Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making

Charvi Rastogi\textsuperscript{1,2}, Yunfeng Zhang\textsuperscript{2}, Dennis Wei\textsuperscript{2}, Kush R. Varshney\textsuperscript{2}, Amit Dhurandhar\textsuperscript{2}, Richard Tomsett\textsuperscript{3}

\textsuperscript{1}Carnegie Mellon University, \textsuperscript{2}IBM Research – T. J. Watson Research Center, \textsuperscript{3}IBM Research – Europe

Abstract

Several strands of research have aimed to bridge the gap between artificial intelligence (AI) and human decision-makers in AI-assisted decision-making, where humans are the consumers of AI model predictions and the ultimate decision-makers in high-stakes applications. However, people’s perception and understanding is often distorted by their cognitive biases, like confirmation bias, anchoring bias, availability bias, to name a few. In this work, we use knowledge from the field of cognitive science to account for cognitive biases in the human-AI collaborative decision-making system and mitigate their negative effects. To this end, we mathematically model cognitive biases and provide a general framework through which researchers and practitioners can understand the interplay between cognitive biases and human-AI accuracy. We then focus on anchoring bias, a bias commonly witnessed in human-AI partnerships. We devise a cognitive science-driven, time-based approach to de-anchoring. A user experiment shows the effectiveness of this approach in human-AI collaborative decision-making. Using the results from this first experiment, we design a time allocation strategy for a resource constrained setting so as to achieve optimal human-AI collaboration under some assumptions. A second user study shows that our time allocation strategy can effectively debias the human when the AI model has low confidence and is incorrect.

1 Introduction

It is a truth universally acknowledged that a human decision-maker in possession of an AI model must be in want of a collaborative partnership. Recently, we have seen a rapid increase in deployment of machine learning models (AI) in decision-making systems, where the AI models serve as helpers to human experts in many high-stakes settings. Examples of such tasks can be found in healthcare, financial loans, criminal justice, job recruiting and fraud monitoring. Specifically, judges use risk assessments to determine criminal sentences, banks use models to manage credit risk, and doctors use image-based ML predictions for diagnosis, to list a few.

The emergence of AI-assisted decision-making in society has raised questions about whether and when we can trust the AI model’s decisions. This has spurred research in crucial directions on various aspects of the human-AI interaction, such as in interpretable, explainable and trustworthy machine learning. One of the main objectives of this research is to bridge the "communication gap" between humans and AI. Research in interpretability and explainability aims to develop AI systems that can effectively communicate the reasoning and justification for its decisions to humans. Research in trustworthy ML and trust calibration in collaborative decision-making aims to appropriately enhance humans’ trust and thereby their receptiveness to communication from AI systems. However, a key component of human-AI communication that is often sidelined is the human decision-makers themselves. Humans’ perception of the communication received from AI is at the core of this communication gap. Research in communication exemplifies the need to model receiver characteristics, thus implying the need to understand and model human cognition in collaborative decision-making.

As a step towards studying human cognition in AI-assisted decision-making, this paper addresses the role of cognitive biases in this setting. Cognitive biases, introduced in the seminal work by Kahneman and Tversky \cite{Tversky1974}, represent a systematic pattern of deviation from rationality in judgment wherein individuals create their own "subjective reality" from their perception of the input. An individual’s perception of reality, not the objective input, may dictate their behavior in the world, thus, leading to distorted and inaccurate judgment.

Our first contribution is to propose a “biased Bayesian” framework to reason about cognitive biases in AI-assisted human decision-making. Viewing a rational decision-maker as being Bayes optimal, we discuss how different
cognitive biases can be seen as modifying different factors in Bayes’ theorem. We illustrate this in Figure 1 based on ideas in (Yeom and Tschantz, 2018). In a collaborative decision-making setting, there are two interactions that may lead to cognitive biases in the perceived space, which represents the human decision-maker. The observed space consists of the feature space and all the information the decision-maker has acquired about the task. The prediction space represents the output generated by the AI model, which could consist of the AI decision, explanation, etc. Through Figure 1 we associate cognitive biases that affect the decision-makers’ priors and their perception of the data available with their interaction with the observed space. Confirmation bias, availability bias, the representativeness heuristic, and bias due to selective accessibility of the feature space are mapped to the observed space. On the other hand, anchoring bias and the weak evidence effect are mapped to the prediction space. In this paper, we provide a model for some of these biases using our biased Bayesian framework.

To focus our work, in the remainder of the paper we study anchoring bias. Anchoring bias is often present in AI-assisted decision making where the decision-maker forms a skewed perception due to an anchor (AI decision) presented to them, which limits the exploration of alternative hypotheses. The anchoring-and-adjustment heuristic, studied widely, suggests that when presented with an anchor, people adjust away from the anchor insufficiently. Building on the notion of bounded rationality, previous work (Lieder et al., 2018) attributes the adjustment strategy to a resource-rational policy wherein the insufficient adjustment is a rational trade-off between accuracy and time. Inspired by this, we conduct an experiment with human participants to study whether allocating more resources — in this case, time — can alleviate anchoring bias. This bias manifests through the rate of agreement with the AI prediction. Thus, by measuring the rate of agreement with the AI prediction on several different carefully designed trials, we are able to show that time indeed is a significant resource that helps the decision-maker sufficiently adjust away from the anchor when needed.

