Computational Brain & Behavior (2022) 5:491–508
https://doi.org/10.1007/s42113-022-00157-y

ORIGINAL PAPER

AI-Assisted Decision-making: a Cognitive Modeling Approach to Infer Latent Reliance Strategies

Heliodoro Tejeda[1] · Aakriti Kumar[1] · Padhraic Smyth[2] · Mark Steyvers[1]

Accepted: 22 September 2022 / Published online: 19 October 2022

Abstract

AI assistance is readily available to humans in a variety of decision-making applications. In order to fully understand the efficacy of such joint decision-making, it is important to first understand the human's reliance on AI. However, there is a disconnect between how joint decision-making is studied and how it is practiced in the real world. More often than not, researchers ask humans to provide independent decisions before they are shown AI assistance. This is done to make explicit the influence of AI assistance on the human's decision. We develop a cognitive model that allows us to infer the *latent* reliance strategy of humans on AI assistance without asking the human to make an independent decision. We validate the model's predictions through two behavioral experiments. The first experiment follows a *concurrent* paradigm where humans are shown AI assistance alongside the decision problem. The second experiment follows a *sequential* paradigm where humans provide an independent judgment on a decision problem before AI assistance is made available. The model's predicted reliance strategies closely track the strategies employed by humans in the two experimental paradigms. Our model provides a principled way to infer reliance on AI-assistance and may be used to expand the scope of investigation on human-AI collaboration.

Keywords AI-assisted decision making · Cognitive modeling · Reliance · Trust · Confidence

Introduction

Over the past decade, there has been an increase in domains where AI is used to assist humans by providing recommendations in the context of a prediction problem. Examples of these AI recommendation systems include making bail decisions in a legal context (Kleinberg et al., 2018), detecting deception in consumer reviews (Ott et al., 2011), making medical decisions in diagnostic imaging (Esteva et al., 2017; Patel et al., 2019; Rajpurkar et al., 2020), recognizing faces in forensic analysis (Phillips et al., 2018), and classifying astronomical images (Wright et al., 2017). Such widespread adoption of AI decision aids has been accompanied by burgeoning interest in investigating the efficacy of AI assistance in collaborative decision-making settings (Yin et al., 2019; Park et al., 2019; Zhang et al., 2021; Poursabzi-Sangdeh et al., 2021; Buçinca et al., 2021; Kumar et al., 2021; Chong et al., 2022; Becker et al., 2022).

To investigate such AI-assisted decision-making, researchers have designed a variety of workflows. Some workflows require the human to provide an independent decision first, then display the AI's advice which the human can then use to update their final decision (Yin et al., 2019; Poursabzi-Sangdeh et al., 2021; Chong et al., 2022). Other workflows present AI advice alongside the prediction problem and the human can decide to follow the advice or ignore it (Rajpurkar et al., 2020; Sayres et al., 2019). Finally, a few studies force individuals to spend time thinking about the decision problem by artificially delaying the presentation of AI advice (Buçinca et al., 2021; Park et al., 2019) or making AI advice available only when it is requested (Kumar et al., 2021; Liang et al., 2022). In this work, we focus on two of the aforementioned workflows of AI-assisted decision-making and refer to them as paradigms; a detailed illustration can be found in Fig. 1. We term the first as a *sequential* paradigm, where AI advice is displayed only after the human provides an independent judgment and the human can choose to revise their initial judgment. We term

✉ Heliodoro Tejeda
htejeda@uci.edu

[1] Department of Cognitive Sciences, University of California, Irvine, USA

[2] Department of Computer Science, University of California, Irvine, USA

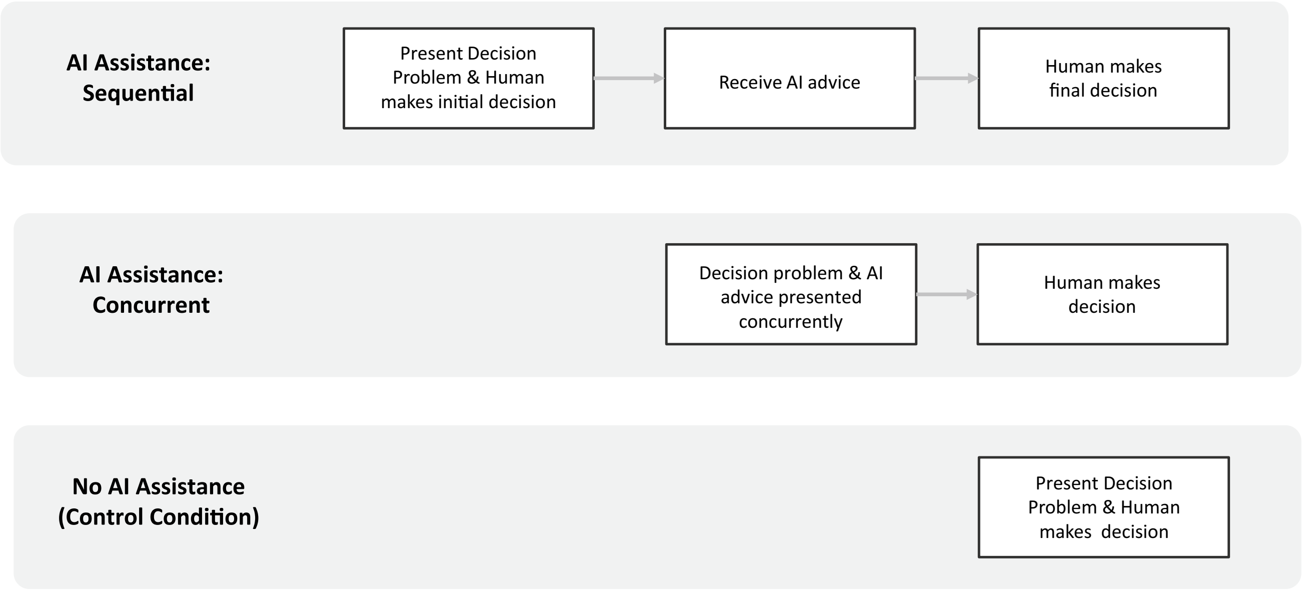

AI Assistance: Sequential
Present Decision Problem & Human makes initial decision
Receive AI advice
Human makes final decision
AI Assistance: Concurrent
Decision problem & AI advice presented concurrently
Human makes decision
No AI Assistance (Control Condition)
Present Decision Problem & Human makes decision

Fig. 1 Illustration of the sequential and concurrent paradigms for AI-assisted decision-making (top two rows). The no-AI assistance paradigm (bottom row) is used as a control condition for the concurrent paradigm

the second as a *concurrent* paradigm where AI advice is displayed concurrently with the prediction problem.

The sequential paradigm provides direct insights about the human's reliance on the AI based on two human judgments: the initial independent judgment and a final judgment after receiving the AI advice. This paradigm makes it easier for experimenters to disentangle the influence of AI advice on the human's decision. However, in many real-world applications, the human user does not independently make a decision before AI assistance is provided since providing the AI's recommendation immediately simplifies the workflow and can save time. The concurrent paradigm offers an alternative setting to study AI-assisted decision-making. One drawback of the concurrent paradigm is the fundamental ambiguity in data interpretation — it is unclear as to how one can assess the usefulness of the AI decision aid to the human user. Since there is no initial human judgment available before AI advice is offered, there is no direct empirical observation about any changes the human is making in their decision-making. Any observed agreement between the human and the AI, in the concurrent paradigm, could arise because the human changed their judgment and took the AI's advice or the human already arrived at the same judgment independent of the AI. How, then, do we assess the impact of AI assistance on the human's decision?

Our research has three main goals. First, we develop a computational cognitive model for AI-assisted decision-making in the concurrent paradigm. The cognitive model provides a principled way to infer the latent reliance of a human on the AI assistant in spite of the fact that there are no direct observations of switching behaviors when a person is presented with the AI advice. We empirically validate the computational model by collecting empirical data from a behavioral study using both the sequential and concurrent paradigms. The data from the sequential paradigm offers a comparison to the concurrent paradigm and provides a test to assess the merit of the computational framework. We demonstrate that the model's predictions of reliance behavior in the concurrent paradigm are qualitatively similar to the reliance behavior observed in the sequential paradigm. In addition, we demonstrate that the model can generalize to held out trials in the concurrent paradigm.

In our second goal, we use the cognitive modeling approach to understand how a human's reliance policy depends on a number of factors related to the human and the AI. Previous research has shown that a human's confidence in their own decision influences the tendency to rely on AI assistance (Lu and Yin, 2021; Pescetelli et al., 2021; Wang et al., 2022). In addition, reliance on the AI is also affected by the AI's confidence in its decision (Zhang et al., 2020). Another contributing factor is the overall accuracy of the AI. In some previous research, only a single AI model with a fixed degree of accuracy was used; for example, an AI model with an accuracy comparable to human performance (Zhang et al., 2020) or above human performance (Lai and Tan, 2019; Pescetelli et al., 2021). A few studies have investigated the effect of varying AI accuracy on reliance strategy (Yin et al., 2019). In our empirical paradigm, we investigate how human reliance varies across multiple levels of AI accuracy. This allows for a more nuanced understanding of the impact of the AI aid's accuracy on the human's reliance behavior. In addition, we investigate how participant confidence and

AI confidence scores affect the trial-by-trial reliance strategy used by participants.

In our third goal, we use the computational model to quantify the effectiveness of the reliance strategies employed by the human. In some instances, people adopt sub-optimal reliance policies when working with an AI. For example, it has been found that people will prefer to use their own (less accurate) forecasts instead of an algorithm if they have seen the algorithm make mistakes (Dietvorst et al., 2015). In another study, people placed too much trust in an automated system (Cummings, 2017). Over- and under-reliance on AI advice may depend on particular task domains and methods of interaction (Promberger and Baron, 2006; Castelo et al., 2019; Logg, 2017). Whereas in these previous studies, the reliance was assessed at the aggregate level, our cognitive modeling approach enables us to estimate the trial-by-trial variations in reliance depending on factors such as the confidence state of the participant and the level of confidence of the AI for particular problem instances. For particular combinations of self- and AI confidence (e.g., low self-confidence and high AI confidence) and particular combinations of human and AI overall accuracy, we can expect joint decision-making accuracy to be better than the human or AI alone (Steyvers et al., 2022). An empirical question is whether participants are able to adopt such a policy. We compare the reliance policies adopted by participants to optimal policies and show that in our experiment, people were quite effective in their adoption of AI advice.

