June 2022

An Empirical Investigation of Reliance on AI-Assistance in a Noisy-Image Classification Task

Heliodoro Tejeda Lemus [a] Aakriti Kumar [a] Mark Steyvers [a]
[a] *Department of Cognitive Science, University of California, Irvine, USA*

Abstract. Humans use AI-assistance in a wide variety of high- and low-stakes decision-making tasks today. However, human reliance on the AI's assistance is often sub-optimal — with people exhibiting under- or over-reliance on the AI. We present an empirical investigation of human-AI assisted decision-making in a noisy image classification task. We analyze the participants' reliance on AI assistance and the accuracy of human-AI assistance as compared to the human or AI working independently. We demonstrate that participants do not show *automation bias* which is a widely reported behavior displayed by humans when assisted by AI. In this specific instance of AI-assisted decision-making, people are able to correctly override the AI's decision when needed and achieve close to the theoretical upper bound on combined performance. We suggest that the reason for this discrepancy from previous research findings is because 1) people are experts at classifying everyday images and have a good understanding of their ability in performing the task, 2) people engage in the metacognitive act of deliberation when asked to indicate confidence in their decision, and 3) people were able to build a good mental model of the AI by incorporating feedback that was provided after each trial. These findings should inform future experiment design.

Keywords. AI-Assistance, Decision-Making, Automation Bias, Human-Subject Experiment, Image Classification

1. Introduction

From making mundane shopping choices to thinking through high-stakes medical scenarios, there has been a sharp increase in the deployment of AI-assistants to help humans make decisions [15, 9, 5, 1]. In line with the old adage, "two minds are better than one", such collaborative human-AI decision-making are expected to increase the efficacy of decisions. However, recent work shows mixed results: while some studies report that decisions made jointly by the human and AI are more effective than either the human or the AI working independently [22, 14, 13, 20], other studies highlight humans' sub-optimal use of AI-advice and explanations [2, 21, 25]. Many empirical investigations of joint human-AI decision-making have indicated that humans are susceptible to biases and errors when working with AI assistance. People may over- or under-rely on the AI's advice leading to sup-optimal performance. Over- or under-reliance on the AI's assistance indicates miscalibrated 'trust' on the part of the human. Trust commensurate to the AI agent's capabilities is critical to effective joint decision-making.

In this paper, we investigate people's reliance on an AI agent's assistance when the AI agent's advice is readily available to them when making a decision. We present data from a behavioral experiment where participants classify noisy images into one of 16 categories. The experiment has two within-subject conditions: one where participants can see an AI assistant's classification and confidence alongside the image, another where participants classify images without the AI classifier's help (control). We varied the accuracy of the AI assistants to look at differences in AI-assisted performance when the AI was better, similar, and worse than humans on the task. Figure 3 shows the experimental interface in both conditions.

Previous work has shown that a human's inclination to seek or incorporate advice is closely tied to their self-confidence in their decision [4, 10]. In our experiment, we capture participants' confidence ratings on each of their classifications. We use classification probabilities of the predicted class provided by the AI as a proxy for the its confidence in the classification [11]. We explore the reliance strategies of humans on the AI agent and accuracy of AI-assisted decision making. To preview the results, we observe that people's reliance on the AI assistant is close to optimal. We show that people do not over-rely on the AI even when their confidence in their own classifications is low. We posit that a useful characterisation of a human's reliance behavior on AI assistance needs to take into account not only the accuracy of the two agents, but also the confidence of the agents in their decisions. In our noisy-image classification task, we see that participants' performance improved when they were paired with any of three AI assistants with different levels of classification accuracy.

The experiment conducted in this study is an extension of our previous work where we combined human predictions and AI predictions using Bayesian model to determine the complementarity between the two individual agents [20]. However, since in most real-world cases there does not exist a third-party overseeing the performance of two separate agents on the same task and combining them together, we created a task in which the participants themselves are doing this integration. This is how it is done in settings such as clinical decision-making where doctors are provided with AI assistance. Doctors may integrate the AI's recommendation with their own decision [15]. For this experiment, we mimic this realistic paradigm, and therefore, our experiments had AI advice readily available.

2. Experiment

In this experiment, participants were tasked with classifying noisy images with the help of an AI assistant on some trials and without help on other trials. AI assistant performance was a between-subject manipulation whereas image noise was a within-subject manipulation. Participants were instructed to use the AI assistant to the best of their abilities.