Given the findings of our first experiment, we formulate a resource allocation problem that accounts for anchoring bias to maximize human-AI collaborative accuracy. The time allocation problem that we study is motivated by real-world AI-assisted decision-making settings. Adaptively determining the time allocated to a particular instance for the best possible judgement can be very useful in multiple applications. For example, consider a (procurement) fraud monitoring system deployed in multinational corporations, which analyzes and flags high-risk invoices (Dhurandhar et al., 2015). Given the scale of these systems, which typically analyze tens of thousands of invoices from many different geographies daily, the number of invoices that may be flagged, even if a small fraction, can easily overwhelm the team of experts validating them. In such scenarios, spending a lot of time on each invoice is not admissible. An adaptive scheme which takes into account the biases of the human and the expected accuracy of the AI model would be highly desirable to produce the most objective decisions. This work would also be useful in the other aforementioned domains, such as in criminal proceedings where judges often have to look over many different case documents and make quick decisions.

As a solution to the time allocation problem above, we propose a simple allocation policy based on AI confidence and identify assumptions under which the policy is theoretically optimal. We conduct a second experiment to evaluate this policy in comparison to baseline policies as well as human-only and AI-only performance. Our results show that while the overall performance of all policies considered is roughly the same, our policy helps the participants adjust away from the AI prediction more when the AI is incorrect and has low confidence.

In summary, we make the following contributions:

• We model cognitive biases in AI-assisted decision-making with a biased Bayesian framework and situate well-known cognitive biases within it, based on their sources.
• Focusing on anchoring bias, we show with human participants that allocating more time to a decision reduces anchoring.
• We formulate a time allocation problem to account for the anchoring-and-adjustment heuristic and maximize human-AI accuracy.
• We propose a confidence-based allocation policy and identify conditions under which it achieves optimal team performance.
• We evaluate the real-world effectiveness of the policy in a second human subject experiment, showing that it helps humans de-anchor from low-confidence and incorrect AI predictions.

2 Related work

Cognitive biases are an important tenet of human decision-making (Barnes JR., 1984; Das and Teng, 1999; Ehrlinger et al., 2016), and have been studied widely in decision-support systems research (Arnott, 2006; Zhang et al., 2015; Solomon, 2014; Phillips-Wren et al., 2019). Cognitive biases also show up in many aspects of collaborative behaviours (Silverman, 1992; Janssen and Kirschner, 2020; Bromme et al., 2010). More specifically, there exists decades old research on cognitive biases (Tversky and Kahneman, 1974) such as confirmation bias (Nickerson, 1998; Klaman, 1995; Oswald and Grosjean, 2004), anchoring bias (Furnham and Boo, 2011; Epley and Gilovich, 2000; 2001), availability bias (Tversky and Kahneman, 1973), etc.

Recently, as AI systems are increasingly embedded into high stakes human decisions, understanding human behavior and reliance on technology has become critical. "Poor partnerships between people and automation will become increasingly costly and catastrophic" (Lee and See, 2004). This concern has sparked crucial research in several directions, such as human trust in algorithmic systems, interpretability and explainability of machine learning models (Arnold et al., 2019; Zhang et al., 2020; Tomsett et al., 2020; Siau and Wang, 2018; Doshi-Velez and Kim, 2017; Lipton, 2018; Adadi and Berrada, 2018; Preece, 2018).

In parallel, significant research in human-AI partnership has considered how to directly optimize the team performance in AI-assisted decision-making [Lai et al., 2020; Lai and Tan, 2019; Bansal et al., 2020, 2019b; Green and Chen, 2019; Okamura and Yamada, 2020]. The perception of the human agents plays an important role in collaborative decision-making [Yeung et al., 2020; Bansal et al., 2019a; Lee, 2018]. These works experiment with several heuristic-driven explanation strategies that only partially take into account the characteristics of the human at the end of the decision-making pipeline. Citing a body of research in psychology, philosophy and cognitive science, Miller (Miller, 2019) argues that the the machine learning community should move away from imprecise, subjective notions of "good" explanations and instead focus on reasons and thought processes that people apply for explanation selection. In the same vein, our work builds on literature in psychology on cognitive biases to inform modeling and effective debiasing strategies. Specifically, (Bansal et al., 2020), in part, provides explanation strategies to address the problem of blind agreement in collaborative decision-making. Our work provides a structured approach to addressing such problems from a cognitive science perspective.

A core component of fruitful collaboration is effective evaluation of collaborators’ accuracy. Research has shown that people do not calibrate their reliance on the AI based on its accuracy (Green and Chen, 2019). Several studies suggest that people are unable to detect algorithmic errors (Poursabzi-Sangdeh et al., 2018), are biased by irrelevant information (Englich et al., 2006), rely on algorithms that are described as having low accuracy, and trust algorithms that are described as accurate but actually present random information (Springer et al., 2018). These behavioural tendencies motivate a crucial research question — how to account for these systemic deviations from rationality, often explained by cognitive biases?

