HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230087

How Displaying AI Confidence Affects Reliance and Hybrid Human-AI Performance

Heliodoro TEJEDA LEMUS [a,1], Aakriti KUMAR [a] and Mark STEYVERS [a]
[a] *Department of Cognitive Sciences, University of California, Irvine*

Abstract. Confidence signals are often used in human interactions to communicate the likelihood of a decision being correct. Similarly, confidence may also be used to indicate the reliability of advice given by an AI. While previous work on explainable AI (XAI) has explored the effect of AI confidence on AI-advice adoption and joint accuracy of the human-AI team, most studies use AI-assistants that exceed human performance. It is unclear how displaying the confidence interacts with the accuracy of the AI. We conduct a comprehensive investigation of the effect of displaying AI confidence on two factors: 1) the accuracy of AI-assisted decision making, and 2) reliance on the AI's assistance. We conduct two behavioral experiments, one where participants were shown AI confidence, and another where no confidence ratings were shown. Our work goes beyond the typical focus on high accuracy AI assistants. In both experiments, participants were assisted by one of three AI classifiers of varying accuracy. Our results demonstrate that displaying AI confidence increases joint accuracy when people are assisted by a classifier that is better than humans on average. Conversely, when assisted by a classifier with performance worse than an average human, joint accuracy was better when no AI confidence was displayed. However, for the adoption of AI advice we observed the opposite pattern: people rely more on a higher accuracy classifier that does not display confidence compared to one that does, and people rely more on a lower accuracy classifier that does display AI confidence compared to one that does not.

Keywords. Human-AI collaboration, Confidence, XAI, AI-assisted Decision Making

1. Introduction

AI systems are increasingly being added to human workflows but human-AI collaboration is often plagued by inefficiencies. Most times, this can be attributed to humans' incorrect assessment of the AI's ability or a lack of understanding of the AI's response. However, humans successfully engage in similar collaborative efforts when working with other humans. When working together, people use verbal and visual cues to signal confidence and communicate the likelihood of a decision being correct [1,2]. For instance, a group of friends playing a trivia game are able to decide who should answer a specific question based on verbal exchanges in addition to an existing mental model of

[1]Corresponding Author: Heliodoro Tejeda, Department of Cognitive Sciences, University of California, Irvine, CA 92697-5100; E-mail: htejeda@uci.edu

each player's expertise. As they take turns answering questions, if one friend says, "I'm pretty sure I know the answer to this one", the others sense her confidence and allow her to respond to that question. The group receives feedback about the friend's expertise and confidence calibration which allows them to update their mental model of their friend. This exchange of confidence signals and turn-taking facilitates their collaboration and success in the game. Inspired by such human interactions, the field of explainable AI (XAI) has looked into ways of conveying an AI's uncertainty in its decisions to signal the reliability of its advice. These uncertainty signals are in the form of confidence intervals for regression tasks or estimated probabilities of correct predictions in classification tasks [3,4,5].

Many studies have investigated the effect of showing AI confidence on AI-advice adoption, trust calibration, and accuracy of AI-assisted decision making [6,7,8]. Recent work [9,5] has also examined the evolution of human confidence in AI and in themselves and found that human self-confidence influences their decision to adopt AI advice. In particular, [5] demonstrated that the confidence differential between AI and the human drives reliance decisions. Humans are more likely to adopt AI advice if the AI indicates high confidence while humans themselves, based on their own independent decision-making process, have low confidence.

In this paper we pursue a comprehensive investigation of the influence of displaying AI confidence on 1) AI-advice adoption, and 2) accuracy of AI-assisted decision making. We conduct a behavioral study in which participants classify images with the assistance of an AI similar to the Judge-Advisor paradigm presented in [5]. In the study participants are tasked with classifying noisy images in two phases. First, participants classified a noisy image without the assistance of the AI. After generating an initial classification, participants are shown advice provided by an AI, for the same image, to aid with their final classification decision. Critically, the experiment varies two main factors. First, we