2.1. Participants

A total of 132 participants were recruited using Amazon Mechanical Turk. Instructions were given to the participants before the start of the experiment to ensure that they understood the interface and what they would have to do. After reading all of the instructions,

the participants were then given a comprehension quiz to solidify their understanding of the task. In order to participate in the study, participants had to pass the comprehension quiz by correctly classifying four of five noisy images with AI help turned off. The participants were given two opportunities to pass the comprehension quiz. This allowed us to weed out participants that were just randomly clicking and ensured that all participants understood both the task and how to use the interface. Participants that successfully passed the comprehension quiz were allowed to proceed to the main experiment.

2.2. Images

The images used for this experiment all came from the ImageNet Large Scale Visual Recognition Challenge (ILSRVR) 2012 validation dataset [16]. We followed Geirhos et al., 2018 to reduce the number of classes in ImageNet from 1000 to 16 classes (airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven, and truck), termed ImageNet-16. After creating the ImageNet-16 dataset, we randomly selected a subset of 256 images to be used for all experiments, 16 images from each of the 16 classes. To increase the difficulty of a classic image classification task, we distorted the images by adding phase noise at each spatial frequency, where the phase noise was uniformly distributed in the interval $[-w, w]$ [7]. A total of eight different phase noise levels, $w = 0, 80, 95, 110, 125, 140, 155, 170$, were applied to the images, a different phase noise level for each unique image. This resulted in 2 unique images per category class having the same phase noise level. An example of phase noise manipulation can be seen in Figure 1.

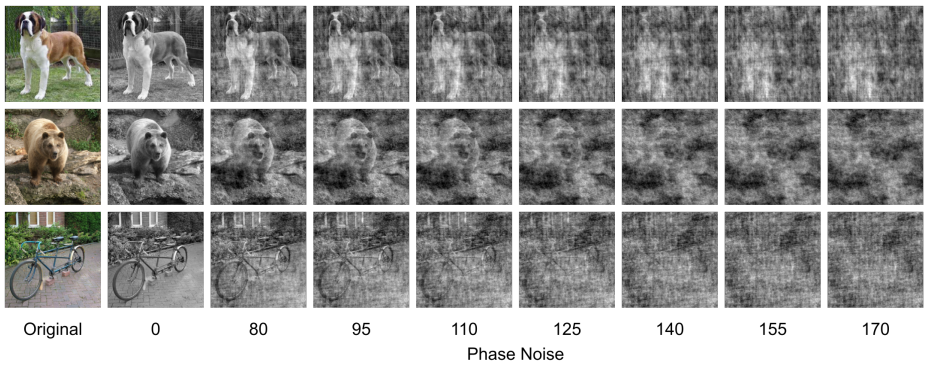


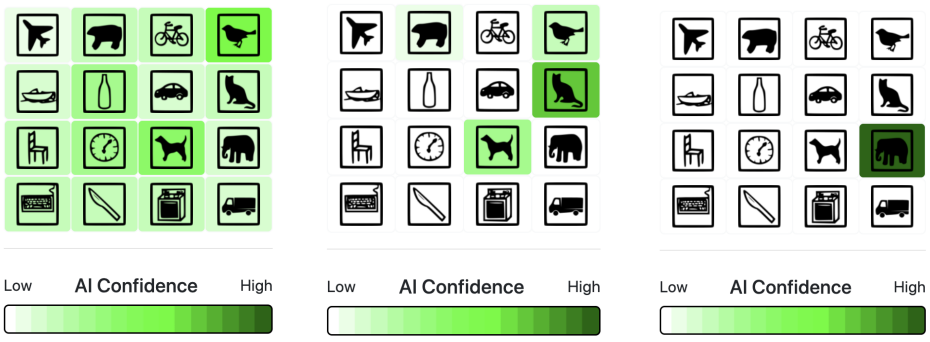
Figure 1. Illustration of three images under different levels of phase noise. Original images (left) were note used in experiments and are shown only for illustrative purposes.

2.3. AI Predictions

We used the VGG-19 architecture that was pretrained on the entire ImageNet dataset as the basis for our AI assistance [17]. To ensure performance on our particular task, we fine-tuned the VGG-19 architecture, creating three different models, on the ImageNet-16 dataset (created using the ILSRVR ImageNet training dataset). All three models were fine-tuned in the same manner, training on all different phase noise levels all at once. However, to generate these different levels of performance, the models were fine-tuned

for different periods of time. One level of performance, classifier A, was set to perform below the baseline level of human performance. The second level of performance, classifier B, was set to be at roughly the baseline level. Finally, the third level of performance, classifier C, was created to be above the baseline level.

Since models have a prediction value for each of the classes it is concerned with, we decided to display this information. To convey model confidence to participants during the experiment, we used a color gradient. The color gradient used ranged from white (hsl(120, 100%, 100%)) to dark green (hsl(120, 100%, 20%)). Each predicted class falls somewhere in the color gradient spectrum and is assigned a particular value based on the predicted probability of the model for that class. The darker the hue, the more confident the model was in its prediction for that particular class. Figure 2 displays how AI confidence can range from low confident (many classes) to high confident (single class).