Work on cognitive biases in human-AI interaction is still rare however. Recently, (Fürnkranz et al., 2020) evaluated a selection of cognitive biases in the very specific context of whether minimizing the complexity or length of a rule will also lead to increased interpretability of machine learning models. Kleiger et al. (Kliegr et al., 2018) review twenty different cognitive biases that can distort interpretation of inductively learned rules in ML models. This work analyses effect of cognitive biases on human understanding of symbolic ML models and associated debiasing techniques. (Baudel et al., 2020) addresses complacency/authority bias induced by algorithmic decision aids introduced in business decision processes. Wang et al. (Wang et al., 2019) propose a conceptual framework for building explainable AI based on literature on cognitive biases. Building on these works, our work takes the step of identifying the role of cognitive biases in AI-assisted decision-making in a theoretically-driven and principled manner, through mathematical modeling. In addition, accounting for subjectivity in human behavior, we inform our approach in a real-world setting based on careful preliminary experiments, thus, paving the pathway for human-oriented research in this setting.

3 Problem setup and modeling

We consider a collaborative decision-making setup, consisting of the machine learning algorithm and the human decision-maker. We first precisely describe our setup and document the associated notation. Following this, we provide a general model for various human cognitive biases induced by this collaborative process.

Our focus in this paper is on the AI-assisted decision-making setup, wherein the objective of the human is to correctly classify the set of feature information available into one of two categories. Thus, we have a
binary classification problem, where the true class is denoted by $y^* \in \{0, 1\}$. To make the decision/prediction, the human is presented with feature information, and we denote the complete set of features available pertaining to each sample by $D$. In addition to the feature information, the human is also shown the output of the machine learning algorithm. Here, the AI output could consist of several parts, such as the machine-generated explanation for its prediction. We express the complete AI output as a function of the machine learning model, denoted by $f(M)$. Finally, we denote the decision made by the human decision-maker by $\tilde{y} \in \{0, 1\}$.

We now describe the approach towards modeling the behavior of human decision-makers when assisted by machine learning algorithms.

### 3.1 Bayesian decision-making

Bayesian models have become increasingly prominent across a broad spectrum of cognitive science (Griffiths and Tenenbaum, 1999; Griffiths and Tenenbaum, 2006; Chater et al., 2006). The Bayesian approach is thoroughly embedded within the framework of decision theory. Its basic tenets are that opinions should be expressed in terms of subjective or personal probabilities, and that the optimal revision of such opinions, in the light of relevant new information, should be accomplished via Bayes’ theorem.

First, consider a simpler setting, where the decision-maker uses the feature information available, $D$, and makes a decision $\tilde{y} \in \{0, 1\}$. Let the decision variable be denoted by $\tilde{Y}$, such that $\tilde{y} = \arg\max_{i \in \{0, 1\}} P(\tilde{Y} = i)$. Based on literature in psychology and cognitive science (Griffiths and Tenenbaum, 2006; Chater et al., 2006), we model a rational decision-maker as Bayes’ optimal. That is, given a prior on the likelihood of the prediction, $P_{pr}(\tilde{Y})$ and the data likelihood distribution $P(D|\tilde{Y})$, then the decision-maker picks the hypothesis/class with the higher posterior probability. Formally, the Bayes’ theorem states that

$$P(\tilde{Y} = i|D) = \frac{P(D|\tilde{Y} = i)P_{pr}(\tilde{Y} = i)}{\sum_{j \in \{0, 1\}} P(D|\tilde{Y} = j)P_{pr}(\tilde{Y} = j)}. \quad (1)$$

where $i \in \{0, 1\}$ and the human decision is given by $\tilde{y} = \arg\max_{i \in \{0, 1\}} P(\tilde{Y} = i)$. Now, in our setting, in addition to the feature information available, the decision-maker takes into account the output of the machine learning algorithm, $f(M)$, which leads to following Bayes’ relation

$$P(\tilde{Y}|D, f(M)) \propto P(D|\tilde{Y})P(f(M)|\tilde{Y})P_{pr}(\tilde{Y}). \quad (2)$$

We assume that the decision-maker perceives the feature information and the AI output independently, which gives

$$P(\tilde{Y}|D, f(M)) \propto P(D|\tilde{Y})P(f(M)|\tilde{Y})P_{pr}(\tilde{Y}), \quad (3)$$

where $P(f(M)|\tilde{Y})$ indicates the conditional probability of the AI output perceived by the decision-maker. This concludes our model for a rational decision-maker assisted by a machine learning model. In reality, the human decision-maker may behave differently from a fully rational agent due to their cognitive biases.

In some studies (Payzan-Le Néstor and Bossaerts 2011, 2012), such deviations have been explained by introducing exponential biases (i.e. inverse temperature parameters), on Bayesian inference because these were found useful in expressing bias levels. We augment the modeling approach in (Matsumori et al., 2018) to a human-AI collaborative setup. Therein we model the biased Bayesian estimation as

$$P(\tilde{Y}|D, f(M)) \propto P(D|\tilde{Y})^\alpha P(f(M)|\tilde{Y})^\beta P_{pr}(\tilde{Y})^\gamma, \quad (4)$$

where $\alpha, \beta, \gamma$ are variables that represent the deviation in the constituent factors of Bayes’ optimal decision-making. This allows us to understand and model several cognitive biases.

1. In anchoring bias, the weight put on AI prediction is high, $\beta > 1$, whereas the weight on prior and data likelihood reduces.
2. By contrast, in confirmation bias the weight on the prior is high, $\gamma > 1$, and the weight on the data and machine prediction reduces in comparison.
3. Selective accessibility is a phenomena used to explain the mechanism of cognitive biases, wherein only the data that supports the cognitive bias of the decision-maker is used by them. This distorts the data likelihood factor in $\beta$. The direction of distortion $\alpha > 1$ or $\alpha < -1$ depends on the source of cognitive bias.
4. The weak evidence effect (Fernbach et al., 2011), causes the decision-maker to disagree with the AI prediction, when the AI provides weak evidence to support its prediction. This effect can be modeled with $\beta < -1$.