Cognitive Model

Before describing the computational model, we note some key aspects of the concurrent advice-taking paradigm in particular that motivate the design of the model. In the experiment, participants have to predict the classification label of a set of images and a confidence level associated with their decision. Each participant alternates between two experimental conditions. In the control (no assistance) condition, participants indicate their predictions without help from the AI. In the AI assistance condition, we follow the concurrent approach; the AI provides a recommended set of predictions by highlighting the class labels according to the AI's confidence scores. The participant can use these recommendations in any way they want to order to maximize their own accuracy (see Fig. 2 for an illustration of the user interface in the experiment). An important aspect of this condition is that the participant's prediction reflects a combination of their own independent decision-making (which is not observable in this paradigm) and the AI prediction. In other words, the policy used by the participant to rely on and integrate AI predictions with their own predictions is not directly observable from their behavior.

The main goal of the computational model is to draw inferences about the latent advice-taking policies. The policy can be determined by a number of factors, such as the confidence state of the participant and the confidence scores of the AI as well as the overall accuracy of the AI. We develop a hierarchical Bayesian model to draw inferences about the policies not only at the population level but also at the level of individual participants. In the first part of the model, a Bayesian Item-Response model (Fox, 2010) is applied to the no-assistance condition to infer individual differences in ability as well as differences in difficulty across items (i.e., prediction problems). In the AI-assistance part of the model, these latent person and item parameters are used to explain the observed prediction from a participant which depends on their (unobservable) unaided prediction and the advice-taking policy that determines the likelihood that a participant switches to the AI prediction or stays with their own prediction. Figure 3 visualizes the graphical model of the computational model that explains the human predictions with and without AI assistance.

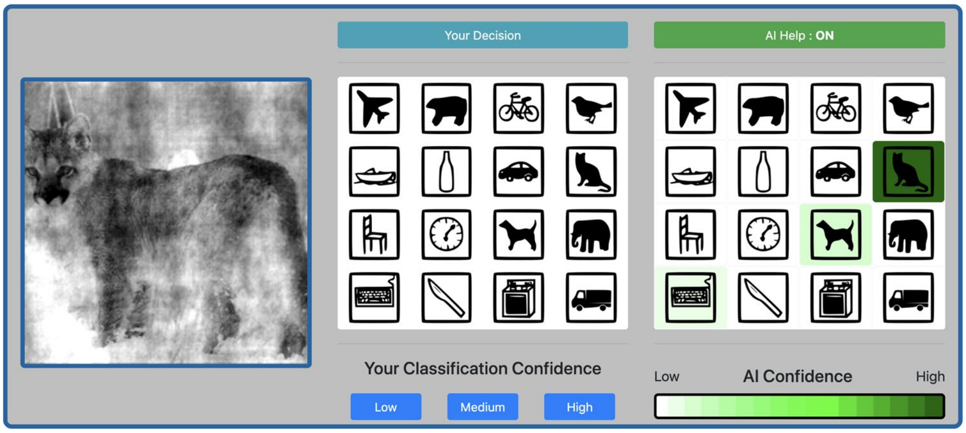


Fig. 2 Illustration of the behavioral experiment interface in the AI assistance condition

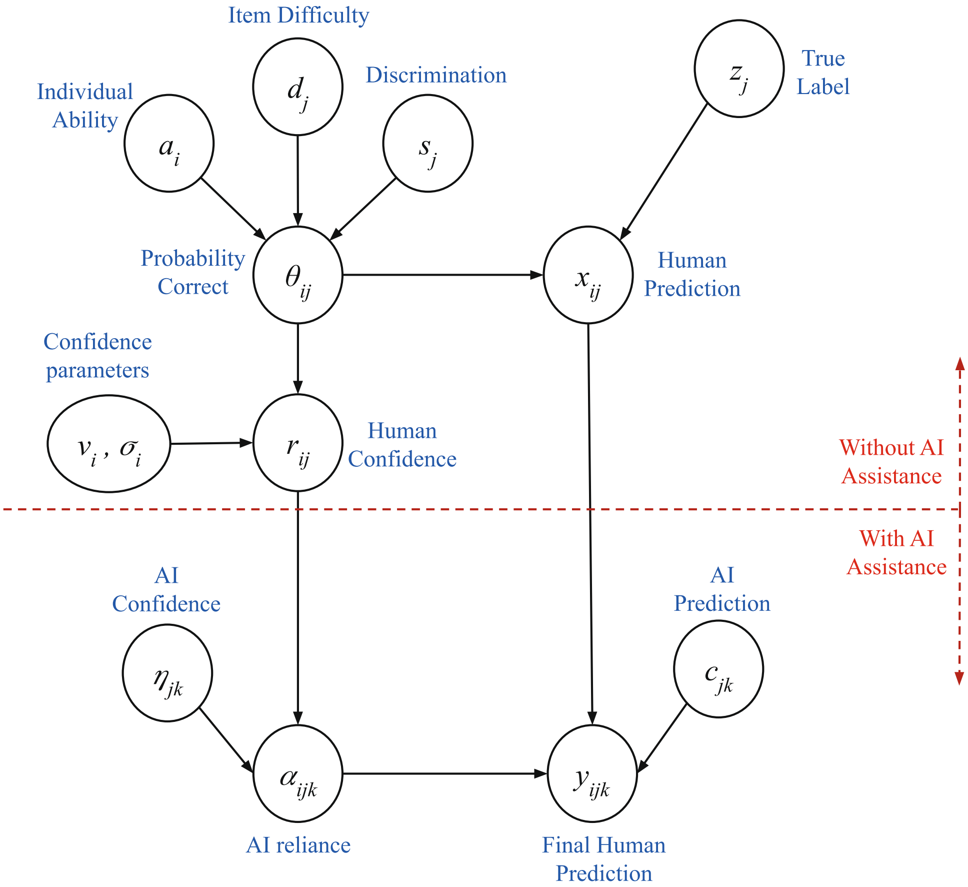


Fig. 3 Graphical model for the AI-assisted decision-making model. In the condition without assistance, r_{ij} and x_{ij} and z_j are observed. In the condition where AI assistance is provided, r_{ij} and x_{ij} are latent and y_{ijk}, z_j, c_{jk}, and η_{jk} are observed. For visual clarity, plate notation is omitted

Modeling Human Decisions Before Assistance

The computational model for human predictions without AI assistance is based on a Bayesian Item-Response model (Fox, 2010). The Item-Response model makes it convenient to model individual differences in accuracy as well as differences in item difficulty (where items refer to the individual images participants have to classify). To model the human predictions, we use a three-parameter IRT model to capture the probability $\theta_{i,j}$ that a correct response is made by person i on item j:

$$\log\left(\frac{\theta_{i,j}}{1-\theta_{i,j}}\right) = s_j a_i - d_j \tag{1}$$

The person parameter a_i is an ability parameter that determines the overall performance of the person across items. The item parameter d_j captures differences in the item difficulty while the item parameter s_j captures discrimination: the tendency of an item to discriminate between high and low ability individuals.

In a typical IRT model, the probability of making a correct response, θ, is used to sample the correctness of an answer. However, for our model, we code the responses from individuals in terms of the predicted label. Let $x_{i,j}$ represent the prediction by person i for item j in the absence of AI assistance. Each prediction involves a choice from a set of L labels, i.e., $x \in \{1, \dots, L\}$. Let z_j represent the true label for item j. We assume that person i produces the correct label z_j on item j with probability $\theta_{i,j}$ and otherwise chooses uniformly from all other labels, as follows:

$$p(x_{i,j} = m) = \begin{cases} \theta_{i,j} & \text{if } z_j = m \\ (1-\theta_{i,j})/(L-1) & \text{if } z_j \neq m \end{cases} \tag{2}$$

Various model extensions could be considered that allow for response biases such that some labels are preferred a priori over other.

Participants not only make a prediction but also express a confidence level, $r_{i,j}$, associated with their prediction. In the experimental paradigm, confidence levels are chosen from a small set of labels, $r_{i,j} \in \{\text{low}, \text{medium}, \text{high}\}$. In the model, we assume that predictions associated with higher accuracy on average lead to higher confidence levels, but that at the item level, the mapping from accuracy to confidence is noisy. To capture the noisy relationship between accuracy and confidence, we use a simple generative model based on an ordered probit model:

$$r_{i,j} \sim \text{OrderedProbit}(\theta_{i,j}, v_i, \sigma_i) \tag{3}$$

In this generative model, normally distributed noise with standard deviation σ_i is added to the probability of being correct $\theta_{i,j}$. The resulting value is then compared against a set of intervals defined by parameters v_i, and the interval which contains the value determines the resulting confidence level. Changes in v_i can lead the participant to different uses of the response scale (i.e., using one particular confidence level relatively often) while σ_i determines (inversely) the degree to which accuracy and confidence are related. Note that the parameters σ and v are person-specific to allow for individual differences in the confidence generating process. Appendix 1 provides more detail on the ordered probit model.

Modeling Human Decisions After Advice

In the model for human decisions in the presence of advice, let $y_{i,j,k}$ represent the observed prediction made by person i on item j after AI advice is considered from AI algorithm k. We include a dependence on the type of algorithm as our empirical paradigm will present AI advice from different algorithms. In the advice-taking model, we assume that the participant initially makes their own prediction $x_{i,j}$ independent of the AI advice but that their final decision $y_{i,j,k}$ can be influenced by the AI advice. Note that in the no-assistance condition, the independent predictions $x_{i,j}$ and associated confidence levels $r_{i,j}$ are directly observable, but they are latent in the AI assistance condition. However, we can use the IRT model in the previous section to simulate the counterfactual situation about the prediction and confidence level that a person would have made if AI advice was not provided. Specifically, we can use the generative model in Eqs. 1–3 to generate predictions for $x_{i,j}$ and $r_{i,j}$ on the basis of information about the participant's overall skill (a) as well as information about the difficulty of the particular item (d_j)[1].