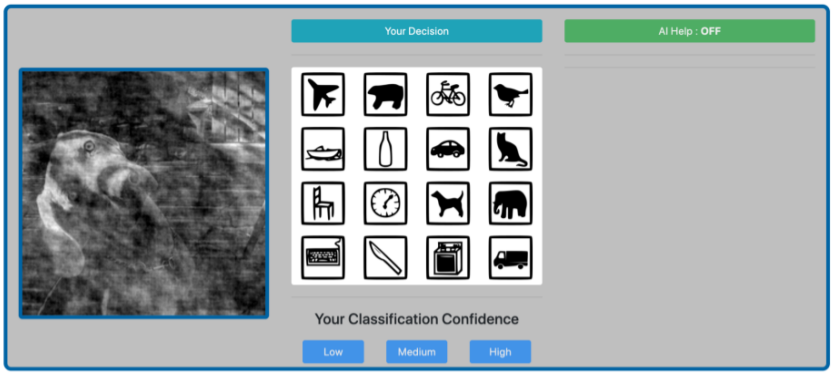


(a) AI Low Confidence (b) AI Medium Confidence (c) AI High Confidence

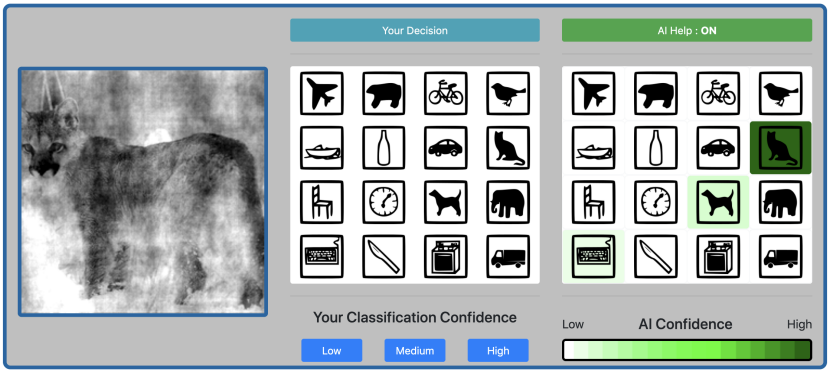
Figure 2. Illustration of the AI confidence.

2.4. Procedure

In the experiment, participants were tasked with classifying a total of 256 noisy images. The 256 trials were split in a block format in which AI assistance was turned on for 48 trials (AI ON condition), followed by 16 trials without AI assistance (AI OFF condition). This process repeated 4 times to give a total of 192 tasks with AI assistance turned on while the remaining 64 were with AI assistance turned off. Participants were instructed to classify all images as best they could and to leverage AI assistance (when provided) to optimize their performance. We use three classifiers of varying levels of accuracy to serve as assistants to the human participants. Each participant was assigned a single classifier level (A, B, or C) at the start of the experiment and would only be presented AI assistance from that particular classifier. Figure 3 displays the experimental interface in both AI assistance conditions. The experiment had a three-column layout in which the leftmost column presented the noisy image that was to be classified. The middle column presented a grid of 16 category buttons for the participant to make their classification as well as three different submission buttons each representing a confidence level (low, medium, and high). Finally, the rightmost column was used for AI assistance. When AI assistance was turned off, this column displayed nothing. However, when AI assistance



(a) AI OFF condition



(b) AI ON condition

Figure 3. Illustration of the behavioral experiment interface in both AI assistance conditions.

was turned on, there would be a grid of the 16 category options. Each of the 16 categories would be highlighted based on a gradient scale associated with the AI classifier prediction of that given category. The darker the hue of the highlighted category, the more confident the classifier was in that selection. For instances in which the classifier was extremely confident in a single category, there would only be one category highlighted with an extremely dark hue. Alternately, in instances where the classifier was not confident in a classification, there would be multiple categories highlighted with low hue levels. Participants were free to use the AI assistance to aid their final classification decision to the best of their abilities so as to optimize their own performance on the task. After every trial, feedback was provided so that participants could develop an understanding of their own abilities on the task as well as the abilities of the AI. Feedback was displayed on the center panel by highlighting the correct response in blue. When participants selected incorrectly, their incorrect response was highlighted in red. For trials where AI assistance was turned on, feedback was provided in the center panel while the AI predictions stayed on screen in the right most panel. One important thing to note is that the same 256 images were presented to all participants in a random order. Having all participants classify the same 256 images allowed us to compare and contrast how participants classified a partic-

ular image with and without AI assistance. A total of 132 participants, 44 per classifier level, were recruited using Amazon Mechanical Turk.