To focus our approach, we consider a particular cognitive bias — anchoring bias, which is specific to the nature of human-AI collaboration and has been an issue in previous works (Lai and Tan, 2019; Springer et al., 2018; Bansal et al., 2020). In the next section, we summarise the findings about anchoring bias in the literature, explain proposed debiasing technique and conduct an experiment to validate the technique.

### 4 Anchoring bias

Anchoring bias is often present in AI-assisted decision-making tasks, where the human is anchored to the AI-
generated decision more than necessary. The anchoring-and-adjustment heuristic, introduced by Kahneman and Tversky in [Kahneman and Tversky, 1974] and studied in [Epley and Gilovich, 2006; Lieder et al., 2018] suggests that after being anchored, humans tend to adjust insufficiently because adjustments are effortful and tend to stop once a plausible estimate is reached. Lieder et al. [Lieder et al., 2018] proposed the resource rational model of anchoring-and-adjustment which explains that the insufficient adjustment can be understood as a rational trade-off between time and accuracy. This is a consequence of the bounded rationality of humans [Simon, 1956, 1972], which entails satisficing, that is, accepting sub-optimal solutions that are good enough, rather than optimizing solely for accuracy. Through user studies, Epley et al. [Epley and Gilovich, 2006] argue that cognitive load and time pressure are contributing factors behind insufficient adjustments.

Informed by the above studies viewing anchoring bias as a problem of insufficient adjustment due to limited resources, our aim is to mitigate the effect of anchoring bias in AI-assisted decision-making, using time as a resource. We use the term de-anchoring to denote the rational process of adjusting away from the anchor. With this goal in mind, we conducted two user studies on Amazon Mechanical Turk. Through the first user study (Experiment 1), our aim is to understand the effect of different time allocations on anchoring bias and de-anchoring in an AI-assisted decision-making task. In Experiment 2, we use the knowledge obtained about the effect of time in Experiment 1 to design a time allocation strategy and test it on the experiment participants.

We now describe Experiment 1 in detail.

4.1 Experiment 1

In this study, we asked the participants to complete an AI-assisted binary prediction task consisting of a number of trials. Our aim is to learn the effect of allocating different amounts of time to different trials on participants with anchoring bias.

To quantify anchoring bias and thereby the insufficiency of adjustments, we use the probability $\mathbb{P}(\hat{y} = \tilde{y})$ that the human decision-maker agrees with the prediction $\hat{y}$, which is easily measured. This measure can be motivated from the biased Bayesian model in (4). In the experiments, the model output $f(M)$ consists of only a predicted label $\hat{y}$. In this case, (4) becomes

$$
\mathbb{P}(\hat{Y} = y | D, f(M)) \propto \mathbb{P}(D | \hat{Y} = y)^{\alpha} \mathbb{P}(\hat{Y} = \tilde{y} | \hat{Y} = y)^{\beta} \mathbb{P}_{pr}(\tilde{Y} = y)^{\gamma}.
$$

(5)

Conditioned on $y = 0$, the middle quantity $\mathbb{P}(\hat{Y} = \tilde{y} | \hat{Y} = y)$ is the true negative/false positive rate (TNR/FPR) of the ML model for $\hat{y} = 0, 1$. Similarly, it is the true positive/false negative rate (TPR/FNR) for $y = 1$ and $\hat{y} = 1, 0$. Let us make the reasonable assumption that the ML model satisfies TNR > FNR and TPR > FPR, i.e., that $\mathbb{P}(\tilde{Y} = \tilde{y} | \hat{Y} = y)$ is larger when $y = \tilde{y}$ than when $y \neq \tilde{y}$. Then as the exponent $\beta$ increases, i.e., as anchoring bias strengthens, the likelihood that $y = \tilde{y}$ maximizes (5) and becomes the human decision $\hat{y}$ also increases. In the limit $\beta \to \infty$, we have agreement $\hat{y} = \tilde{y}$ with probability 1. Conversely, for $\beta = 1$, the two other factors in (5) are weighed appropriately and the probability of agreement assumes a natural baseline value. We conclude that the probability of agreement is a measure of anchoring bias. It is also important to ensure that this measure is based on tasks where the human has reason to choose a different prediction.

Thus, given the above relationship between anchoring bias and agreement probability (equivalently disagreement probability), we tested the following hypothesis to determine whether time is a useful resource in mitigating anchoring bias:

- **Hypothesis 1 (H1):** Increasing the time allocated to a task alleviates anchoring bias, yielding a higher likelihood of sufficient adjustment away from the AI-generated decision when the decision-maker has the knowledge required to provide a different prediction.

Participants. We recruited 47 participants from Amazon Mechanical Turk for Experiment 1, limiting the pool to subjects from within the United States with a prior task approval rating of at least 98% and a minimum of 100 approved tasks. 10 participants were between Age 18 and 29, 26 between Age 30 and 39, 6 between Age 40 and 49, and 5 over Age 50. The average completion time for this user study was 27 minutes, and each participant received a compensation of $4.5 (roughly equals an hourly wage of $10). The participants received a base pay of $3.5 and bonus of $1 (to incentivize accuracy).

