Title
Explaining Algorithm Aversion with Metacognitive Bandits

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN
1069-7977

Authors
Kumar, Aakriti
Patel, Trisha
Benjamin, Aaron S
et al.

Publication Date
2021

Peer reviewed

Explaining Algorithm Aversion with Metacognitive Bandits

Aakriti Kumar (aakritk@uci.edu)
Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697 USA

Trisha Patel (tpatel65@illinois.edu)
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA

Aaron S Benjamin (asbenjam@illinois.edu)
Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820 USA

Mark Steyvers (mark.steyvers@uci.edu)
Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697 USA

Abstract

Human-AI collaboration is an increasingly commonplace part of decision-making in real world applications. However, how humans behave when collaborating with AI is not well understood. We develop metacognitive bandits, a computational model of a human's advice-seeking behavior when working with an AI. The model describes a person's metacognitive process of deciding when to rely on their own judgment and when to solicit the advice of the AI. It also accounts for the difficulty of each trial in making the decision to solicit advice. We illustrate that the metacognitive bandit makes decisions similar to humans in a behavioral experiment. We also demonstrate that algorithm aversion, a widely reported bias, can be explained as the result of a quasi-optimal sequential decision-making process. Our model does not need to assume any prior biases towards AI to produce this behavior.

Keywords: Algorithm aversion; Human-AI interaction; Bandit problems; Cognitive modelling; Bayesian modeling, Metacognition

Introduction

In an era in which AI is influencing areas of decision-making that were historically the province of human subjectivity and expertise, fostering effective teamwork between humans and agents is increasingly important. AI now assists doctors who look towards binary classifiers to decide which patients to send to outpatient programs (Kamar, 2016) and courts who use risk assessment tools to predict recidivism (Green & Chen, 2019), among many other examples. As part of the collaborative process, humans are increasingly reliant on complex and often opaque algorithms to support their decision making. The shift to hybrid human-AI decision making has been accompanied by a growing body of work that investigates the dynamics of AI-assisted decision making. In this paper, we present a cognitive science perspective on how humans decide when to solicit an AI's advice as opposed to relying on their own judgement.

Resistance to outside advice is not unique to human-machine teams: humans discount advice from peers and tend to rely on their own judgment, even when that judgment is inexpert (Bonaccio & Dalal, 2006). Humans also exhibit excessive and unwarranted confidence in their own judgments relative to those of their peers (Gino & Moore, 2007). Recent work suggests that a number of similar behaviors might be at work when humans collaborate with AI. These human biases can lead to suboptimal outcomes.

Two widely discussed biases are *algorithm aversion* and *algorithm appreciation*. Algorithm appreciation is the tendency of a human to prefer algorithmic help over another human's help (Logg, Minson, & Moore, 2019). In contrast, algorithm aversion has been described as the tendency of a human to disregard the recommendations of a machine after observing that it made a mistake. This can occur even when the algorithm can be beneficial to the human decision maker on average (Dietvorst, Simmons, & Massey, 2015). Human behavior consistent with these biases is often reported as inappropriate reliance by the human on the AI. One proposed explanation for algorithm aversion is that early errors by the algorithm lead to a loss of trust, and consequently to an inadequate exploration of an algorithm's capability by the human partner. This finding is consistent with human factors research showing that initial interactions and negative interactions are known to have a greater impact on trust in AI than interactions later in the exchange (Logg, 2017). This loss of trust may result in reduced reliance on the AI.

In this paper, we present the first (to our knowledge) cognitive models for human-AI interaction. Our starting point is that humans compute the utility of advice from an algorithm and that they behave like quasi-ideal observers, performing Bayesian inference to decide when to ask for AI assistance. The computational model combines two different cognitive processes: *explore/exploit sequential decision-making* and *metacognition*. On each trial, the human needs to decide whether to solicit the advice of the AI or rely on their own judgement. This decision requires a form of metacognition — thinking about the relative abilities of oneself and the AI. In our setting, the human engages metacognition to infer and compare the utility of making one's own decision with the utility of seeking the advice of an AI. We model the sequential decision-making problem of soliciting advice on each trial as an explore/exploit problem.