We use the classification probabilities of the predicted class (highest probability class) as confidence values for the AI assistant. To simplify our analysis, we discretize these confidence values to match the human labels of low, medium, and high confidence. We set the AI confidence cutoffs of low, medium, and high to the intervals of (0.00-0.33], (0.33, 0.66], and (0.66, 1.00) respectively.

3. Empirical Results

3.1. Is AI-assisted decision making more accurate than Human or the AI working independently?

Figure 4 shows the overall performance of the AI alone (classifier), humans alone (AI OFF), and the AI-assisted performance (AI ON) in the three conditions of the experiment. The second row of Figure 4 shows the change in accuracy in the AI ON condition versus the AI OFF condition. We see that humans are able to improve their performance across classifiers when aided by the AI. By looking at the second row of Figure 4, difference, we can more clearly see these improvements in accuracy across noise levels and how improvement grows with improving classifier accuracy. Figure 4 (A) is especially interesting because it indicates that humans are able to appropriately rely on the AI and improve their performance even when aided by an AI that has has worse accuracy than humans on average. This trend of improved accuracy of the human-AI assisted condition is consistent across the three classifiers. Participants in our experiment show appropriate reliance on the AI assistant and hence are able to improve their performance.

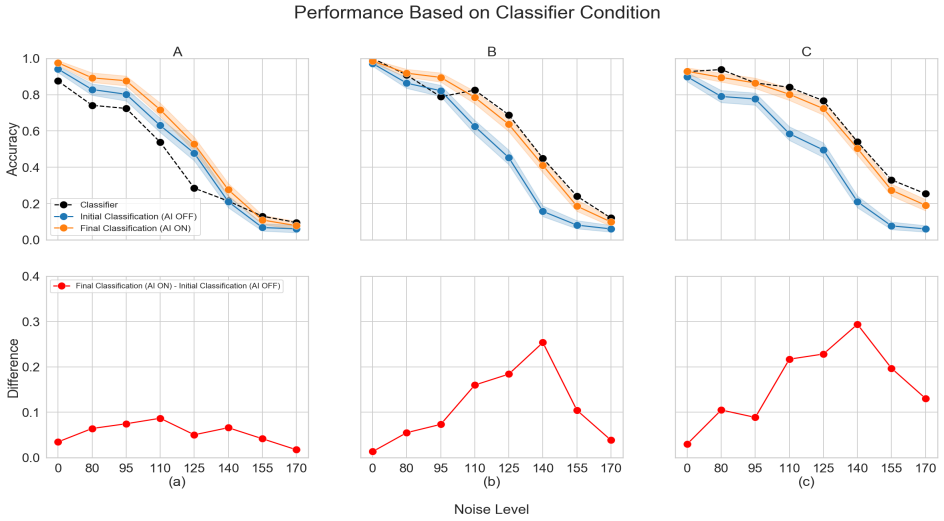


Figure 4. Participant performance across noise levels. Row one shows performance on the task. Row two shows the difference in performance between AI ON - AI OFF conditions. Columns represents classifier levels (A, B, and C).

3.2. How good are people's reliance strategies?

We compare the observed accuracy of participants in the AI ON condition to the expected accuracy when the AI and human's classifications are combined optimally post-hoc. This gives us a theoretical upper bound on performance if human and AI decisions were combined optimally. This is not a strategy that a human could have implemented but instead gives us the maximum performance achievable if the human had perfect information and knew when to rely on their own decision vs the AI's decision. We simulate this 'optimal' accuracy by marking the image as being classified correctly if either the human or the AI classify the image correctly.