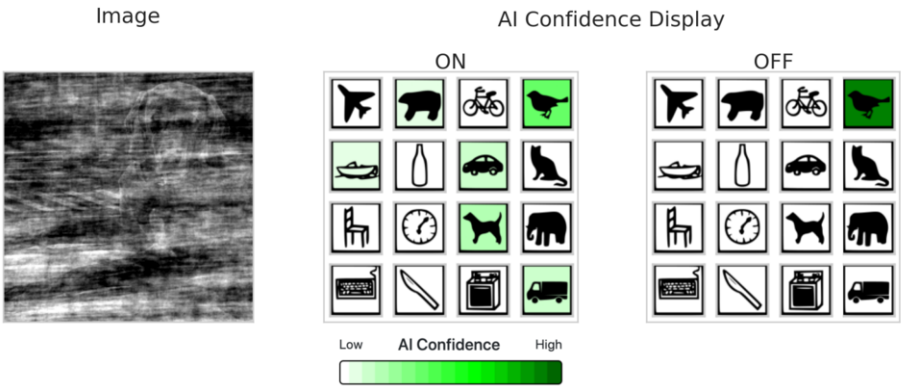


Figure 1. Illustration of the different experimental conditions, confidence displayed versus not displayed. The left panel shows an example of a noisy image. The center panel displays AI advice for the example image in the confidence display ON condition (note that multiple options are highlighted). The right panel shows the AI advice for the same image in the AI confidence display OFF condition. In this condition, no AI confidence information is displayed. Instead, a single option corresponding to the most likely class is displayed at a fixed green hue.

manipulate (between participants) whether or not AI confidence is displayed. When AI confidence is not shown, the AI advice is shown by highlighting only the AI's top predicted classification decision (see Figure 1 for an illustrative example). When AI confidence displays are shown, the full set of AI's predictions were visualised using 'hues' corresponding to the confidence of the AI. Darker hues corresponded to higher confidence in the prediction. The second main factor in the experiment is the overall accuracy of the AI. We used different machine classifiers that performed at three levels of accuracy: well below human accuracy (Classifier A), similar to human accuracy (Classifier B), and better than human accuracy (Classifier C). By manipulating the AI accuracy, we investigate whether the reliance on AI interacts with overall AI accuracy.

2. Methods

2.1. Participants

For this study, a total of 135 participants completed the study using Amazon Mechanical Turk. Before the start of the experiment, participants were given instructions explaining the user interface and guiding them through the experimental procedure. Once all instructions pages had been read, a comprehension quiz consisting of five noisy images to be classified (without AI assistance) was given to participants to ensure their understanding of the experiment. In order to pass the comprehension quiz and ensure that participants were not randomly clicking on options, the participants needed to correctly classify four of the five images they were shown during the quiz. There were two opportunities given to participants to pass the comprehension quiz and re-enforce that they fully understood the user interface and the experiment. Upon successfully passing the comprehension quiz, participants were then granted access to the main experiment. Participants that completed the study were compensated $7 USD for their time. The experimental protocol was approved by the University of California, Irvine Institutional Review Board.

2.2. Images

All of the images used for this experiment come from the ImageNet Large Scale Visual Recognition Challenge (ILSRVR) 2012 validation dataset [10]. This study followed the

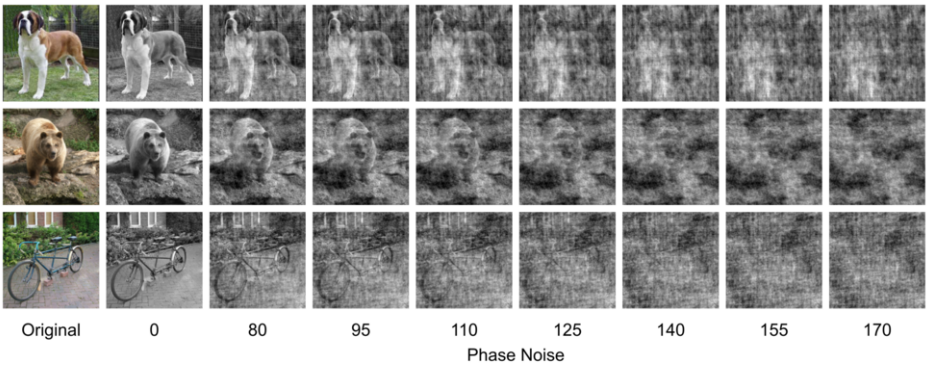


Figure 2. Illustration of three images under different levels of phase noise. Original images (left) were not used in experiments and are shown only for illustrative purposes.