On the one hand, the human can *explore* by choosing to solicit the advice of the AI. This action is risky, since the AI has an unknown capacity and the action to solicit advice is associated with time costs associated with soliciting, processing, and integrating the advice with one's own judgment). The solicitation pays off if the utility of AI advice exceeds the utility of making an independent decision. On the other hand, the human can *exploit* by choosing to forge ahead with an independent judgment. This choice is less risky when confidence in one's decision is high. We will show that the proposed computational models produce behavior consistent with a wide range of behaviors that humans display when presented with such a choice, including algorithm aversion. Consistent with the standard interpretation of algorithm aversion, early errors lead the (simulated) human to under-utilize the AI and in some cases to completely disregard it. However, the model also predicts a more general pattern of algorithm aversion when the AI utility is not only based on perceived accuracy but also includes temporal factors related to time to receive the AI advice and cognitive efforts to process the advice. The computational model predicts that humans can abandon the AI advice even in the absence of early errors if the perceived accuracy advantages of AI advice do not make up for the perceived temporal costs. Finally, the model also provides a framework for incorporating the difficulty of a trial into the decision of soliciting advice.

The paper is organised as follows. We first report the qualitative results from an experiment on AI advice solicitation. We then introduce the basic metacognitive bandit model that can account for several important qualitative patterns in the data. In addition, we introduce model extensions that can account for variation in task difficulty across trials. We end with a discussion of the implications of this metacognitive framework for human-AI team collaborations and suggestion some future directions for research.

Experimental Data

We provide a brief description here from one of a series of experiments on AI advice solicitation. In the experiment, 40 participants first made an independent judgment on a perceptual decision-making task and were then given the option to solicit the advice from an AI agent. A key feature of the experiment is that information about the accuracy of the AI is only evident when its advice is solicited.

Methods

40 participants were recruited through an online subject pool at the University of Illinois at Urbana-Champaign and received assignment credit for participation. Participants were first shown a fixation point for 500 ms followed by a random-dot kinematogram (See Figure 1(a)) for 500 ms. Participants were tasked with identifying the dominant direction of movement in the kinematogram (left or right). The coherence (randomness) of the kinematograms varied between 0 and .3 —

a coherence of 0 corresponds to maximum randomness or highest difficulty and difficulty decreases with increase in coherence value. Participants were presented with 240 trials. The sequence of events in the experiment as shown in Figure 1(b) were as follows. Participants were shown a kinematogram and were asked to submit an initial response. After submitting their response, they were asked to rate their confidence (low, medium or high) in their decision. Next, they were given the option to solicit the advice of an AI agent. If they chose to solicit advice, they were shown the AI recommendation. If not, they were shown feedback (correct/incorrect) on their original answer.

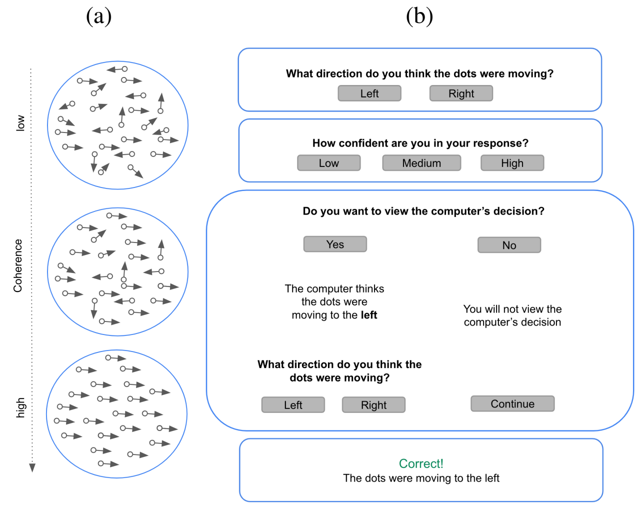

Figure 1: Experimental setup: (a) Kinematograms with varying coherence levels (inversely related to difficulty) were used as stimuli. (b) Sequence of events in the task.

If they solicited the AI's advice, they were allowed to change their answer after viewing the AI's recommendation. The AI advice did not include a confidence rating. AI advice was simulated by the experimenters such that AI accuracy decreased as a function of coherence. Participants submitted their answer after taking into account the AI's advice. This was followed by feedback (correct/incorrect) on their final response.