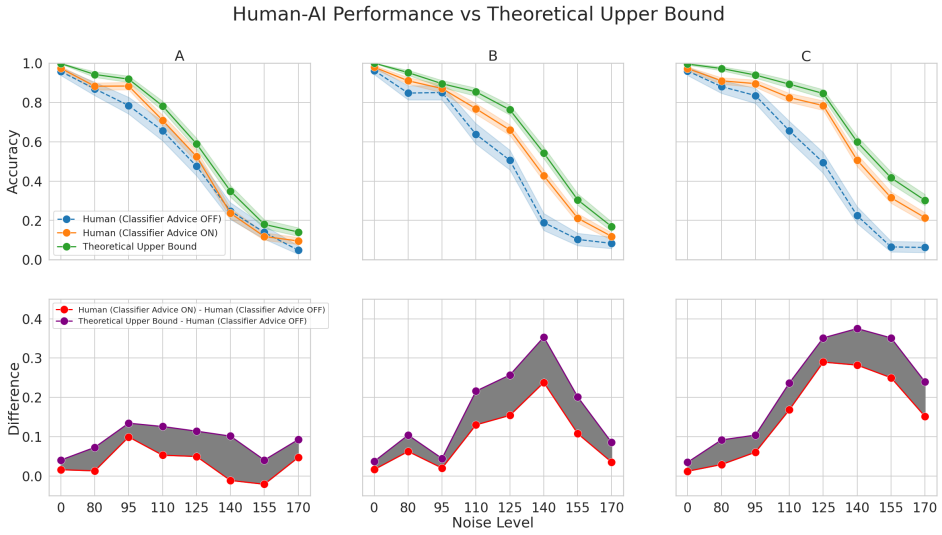


Figure 5. Participant performance and theoretical upper bound across all noise levels. Row one shows participant performance on the task as well as the theoretical upper bound calculated post-hoc. Row two shows the difference between the true AI integration strategy employed by participants minus the human only performance (red line) and the difference between the theoretical upper bound minus the human only performance (purple line). The area shaded between the two difference lines shows how far from optimal the true strategy participants employed was from optimal. Columns represent classifier levels (A, B, and C).

Figure 5 displays overall performance on the task for each classifier level (A, B, and C) and includes the theoretical upper bound (green line). Human only performance is represented by the dashed line in blue and is used as the basis in comparing the true strategy employed by participants in the task (AI ON condition, orange line) and the theoretical upper bound that was calculated post-hoc. The second row of Figure 5 displays the difference between the true strategy and the theoretical upper bound when compared to human only performance. The shaded region between both of the difference lines represents how far from optimal people were when integrating the AI advice into their own decision making. As we can see, in each of the three columns, people perform near the optimal theoretical upper bound that was computed post-hoc. This indicates, that the strategy participants are using to integrate AI assistance with their decision making is nearly optimal.

In Figure 6 we further break down the strategy employed by participants when integrating AI assistance by diving into the confidence levels of the participants' decisions. The top row of Figure 6 shows the potential accuracy when AI and human decisions are combined optimally, the middle row shows the observed accuracy in the AI ON condition, and the bottom row shows the difference in accuracy when AI and human decisions are combined optimally as compared to the observed accuracy in our empirical data in the AI ON condition. We see that on average, people's observed accuracy is close to the expected accuracy when AI and human decisions are combined optimally. This trend is consistent across different levels of human and AI confidence, and across the different AI assistants. This suggests that people were able to incorporate AI advice in a close to optimal manner when assisted by AI of varying accuracy.