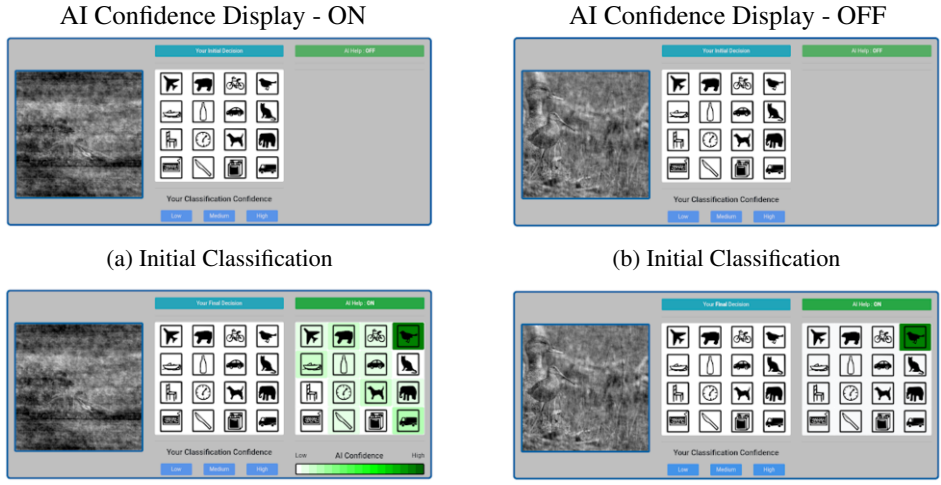
procedure designed by [11] in which the number of ImageNet classes was reduced from 1000 to 16 (airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven, and truck), which we have termed ImageNet-16. We then randomly selected 256 unique images from our ImageNet-16 dataset to serve as the images to be used for this experiment. To ensure that the classes were evenly balanced in our experimental dataset, during our random selection we ensured to select 16 unique images from each of the 16 ImageNet-16 class categories. Image distortion was then implemented for the entire experimental dataset (256 images) by uniformly distributing phase noise in the interval $[-w, w]$ at each spatial frequency [11]. Each image was randomly assigned one of the eight phase noise levels, $w = 0, 80, 95, 110, 125, 140, 155, 170$, resulting in each category class having two unique images with the same phase noise level. Examples of the phase noise manipulation can be seen in Figure 2.

2.3. AI Models - VGG-19

We used the VGG-19 architecture, a convolutional neural network (CNN), that was pre-trained on the ImageNet dataset as the basis for our classifiers providing AI assistance [12]. We trained three different versions of the VGG-19 model to provide variability in the overall performance of the classifiers providing AI advice. We will refer to these classifiers, in order of performance as Classifier A, Classifier B, and Classifier C. The three VGG-19 classifiers were trained by fine-tuning to differing degrees on the ImageNet-16 dataset (created using the ILSRVR ImageNet training dataset). Classifier A did not undergo any fine-tuning; rather the weights from a pre-trained VGG-19 model (trained on ImageNet) were used. This lead to Classifier A having a performance well below a baseline of human performance on our task. Classifiers B and C were fine-tuned using the following procedure: loading the pre-trained VGG-19 model weights, adding a final layer to the model to output 16 classes, followed by training on all different phase noise levels all at once. Classifier B was fine-tuned for less than one epoch (10% of batches of the first epoch) and provides a performance which is roughly around baseline human performance. Classifier C was fine-tuned for 10 epochs and provides a performance level above baseline human performance. Baseline human performance was determined by a pilot study (without AI advice) of 145 participants.