Task and AI model. We designed a performance prediction task wherein a participant was asked to predict whether a student would pass or fail a class, based on the student’s characteristics, past performance and some demographic information. The dataset for this task was obtained from the UCI Machine Learning Repository, published as the Student Performance Dataset (Cortez and Silva, 2008). This dataset contains 1044 instances of students’ class performances in 2 subjects (Mathematics and Portuguese), each described by 33 features. To prepare the dataset, we binarized the target labels (‘pass’, ‘fail’), split the dataset into training and test sets (70/30 split). To create our AI, we trained a logistic regression model on the standardized set of features from the train-
ing dataset. Based on the feature importance (logistic regression coefficients) assigned to each feature in the dataset, we retained the top 10 features for the experiments. These included — mother’s and father’s education, mother’s and father’s jobs, hours spent studying weekly, interest in higher education, hours spent going out with friends weekly, number of absences in the school year, enrolment in extra educational support and number of past failures in the class.

Study procedure Since we are interested in studying decision makers’ behavior when humans have prior knowledge and experience in the prediction task, we first trained our participants before collecting their decision data for analysis. The training section consists of 15 trials where the participant is first asked to provide their prediction based on the student data and is then shown the correct answer after attempting the task. These trials are the same for all participants and are sampled from the training set such that the predicted probability (of the predicted class) estimated by the AI model is distributed uniformly over the intervals $[0.5, 0.6], [0.6, 0.7], \ldots, (0.9, 1]$. Taking predicted probability as a proxy for difficulty, this ensures that all levels of difficulty are represented in the task. To help accelerate participants’ learning, we showed barcharts that display the distributions of the outcome across the feature values of each feature. These barcharts were not provided in the testing section to ensure stable performance throughout the testing section and to emulate a real-world setting.

To induce anchoring bias, the participant was informed at the start of the training section that the AI model was 85% accurate (we carefully chose the training trials to ensure that the AI was indeed 85% accurate over these trials), while the model’s actual accuracy is 70.8% over the entire training set, and 66.5% over the test set. Since our goal is to induce anchoring bias and the training time is short, we stated a high AI accuracy. Moreover, this disparity between stated accuracy (85%) and true accuracy (70.8%) is realistic if there is a distribution shift between the training and the test set, which would imply that the humans’ trust in AI is misplaced. In addition to stating AI accuracy at the beginning, we informed the participants about the AI prediction for each training trial after they have attempted it, so that they can learn about AI’s performance first-hand.

The training section is followed by the testing section which consists of 36 trials sampled from the test set, and were kept the same (in the same order) for all participants. In this section, the participants were asked to make a decision based on both the student data and the AI prediction. They were also asked to describe their confidence level in their prediction as low, medium, or high.

To measure the de-anchoring effect of time, we included some trials where the AI is incorrect but the participants have the requisite knowledge to adjust away from the incorrect answer. That is, we included trials where the participants’ accuracy would be lower when they are anchored to the AI prediction than when they are not. We call these trials — probe trials, which help us probe the effect of time on de-anchoring. On the flip side, we could not include too many of these trials because participants may lose their trust in the AI if exposed to many apparently incorrect AI decisions. To achieve this balance, we sampled 8 trials of medium difficulty where the AI prediction is accurate (predicted probability ranging from 0.6 to 0.8) and flip the AI prediction shown to the participants. The remaining trials, termed unmodified trials are sampled randomly from the test set while maintaining a uniform distribution over the AI predicted probability (of the predicted class). Here, again, we use the predicted probability as a proxy for the difficulty of the task, as evaluated by the machine learning model. We note that the accuracy of the AI predictions shown to the participants is 58.3% which is far lower than the 85% accuracy shown in the training section.

Time allocation. To investigate the effect of time on the anchoring-and-adjustment heuristic in AI-assisted decision-making, we divide the testing section into four blocks for each participant based on the time allocation per trial. To select the allocated time intervals, we first conducted a shorter version of the same study to learn the amount of time needed to solve the student performance prediction task, which suggested that the time intervals of 10s, 15s, 20s and 25s captured the range from necessary to sufficient. Now, with these four time intervals, we divided the testing section into four blocks of 9 trials each, where the time allocated in each block followed the sequence $[t_1, t_2, t_3, t_4]$ and for each participant this sequence was a random permutation of $[10, 15, 20, 25]$. The participants were not allowed to move to the next trial till the allocated time ran out. Furthermore, each block was comprised of 2 probe trials and 7 unmodified trials, randomly ordered. Recall that each participant was provided the same set of trials in the same order.

Now, with the controlled randomization of the time allocation, independent of the participant, their performance and the sequence of the tasks, we are able to identify the effect of time on de-anchoring. It is possible that a participant that disagrees with the AI prediction often in the first half, is not anchored to the AI prediction in the latter half. Our study design allows us to average out such participant-specific effects, through the randomization of time allocation interval sequences across participants.
4.2 Results

The main results of Experiment 1 are illustrated in Figure 2. We see that the probe trials served their intended purpose, since the average disagreement is much higher for probe trials compared to unmodified trials for all time allocations. This suggests that the participants had learned to make accurate predictions for this task, otherwise they would not be able to detect the AI’s errors in the probe trials, more so in the 10-second condition. We also observe that the likelihood of disagreement for unmodified trials is low (close to 0.1) for all time allocations. This suggests that the participants’ knowledge level in this task is roughly similar to or less than that of the AI since the participants are unable to offer any extra knowledge in the unmodified trials.