In the advice-taking model, we assume that the participant will stay with their original decision $x_{i,j}$ if it agrees with the AI's recommendation, denoted by $c_{j,k}$. However, when the original decision is not the same as the AI's recommendation, we assume the participant switches to the AI's recommendation with probability $\alpha_{i,j,k}$. Therefore, we can model the probability that the participant chooses label m for their final prediction as follows:

$$p(y_{i,j,k}=m)=\begin{cases}\alpha_{i,j,k} & \text{if } x_{i,j}\neq m \wedge c_{j,k}=m\\ 1 & \text{if } x_{i,j}=m \wedge c_{j,k}=m\\ 0 & \text{if } x_{i,j}\neq m \wedge c_{j,k}\neq m\end{cases} \tag{4}$$

The variable $\alpha_{i,j,k}$ determines the tendency of participant i to trust the AI advice from algorithm k related to item j. In the next section, we describe how this latent variable can depend on factors such as the confidence state of the participant as well as the confidence score of the AI.

Note that in this model, when the participant is provided with AI assistance, the independent prediction $x_{i,j}$ is latent in our experimental paradigm. Instead of explicitly simulating the process of first sampling an independent prediction $x_{i,j}$ and then a final prediction $y_{i,j,k}$, we can simplify the generative process by marginalizing out $x_{i,j}$:

$$p(y_{i,j,k}=m)=\begin{cases}\theta_{i,j}+(1-\theta_{i,j})\alpha_{i,j,k} & \text{if } z_j=m \wedge c_{j,k}=m\\ \frac{1-\theta_{i,j}}{L-1}+\left(1-\frac{1-\theta_{i,j}}{L-1}\right)\alpha_{i,j,k} & \text{if } z_j\neq m \wedge c_{j,k}=m\\ \frac{1-\theta_{i,j}}{L-1}(1-\alpha_{i,j,k}) & \text{if } z_j\neq m \wedge c_{j,k}\neq m\end{cases} \tag{5}$$

In this equation, the probability that the participant selects label m is split into three different cases. The first case reflects the probability that the participant makes the correct decision independently (which happened to agree with the AI recommendation) or makes an incorrect decision initially but then adopts the correct AI advice. The second case reflects the probability that the participant initially selects an incorrect decision (which happened to agree with the AI recommendation) or makes another decision different from the AI but then adopts the incorrect AI advice. The third case reflects the probability that the participant makes an incorrect independent decision and decides not to switch to the AI's recommendation.

Modeling Individual Differences in Advice-Taking

The key latent variable of interest in the model is $\alpha_{i,j,k}$, which determines the willingness of the participant per item to switch to the AI's recommended prediction if it differs from their own prediction. Generally, $\alpha_{i,j,k}$ can depend on many characteristics related to the person, item, and classifier. Here, we will consider functions where α depends on the confidence state of the participant for item j ($r_{i,j}$), the AI confidence score associated with item j ($\eta_{j,k}$), and the type of classifier k:

$$\alpha_{i,j,k}=f(r_{i,j},\eta_{j,k},k) \tag{6}$$

One way to specify function f is based on a linear model that captures main effects as well as interaction between the two putative factors. However, to avoid specifying the exact functional form of f, we will instead simplify the model and treat function f as a lookup table that specifies the α values based on a small number of combinations of participant confidence, AI confidence, and classifier type. Specifically, we create $3 \times 4 \times 3$ lookup table that specifies the α value based on 3 levels of participant confidence ("low," "medium,"

[1] Note that in empirical paradigm, each image is presented in both the control condition as well as the AI assistance condition to allow for the estimation of item difficulty parameters for each image.

"high"), 4 levels of AI confidence, and 3 types of classifiers (k). We use a hierarchical Bayesian modeling approach to estimate individual differences in the policy α (see Appendix 2 for details).

Experiments

To validate our cognitive model, we investigated human performance with and without AI assistance in two paradigms: the concurrent and sequential paradigm. We will apply the cognitive model to the concurrent paradigm to infer the AI reliance strategies by individual participants. The results from the sequential paradigm serve as a means to validate our cognitive model, as the sequential paradigm allow us to empirically analyze participant strategies when integrating AI assistance.

In both paradigms, participants have to classify noisy images into 16 different categories (see Fig. 2 for an example of the user interface). There were two experimental manipulations. First, the image noise was varied to produce substantial difference in classification difficulty (Fig. 4). Second, we varied the overall accuracy of the AI predictions across three conditions: classifier A, classifier B, and classifier C. Classifier A was designed to produce predictions that are, on average, less accurate than human performance. Classifiers B and C were designed to produce predictions that are, on average, as accurate and more accurate than human performance. Each participant was paired with one type of classifier.

The main difference between the two paradigms is that in the concurrent paradigm, participants alternated between blocks of trials where AI assistance was or was not provided. In the sequential paradigm, there were no alternating blocks. On each trial, the participant first made an independent prediction for a image classification problem and was then given an opportunity to revise their prediction after AI assistance was provided.

Methods

Participants

A total of 60 and 75 participants were recruited using Amazon Mechanical Turk for the concurrent and sequential experiments respectively. To ensure that participants understood the task, they were given a set of instructions describing the experiment and what they would have to do. Upon reading all of the instructions, participants were then tasked with a comprehension quiz to ensure they fully understood the task. The quiz consisted of having participants classify five different noisy images with AI help turned off. In order to participate in the study, participants had to correctly classify four of the five images in the quiz. Participants were given two opportunities to pass the quiz. Successful participants were then allowed to proceed with the rest of the experiment.

Images

All images used for this experiment come from the ImageNet Large Scale Visual Recognition Challenge (ILSRVR) 2012 validation dataset (Russakovsky et al., 2015). Following (Geirhos et al., 2019), a subset of 256 images was selected divided equally among 16 classes (chair, oven, knife, bottle, keyboard, clock, boat, bicycle, airplane, truck, car, elephant, bear, dog, cat, and bird). To manipulate the classification difficulty, images were distorted by phase noise at each spatial frequency, where the phase noise is uniformly distributed in the interval $[-\omega, \omega]$ (Geirhos et al., 2019). Eight levels of phase noise, $\omega = \{0, 80, 95, 110, 125, 140, 155, 170\}$, were applied to the images, a different noise level for each unique image, resulting in 2 unique images per category per noise level (see Fig. 4 for examples of the phase noise manipulation).

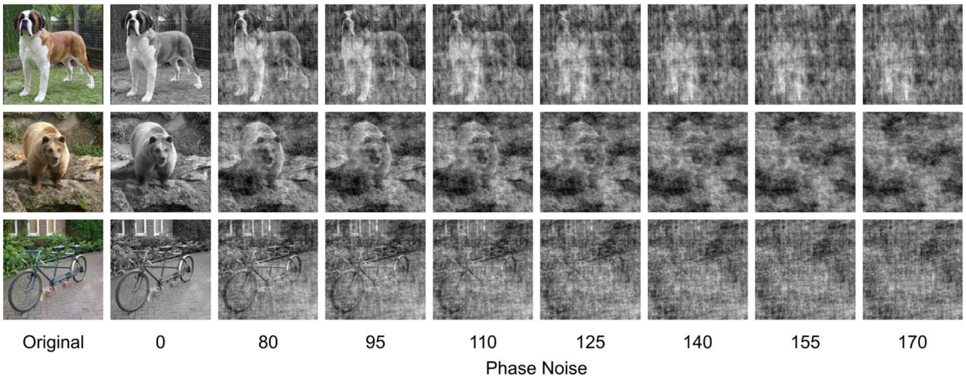


Fig. 4 Illustration of three images under different levels of phase noise. Original images (left) were not used in experiments and are shown only for illustrative purposes

AI Predictions

We used a convolutional neural network (CNN), based on the VGG-19 architecture (Simonyan and Zisserman, 2014), pretrained on the ImageNet dataset as the basis for the AI assistance. Our choice of VGG-19 was motivated by previous experiments (Steyvers et al., 2022) that showed that the performance of the VGG-19 model could be manipulated to produce above-human performance for the challenging image noise conditions in the experiment.

Three different levels of classifier performance were created by differentially fine-tuning the VGG-19 architecture to the phase noise used in our experiment. All models were trained with all levels of phase noise. However, to generate these different levels of performance, the models were fine-tuned for different periods of time. We used a pilot experiment with 145 participants to assess human performance at the different noise levels. Classifier A was produced by fine-tuning for less than one epoch (10% of batches of the first epoch) and produced a performance level that was on average below human performance. Classifier B was produced by fine-tuning for the entirety of one epoch and produced a performance level that was on average near human performance. Classifier C was fine-tuned for 10 epochs and produced a performance level above average human performance.

Procedure

In both the concurrent and sequential paradigms, participants were instructed to classify images as best as possible and to leverage AI assistance, when provided, to optimize performance. Each participant was assigned to a single classifier level (A, B, or C) at the start of the experiment and each was only presented with AI assistance from that particular classifier; 20 participants were assigned to each classifier level in concurrent paradigm, and 25 participants to each classifier level in the sequential paradigm. Participants were given no information about the accuracy of the classifier.

Concurrent paradigm In the concurrent paradigm, there were 256 trials total. Each trial presented a unique image randomly selected from the set of 256 images. The classification trials were separated into 4 blocks where each block consisted of 48 consecutive trials in which AI assistance was turned on, and 16 consecutive trials without AI assistance. The larger number of trials with AI assistance was used to better assess participants AI reliance strategies under different levels of AI confidence. Because of the random ordering of images across participants, a particular image was shown for some participants in the AI assistance condition and for other participants in the control condition without AI assistance. Each unique image was shown to a median of 15 participants in the control condition and 45 participants in the AI assistance condition.

