence for participants with high animal domain knowledge.

In the regression models reported in Tables 2 and 3, AI recommendations significantly increased participants' accuracy and confidence even when accounting for domain knowledge. The effect on accuracy is stronger for plants than animal images, likely because participants relied on the AI more for plants due to lower domain knowledge. Conversely, the effect on confidence is stronger for the animal than plant images, likely because higher domain knowledge allowed participants to gain more metacognitive benefits from the AI recommendations. As expected from Figure 4, there was a significant interaction between domain knowledge and AI for animal stimuli in Table 3 Model 3. This suggests that AI recommendations helped reduce overconfidence, potentially by reminding participants of the limits of their domain knowledge.

Interaction of domain knowledge and uncertainty information

As shown in Figures 5 and 6, uncertainty information did not systematically increase accuracy and confidence across levels of domain knowledge. None of the participants randomly assigned to receive uncertainty information reported high plant domain knowledge (≥3.6), so the data do not span the entire scale range. However, there is evidence of an interaction between domain knowledge and confidence for both plant and animal images. In Figure 6a and b, confidence increases as domain knowledge increases when participants do not receive uncertainty information. However, with uncertainty information, confidence does not vary with domain knowledge. As shown in Figure 6b, uncertainty information increases the confidence of participants with low domain knowledge and decreases the confidence of participants with high domain knowledge. This suggests that

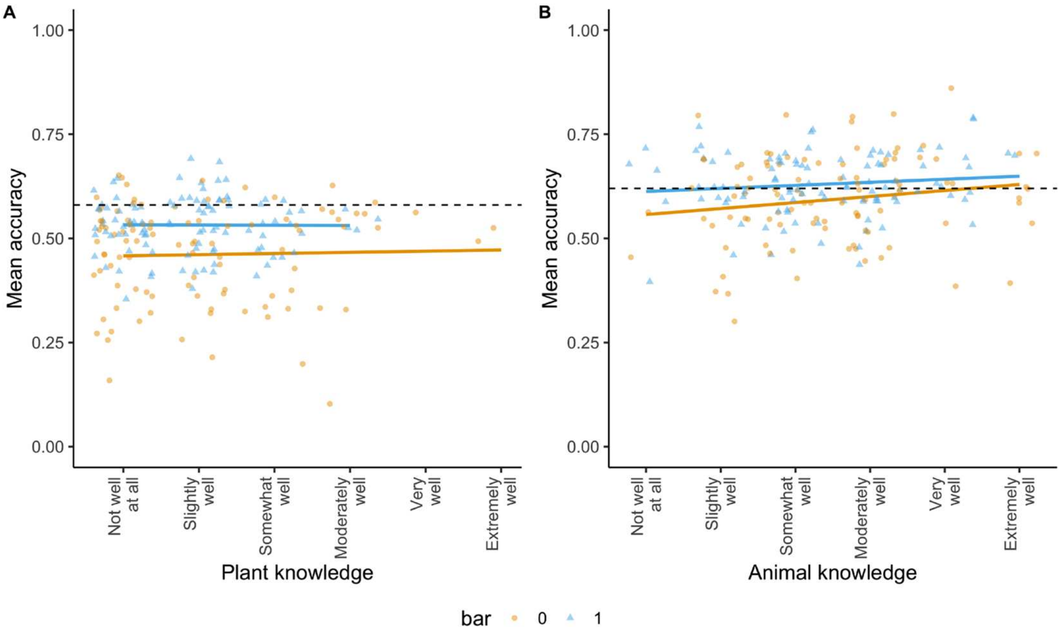

Figure 5. Comparison of with (blue triangle) and without (yellow circle) uncertainty information on accuracy separated by (a) plant and (b) animal domain knowledge. The dotted line represents the performance of the AI.