Empirical Results

For the purpose of this paper, we focus on four qualitative findings related to the initial decision in the perceptual task and the decision to solicit advice. First, human performance was on average poorer than that of the AI. Participants were correct 69% of the time on their first judgment, and the AI was correct 81% of the time. Therefore, on average, participants should be able to increase performance by soliciting and adopting the advice of the AI. Second, there was substantial variation in the degree of soliciting AI advice across participants and trials. Figure 3(a) and 4(a) show that the tendency to solicit AI advice decreased over time. In addition, some individuals stopped soliciting advice after only a few trials, whereas other individuals kept soliciting advice across the entire experiment. Third, Figure 6(a) shows that confidence in the initial decision is related to accuracy, suggesting that participants have accurate metacognitive awareness of the difficulty of that particular trial and the associated level of uncertainty in their decision. Finally, figure 6(b) shows a final important result: the AI was solicited more often when the participant was less confident.

Computational Modeling Approach

The computational problem associated with the decision to solicit advice from an AI can be formulated as an optimal exploration effort: Humans need to infer the relative utility of relying on themselves or the AI assistant to inform future decisions of when to seek help. This problem can be elegantly captured using a multi-armed bandit framework. Bandit problems have been widely used to study sequential decision-making when there is uncertainty about the rewards associated with decisions (or arms). In a machine learning context, multi-armed bandits have been used to efficiently choose between different sources of information, such as crowd workers and/or machine learning models (Tran-Thanh, Stein, Rogers, & Jennings, 2014) and active assessment of machine classifiers (Ji, Logan IV, Smyth, & Steyvers, 2021). In cognitive science, multi-armed bandits have been used to model human sequential decision behavior in reward and information seeking environments (Steyvers, Lee, & Wagenmakers, 2009; Speekenbrink & Konstantinidis, 2015; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018). The objective of a bandit problem is to maximize the expected value of the cumulative rewards received during the decision-making process. This requires a balance between exploring all available arms and exploiting the best possible arm at any time.

We specify the decision to seek help from AI as a pull of one of two arms: self and AI. However, note that the decision to select an arm is a metacognitive one: the human needs to evaluate their own performance (which will reflect the subjective difficulty of the current problem) as well as learn about the AI arm's utility. This is different from a traditional bandit setting in which the evaluation of arms corresponds to competing external events. We now describe the details of the models we consider. Note that the goal is to capture the important qualitative trends observed in the empirical data; we plan to pursue quantitative modelling in future work.

Metacognitive Bandit

The metacognitive bandit captures the metacognitive process employed by a human to decide whether to seek AI help on an individual trial. The human relies on the performance history of both arms (AI and self) to inform the decision of arm selection. We use the framework of upper confidence bound (UCB) bandit models to model this process. Specifically, we use the Bayesian UCB framework proposed by Pavlidis, Tasoulis, and Hand (2008) as a solution to this metacognitive task. In this framework, the decision-maker constructs a $100(1-\lambda)\%$ credible interval for the expected reward from each action at each trial and greedily chooses the action with the highest upper bound of the credible interval. It favors the exploration of actions with high uncertainty that have the potential to produce favorable outcomes.

On each trial, the human infers a utility for soliciting the AI's help and a utility for coming up with a solution on one's own. This utility is a combination of expected accuracy and time costs associated with both arms. Since the AI has unknown accuracy, the human does not have a good estimate of its ability and the choice to solicit advice is risky. There are also time costs associated with the decision to ask for help. The human may also be uncertain about their own ability and may need to rely on trial feedback to update their own estimates of ability. The human then picks the arm that has the highest upper bound at each time step.