On each trial, participants were shown an interface as illustrated in Fig. 2. Participants classified images into 16 categories by pressing the response buttons that represented the categories with visual icons as well as labels (when the participant hovers the mouse over the button). For each classification, the participant provided a discrete confidence level (low, medium, and high). Finally, the rightmost column of the interface was used for AI assistance. When AI assistance was turned off, this column displayed nothing. However, when AI assistance was turned on, a grid of the 16 category options was shown with the same layout as the participant response options. Each of the 16 categories would be highlighted based on a gradient scale associated with the probability that the AI classifier assigned to the category. The darker the hue of the highlighted category, the more confident the classifier was in that selection. Instances in which the classifier was extremely confident in a single category, there would only be one category highlighted with an extremely dark hue. However, in instances where the classifier was not confident in a classification, there would be multiple categories highlighted with low hue levels. Participants were to use the AI assistance to aid their classification decision so as to optimize their own performance on the task. At the end of each trial, feedback was provided to enable the participant to develop an AI reliance strategy tailored to the particular AI algorithm they were paired with. In the feedback phase, the correct response option was highlighted in blue. If the participant was incorrect, the incorrect response was highlighted in red.

Sequential paradigm In the sequential paradigm, there were 192 trials total. Each trial presented a unique image randomly selected from the set of 256 images. On each trial, participants were first tasked with classifying an image on their own and were shown the interface as displayed in Fig. 2 but without AI assistance (the third column showing AI assistance was completely blank). After selecting their initial classification decision and submitting their response by selecting a confidence level, participants then were provided with AI assistance. The user interface at this stage looked exactly like Fig. 2 and the procedure for displaying AI confidence was the same as in the concurrent procedure. With AI assistance turned on, participants then made a final classification decision for the image shown and submitted their response by selecting their confidence level. Once a final classification was made, participants were provided feedback for 3 s.

Results

Figure 5 shows the average accuracy across noise levels, AI classifier accuracy levels, AI assistance conditions, and the concurrent and sequential advice-taking paradigms. In both the concurrent and sequential procedures, substantial performance differences are observed as the level of image noise varies, ranging from near ceiling performance at the zero noise level to close to chance-level performance (i.e., 1/16 = 0.0625) at the highest noise level. Across all classifier conditions, human performance improves with AI assistance, especially at intermediate levels of noise, as illustrated in Fig. 6. For classifiers B and C, the AI assistance produces performance levels comparable to the AI alone. For classifier A, the AI assistance improves human performance even though the AI assistance's accuracy is below human performance, *on average*. Note that this result is possible when participants rely on AI assistance on select trials when participants are in a low confidence state and the classifier is in a relatively high confidence state (see Appendix 5 for an analysis of the relationship between human and AI confidence). Overall, these results show that participants are able to rely on AI assistance to produce complimentarity — the joint human-AI accuracy is equal to or better than either the human or the AI alone.

The results are very similar across the concurrent and sequential paradigms. The average human accuracy with AI assistance for classifiers A, B, and C is 57%, 62%, and 68% respectively in the concurrent paradigm and 56%, 61%, and 65% respectively in the sequential paradigm. A Bayesian

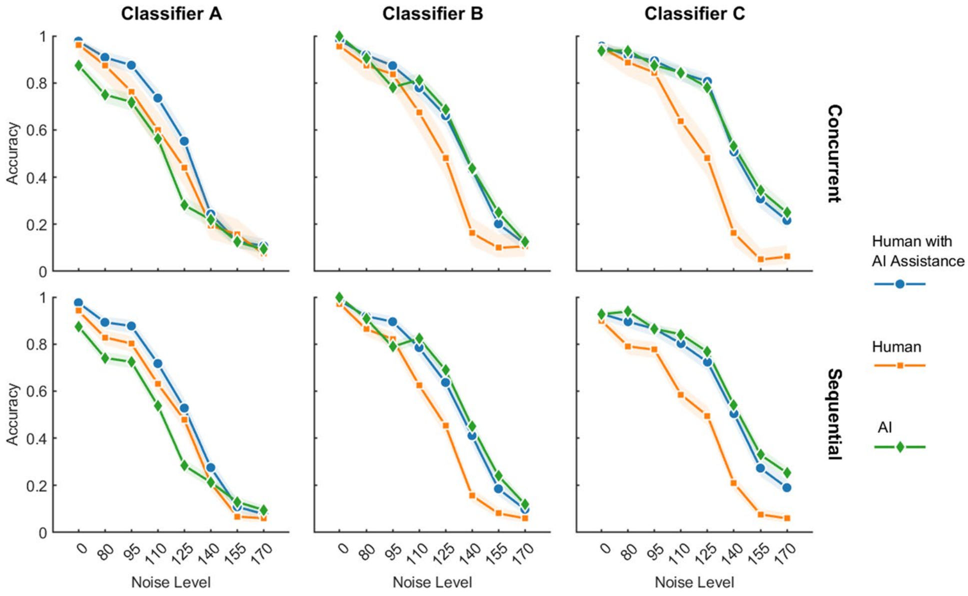


Fig. 5 Human accuracy with and without AI assistance as well as AI accuracy as a function of noise level (horizontal axis) across the concurrent and sequential paradigms (rows). Columns show different types of AI classifiers: classifier A's accuracy is below average human accuracy, classifier B's accuracy is comparable to average human accuracy, and classifier C's accuracy is above average human accuracy. Error bars reflect the 95% confidence interval of the mean based on a binomial model

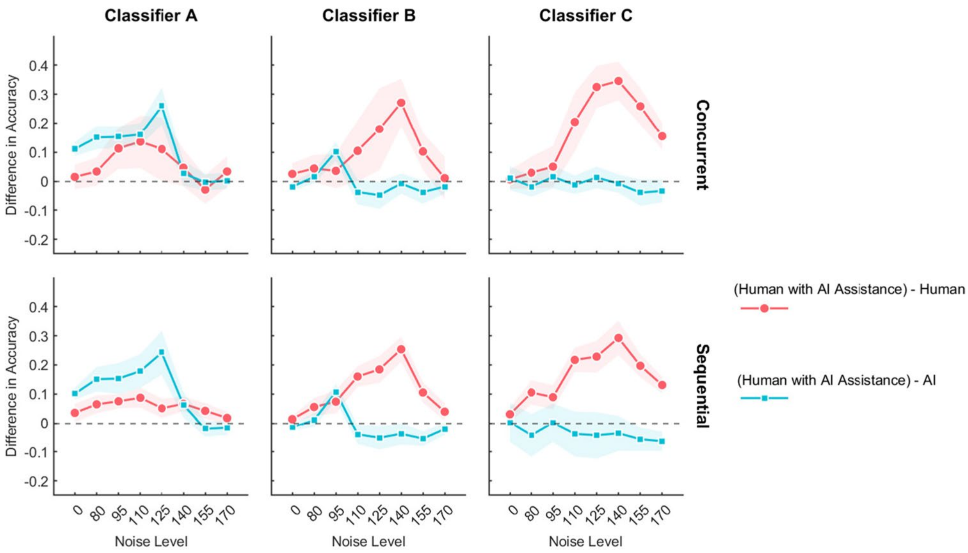


Fig. 6 Differences in accuracy with AI assistance relative to no AI assistance and AI only. Results are shown as a function of noise level (horizontal axis) and type of AI (columns) across the concurrent and sequential advice-taking paradigms. Error bars reflect the 95% confidence intervals

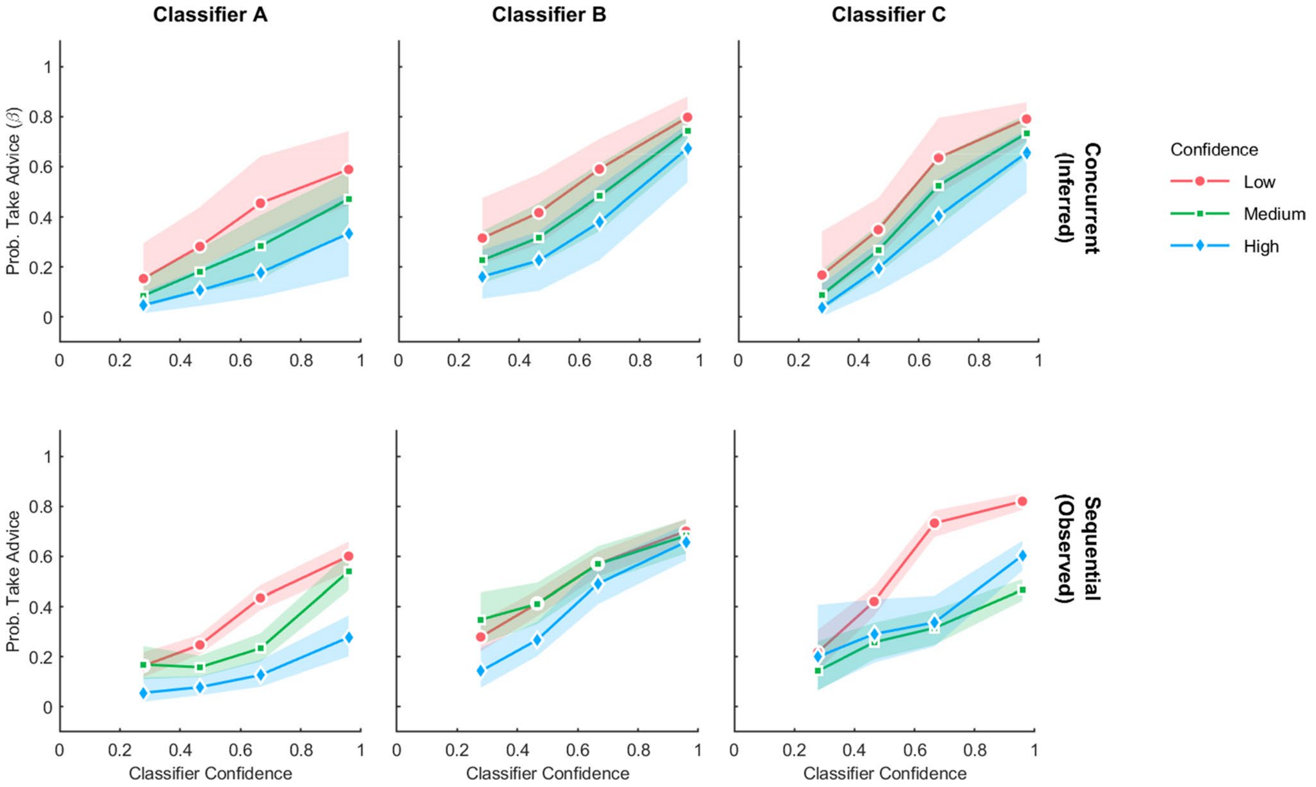


Fig. 7 Advice-taking policies inferred from the advice-taking behavior in the concurrent paradigm (top row) and observed in the sequential paradigm (bottom row). The policy determines the probability of taking the AI advice as a function of human confidence (colors), classifier confidence (horizontal axis), and type of classifier (columns). The colored areas in the top row show 95% posterior credible intervals. The colored areas in the bottom row reflect the 95% confidence interval of the mean based on a binomial model. The inferred advice-taking parameters (β) are converted from log-odds to probabilities in this visualization

independent samples *t*-test showed no evidence for a difference in performance for any the classifiers (i.e., all Bayes Factors < 1)[2]. That these results are consistent and very similar in both the concurrent and sequential experiments suggests that the experimental advice-taking paradigm does not produce important differences in how humans rely on and integrate AI assistance.