2.4. Experiment Procedure

Participants were tasked with classifying a total of 192 noisy images drawn from a uniform distribution over 256 unique noisy images. Each of the 192 trials followed the Judge-Advisor sequential paradigm [5] in which two classifications are made for the same image. The initial classification is made by the participant alone, while the final classification is made after receiving advice from an AI assistant. Participants were instructed to classify each image to the best of their abilities then to leverage the AI assistant (for their final decision) to optimize their performance. The three classifiers (A, B, C) with varying performance accuracy were used as the AI assistants for the participants. Additionally, AI assistance was displayed in two different formats, one which provided AI confidence displays for each of the classification options and the other which only provided a singular classification suggestion (see Figure 1 for illustrative examples). The performance level of the AI assistant as well as whether or not the AI assistant provided



(a) Initial Classification (b) Initial Classification

(c) Final Classification (d) Final Classification

Figure 3. Illustration of the behavioral experiment interface in both AI assistance conditions. The left column (a, c) displays the experimental interface when AI confidence is displayed to participants

confidence with its advice or not were between-subject manipulations. At the start of the experiment, each participant was assigned into either the AI confidence displayed condition or the AI no confidence displayed condition and assigned a single classifier performance level, classifier A, B, or C, and would only receive advice from that AI assistant in that particular display manner for the entirety of the experiment. A total of 66 participants were assigned to the AI confidence displayed condition (23 for classifier A, 23 for classifier B, and 20 for classifier C) and 69 participants were assigned to the AI no confidence displayed condition (24 for classifier A, 24 for classifier B, and 21 for classifier C). Figure 3 displays the experimental interface in both AI assistance conditions.

User Interface: The experiment has a three-column layout with which participants interact with. The leftmost column displays the noisy image to be classified. The middle column holds a button grid for each of the 16 categories that can be selected as well as three submission buttons indicative of a participants confidence in their classification (*low, medium, high*). Finally, the rightmost column is where AI advice is displayed (only after an initial classification has been made). Figure 3c and Figure 3d demonstrate the difference between the AI confidence displayed versus the no AI confidence displayed between-subject manipulation.

In the AI confidence display condition, a color gradient was used to convey model confidence across the 16 classes. The AI predictions were binned by their probabilities of class prediction. Each bin corresponded to a color gradient ranging from white (hsl(120, 100%, 100%)) no confidence in this prediction, to dark green (hsl(120, 100%, 20%)) extremely confident in this prediction. A color bar was provided to participants to display the mapping of color hue to AI confidence. For the no confidence display experimental condition, only the maximum prediction value was displayed using a green hue (hsl(120, 100%, 25%)). Figure 1 displays the difference between the two experimental conditions

of AI confidence displayed versus no confidence displayed.

As can be seen, the AI advice in Figure 3c has multiple options highlighted at different hue intensities. The darker the hue of the highlighted category, the more confident the AI is in that particular prediction, as indicated by the AI Confidence hue bar at the bottom of the rightmost column. In comparison, Figure 3d displays AI advice without confidence being displayed. All AI advice is presented with a fixed hue of green highlighting a singular (max) prediction.

Feedback: After submitting a final classification for each trial, feedback is provided to aid participants in understanding their own abilities as well as the abilities of the AI. Feedback was displayed in the middle panel by highlighting the true image classification button in blue while highlighting an incorrect participant classification in red.

3. Results

3.1. Effect of AI Assistance on Classification Accuracy

Figure 4 shows classification accuracy for both AI confidence display experimental conditions across the three different classifers (A, B, and C). The blue line displays initial classification performance (without AI assistance), the orange line displays final classification performance (after receiving AI assistance), and the black dashed lines display the AI classifier performance. The results show that classifiers A, B, and C perform below average (initial) human performance, around average human performance, and above average human performance respectively. We can see that irrespective of AI confidence display condition and classifier performance level, participants increase (on average) their accuracy from their initial classification to their final classification.