Let θ and ϕ denote the latent accuracy of the self arm (S) and the AI arm (AI) respectively. Let x_t denote the reward observed at each trial t for arm S and y_t denote the reward observed at each trial t for arm AI. The reward is 1 when an arm gives a correct response and 0 when the response is incorrect. Let a_t denote the action taken by the human where $a_t = 1$ if the AI was solicited on trial t and $a_t = 0$ if the AI was not solicited (i.e. the self arm was selected). We assume that the human always observes the reward for the self arm. However, for the AI arm, the correct and incorrect responses can only be observed for those trials when the arm was selected (i.e., a_t=1). We assume that the observed rewards of the two arms are generated independently and identically from two unknown Bernoulli distributions. In this Bayesian model, the human updates the posterior of latent accuracy θ_t and ϕ_t at trial t based on the history of the rewards $x_{1:t-1}$ and $y_{1:t-1}$ according to:

$$
\begin{aligned}
&\theta_t|x_1,\ldots,x_{t-1} \sim Beta(\alpha+\sum_{j=1}^{t-1} x_j, \beta+\sum_{j=1}^{t-1}(1-x_j)) \\
&\phi_t|y_1,\ldots,y_{t-1},a_1,\ldots,a_{t-1} \sim \\
&\qquad Beta(\gamma+\sum_{j=1}^{t-1} y_j a_j, \delta+\sum_{j=1}^{t-1}(1-y_j a_j))
\end{aligned} \tag{1}
$$

where (α,β) and (γ,δ) encode prior beliefs of the human about their own accuracy (arm S) and AI respectively. In essence, this model suggests that humans keep a count of the number of observed correct and incorrect responses by both arms.

The posterior latent ability distribution serves as a proxy for expected accuracy of the arms: higher perceived ability of an arm corresponds to a higher probability of the arm giving a correct response. In the UCB decision process, at each trial t, the human compares the upper confidence bounds of both arms to pick the arm a with the higher inferred utility:

$$
a_t = \begin{cases} 1 & \text{if } UCB(\theta_t,\lambda) > UCB(\phi_t,\lambda) - c \\ 0 & \text{otherwise} \end{cases} \tag{2}
$$

where $UCB(\theta_t,\lambda)$ is the $100(1-\lambda)\%$ upper confidence bound for the posterior distribution for θ and c is a time cost associated with the AI arm. This cost includes the additional cognitive effort and time assdociated with asking for advice and incorporating it into the final decision.

To illustrate the model, Figure 2 shows some predicted actions by the model based on different performance histories.

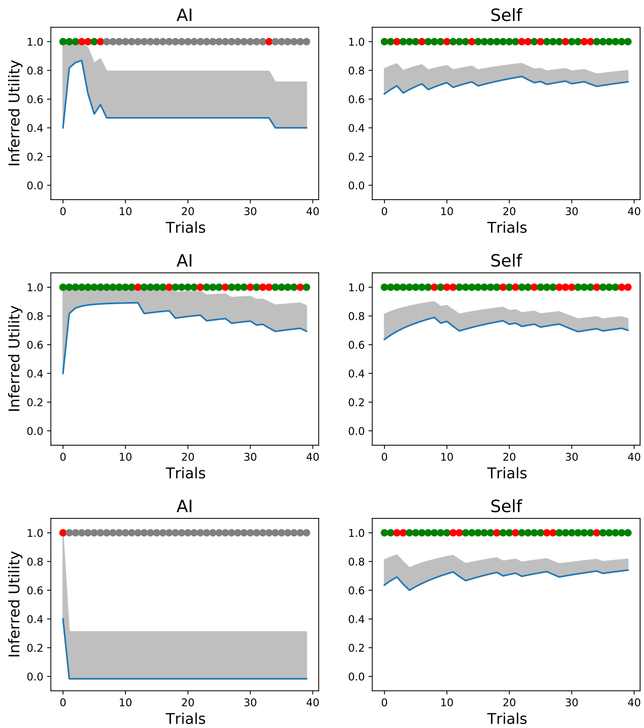

Figure 2: Illustrative results for the basic metacognitive bandit for three participants (rows) with inferred utilities for the AI arm (left column) and self arm (right column). The shaded region shows the posterior uncertainty associated with the inferred utility. Dots correspond to the rewards observed for each arm pull. The panels on the left show the human's inferred utility of asking the AI for help. Gray dots correspond to the AI arm not being pulled, hence no reward was observed for those trials. Green and red dots corresponds to correct and incorrect responses respectively. The panels on the right show the human's inferred utility of relying on own judgment and the associated reward sequence. The last row is an example of behavior that exemplifies the pattern often associated with algorithm aversion.