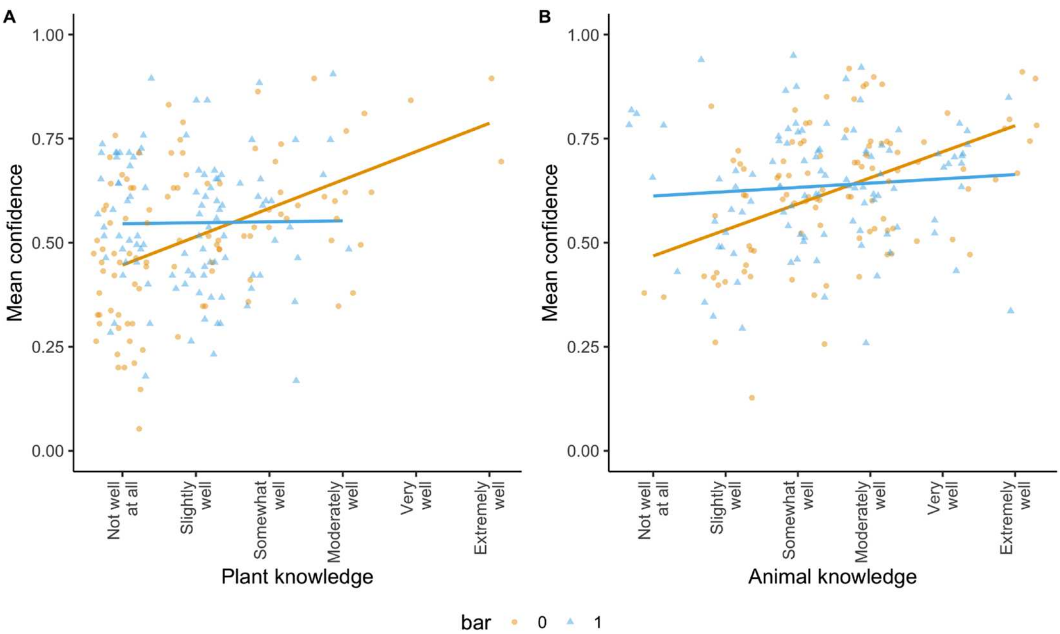

Figure 6. Comparison of with (blue triangle) and without (yellow circle) uncertainty information on confidence separated by (a) plant and (b) animal domain knowledge.

participants who received the uncertainty information tended to have similar levels of confidence irrespective of their domain knowledge and more consistent with their actual performance.

In the regression models reported in Table 4, participants had higher accuracy when provided uncertainty information for plants (Model 2), but there was no effect of uncertainty information for animals (Model 3). This suggests that participants may have relied on the uncertainty information more when they had lower domain knowledge in general. In contrast, in Table 5, participants had higher confidence when provided uncertainty information for animals, but there was no effect for plants. This suggests that participants may have used the AI recommendations for confirmation, increasing their confidence. However, this did not work for plants where there was low domain knowledge. In addition, there was a significant interaction between domain knowledge and uncertainty information for predicting the confidence of animal images (see Table 5 Model 3). This suggests that as domain knowledge increased, the effect of uncertainty information on confidence decreased. This interaction also approached significance ($\alpha = 0.01$) for plant images.

Perceived AI usefulness was associated with higher accuracy and confidence for animals (see Model 3 in Tables 4 and 5). For plant stimuli, this effect was observed for confidence and approached significance for accuracy. Similarly, higher perceived task difficulty was associated with higher accuracy for plants (see Table 4, Model 2) and lower confidence for animals (see Table 5, Model 3). Participants who perceived the AI to be more useful or the task to be more difficult may have relied on the AI more.

Conclusion

In a mixed-subject design, participants performed an image recognition task with and without AI recommendations (within-subjects). For the AI recommendations, participants were randomly assigned to receive uncertainty information or not (between-subjects). Based on the results, we have three primary findings, (1) AI recommendations increased accuracy and confidence, (2) uncertainty information increased accuracy but did not affect confidence, and (3) AI recommendations and uncertainty information reduced overconfidence (by reducing confidence) when there was high domain knowledge.

In this study, AI recommendations increased users' accuracy and confidence across all the models for both plant and animal stimuli. None of the other covariates were significant, suggesting the increase was primarily due to AI recommendations. On average, the human-AI team did not perform better than the AI alone, consistent with other studies that relied on online survey platforms (Green and Chen 2019; Grgić-Hlača, Engel, and Gummadi 2019; Lin et al. 2020) but unlike studies that used professional experts (Bai et al. 2020; Bien et al. 2018; Lakhani and Sundaram 2017; Patel et al. 2019). However, some individuals were able to out-perform the AI.