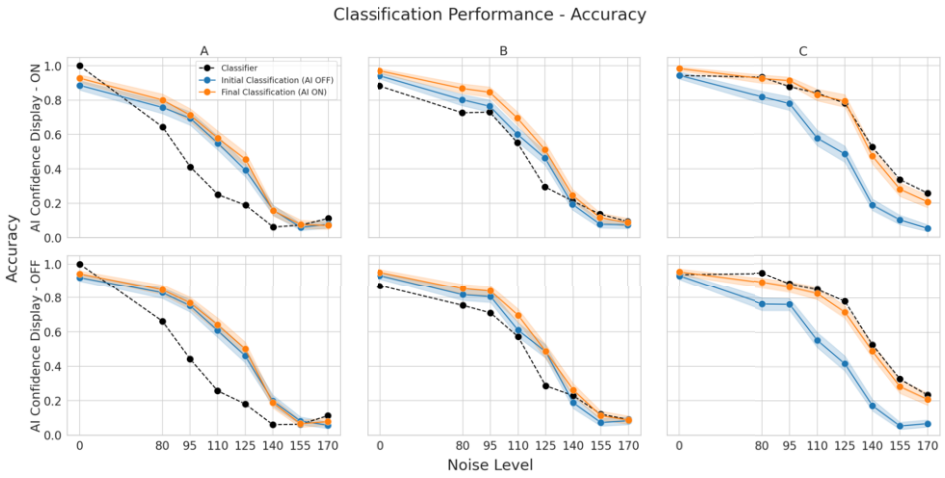


Figure 4. Accuracy on the image classification task across the eight different noise levels. Each column represents a particular classifier level: A (worst classifier), B, and C (best classifier). The dashed black line is the classifiers accuracy. The two rows correspond to the different AI confidence display conditions. The first row displays accuracy for the AI confidence display - ON condition while the second row displays accuracy for the AI confidence display - OFF condition.

3.2. Effect of Displaying AI Confidence on Final Classification Accuracy

A regression analysis was run to determine the effect of displaying AI confidence on classification accuracy (comparing the orange lines in Figure 4). The results reveal small effects of displaying AI confidence on accuracy. To determine the significance of this effect, logistic regression was used to analyze the relationship between showing confidence displays on the probability of making a correct final decision while using noise level as a covariate. We performed this analysis separately for each of the three model performance levels. For model A, holding noise level constant, the odds of a correct final decision decreased by 23% (95% CI [.10, .36], $p < 0.001$) when the confidence display was presented. For model C, the same analysis showed that the odds of correct final decision increased by 14% (95% CI [.02, .28], $p < 0.02$) when the confidence display was presented. There was no significant effect for model B. The key takeaway from our experiments is that there are small but significant differences in classification accuracy when participants are shown both AI classification and confidence compared to when they are shown only the top class predicted by the AI.

3.3. Effect of Displaying AI Confidence on Reliance

Figure 5 displays the probability of a participant switching their initial classification after being presented with AI advice. As seen in the top row, participants are more likely to switch their initial classification as noise level (difficulty) increases. We used logistic regression to analyze the effect confidence displays on the probability of switching classification with noise level as a covariate. For model A and B, the odds of switching

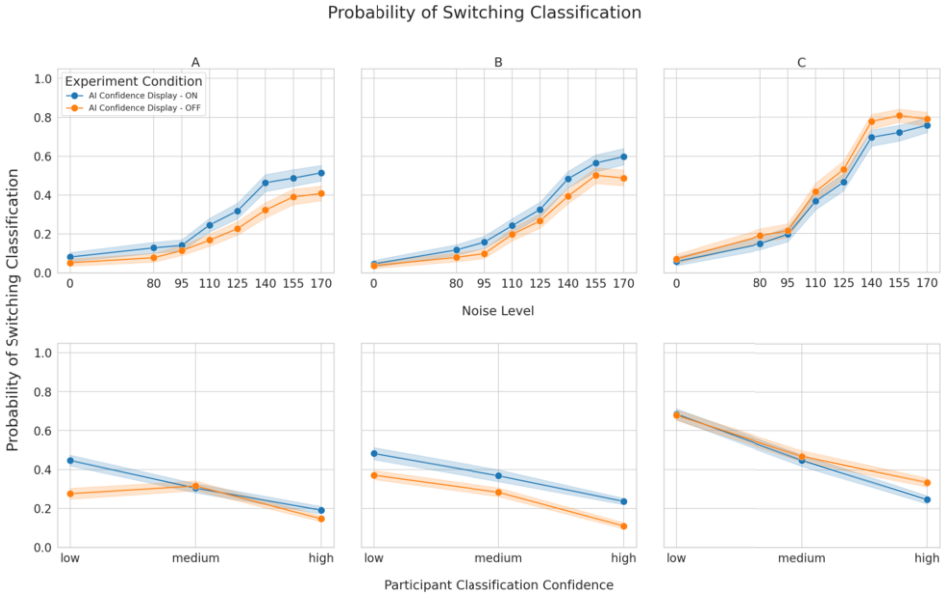


Figure 5. Probability of switching initial classification after receiving AI advice for three different classifier levels: A (worst classifier), B and C (best classifier). The first row displays the probability a participant switches their initial classification as a function of image noise level. The second row displays the probability of a participant switching their initial classification as a function of their initial decision confidence.

classification increased by 56% (95% CI [.42 .73], $p < 0.001$) and 41% (95% CI [.27 .56], $p < 0.001$) respectively when confidence displays were presented. For model C, the odds of switching classification decreased by 30% (95% CI [.17 .45], $p < 0.001$) when confidence displays were presented. The bottom row of Figure 5 shows that participants are less likely to switch their initial classification the more confident they are in their own initial decision.