Anchoring-and-adjustment. The results on the probe trials in Figure 2 suggest that the participants’ likelihood of sufficiently adjusting away from the incorrect AI prediction increased as the time allocated increased. This strengthens the argument that the anchoring-and-adjustment heuristic is a resource-rational trade-off between time and accuracy (Lieder et al., 2018). Specifically, we observe that the average disagreement percentage in probe trials increased from 48% in the 10-second condition to 67% in the 25-second condition. We used the bootstrap method with 5000 re-samples to estimate the coefficient of a linear regression fit on average disagreement vs. time allocated for probe trials. This resulted in a significantly positive coefficient of 0.01 (bootstrap 95% confidence interval [0.001, 0.018]). This result is consistent with our Hypothesis 1 (H1) that increasing time for decision tasks alleviates anchoring bias. We note that the coefficient is small in value because the scales of the independent and dependent variables of the regression (time and average disagreement) have not been adjusted for the regression, so the coefficient of 0.01 yields a 0.15 increase in average disagreement between the 10s and the 25s time condition.

Time adherence. Figure 3 suggests that the participants adhere reasonably to the four different time conditions used. We note that this time reflects the maximum time taken to click the radio button (in case of multiple clicks), but the participants may have spent more time thinking over their decision. In the survey at the end of the study, we asked the participants how often they used the entire time available to them in the trials, and obtained the following distribution of answers — Frequently 15, Occasionally 24, Rarely 6, Never 2.

5 Optimal resource allocation in human-AI collaboration

In Section 4, we see that time is a useful resource for de-anchoring the decision-maker. More generally, there are many works that study debiasing techniques to address the negative effects of cognitive biases. These debiasing techniques require resources such as time, computation and explanation strategies. Thus, in this section, we model the problem of mitigating the effect of cognitive biases in the AI-assisted decision-making setting as a resource allocation problem, where our aim is to efficiently use the resources available and improve human-AI collaboration accuracy.

5.1 Resource allocation problem

From Experiment 1 in Section 4.1, we learnt that given more time, the decision-maker is more likely to adjust
away from the anchor (AI decision) if the decision-maker has reason to believe that the correct answer is different. This is shown by change in their probability of agreement with the AI prediction, denoted by $P_a = P(\hat{y} = \hat{y})$. The results of Experiment 1 indicate that, ideally, decision makers should be provided with ample time for each decision. However, in practice, time is a limited resource that we want to use efficiently. Thus, we formulate a resource allocation problem that captures the trade-off between time and accuracy. More generally, this problem suggests a framework for optimizing human-AI team performance using constrained resources to debias the decision-maker.

In our setup, the human decision-maker has to provide a final decision $\hat{y}_i$ for $N$ total trials with AI assistance, specifically in the form of a predicted label $\hat{y}_i$. The objective is to maximize the average accuracy over the trials, denoted by $R$, of the human-AI collaboration:

$$\mathbb{E}[R] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[R_i],$$

where $R_i$ is an indicator of human-AI correctness in trial $i$.

We first relate collaborative accuracy $\mathbb{E}[R]$ to the anchoring-and-adjustment heuristic. Intuitively, if we know the AI to be incorrect in a given trial, we should aim to facilitate adjustment away from the anchor as much as possible, whereas if AI is known to be correct, then anchoring bias is actually beneficial. Based on this intuition, $\mathbb{E}[R_i]$ in (6) can be rewritten by conditioning on AI correctness/incorrectness as follows:

$$\mathbb{E}[R_i] = \mathbb{P}(\hat{y}_i = y_i | \hat{y}_i = y_i^*) \mathbb{P}(\hat{y}_i = y_i^*$$

$$+ (1 - \mathbb{P}(\hat{y}_i = y_i | \hat{y}_i = y_i^*)) (1 - \mathbb{P}(\hat{y}_i = y_i^*)).$$

We see therefore that human-AI correctness depends on the probability of agreement $P_{a_i}$ conditioned on AI being correct and the probability of agreement $P_{a_i}$ conditioned on AI being incorrect. Recalling from Section 4 the link established between agreement probability and anchoring bias, (7) shows the effect of anchoring bias on human-AI accuracy. Specifically in the case of (7), the effect is through the two conditional agreement probabilities $P_{a_i}$ and $P_{a_i}$.

We consider time allocation strategies to modify agreement probabilities and thus improve collaborative accuracy, based on the relationship established in Experiment 1.

We denote the time used in trial $i$ as $T_i$, which impacts correctness $R_i$ (7) as follows:

$$\mathbb{E}[R_i | T_i] = P_{a_i}(T_i) \mathbb{P}(\hat{y}_i = y_i^*) + P_{a_i}(T_i) (1 - \mathbb{P}(\hat{y}_i = y_i^*)).$$

The allocation of time affects only the human decision-maker, making the agreement probabilities functions of $T_i$, whereas the probability of AI correctness is unaffected. Given a finite resource budget $T$, we also have the constraint $\sum_{i=1}^{N} T_i = T$. The resource allocation problem would then be to maximize the average of (8) over trials (as in (6)) subject to the budget constraint.