Model-Based Analysis

The empirical results showed that the concurrent and sequential advice-taking paradigms produce similar levels of accuracy across all experimental manipulations. In this section, we report the results of applying the cognitive model to the data from the concurrent paradigm.

We used a Markov chain Monte Carlo (MCMC) procedure to infer model parameters for the graphical model as illustrated in Fig. 3 (see Appendix 2 for details). Generally, the model is able to capture all the qualitative trends in the concurrent paradigm (see Appendix 4 on an out-of-sample assessment of model fit). We focus our analysis on two key parameters estimated by the model: β, the advice-taking policy at the population level, and α, the advice-taking policy for individual participants. In the next sections, we illustrate the inferred policies and compare the results against the empirically observed strategies from the sequential advice-taking paradigm. In addition, we analyze how effective the policies are relative to the set of all possible policies that participants could have adopted, ranging from the worst to best policies.

Inferred Advice-Taking Policies

Figure 7, top row, shows the inferred advice-taking policy β as a function of classifier confidence, participant confidence and classifier. These policies represent the behavior of an average participant at the population level of the model. Figure 8 shows examples of inferred advice-taking policies (α) from a subset of individual participants. Overall, the probability of taking AI advice differs

[2] Bayes factors were computed using JASP (JASP Team , 2022) with the default priors that came with the software.

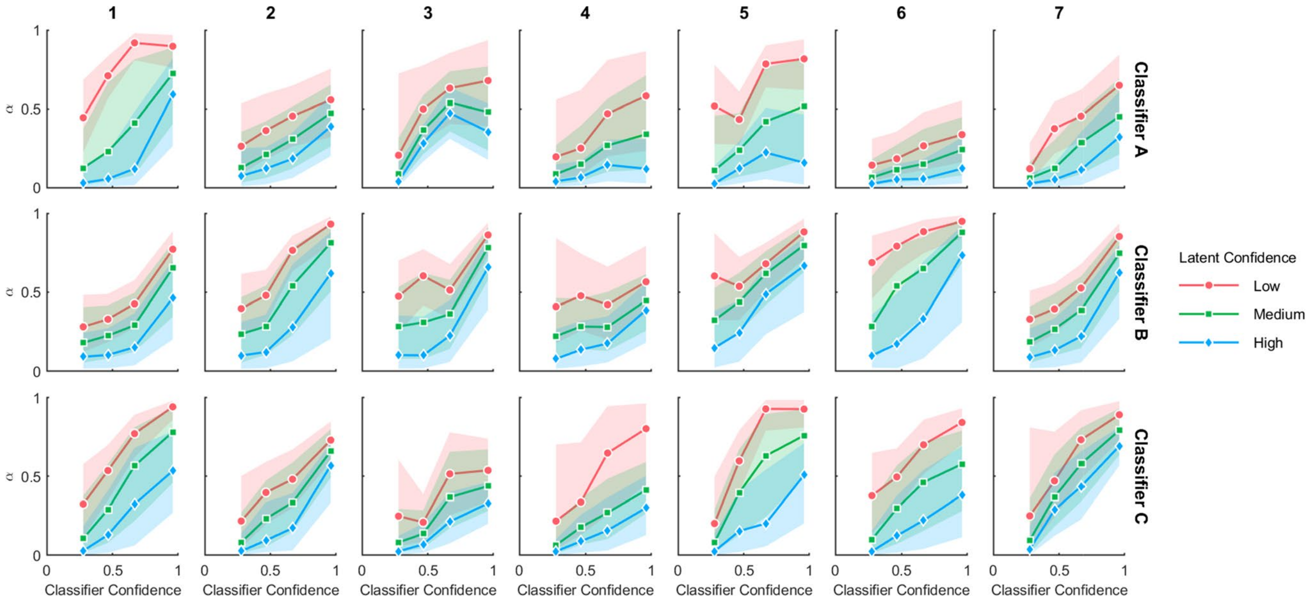


Fig. 8 Inferred advice-taking policies for a subset of 7 individual participants in the concurrent paradigm. The policy determines the probability (α) of taking the classifier advice as a function of human confidence (colors), classifier confidence (horizontal axis), and type of classifier (rows). Colored areas show 95% posterior credible intervals

substantially across classifiers. Advice is more likely to be accepted when the participant is in a low confidence decision-state and the classifier provides high confidence recommendations. In addition, across the different levels of classifier accuracy, advice is more likely to be accepted from high accuracy classifiers. Overall, these results show that the advice-taking behavior depends on a number of factors and is not based on simple strategies that rely solely on the confidence level of the AI or the confidence level of the participant. In addition, the results show that the advice-taking behavior is adjusted when the AI assistance becomes more accurate, from classifier A to classifier C, showing that participants are sensitive to AI accuracy.

Figure 7, bottom row, shows the empirically observed reliance strategies for the sequential paradigm. This analysis focuses on the subset of trials where the initial prediction from the participant differs from the AI prediction (which is not yet shown) and then calculating the proportion of trials where the participant switches to the AI prediction. Importantly, even though there are some quantitative differences that can be observed between the reliance strategies in the two paradigms, the qualitative patterns are the same. Thus, the results from the sequential paradigm provide a key validation of the cognitive model. The latent strategies uncovered by the cognitive model in the concurrent paradigm are very similar to those observed in the sequential paradigm.

Effectiveness of the Advice-Taking Policies

We now address the question of how effective are the participants' advice-taking policies. How much better (or worse) could participants have performed if they changed their advice-taking strategy? Figure 9 shows the range of

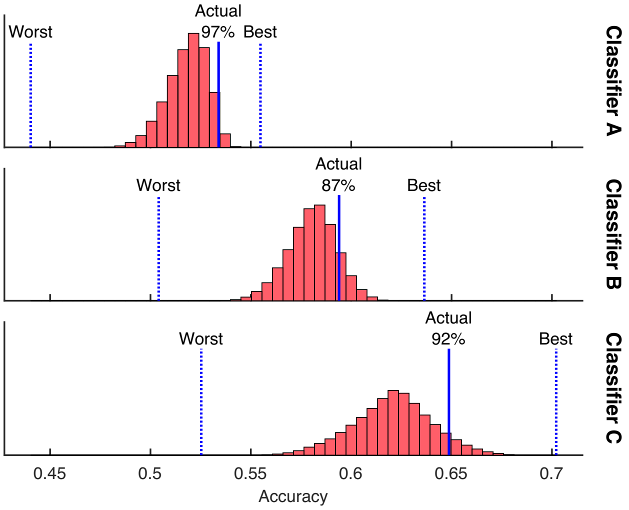


Fig. 9 Accuracy of the advice-taking policy at the population level relative to the best and worst possible advice-taking policies. The distributions show the accuracy of randomly sampled advice-taking policies. To quantify the participants' performance levels, percentages show the percentile rank of their performance relative to the accuracy distribution over all possible policies

all possible outcomes across different instantiations of the advice-taking policies. The accuracies of the worst and best possible advice-taking strategies were inferred by an analysis that optimizes performance conditional on the performance of the participants (Appendix 3). Note that the worst to best accuracies span the range of all possible outcomes. To understand how effective the average participants' policies (β) are on this range, we used a Monte Carlo sampling procedure to derive the accuracy distribution over all strategies (see Appendix 3 for details) and compute the percentile rank of the participant strategies in this distribution. These results show that the actual policies adopted by participants were highly effective, scoring in or near the top 10% of all possible strategies. Figure 10 shows the percentile rank for all individual participants when the effectiveness analysis is applied to the individual participant data. While a small subset of participants used suboptimal reliance strategies, the majority of participants used highly effective strategies.