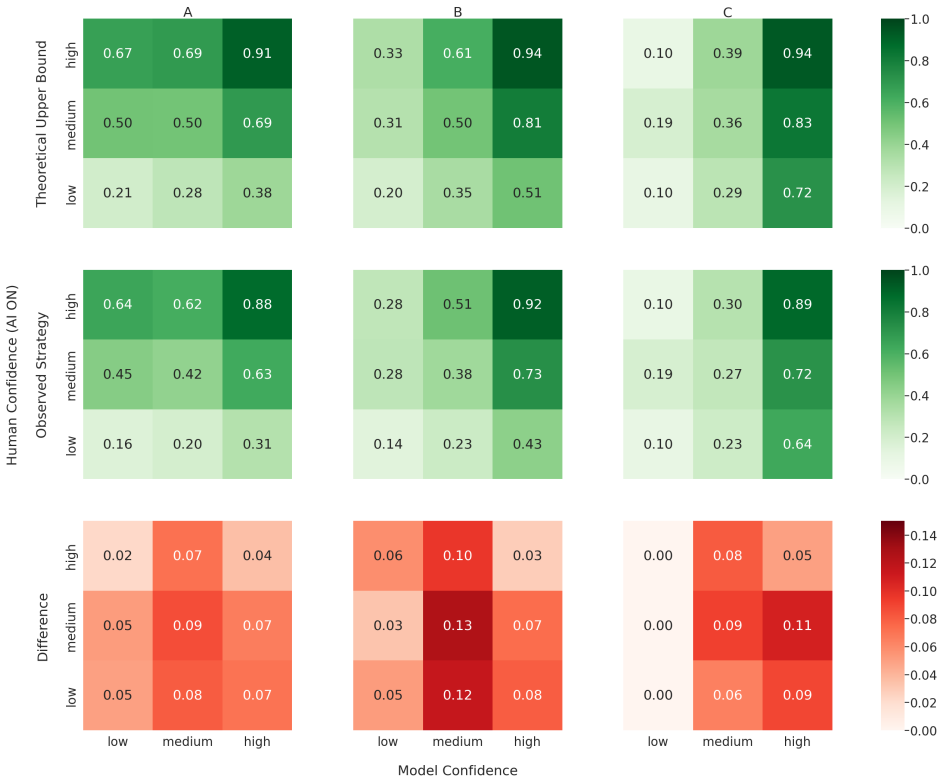


Figure 6. Accuracy when adopting the (top row) optimal policy, (middle row) the observed policy; and (bottom row) the difference in accuracy when adopting the optimal policy versus the observed policy (Optimal - Observed) . Columns correspond to the three different classifiers A, B and C. Within each grid, we have human confidence (low, medium, high) on the y-axis and AI confidence (low, medium, high) on the x-axis.

3.3. Is people's behavior consistent with automation bias?

Automation bias is described as a human's tendency to over-rely on machine recommendations [8]. It has been widely reported as a bias displayed by humans [8, 24, 12]. In our experiment, the path of least resistance for a participant would be to agree with the AI's

decision as it is always available in the AI ON condition trials. Hence, it is necessary to check for automation bias. We compare the AI ON condition's actual performance to a strategy that would describe automation bias in our system. Such a strategy would always select the AI's classification decision. Figure 7 we see the overall performance

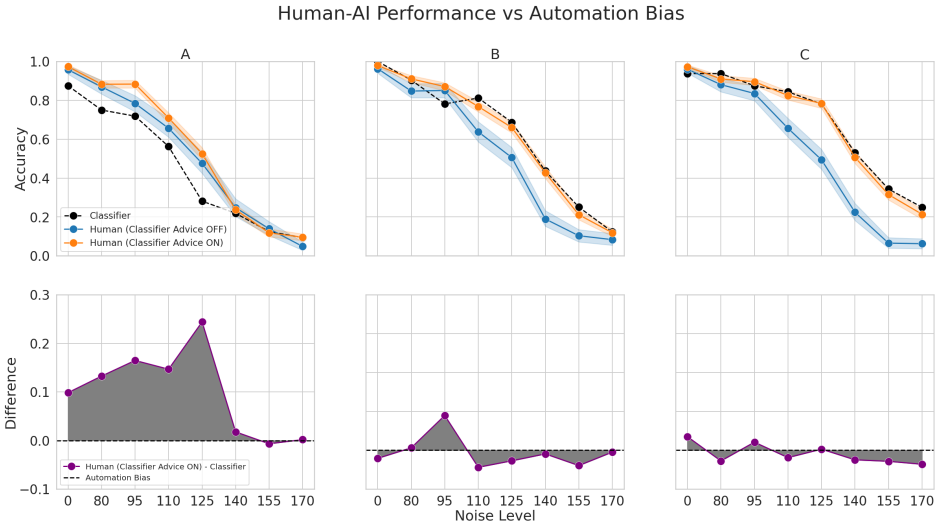


Figure 7. Participant performance and classifier performance across all noise levels. Row one shows participant performance on the task as well as the classifier performance (automation bias would lead to this performance). Row two shows the difference between the true AI integration strategy employed by participants minus the classifier performance (purple line) and the automation bias line (black dashed line). The area shaded in gray indicates that participant performance was wither above or below the automation bias line. Columns represent classifier levels (A, B, and C).

of individuals on the task and the classifier performance (which is how we describe automation bias). Given that we describe automation bias as always selecting the AI's classification condition, automation bias would then lead to the same performance as the AI. Automation bias is indicated in the figure by the black dashed line in row two. We can see that in column A, participants outperformed the classifier, indicating that there was no automation bias. In both columns B and C, we see that participant performance is closer to model performance, however, there are subtle differences indicating that participants are not just selecting whatever class the AI assistant is recommending.

In Figure 8 we further break down the strategy employed by participants and compare automation bias at different levels of confidence. Bottom row of Figure 8 shows the expected difference in accuracy if participant's were to adopt a policy consistent with automation bias as compared to the policy observed in the AI ON condition of our empirical data. We see that people's policy is easily distinguishable from automation-bias. In cases where the human's confidence is low, we see that the accuracy as observed in the AI ON condition far exceeds the expected accuracy of the automation bias policy. This indicates that the images where the human had low confidence were also regions of low accuracy for the classifier. Humans still employed a strategy better than over-relying on the AI. Alternately, in instances where the human's confidence is high, we see that the accuracy as observed in the AI ON condition is comparable to the expected accuracy of the automation bias policy.