Uncertainty information improved accuracy but did not affect confidence, consistent with Gkatzia, Lemon, and Rieser (2016). Improvements in accuracy were largely driven by the plant stimuli as uncertainty information did not significantly influence accuracy for animal stimuli. However, uncertainty information did significantly increase confidence for animal, but not plant, stimuli. This may be related to how people used the uncertainty information for the different stimuli. Results indicate that participants who found the AI information useful also performed significantly better on animal stimuli and had significantly higher confidence when responding to both animal and plant stimuli. The plant stimuli were more difficult so participants may have relied on the AI more, increasing accuracy but not changing post-AI confidence (Antifakos et al. 2005). In contrast, participants may have used the uncertainty information as a second opinion for the animal stimuli, increasing confidence but not changing post-AI accuracy.

Overall, participants reported higher domain knowledge for animals than plants, which was consistent with their performance as well as the AI's performance. Participants with higher domain knowledge consistently had higher confidence, but not necessarily accuracy. Participants with higher animal domain knowledge had higher accuracy for animal stimuli but this was not consistent across modeling paradigms. This suggests that self-reported domain knowledge may not be a reliable measure of expertise. In addition, there is a risk of overconfidence when participants perceive they have higher domain knowledge. In the context of human-AI teams, this is when it is critical to clearly communicate when it is or is not appropriate to rely on the AI.

There was some evidence that domain knowledge interacted with the presence of AI recommendations to increase accuracy and confidence. For the animal stimuli, participants with high domain knowledge were more accurate and more confident when there were AI recommendations. However, for the plant stimuli, participants with high domain knowledge were more confident, but not more accurate when there were AI recommendations. This may be related to the fact that there was lower domain knowledge related to plants overall. Overall, this suggests that AI recommendations can improve performance, but as discussed above, there is a risk of inducing overconfidence, particularly when participants have high perceived domain knowledge.

Providing uncertainty information may be an effective strategy for minimizing overconfidence in human-AI teams. For the animal stimuli, participants with high domain knowledge were less confident, but there was no effect on accuracy when there was uncertainty information. For the plant stimuli, there was no significant interaction between domain knowledge and uncertainty information. This suggests that uncertainty information helps calibrate users and reduces overconfidence.

This study has three primary limitations. First, participants were recruited from Prolific and did not necessarily include true experts at identifying plants and animals. This limited our ability to make inferences, particularly for the plant stimuli where there was lower overall domain knowledge. Second, the domain knowledge scores were self-reported and may have been influenced by their perception of task performance. Future work should use objective measures of domain knowledge to better understand how this factor influences the way people use and perceive AI recommendations (Greis et al. 2017; Zhou et al. 2015). Third, confidence was measured on a Likert scale, which reduced the resolution of the measurement. However, this makes the test more conservative and we still observed effects while minimizing participant burden.

AI recommendations have the potential to improve performance across a wide range of domains. However, users may interact with those recommendations differently depending on their knowledge and experience in the domain. This study suggests that recommendations that include uncertainty may support efforts to help users calibrate and integrate their domain knowledge with the AI input. In particular, communicating uncertainty supports efforts to increase accuracy without inducing overconfidence.

Acknowledgments

We thank Cihan Dagli, Krista Lentine, Mark Schnitzler, and Henry Randall for their insights on the design of AI decision support systems.

Disclosure statement

The authors report that there are no competing interests to declare.

Funding

This work was supported by a National Science Foundation Award #2026324.

References

Abdar, Moloud, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, et al. 2021. “A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges.” *Information Fusion* 76: 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Antifakos, Stavros, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. "Towards Improving Trust in Context-Aware Systems by Displaying System Confidence." In *MobileHCI '05: Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services*, 9–14. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1085777.1085780

Ashktorab, Zahra, Q. Vera Liao, Casey Dugan, James Johnson, Qian Pan, Wei Zhang, Sadhana Kumaravel, and Murray Campbell. 2020. "Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes." *Proceedings of the ACM on Human-Computer Interaction* 4 (CSCW2): 1–20. https://doi.org/10.1145/3415167

Bai, Harrison X., Robin Wang, Zeng Xiong, Ben Hsieh, Ken Chang, Kasey Halsey, Thi My Linh Tran, et al. 2020. "Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT." *Radiology* 296 (3): E156–E165. https://doi.org/10.1148/radiol.2020201491