In this simulation, we set λ to .1 such that the UCB action selection is based on the 90th quantile of the latent abilities of the two arms. We set c to .1 to impose a small cost associated with the action of soliciting advice. We set (α, β) to (7,4) reflecting a moderately strong prior belief that the human agent's own accuracy is above chance and (γ, δ) to (.1,.1) corresponding to an absence of any belief about the AI performance in the task. Figures 3(b) and 4(b) show the advice seeking behavior across 240 trials for 40 simulated participants. We observe that the basic metacognitive bandit model produces behavior that is consistent with a variety of advice-seeking behaviors, including algorithm aversion. For example, for the third simulated participant (bottom row of Figure 2), advice was solicited on the first trial and proved to be incorrect. That single negative observation is sufficient for the model to no longer seek the AI advice for the remaining trials.

We should note that the substantial differences in behavior across the simulated participants occur despite the fact that no individual differences are built into the model. Each simulated human participant starts with the same prior beliefs and assumptions about time costs. Despite this homogeneity, there are strongly diverging patterns of advice-seeking behavior that are determined entirely by the feedback history.

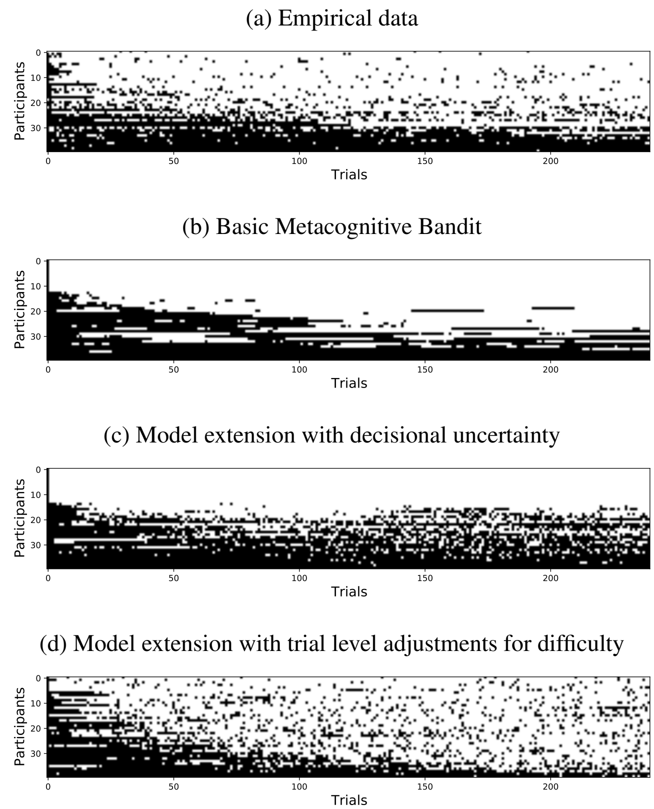

Figure 3: Advice soliciting behavior for actual and simulated participants on 240 trials. Participants are sorted in increasing order of proportion of trials on which advice is solicited. White corresponds to trials where a participant did not solicit AI advice. (a) Empirical data (b) predictions of basic metacognitive bandit model with deterministic arm selection. (c) predictions from model extension with stochastic arm selection (d) predictions from model with adjustments for perceived trial difficulty.

Model Extension: Adding Decisional Uncertainty

The basic metacognitive bandit model suggests that humans engage in perfectly optimal decision-making at each trial. However, this is a strong assumption. In the first extension to our model, we allow for some stochasticity in decision making. We assume that humans employ the metacognitive bandit to choose the arm with the highest utility most frequently, but occasionally deviate from optimal behavior. The softmax action selection function is widely used to model uncertainty in human decision-making and gives us an elegant way to incor-