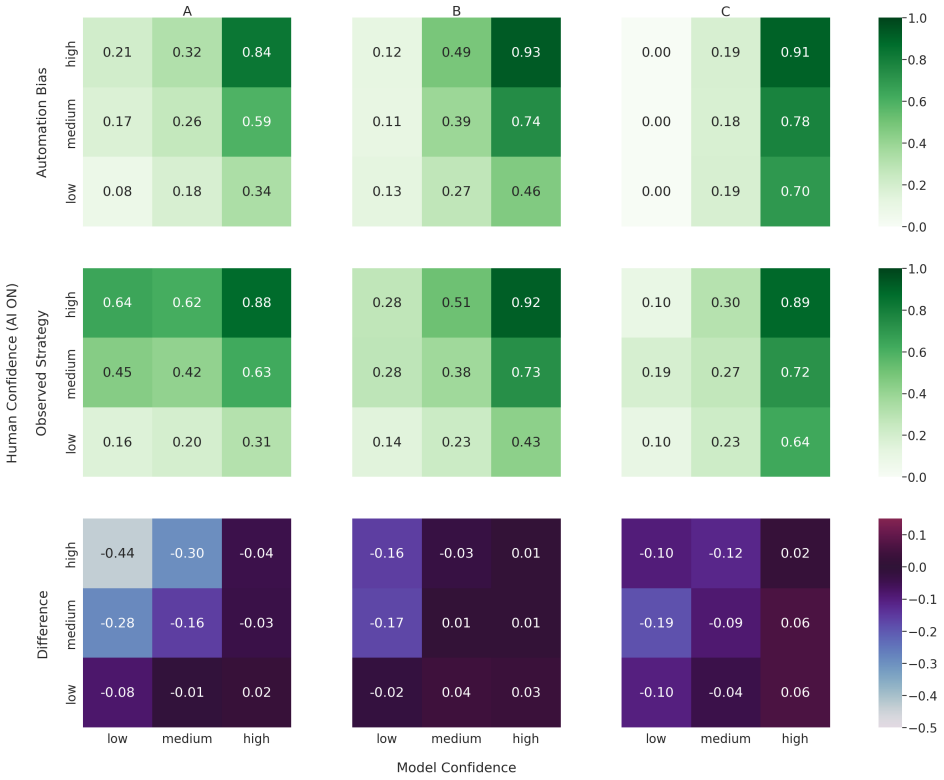


Figure 8. Accuracy when adopting (top row) a policy consistent with automation bias, (middle row) the observed policy (AI ON); and (bottom row) the difference in accuracy when adopting the automation bias policy versus the observed policy in the AI ON condition (Automation Bias - Observed). Columns correspond to the three different classifiers A, B and C. Within each grid, we have human confidence (low, medium, high) on the y-axis and AI confidence (low, medium, high) on the x-axis.

4. Discussion

This paper adds to a growing body of literature that investigates AI-assisted decision-making. Our empirical results reveal that in this image classification task, where people have a good understanding of their own ability and confidence on each trial, they are close to optimal in their adoption of the AI's advice. We show that performance in the AI ON condition does not deteriorate even in cases where the AI assistant has worse accuracy on the task than the human. We also find that people's reliance strategy on the AI in this task is not consistent with behavior that is associated with automation bias.

Most recent investigations of human decision-making with AI-assistance follow a judge-advisor system [19, 18] where humans are required to independently solve the task at hand before they are shown an AI assistant's recommendation [4, 10]. Once shown the AI's advice, humans may update their final decision. Such experimental paradigms provide direct insights about a human's reliance behavior and make it easier for experimenters to disentangle the influence of the AI's advice on the final decision of the human. Also, deliberation and independent assessment of a problem have shown to help decrease over reliance on the AI [3]. However, this setup is somewhat artificial and in-

compatible with how AI assistants work in the real world. The ultimate aim of providing AI-assistance is to reduce the workload of humans and improve the accuracy of the joint decisions. We argue that our setup is a more natural way of incorporating AI assistance into everyday workflows. In our experiment, in the AI ON condition, the AI assistant's advice and its confidence in the advice is available to the human as soon as the task is presented. While this format of providing assistance may raise concerns about automation bias, we show that people adopted close to optimal strategies in our task. The AI OFF condition gives us information about the humans' independent classification judgement and confidence rating which we use to indirectly assess the influence of advice.