3.4. Summary of Results

Overall, the results show that the effect of showing AI confidence displays depends on classifier performance. For Classifier C (the best performing classifier), presenting AI confidence displays increased joint human-AI accuracy and reduced reliance when human confidence was high. However, for Classifier A (the worst performing classifier), presenting AI confidence displays decreased joint human-AI accuracy and led to an increase in reliance regardless of human confidence.

4. Discussion

Explainable AI has the difficult task of ensuring transparency and explainability without overwhelming humans who are assisted by AI agents. Achieving this goal requires striking a delicate balance between decreasing cognitive load and optimizing information gain [13]. In this study, we investigated the role of displaying AI confidence, a commonly used explainability technique, in improving joint accuracy and reliance on AI-assistance. Our results showed an interaction between accuracy of the AI and displaying AI confidence. For classifier C (the best performing classifier), displaying AI confidence was beneficial as it improved joint accuracy and reduced over-reliance on assistance. We hypothesize that for this classifier, displaying AI confidence provided participants additional context to adjust their reliance especially when the participants themselves were highly confident (e.g., they could rely less on the AI in case the AI is not confident). In the absence of AI confidence scores, participants might rely on more generic reliance strategies that relate to overall AI accuracy, leading to an overreliance on AI when the participant is highly confident.

On the other hand, for classifier A (the worst performing classifier), we observe the opposite effect – showing AI confidence scores reduces accuracy. We hypothesize that factors other than overall accuracy may have contributed to this result. In particular, one critical factor might be the *calibration error* which relates the confidence scores to the empirical accuracy (e.g., as assessed by measures such as the expected calibration error metric [14]). The calibration error is the highest for classifier A and therefore, displaying confidence scores might have misled participants. Overall, if our hypotheses are true, it is better not to show confidence scores in cases where the calibration error is high. However if the calibration error is sufficiently low, confidence provides additional context that people may use to adjust their reliance on a trial-by-trial basis. Currently, our results do not allow us to disentangle the effects of AI accuracy and confidence calibration when displaying AI confidence to humans. Hence, an important direction for future research is to examine AI-assisted decision making where calibration and accuracy are manipulated independently.

References

[1] J. Navajas, T. Niella, G. Garbulsky, B. Bahrami, and M. Sigman, "Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds," *Nature Human Behaviour*, vol. 2, no. 2, pp. 126–132, 2018.

[2] B. Bahrami, K. Olsen, P. E. Latham, A. Roepstorff, G. Rees, and C. D. Frith, "Optimally interacting minds," *Science*, vol. 329, no. 5995, pp. 1081–1085, 2010.

[3] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo, *et al.*, "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

[4] M. Jacobs, M. F. Pradier, T. H. McCoy Jr, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos, "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection," *Translational psychiatry*, vol. 11, no. 1, p. 108, 2021.

[5] H. Tejeda, A. Kumar, P. Smyth, and M. Steyvers, "Ai-assisted decision-making: A cognitive modeling approach to infer latent reliance strategies," *Computational Brain & Behavior*, pp. 1–18, 2022.

[6] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305, 2020.

[7] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, and M. Ghassemi, "Do as ai say: susceptibility in deployment of clinical decision-aids," *NPJ digital medicine*, vol. 4, no. 1, p. 31, 2021.

[8] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-ai team performance," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, pp. 2–11, 2019.

[9] L. Chong, G. Zhang, K. Goucher-Lambert, K. Kotovsky, and J. Cagan, "Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice," *Computers in Human Behavior*, vol. 127, p. 107018, 2022.

[10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[11] R. Geirhos, C. R. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[14] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, 2015.