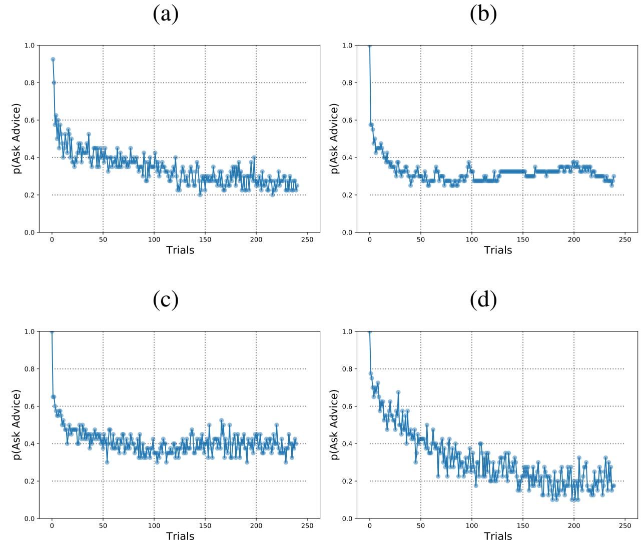

Figure 4: Probability of seeking AI advice across participants on 240 trials (a) Empirical data (b) Predictions of basic metacognitive bandit model with deterministic arm selection. (c) Predictions from model extension with stochastic arm selection (d) Predictions from model with adjustments for perceived trial difficulty

porate stochasticity in our model. The probability of choosing the AI arm according to the softmax function is:

$$p(a_t = 1) = \frac{1}{1 + \exp(-\frac{UCB(\theta_t, \lambda) - (UCB(\phi_t, \lambda) - c)}{\tau})} \quad (3)$$

where τ is temperature that modulates the level of stochasticity; low values of tau corresponding to lower stochasticity in decision making behavior. Figures 3(c) and 4(c) show the predictions from this model where $\tau = .01$. Note that the key difference from the basic metacognitive bandit model is the added noise in the decision-making process where occasionally the arm with a lower UCB bound is chosen.

Model Extension: Adjusting for Subjective Trial Difficulty

We now extend our basic metacognitive bandit to a *subjective trial difficulty adjusted* metacognitive bandit that captures the metacognitive process of inferring overall ability while also accounting for the difficulty of the trials. We posit that humans estimate a probability of being correct on the current stimulus without the AI's aid based on their inferred ability and the *subjective difficulty* of the stimulus at trial t. We use the term 'subjective difficulty' to draw attention to the possibility that an objectively easy trial can be perceived as a difficult trial by a human. This may happen because the human was not paying attention, or because the human doesn't have enough context or prior knowledge about a trial. The human infers a subjective difficulty for each stimulus presented. Let d_t denote the subjective difficulty of the stimulus presented at trial t.

The Rasch framework is a natural choice to incorporate trial-by-trial variation in difficulty while inferring ability. The Rasch model is a classic psychometric model used to analyse participant responses to questions as a function of the participant's latent ability and the question's difficulty. In contrast to the standard Rasch model application, in which ability and difficulty combine to produce performance, we adopt the Rasch model as a component of the metacognitive process underlying the decision to solicit advice. We hypothesize that the human participant employs the Rasch model to assess their own ability and decide when to seek advice based on the current estimate of their ability, the current estimate of the ability of the AI, and the perceived difficulty of the current trial. The probability of being correct without the AI's help as estimated by the human is based on a Rasch model:

$$P(x_t = 1 | \theta, d_t) = \frac{1}{1 + \exp(-(\theta - d_t))} \quad (4)$$

We use the sigmoid function to transform the value $(\theta - d_t)$ to a probability value between 0 and 1. This transformed density of the latent ability serves as the distribution of expected accuracy for the self arm. In this model, the likelihood of observing a sequence of trial outcomes (i.e., runs of successes and failures) is:

$$p(X = x_{1:t-1} | \theta, d_{1:t-1}) = \prod_{j=1}^{t-1} \frac{\exp(x_j(\theta - d_j))}{1 + \exp(\theta - d_j)} \quad (5)$$

We assume that the human participant engages in an inference process about their own overall ability θ. Using Bayes's rule, the posterior over θ is:

$$p(\theta | X = x_{1:t-1}, d_{1:t-1}) \propto p(X_{1:t-1} | \theta, d_{1:t-1}) p(\theta) \quad (6)$$

where we assume the prior $p(\theta) \sim N(\mu, \sigma^2)$. Since calculating the posterior exactly is intractable, we adopt an approximate inference technique to simulate the human's assessment of their own ability. We implement a Hamiltonian Monte Carlo algorithm to draw samples from the posterior of θ. The samples from the posterior are then used in equation 2 to infer the probability of being correct which adjusts for the difficulty of each particular trial.