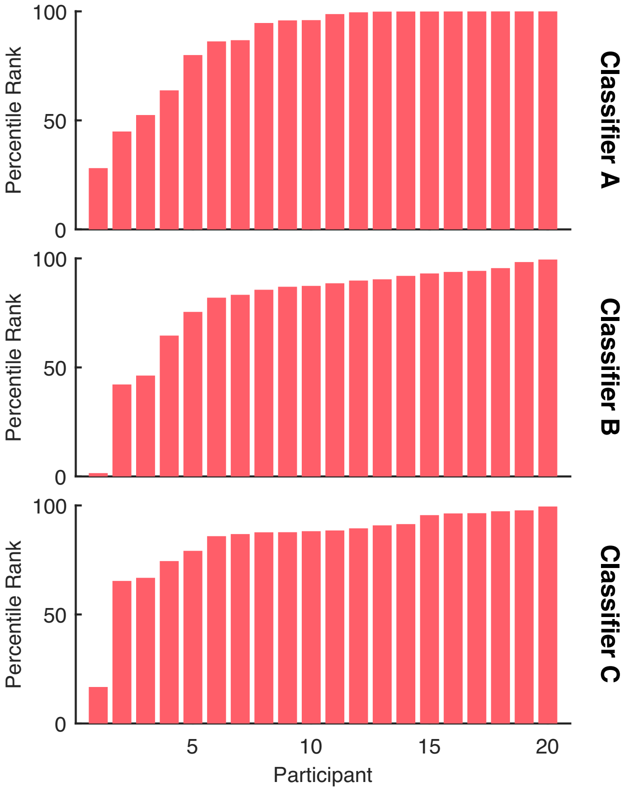


Fig. 10 Individual differences in the effectiveness of advice-taking strategies as assessed by the percentile rank relative to the distribution of all possible advice-taking policies

Discussion

Appropriate reliance on AI advice is critical to effective collaboration between humans and AI. Most research on AI-assisted decision-making has focused on gaining insight into the human's reliance on AI though empirical observations based on trust ratings and comparisons of observed accuracy and final decisions by humans and AI. For instance, in work that uses trust as a proxy for reliance, individuals are required to report their trust in the AI assistant (Lee and See, 2004). However, self-reported trust is not a reliable indicator of trust (Schaffer et al., 2019). Researchers have also compared the accuracy of the human-AI team when AI assistance is provided to the accuracy without assistance (Lai and Tan, 2019). However, this difference in accuracy is directly correlated with the performance of the AI. Another method used to investigate reliance is based on analyzing the agreement between the human's final decision and the AI's prediction (Zhang et al., 2020). This approach is problematic when used in the concurrent paradigm — while agreement can occur because of an individual's trust in the AI, it can also occur because the individual might have arrived at the same prediction as the AI even without the AI's assistance. Finally, in experiments using the sequential paradigm, reliance can be assessed by the propensity of individuals to switch to the AI's recommendation for those cases where their initial independent decision differs from the AI (Zhang et al., 2020; Yin et al., 2019). While this is a simple and straightforward procedure to gain insight into a reliance strategy, it cannot be applied to the concurrent paradigm as the individual's independent response is inherently unobservable.

Instead of using empirical measures to assess reliance, we developed a cognitive modeling approach that treats reliance as a latent construct. The modeling framework provides a principled way to reveal the latent reliance strategy of the individual by using a probabilistic model of the advice-taking behavior in the concurrent paradigm. It can be used to infer the likelihood that a human would have made a correct decision for a particular item independently even when their independent decision is not directly observed. The model is able to make this inference because it assumes that people, at the same levels of skill, will likely make the same prediction. The model allows us to investigate the difference between agreement with the AI and switching to AI advice (two metrics often used to assess trust) without explicitly asking the human to respond independently to each problem. In order to apply the model, empirical observations are needed that assess people's independent decisions without the assistance of an AI.

We showed that the AI reliance strategy inferred by the cognitive model on the basis of the concurrent paradigm is qualitatively similar to the AI reliance strategy observed in the sequential paradigm. Therefore, this demonstrates that a latent modeling approach can be used to investigate AI-assisted decision-making. The reliance strategy estimated by the model showed that participants discriminatively relied on the AI and varied their reliance from problem to problem. Participants were more likely to rely on the AI if they were less confident in their own decisions or when the AI was relatively confident. In addition, participants relied more heavily on AIs that were more accurate overall. This finding is consistent with (Liang et al., 2022) who showed that people rely on AI assistance more when the task is difficult and when they were given feedback about their performance and the AI's performance.

The results also showed that participants were able to build very effective reliance strategies compared to the optimal reliance strategy. We believe that participants were able to achieve this because of the following reasons. First, this is a simple image classification task and most people are experts at identifying everyday objects from images. This enables people to have a good understanding of their own expertise and confidence on any presented image. Second, 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. This feedback allowed people to build reasonable mental models of the AI assistant when paired with any of the three classifiers.

Finally, our results showed that the concurrent and sequential AI assistance paradigms led to comparable accuracy. Some researchers have argued that the sequential paradigm is superior to the concurrent paradigm because the initial unassisted prediction encourages independent reflection which could lead to retrieval of additional problem-relevant information (Green and Chen, 2019). However, consistent with our study, other studies have found no difference in overall performance between the concurrent and sequential paradigm (Buçinca et al., 2021). Another factor that could be relevant is the timing of AI assistance. The AI advice can be presented after some delay which provides the decision-maker additional time to reflect on the problem and improve their own decision-making accuracy (Park et al., 2019). Another possibility is to vary the amount of time available for people to process the AI prediction after it is shown making it more likely for people to detect AI errors (Rastogi et al., 2022). Overall, more research is needed to understand the effects of soliciting independent human predictions and varying the timing of the AI recommendation.

Our empirical and theoretical work comes with a number of limitations. First, we provided trial-by-trial feedback to help participants with the task of building a suitable mental model of AI performance. However, feedback is not always possible in real-world scenarios (Lu and Yin, 2021). Future research should investigate modeling extensions that model the cognitive process when participants do not receive feedback at all or receive feedback after a delay. Second, while the cognitive model captured the general process of advice taking based on a latent reliance policy, it did not model the process of establishing the reliance policy over time. Therefore, one important model extension, which we leave for future research, is to model the trial-by-trial adjustments of the reliance policy as a function of beliefs held a priori by participants about the accuracy of AI algorithms and external signals of AI confidence and accuracy as well as internally generated confidence signals.

Appendix

Appendix 1. The ordered probit model

The ordered probit model, $r \sim \text{OrderedProbit}(\theta, v, \sigma)$ is a generative model that maps a (latent) value θ to one of $R + 1$ discrete scores $r \in \{0, \dots, R\}$. In this process, noise is added to the latent value resulting in a new latent value, $\theta' = \theta + \epsilon$, where $\epsilon \sim \text{N}(0, \sigma)$ and the resulting discrete score is determined by the interval where θ' lies:

$$r = \begin{cases} 0 \;\; \text{if } \theta' \le v_1 \\ 1 \;\; \text{if } v_1 < \theta' \le v_2 \\ 2 \;\; \text{if } v_2 < \theta' \le v_3 \\ R \;\; \text{if } \theta' > v_R \end{cases} \tag{7}$$

The ordered vector $v = [v_1, \dots, v_R]$ represents the transition points between different discrete scores. With this construction, the probability of producing a score $r = m$ conditional on the latent value θ is:

$$P(r = m|\theta, \sigma) = \Phi((v_{m+1} - \theta)/\sigma) - \Phi((v_m - \theta)/\sigma) \tag{8}$$

where Φ is the cumulative standard normal distribution and $v_0 = -\infty$.

For our empirical (concurrent) paradigm, we use the ordered probit model to map the latent probability correct, θ to three different levels of confidence, $r \in \{Low, Medium, High\}$. Figure 11 shows an example of how the latent scores are mapped to confidence levels. Note that the higher value of the parameter σ (top panel) results in a noisier mapping of latent probabilities to discrete scores.

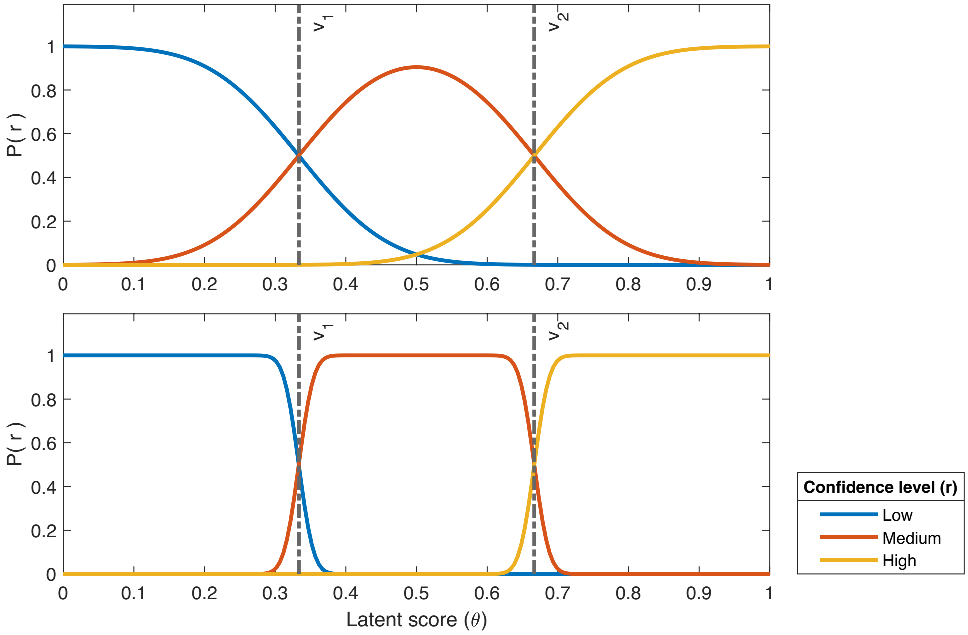


Fig. 11 Illustration of the ordered probit model to produce three levels of confidence. Top and bottom panels are produced with $\sigma = 1/10$ and $\sigma = 1/60$ respectively

Appendix 2. Details on model inference

We used Markov chain Monte Carlo (MCMC) to infer model parameters and obtain samples from the posterior distribution, conditioned on the observed data. We chose JAGS for posterior inference (Plummer et al. , 2003). To facilitate posterior inference, the inference procedure was separated into two stages. In the first stage, the observed data $x_{i,j}$, z_j, and $r_{i,j}$ from the no AI assistance condition was used to infer all model parameters related to person and item differences (a_i,d_j,s_j) and confidence generating process (σ_i, v_i). As a result of this stage, we computed posterior predictive distributions for the latent (independent) decisions $x_{i,j}$ and associated confidence levels $r_{i,j}$ for the AI assistance condition. In the second inference stage, the posterior modes of $x_{i,j}$ and $r_{i,j}$ were used as observed data, along with $y_{i,j,k}$, $c_{j,k}$, z_j, and $\eta_{j,k}$ to infer the advice-taking model parameters $\alpha_{i,j,k}$. In theory, one does not need to separate the first and second stage of inference and model parameters can be estimated in one joint procedure. We followed this two-stage inference process to facilitate the comparison with the optimization experiments (described in the next section). For both the first and second stage inference process, we ran the sampler with 8 chains with a burn-in of 1000 iterations before taking 50 samples per chain. The chains mixed appropriately. For prior distributions, we used normal priors for the ability and discrimination IRT parameters, consistent with previous Bayesian IRT modeling (Fox , 2010): $a_i \sim \mathcal{N}(0, 1)$, $s_j \sim \mathcal{N}(1, 1)I(0,)$, where $I(0,)$ denotes truncation a values below zero. Because of the large item differences in the classification task, we use a uniform prior spanning a large range of item differences, $d_j \sim \text{Uniform}(-10, 10)$. For the generative process of the confidence levels, we used $\tau_i \sim \text{Uniform}(0, 15), \sigma_i = 1/\tau_i$. In addition, we used uniform priors on the two cutpoints needed to produce three levels of confidence, $v_{i,1} \sim \text{Uniform}(0, 1)$, $v_{i,2} \sim \text{Uniform}(0, 1)$, with the constraint that the cutpoints are ordered (i.e., $v_{i,1} < v_{i,2}$).