The challenge with formulation (8) is that it requires identifying the true probability of AI correctness, which is a non-trivial task (Guo et al., 2017). Instead, we operate under the more realistic assumption that the AI model can estimate its probability of correctness from the class probabilities that it predicts (as provided for example by a logistic regression model). We refer to this estimate as AI confidence and denote it as $\hat{C}_i$. We may then consider a decomposition of human-AI correctness as in (7), (8) but conditioned on $\hat{C}_i$. In keeping with the two cases in (7), (8) and to simplify the allocation strategy, we binarize $\hat{C}_i$ into two intervals, low confidence $\hat{C}_i \in \hat{C}_L$, and high confidence $\hat{C}_i \in \hat{C}_H$. The time allocated is then $T_i(\hat{C}_i) = t_L$ for $\hat{C}_i \in \hat{C}_L$ and $T_i(\hat{C}_i) = t_H$ for $\hat{C}_i \in \hat{C}_H$. Thus we have

$$\mathbb{E}[R_i] = \mathbb{P}(\hat{C}_i \in \hat{C}_L) \mathbb{E}[R_i | \hat{C}_i \in \hat{C}_L, T_i = t_L]$$

$$+ \mathbb{P}(\hat{C}_i \in \hat{C}_H) \mathbb{E}[R_i | \hat{C}_i \in \hat{C}_H, T_i = t_H].$$

The quantities $\mathbb{E}[R_i | \hat{C}_i \in C, T_i], C = \hat{C}_L, \hat{C}_H$, are not pure agreement probabilities as in (8) because the low/high-
confidence events $\hat{C}_i \in C_L$, $\hat{C}_i \in C_H$ generally differ from the correctness/incorrectness events $\hat{y}_i = y^*_i$, $\hat{y}_i \neq y^*_i$. Nevertheless, since we expect these events to be correlated, $\mathbb{E}[R_i | \hat{C}_i \in C, T_i]$ is related to the agreement probabilities in $\bar{\mathbb{E}}$.

Figure 4 presents an ideal scenario that one hopes to attain in $(9)$. In presence of anchoring bias, our aim is to achieve the human-AI team accuracy shown. This approach capitalises on human expertise where AI accuracy is low. Specifically, by giving human decision makers more time, we encourage them to rely on their own knowledge (de-anchor from the AI prediction) when the AI is less confident, $\hat{C}_i \in C_L$. Usage of more time in low AI confidence tasks, implies less time in tasks where AI is more confident, $\hat{C}_i \in C_H$, where anchoring bias has lower negative effects and is even beneficial. Thus, this two-level AI confidence based time allocation policy allows us to mitigate the negative effects of anchoring bias and achieve the “best of both worlds”, as illustrated in Figure 4.

We now identify an assumption under which the optimal time allocation policy is straightforward to see. **Assumption 1.** For any $t_1, t_2 \in \mathbb{R}^+$, if $t_1 < t_2$, then

$$\mathbb{E}[R_i | \hat{C}_i \in C_L, T_i = t_1] \leq \mathbb{E}[R_i | \hat{C}_i \in C_L, T_i = t_2],$$

and

$$\mathbb{E}[R_i | \hat{C}_i \in C_H, T_i = t_1] \geq \mathbb{E}[R_i | \hat{C}_i \in C_H, T_i = t_2].$$

We refer the reader to Appendix A for a proof showing exactly where the confidence-based policy in Figure 4 is most advantageous. At $t = 0$, $\mathbb{E}[R_i | \hat{C}_i \in C, T_i = 0]$ is equal to AI accuracy conditioned on $C = C_L, C_H$, and by giving the human more time to de-anchor, we might expect $\mathbb{E}[R_i | \hat{C}_i \in C_L, T_i = t]$ to increase and $\mathbb{E}[R_i | \hat{C}_i \in C_H, T_i = t]$ to decrease. A second way to understand Assumption 1 is to break down the conditional accuracy into two parts:

$$\mathbb{E}[R_i | \hat{C}_i \in C, T_i = t] =$$

$$\mathbb{P}(\hat{y}_i = \hat{y}_i, \hat{C}_i \in C, T_i = t)\mathbb{P}(\hat{y}_i = y^*_i | \hat{C}_i \in C) + \mathbb{P}(\hat{y}_i \neq y^*_i | \hat{C}_i \in C, T_i = t)\mathbb{P}(\hat{y}_i \neq y^*_i | \hat{C}_i \in C).$$

for $C = C_L, C_H$. The results of Section 4.1 indicate that the disagreement probability in the second RHS term in $(11)$ increases or stays the same with time $t$, and the agreement probability in the first term decreases or stays the same with $t$. For $C = C_L$, assuming positive correlation between low confidence $\hat{C}_i \in C_L$ and low accuracy (not necessarily the perfect correlation in Figure 4), the probability $\mathbb{P}(\hat{y}_i \neq y^*_i | \hat{C}_i \in C_L)$ tends to be larger and the second term dominates, resulting in the LHS increasing with $t$. For $C = C_H$, the first term tends to dominate, leading to decrease with $t$.

Under Assumption 1, the optimal strategy is to maximize time for low-AI-confidence trials and minimize time for high-confidence trials, as is stated formally below.