We assume that the human only adjusts for trial difficulty when inferring their own ability on a particular trial. As a simplifying assumption, we assume the human's inference about the AI's ability is independent of difficulty (as the human does not know what the AI finds difficult). The inference of the AI's ability is the same as the beta update in Equation 1.

After the adjustment for trial difficulty by the human, the probability of choosing the AI arm is again evaluated using the softmax function:

$$p(a_t = 1) = \frac{1}{1 + \exp(-\frac{UCB(\sigma(\theta_t - d_t), \lambda) - (UCB(\phi_t, \lambda) - c)}{\tau})} \quad (7)$$

where σ is the sigmoid function. Note that θ is conditioned on the history of the rewards $x_{1:t-1}$ accumulated by the human and the associated perceived difficulties $d_{1:t-1}$ of the trials, while ϕ is conditioned only on the history of the rewards $y_{1:t-1}$ accumulated by pulling the AI arm.

We simulated the metacognitive bandit adjusted for perceived trial difficulty by conditioning on the same true coherence and reward sequence as in the experimental data. We set λ to .1 such that the UCB action selection is based on the 90th quantile of the latent abilities of the two arms. We set c to .1 to impose a small cost associated with the action of soliciting advice. The experiment didn't ask participants for subjective difficulty on each trial. Instead, we use a noisy transformation of the true coherence of the stimuli used in the experiment to simulate subjective difficulty. Let C_t be the true coherence level at time t. Perceived coherence ω_t is a sample from a normal distribution centered at C_t and standard deviation .2. We then impose an inverse transformation to estimate a subjective difficulty based on the true coherence of a trial, $d_t = k/(\omega_t + \varepsilon)$, where ε is a small value added to the denominator (set to .001 in our simulation) to avoid numerical issues. k is a proportionality constant set to .02. This equation gives us a way to estimate trial-level subjective difficulty for our experiment. This is substituted in equation 5 to calculate the probability of being correct on each trial. Figures 3(d) and 4(d) show the advice seeking trend across the simulated population.

Confidence Ratings

Another important feature of the metacognitive bandit adjusted for subjective item difficulty is that it allows us to generate confidence ratings based on the learner's inferred probability of being correct on each trial. In the data, we observe that participants tend to give lower confidence ratings on trials that they are likely to get wrong. This is another way of saying that metacognition in this task is accurate—participants are able to judge the likelihood of answering correctly on a particular trial. This metacognitive awareness can also be seen in the advice-seeking behavior: participants are more likely to seek advice on trials for which they have low confidence.

We use the estimated perceived coherence of a trial to simulate the response and confidence of the human on that trial. Figure 5 shows the correspondence between the coherence value and the confidence of the human. We expect the human to have high confidence when the absolute value of coherence is high (between .16 and .3) and the direction of movement of the stimuli is highly discernible. We expect the human to have medium confidence when the absolute value of coherence is between .06 and .16 and low confidence when the absolute value of coherence is less than .6. If the human's perceived coherence has the same sign as the true coherence, we predict that the human can correctly guess the dominant direction of movement in the stimulus. Figures 6 (a) and (b) demonstrate that the model is able to capture the qualitative relationship between confidence, accuracy and the probability of seeking AI advice.