Finally, for the advice-taking process, the AI reliance parameter α is treated as a $3 \times 4 \times 3$ lookup table for each individual i where entries are determined by the three confidence levels of the participant ("low," "medium," and "high"), 4 classifier confidence levels (0.00–0.35, 0.35–0.57, 0.57–0.78, 0.78–1.00), and 3 AI classifiers (A, B, and C). The classifier confidence levels were chosen to evenly distribute the observations across bins. Changing notation, the AI reliance parameters can be represented by $\alpha_{i,r,\eta,k}$ where r indexes the participant confidence level and η is the (discretized) AI confidence level. We use a hierarchical Bayesian approach to estimate the individual differences in reliance policies by assuming that these are sampled from a normal distribution on the log-odds scale $\log\left(\frac{\alpha_{i,r,\eta,k}}{1-\alpha_{i,r,\eta,k}}\right) \sim \mathcal{N}(\beta_{r,\eta,k}, \phi)$. The parameter β represents the advice-taking policy at the population level, the tendency across participants to accept AI advice. The standard deviation ϕ captures the spread in individual differences. For priors, we use $\beta_{r,\eta,k} \sim \mathcal{N}(0, 3)$. In addition, because there are relatively few "medium" confidence levels, we imposed an order constraint, $\beta_{1,\eta,k} \leq \beta_{2,\eta,k}, \beta_{2,\eta,k} \leq \beta_{3,\eta,k}$ for $\eta = 1, .., 4,$ and $k = 1, ...3$.

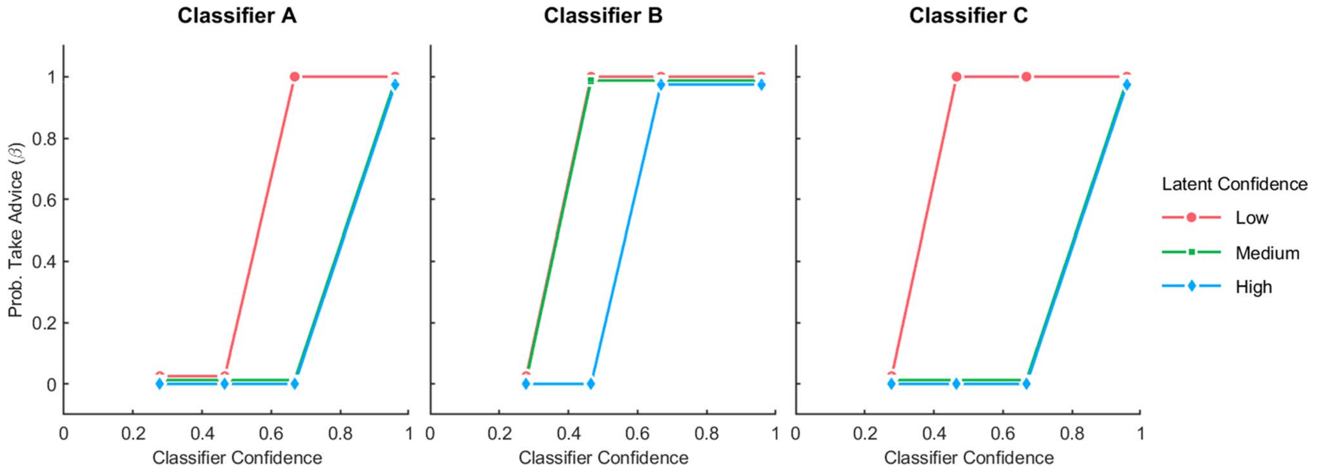


Fig. 12 Advice-taking policies derived from an optimality analysis to maximize accuracy conditional on human confidence and accuracy. The policy determines the probability of taking the AI advice as a function of human confidence (colors), classifier confidence (horizontal axis), and type of classifier (columns)

Appendix 3. Optimality analysis

The inferred advice-taking policies show that participants use a number of different factors in sensible ways. A natural question is what could participants have done differently in order to optimize their performance. We conducted an optimality analysis to identify the best possible policy to better understand the strategies that would have maximized accuracy in the AI-assistance condition. Importantly, in this analysis, the goal is to identify the most effective policy *conditional on* the performance of the participants and confidence states before AI assistance is provided. In other words, in the optimality analysis, we are not changing any assumptions about the ability of participants to classify images and express their confidence in their prediction — we are only considering scenarios where participants might have utilized the AI assistance in different ways. Also, we are not considering how participants have to learn about the effectiveness of their AI reliance policy over the course of the experiment.

Specifically, we start the analysis with the inferred confidence state of the participant (r) and accuracy (θ) of the predictions before AI assistance is considered. We then find out what strategy (at the individual level, α, or at the population level, β) would have maximized the accuracy of the final predictions (y) after AI assistance. The analysis is applied separately for each classifier A, B, and C.

The goal of the optimality analysis is to identify the best and worst possible advice-taking policy ($\alpha_{i,j,k}$) conditional on inferred accuracy ($\theta_{i,j}$), confidence state ($r_{i,j}$) of the participant for individual items before AI assistance is provided, and classifier type k. For the inferred accuracy $\theta_{i,j}$, we took the posterior mean for each item based on our MCMC procedure. For the confidence state $r_{i,j}$, we took the posterior mode for each item.

The analysis was based on a brute-force search conducted separately for the three types of classifier. In this search, the parameter values α were discretized into 80 equally spaced values between 0 and 1, and then searching the space across 3 levels of DM confidence ($r_{i,j}$) and 4 levels of AI confidence (η_j). We then applied Eqs. 6–5 to identify the α policy that produced the highest as well as the lowest accuracy of predictions $y_{i,j,k}$ in Eq. 5. The parameters were subject to an order constraint identical to the order constraint imposed in the MCMC procedure: α should be monotonically increasing for higher levels of participant confidence.

Figure 12 shows the resulting optimized policies (β) at the average participant level. These policies only take on extreme values such that advice is either always taken ($\beta = 1$) or never taken (β=0) for particular combinations of participant and classifier confidence. Similar to the participant strategies, the results show that classifier advice should more readily be adopted when the participant is in a low confidence state and the classifier is in a high confidence state[3].

[3] Note that the optimal policy for classifier B shows that advice should be accepted more often than for classifier C even though classifier B performs worse on average than classifier C. This result can be attributed to between-group differences in classifiers B and C as well as differences in classifier calibration.

Deriving the percentile rank of participants policies

We conducted a Monte Carlo procedure to estimate the percentile rank of the accuracy of participants' policies relative to accuracy achieved by random strategies. In this Monte Carlo procedure, we sampled $\alpha_{i,j,k}$ from a uniform (0,1) distribution separately for the 3 levels of DM confidence ($r_{i,j}$) and 4 levels of AI confidence (η_j). We computed the expected accuracy for each of the $\alpha_{i,j,k}$ samples. We next computed the percentile rank of the actual participants' policy relative to this distribution.

Appendix 4. Out-of-sample model predictions

To assess model fit of the concurrent experiment, we used a 10-fold cross-validation approach to compute out-of-sample model predictions for the human decisions and confidence levels. For each individual, the observed data from the AI assistance and no AI Assistance condition was randomly partitioned into 10 disjoint test sets. For the no AI assistance condition, model parameters were inferred on the basis of observed predictions x and confidence levels r for each training fold. For the test set, we used the MCMC inference procedure described in Appendix 2 to infer the predictions x and confidence levels r for the test set. For the AI assistance condition, the model has to predict the withheld data on y, the decisions made with the aid of the AI.

Figure 13 shows the out-of-sample model predictions and observed data. One point of deviation is that the model somewhat underpredicts the size of the assistance effect (bottom row). However, the model captures all qualitative trends in the data.