We believe that participants in our experiment were able to build close to optimal reliance strategies because of the following reasons. First, this is a simple task and most people are experts at identifying everyday objects. This enables people to have a good understanding of their own expertise and confidence on any presented image. Second, indicating confidence in a decision requires humans to employ a second-order metacognitive computation to evaluate their decision [6]. While this may not be an obvious 'cognitive forcing function' [3], prior work indicates that the mechanism humans employ to generate confidence ratings requires metacognitive deliberation about the strength of evidence available to make the decision [23]. Finally, in our experiment, people received feedback after each trial, which gave them the opportunity to learn about the AI assistant's accuracy and confidence calibration. We show that people were able to use feedback and build reasonable mental models of the AI assistant when paired with any of the three classifiers of varying levels of accuracy. We believe that participants track more than just the accuracy of the AI as they are able to improve their performance even when the AI is less accurate than them on average. However, immediate feedback is not always possible in real-world scenarios. The impact of delayed feedback on the reliance behavior must be investigated in isolation.

In future work, we will present a cognitive model that can predict a human's classification decision and confidence rating on each trial based on information gathered about the human during the AI OFF condition. Such a model would allow us to infer the human's 'latent' decision to switch to the AI's recommendation without explicitly asking the human to provide an independent judgement. We believe that this work has important implications in designing better AI assistants and explanations. A key limitation of our work is that we used a low-stakes image classification task. Further work is needed to understand how performance varies with varying confidence of the agents and how this generalizes to other tasks.

References

[1] Suresh Kumar Annappindi. *System and method for predicting consumer credit risk using income risk based credit score*. US Patent 8,799,150. Aug. 2014.

[2] Gagan Bansal et al. "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*. 2021, pp. 1–16.

[3] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. "To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–21.

[4] Leah Chong et al. “Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice”. In: *Computers in Human Behavior* 127 (2022), p. 107018. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2021.107018. URL: https://www.sciencedirect.com/science/article/pii/S0747563221003411.

[5] Steven E Dilsizian and Eliot L Siegel. “Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment”. In: *Current cardiology reports* 16.1 (2014), pp. 1–8.

[6] Stephen M Fleming and Nathaniel D Daw. “Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation.” In: *Psychological review* 124.1 (2017), p. 91.

[7] R Geirhos et al. “Generalisation in humans and deep neural networks”. In: *Thirty-second Annual Conference on Neural Information Processing Systems (NeurIPS 2018)*. Curran. 2019, pp. 7549–7561.

[8] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. “Automation bias: a systematic review of frequency, effect mediators, and mitigators”. In: *Journal of the American Medical Informatics Association* 19.1 (2012), pp. 121–127.

[9] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. “Human decision making with machine assistance: An experiment on bailing and jailing”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–25.

[10] Aakriti Kumar et al. “Explaining Algorithm Aversion with Metacognitive Bandits”. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 43. 43. 2021.

[11] Vivian Lai et al. “Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies”. In: *arXiv preprint arXiv:2112.11471* (2021).

[12] Raja Parasuraman, Robert Molloy, and Indramani L Singh. “Performance consequences of automation-induced’complacency’”. In: *The International Journal of Aviation Psychology* 3.1 (1993), pp. 1–23.

[13] Bhavik N Patel et al. “Human–machine partnership with artificial intelligence for chest radiograph diagnosis”. In: *NPJ Digital Medicine* 2.1 (2019), pp. 1–10.

[14] P Jonathon Phillips et al. “Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms”. In: *Proceedings of the National Academy of Sciences* 115.24 (2018), pp. 6171–6176.

[15] Pranav Rajpurkar et al. “CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV”. In: *npj Digital Medicine* 3.1 (Sept. 2020). ISSN: 2398-6352. DOI: 10.1038/s41746-020-00322-2. URL: https://doi.org/10.1038/s41746-020-00322-2.

[16] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.

[17] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[18] JA Sniezek and Timothy Buckley. “Social influence in the advisor-judge relationship”. In: *Annual meeting of the Judgment and Decision Making Society, Atlanta, Georgia*. 1989.

[19] Janet A Sniezek, Gunnar E Schrah, and Reeshad S Dalal. "Improving judgement with prepaid expert advice". In: *Journal of Behavioral Decision Making* 17.3 (2004), pp. 173–190.

[20] Mark Steyvers et al. "Bayesian Modeling of Human-AI Complementarity in Image Classification". In: *Proceedings of the National Academy of Sciences* (2022).

[21] Sarah Tan et al. "Investigating human+ machine complementarity for recidivism predictions". In: *arXiv preprint arXiv:1808.09123* (2018).

[22] Darryl E Wright et al. "A transient search using combined human and machine classifications". In: *Monthly Notices of the Royal Astronomical Society* 472.2 (2017), pp. 1315–1323.

[23] Nick Yeung and Christopher Summerfield. "Metacognition in human decision-making: confidence and error monitoring". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594 (2012), pp. 1310–1321.

[24] Guanglu Zhang et al. "A cautionary tale about the impact of AI on human design teams". In: *Design Studies* 72 (2021), p. 100990.

[25] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*. 2020, pp. 295–305.