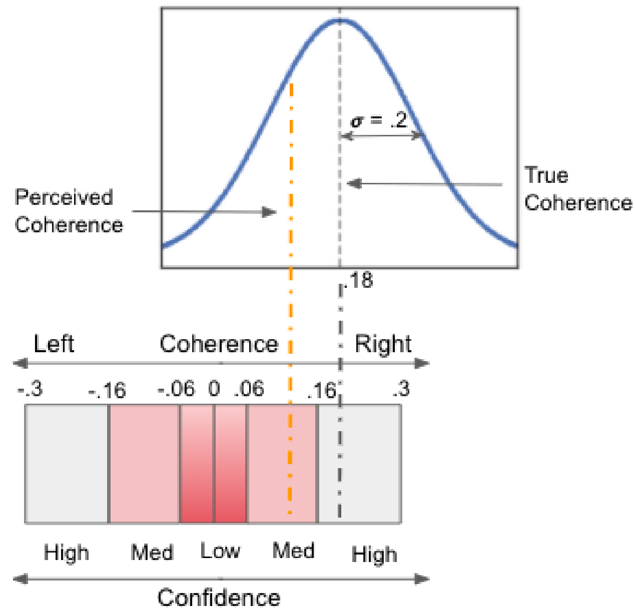

Figure 5: Proposed generative model for human response and confidence: True coherence is sampled from a uniform distribution between $-.3$ and .3. Perceived coherence is a noisy sample from a normal centered at the true coherence and is used to determine the accuracy and confidence of the human on a trial.

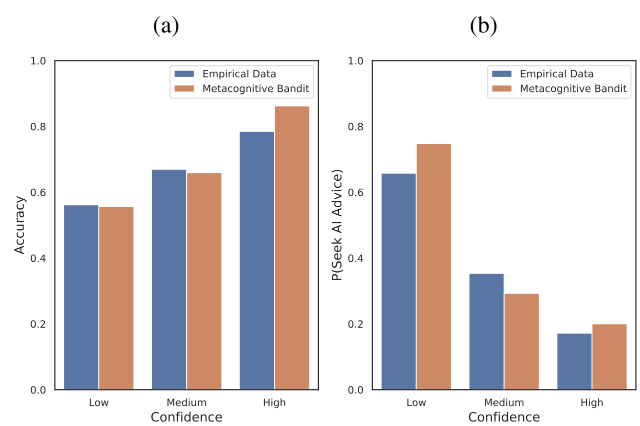

Figure 6: Relationship between the reported confidence of participants in their response and (a) the accuracy of response, and (b) probability of soliciting AI advice.

Discussion

Work on human-AI interaction often describes algorithm aversion as an unwarranted bias displayed by humans who become overly skeptical of an algorithm's ability. Through a series of metacognitive bandit models, we demonstrate that algorithm aversion can arise as a consequence of quasi-optimal decision making by humans that factors in not only an assessment of accuracy differences between the human and the AI, but also temporal factors such as the time and cognitive effort required to process the AI advice. Our model describes how humans update their beliefs about their ability relative to the AI and use it to decide when to seek advice from the AI. However, we look at a very specific behavioral paradigm and use simulated AI advice. An important future direction is to look at more naturalistic decision-making settings while using a real AI in the loop.

We also note that currently our model only qualitatively captures trends in the data. To get a complete picture, we need to do more quantitative model fitting. Our model also does not yet explain how advice is integrated into the decision by the human, but the same framework that provides for advice solicitation can also support a weighting of evidence from multiple sources. Understanding how AI advice factors into human judgment is another direction we plan to pursue. Ultimately, a model of human-agent interaction will be critical for understanding human behavior in hybrid teams and also for understanding how to design agent partners in a way that humans can use most effectively.

References

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, *101*(2), 127–151.

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*(1), 114.

Gino, F., & Moore, D. A. (2007). Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, *20*(1), 21–35.

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 90–99).

Ji, D., Logan IV, R. L., Smyth, P., & Steyvers, M. (2021). Active bayesian assessment for black-box classifiers. In *35th aaai conference on artificial intelligence.*

Kamar, E. (2016). Directions in hybrid intelligence: Complementing ai systems with human intelligence. *IJCAI*, pp. 4070-4073.

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

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103.

Pavlidis, N. G., Tasoulis, D. K., & Hand, D. J. (2008). Simulation studies of multi-armed bandits with covariates. In *Tenth international conference on computer modeling and simulation (uksim 2008)* (pp. 493–498).

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in cognitive science*, *7*(2), 351–367.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179.

Tran-Thanh, L., Stein, S., Rogers, A., & Jennings, N. R. (2014). Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, *214*, 89–111.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature human behaviour*, *2*(12), 915–924.