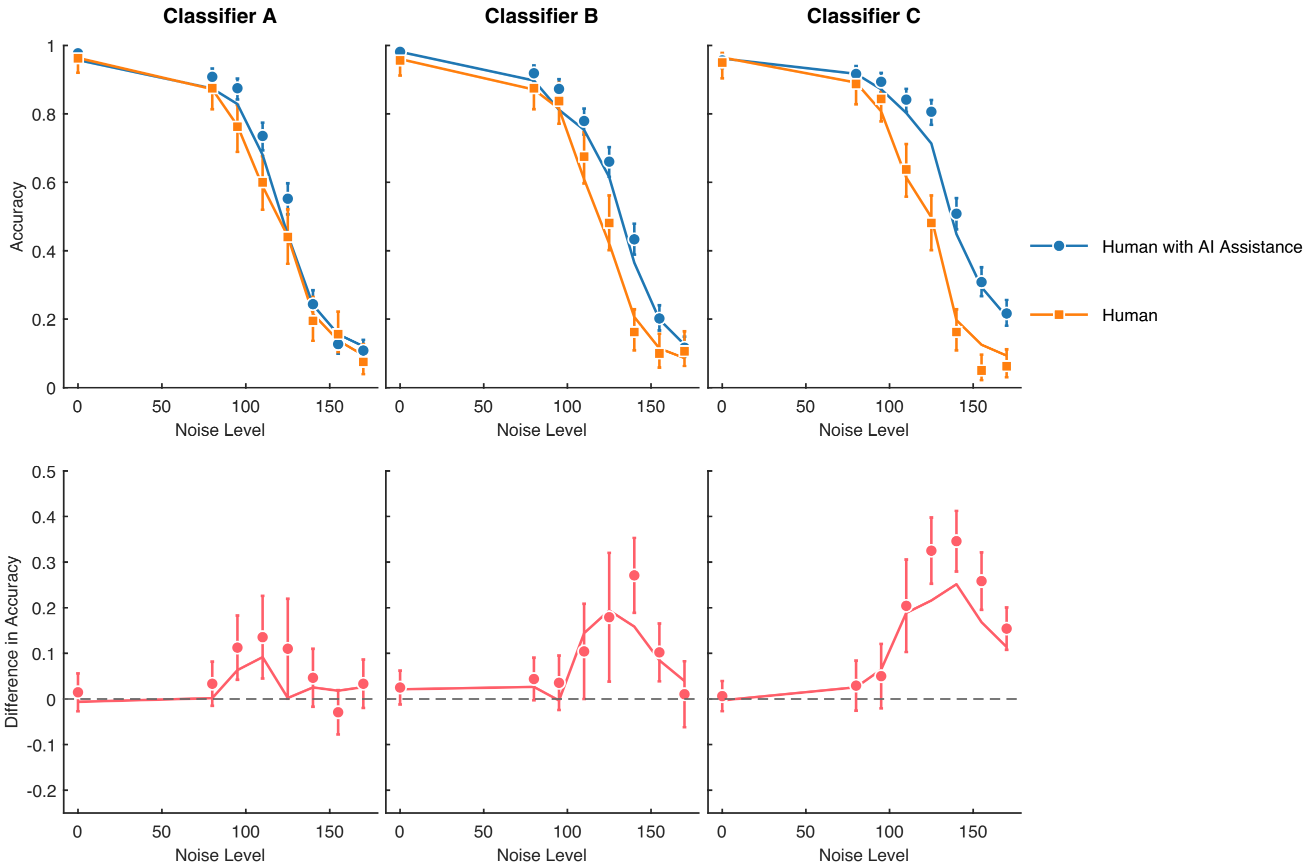
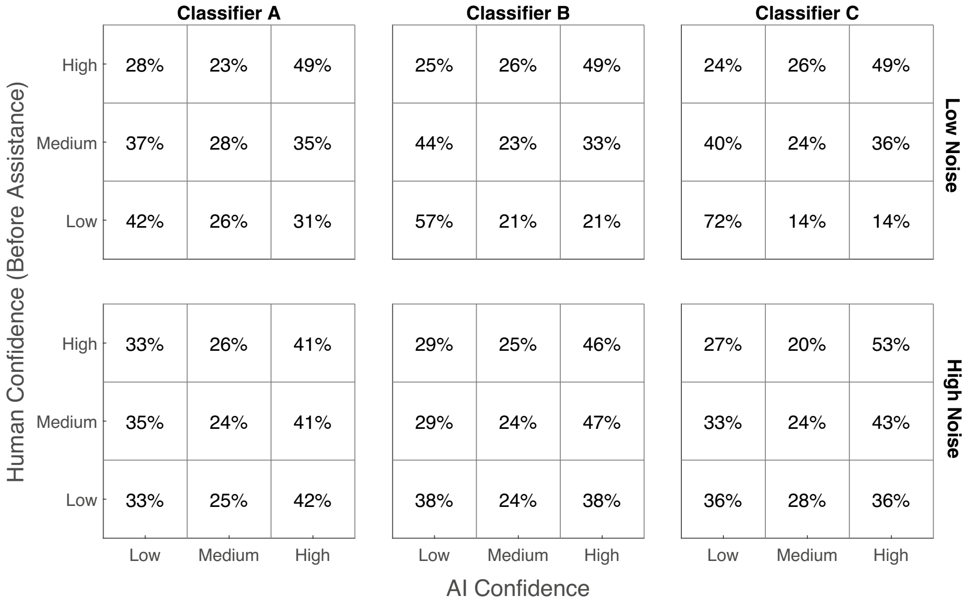


Fig. 13 Model predictions and observed data for human performance with and without AI assistance in the concurrent paradigm. Model predictions and data are shown with lines and points respectively. Error bars reflect the 95% confidence interval of the mean of the observed data based on a binomial model

Appendix 5. Relationship between human and AI confidence

Prior to AI assistance, human confidence levels are correlated with AI confidence scores, with Spearman's rank correlations of 0.28, 0.43, and 0.47 for AI classifiers A, B, and C respectively (in this analysis, we are combining results across the sequential and concurrent conditions). Therefore, what is a difficult problem for the human participant (e.g., a high noise classification problem) tends to be challenging for the classifier as well. Figure 14 provides more detailed information about the relationship broken down by classifier and degree of image noise. For ease of interpretation, AI confidence scores were discretized into three labels "Low," "Medium," and "High" where the cutoffs to define the labels were chosen such that the marginal frequencies of the labels matched the marginal frequencies of human confidence ratings (note that in the experiment, the participants did not see these discrete AI confidence labels).

For low noise conditions (phase noise levels at 110 or lower), there is a stronger correspondence between human and AI confidence, such that there are few cases (fewer than 14% for classifier C) where the human is in a low confidence state and the AI is in a high confidence state (or vice versa). However, for the more challenging high-noise classification problems (phase noise levels above 110), the correspondence between human and AI confidence is reduced and in roughly a third of cases, the human and AI are in opposite confidence states.



Human Confidence (Before Assistance)	Classifier A			Classifier B			Classifier C			
AI Confidence	Low	Medium	High	Low	Medium	High	Low	Medium	High	
High	28%	23%	49%	25%	26%	49%	24%	26%	49%	Low Noise
Medium	37%	28%	35%	44%	23%	33%	40%	24%	36%	
Low	42%	26%	31%	57%	21%	21%	72%	14%	14%	
High	33%	26%	41%	29%	25%	46%	27%	20%	53%	High Noise
Medium	35%	24%	41%	29%	24%	47%	33%	24%	43%	
Low	33%	25%	42%	38%	24%	38%	36%	28%	36%	

Fig. 14 Relationship between AI and human confidence scores prior to AI assistance across AI classifiers and noise levels. Percentages in each row show the relative number of AI discretized confidence levels given a particular level of human confidence. AI confidence levels were discretized into three labels to match the marginal frequencies of the human label frequencies (34%, 25%, and 41% for "Low," "Medium," and "High" ratings). The results are combined across the concurrent and sequential conditions without AI assistance. Low noise (top row) includes the 0, 80, 95, and 110 phase noise levels whereas high noise (bottom row) includes the 125, 140, 155, and 170 phase noise levels

Author Contribution M.S., H.T., and P.S. designed research; M.S. and H.T. performed research; M.S., H.T., and A.K. analyzed data; and M.S., A.K., H.T., and P.S. wrote the paper.

Funding Information This work was supported by the National Science Foundation (Grant Numbers: 1900644 and 1927245). Authors P.S. and M.S. received the research support from the NSF.

Data Availability The original and preprocessed versions of the data can be accessed at https://osf.io/9xbpu/?view_only=f8d57ea4ec1940a3bc16bebc17a8ce51.

Code Availability Code can be accessed at https://osf.io/9xbpu/?view_only=f8d57ea4ec1940a3bc16bebc17a8ce51.

Declarations

Ethics Approval The experimental protocol was approved by University of California, Irvine Institutional Review Board.

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Consent for Publication Consent to publish is not required as no individual subject information is presented in this study.

Competing Interests The authors declare no competing interests.

References

Becker, F., Skirzyński, J., van Opheusden, B., & Lieder, F. (2022). Boosting human decision-making with AI-generated decision aids. arXiv preprint arXiv:2203.02776

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). 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, 5,* 1–21.

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*, 809–825.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (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.

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In *Decision Making in Aviation* (pp. 289–294). Routledge.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*, 114.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*, 115–118.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Geirhos, R., Medina Temme, C., Rauber, J., Schütt, H., Bethge, M., & Wichmann, F. (2019). Generalisation in humans and deep neural networks. In *Thirty-second Annual Conference on Neural Information Processing Systems (NeurIPS 2018)* (pp. 7549–7561). Curran.

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–24.

JASP Team (2022). JASP (Version 0.16.2)[Computer software].

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*, 237–293.

Kumar, A., Patel, T., Benjamin, A. S., & Steyvers, M. (2021). Explaining algorithm aversion with metacognitive bandits. In *Proceedings of the Annual Meeting of the Cognitive Science Society*. volume 43.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 29–38).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50–80.

Liang, G., Sloane, J. F., Donkin, C., & Newell, B. R. (2022). Adapting to the algorithm: How accuracy comparisons promote the use of a decision aid. *Cognitive Research: Principles and Implications, 7,* 1–21.

Logg, J. M. (2017). Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086*.

Lu, Z., & Yin, M. (2021). Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557

Park, J. S., Barber, R., Kirlik, A., & Karahalios, K. (2019). A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction, 3*, 1–15.

Patel, B. N., Rosenberg, L., Willcox, G., Baltaxe, D., Lyons, M., Irvin, J., Rajpurkar, P., Amrhein, T., Gupta, R., Halabi, S., et al. (2019). Human-machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine, 2*, 1–10.

Pescetelli, N., Hauperich, A.-K., & Yeung, N. (2021). Confidence, advice seeking and changes of mind in decision making. *Cognition, 215*, 104810.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences, 115*, 6171–6176.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–52).

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making, 19*, 455–468.

Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R. L., Mendelson, M., Maartens, G., van Hoving, D. J., Griesel, R., Ng, A. Y., Boyles, T. H., & Lungren, M. P. (2020). CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digital Medicine*, *3*.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 6,* 1–22.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision, 115,* 211–252.

Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., et al. (2019). Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology, 126*, 552–564.

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 240–251).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences, 119*, e2111547119.

Wang, X., Lu, Z., & Yin, M. (2022). Will you accept the AI recommendation? Predicting human behavior in AI-assisted decision making.

Wright, D. E., Lintott, C. J., Smartt, S. J., Smith, K. W., Fortson, L., Trouille, L., Allen, C. R., Beck, M., Bouslog, M. C., Boyer, A., et al. (2017). A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society, 472*, 1315–1323.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–12).

Zhang, G., Raina, A., Cagan, J., & McComb, C. (2021). A cautionary tale about the impact of AI on human design teams. *Design Studies, 72*, 100990.

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