Bansal, Gagan, Tongshuang Wu, and Joyce Zhou. 2021. "Does the Whole Exceed Its Parts? The Efect of Ai Explanations on Complementary Team Performance." In *Conference on Human Factors in Computing Systems - Proceedings*. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445717

Bhatt, Umang, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, et al. 2021. "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty." In *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3461702.3462571

Bien, Nicholas, Pranav Rajpurkar, Robyn L. Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, et al. 2018. "Deep-Learning-Assisted Diagnosis for Knee Magnetic Resonance Imaging: Development and Retrospective Validation of MRNet." *PLoS Medicine* 15 (11): e1002699. https://doi.org/10.1371/journal.pmed.1002699

Bles, Anne Marthe, van der, Sander, van der Linden , Alexandra L. J. Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval, and David J. Spiegelhalter. 2019. "Communicating Uncertainty about Facts, Numbers and Science." *Royal Society Open Science* 6 (5): 181870. https://doi.org/10.1098/rsos.181870

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

Bussone, Adrian, Simone Stumpf, and Dympna O'Sullivan. 2015. "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems." In *2015 International Conference on Healthcare Informatics*, 160–169. https://doi.org/10.1109/ICHI.2015.26

Chong, Leah, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. "Human Confidence in Artificial Intelligence and in Themselves: The Evolution and Impact of Confidence on Adoption of AI Advice." *Computers in Human Behavior* 127: 107018. https://doi.org/10.1016/j.chb.2021.107018

Dastin, Jeffrey. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." In *Ethics of Data and Analytics*, 296–299. New York, NY: Auerbach Publications.

Elder, H., C. Canfield, D. B. Shank, T. Rieger, and Casey Hines. 2022. "Knowing When to Pass: The Effect of AI Reliability in Risky Decision Contexts." *Human Factors* [online Online ahead of print on May 21, 2022]. https://doi.org/10.1177/001872082211006.

Feng, Shi, and Jordan Boyd-Graber. 2019. "What Can Ai Do for Me? Evaluating Machine Learning Interpretations in Cooperative Play." In *International Conference on Intelligent User Interfaces, Proceedings IUI, Part F147615*, 229–239. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3301275.3302265

Fernandes, Michael, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. "Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making." In *CHI '18*: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3173574.3173718

Ferrario, Andrea, Michele Loi, and Eleonora Viganò. 2020. "Trust Does Not Need to Be Human: It is Possible to Trust Medical AI." *Journal of Medical Ethics* 47 (6): 437–438. https://doi.org/10.1136/medethics-2020-106922

Gkatzia, Dimitra, Oliver Lemon, and Verena Rieser. 2016. "Natural Language Generation Enhances Human Decision-Making with Uncertain Information." CoRR *abs/1606.0*. http://arxiv.org/abs/1606.03254.

Green, Ben, and Yiling Chen. 2019. "Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments." In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287563

Greis, Miriam, Emre Avci, Albrecht Schmidt, and Tonja Machulla. 2017. "Increasing Users' Confidence in Uncertain Data by Aggregating Data from Multiple Sources." In *CHI '17*: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 828–840. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3025453.3025998

Grgić-Hlača, Nina, Christoph Engel, and Krishna P. Gummadi. 2019. "Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing." *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW): 1–25. https://doi.org/10.1145/3359280

Gunning, David, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang Zhong Yang. 2019. "XAI—Explainable Artificial Intelligence." *Science Robotics* 4 (37): eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2017. "ImageNet Classification with Deep Convolutional Neural Networks." *Communications of the ACM* 60 (6): 84–90. https://doi.org/10.1145/3065386

Lakhani, Paras, and Baskaran Sundaram. 2017. "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks." *Radiology* 284 (2): 574–582. https://doi.org/10.1148/radiol.2017162326

Lin, Zhiyuan "Jerry", Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. "The Limits of Human Predictions of Recidivism." *Science Advances* 6 (7): eaaz0652. https://doi.org/10.1126/sciadv.aaz0652

Lipkus, Isaac M., and J. G. Hollands. 1999. "The Visual Communication of Risk." *Journal of the National Cancer Institute. Monographs* 1999 (25): 149–163. https://doi.org/10.1093/oxfordjournals.jncimonographs.a024191

Logg, Jennifer M., Julia A. Minson, and Don A. Moore. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* 151: 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Maadi, Mansoureh, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. 2021. "A Review on Human–AI Interaction in Machine Learning and Insights for Medical Applications." *International Journal of Environmental Research and Public Health* 18 (4): 2121. https://doi.org/10.3390/ijerph18042121

Madsen, Maria, and Shirley D. Gregor. 2000. "Measuring Human-Computer Trust." In *Proceedings of the 11th Australasian Conference on Information Systems*, p. 53. Brisbane, Australia: Information Systems Management Research Centre.

Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. "An Integrative Model of Organizational Trust." *The Academy of Management Review* 20 (3): 709–734. https://doi.org/10.2307/258792

Moore, Don, and Paul Healy. 2008. "The Trouble with Overconfidence." *Psychological Review* 115 (2): 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Patel, Bhavik N., Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, et al. 2019. "Human–Machine Partnership with Artificial Intelligence for Chest Radiograph Diagnosis." *Npj Digital Medicine* 2 (1): 111. https://doi.org/10.1038/s41746-019-0189-7

Peer, Eyal, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70: 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Piehlmaier, Dominik M. 2022. "Overconfidence and the Adoption of Robo-Advice: Why Overconfident Investors Drive the Expansion of Automated Financial Advice." *Financial Innovation* 8 (1): 14. https://doi.org/10.1186/s40854-021-00324-3

Rosenberg, Louis, Matthew Lungren, Safwan Halabi, Gregg Willcox, David Baltaxe, and Mimi Lyons. 2018. "Artificial Swarm Intelligence Employed to Amplify Diagnostic Accuracy in Radiology." In *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 1186–1191. https://doi.org/10.1109/IEMCON.2018.8614883

Snow, Rion, Brendan O'connor, Dan Jurafsky, and Y. Ng Andrew. 2008. "Cheap and Fast—But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks." In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254–263. Honolulu, Hawaii: Association for Computational Linguistics.

Sohn, Emily. 2023. "The Reproducibility Issues That Haunt Health-Care AI." *Nature* 613 (7943): 402–403. https://doi.org/10.1038/d41586-023-00023-2

Subramanian, Harishankar V., Casey I. Canfield, Daniel Burton Shank, Luke Andrews, and Cihan H. Dagli. 2020. "Communicating Uncertain Information from Deep Learning Models in Human Machine Teams." In *Proceedings of the American Society for Engineering Management 2020 International Annual Conference*, 167. Huntsville, AL: American Society for Engineering Management (ASEM).

Tomsett, Richard, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. "Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI." *Patterns (New York, N.Y.)* 1 (4): 100049. https://doi.org/10.1016/j.patter.2020.100049

Venkatesh, Viswanath, and Fred D. Davis. 2000. "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies." *Management Science* 46 (2): 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Wang, Jinping, Maria D. Molina, and S. Shyam Sundar. 2020. "When Expert Recommendation Contradicts Peer Opinion: Relative Social Influence of Valence, Group Identity and Artificial Intelligence." *Computers in Human Behavior* 107: 106278. https://doi.org/10.1016/j.chb.2020.106278

Wang, Weiquan, and Izak Benbasat. 2013. "Research Note—A Contingency Approach to Investigating the Effects of User-System Interaction Modes of Online Decision Aids." *Information Systems Research* 24 (3): 861–876. https://doi.org/10.1287/isre.1120.0445

Wang, Xinru, and Ming Yin. 2021. "Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making." In *26th International Conference on Intelligent User Interfaces*, 318–328. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3397481.3450650

Zhang, Yinjunjie, Zhicheng Xu, and Marco A. Palma. 2019. "Conveniently Dependent or Naively Overconfident? An Experimental Study on the Reaction to External Help." *PloS One* 14 (5): e0216617. https://doi.org/10.1371/journal.pone.0216617

Zhang, Yunfeng, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. "Efect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making." In *FAT* 2020: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372852

Zhou, Jianlong, Constant Bridon, Fang Chen, Ahmad Khawaji, and Yang Wang. 2015. "Be Informed and Be Involved: Effects of Uncertainty and Correlation on User's Confidence in Decision Making." In *CHI EA '15: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 923–928. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/2702613.2732769