**Proposition 1.** Consider the AI-assisted decision-making setup discussed in this work with $N$ trials where the total time available is $T$. Suppose Assumption 1, stated in $(10)$, holds true for human-AI accuracy. Then, the optimal confidence-based allocation is as follows,

$$T_i = \left\{ \begin{array}{ll} t_H = t_{\min} & \text{if } \hat{C}_i \in C_H \\ t_L = t_{\max} & \text{if } \hat{C}_i \in C_L, \end{array} \right.$$

where $t_{\min}$ is the minimum allowable time, and $t_{\max}$ is the corresponding maximum time such that $t_{\min} \mathbb{P}(\hat{C}_i \in C_L) + t_{\max} \mathbb{P}(\hat{C}_i \in C_H) = T$.

Proposition 1 gives us the optimal time allocation strategy by efficiently allocating more time for adjusting away from the anchor in the tasks that yield a lower probability of accuracy if the human is anchored to the AI predictions. We note that, although in an ideal scenario as shown in Figure 4, we should set $t_{\min} = 0$, in real world implementation $\hat{C}_i$ is an approximation of the true confidence, and hence, it would be helpful to have human oversight with $t_{\min} > 0$ in case the AI confidence is poorly calibrated.

To further elaborate on the optimality of the confidence-based time allocation strategy, we compare it with two baseline strategies that obey the same resource constraint, namely, Constant time and Random time strategies, defined as

- **Constant time**: For all $i$, $T_i = \frac{T}{N}$.
- **Random time**: Out of the $N$ trials, $N \times \mathbb{P}(\hat{C}_i \in C_L)$ trials are selected randomly and allocated time $t_L$. The remaining trials are allocated time $t_H$.

Constant time is the most natural baseline allocation, while Random time assigns the same values $t_L$ and $t_H$ as the confidence-based policy but does so at random. Both are evaluated in the experiment described in Section 5.2. While Corollary 1 below is implied by Proposition 1, its proof shows exactly where the confidence-based policy has higher accuracy.

**Corollary 1.** Consider the two time allocation strategies defined above - (1) Constant time and (2) Random time, in AI-assisted decision-making where we have total $N$ trials and time $T$. Suppose Assumption 1 stated in $(10)$ holds. Then the human-AI accuracy of the confidence-based time allocation policy is greater than or equal to the accuracy of the constant time allocation and random time allocation strategies.
Proof. Let the two-level confidence based allocation policy be denoted by $\pi$. Now, we have that the accuracy for each round under this policy,

$$E_\pi[R_t] = \mathbb{E} \left[ \mathbb{E}[R_t \mid \hat{C}_i, T_i = \pi(\hat{C}_i)] \right]$$

$$= \mathbb{P}(\hat{C}_i \in C_L)\mathbb{E}[R_t \mid \hat{C}_i \in C_L, T_i = t_L] + \mathbb{P}(\hat{C}_i \in C_H)\mathbb{E}[R_t \mid \hat{C}_i \in C_H, T_i = t_H]. \quad (13)$$

For the constant allocation policy, we get

$$E_{\text{const}}[R_t] = \mathbb{E}[R_t \mid T_i = \frac{T}{N}]$$

$$= \mathbb{P}(\hat{C}_i \in C_L)\mathbb{E}[R_t \mid \hat{C}_i \in C_L, T_i = \frac{T}{N}] + \mathbb{P}(\hat{C}_i \in C_H)\mathbb{E}[R_t \mid \hat{C}_i \in C_H, T_i = \frac{T}{N}]. \quad (14)$$

Now, according to assumption 1 in [10], we have $\mathbb{E}[R_t \mid \hat{C}_i \in C_L, T_i = t_L] \geq \mathbb{E}[R_t \mid \hat{C}_i \in C_L, T_i = \frac{T}{N}]$, and $\mathbb{E}[R_t \mid \hat{C}_i \in C_H, T_i = t_H] \geq \mathbb{E}[R_t \mid \hat{C}_i \in C_H, T_i = \frac{T}{N}]$. Thus, $E_\pi[R_t] \geq E_{\text{const}}[R_t]$. Similarly for random allocation, we have

$$E_{\text{rand}}[R_t]$$

$$= \mathbb{P}(T_i = t_L)\mathbb{E}[R_t \mid T_i = t_L] + \mathbb{P}(T_i = t_H)\mathbb{E}[R_t \mid T_i = t_H]$$

$$= \mathbb{P}(\hat{C}_i \in C_L)\mathbb{P}(T_i = t_L)\mathbb{E}[R_t \mid \hat{C}_i \in C_L, T_i = t_L] + \mathbb{P}(\hat{C}_i \in C_H)\mathbb{P}(T_i = t_L)\mathbb{E}[R_t \mid \hat{C}_i \in C_H, T_i = t_L]$$

$$+ \mathbb{P}(\hat{C}_i \in C_L)\mathbb{P}(T_i = t_H)\mathbb{E}[R_t \mid \hat{C}_i \in C_L, T_i = t_H] + \mathbb{P}(\hat{C}_i \in C_H)\mathbb{P}(T_i = t_H)\mathbb{E}[R_t \mid \hat{C}_i \in C_H, T_i = t_H]. \quad (15)$$

Using the assumptions [10] as stated for constant allocation, we prove that $E_\pi[R_t] \geq E_{\text{rand}}[R_t]$. \qed

5.2 Experiment 2: Dynamic time allocation for human-AI collaboration

In this experiment, we implement our confidence-based time allocation strategy for human-AI collaboration in a user study deployed on Amazon Mechanical Turk. Based on the results of Experiment 1 (Figure 2), we assign $t_L = 25s$ and $t_H = 10s$. Through this experiment, we test the following hypotheses.

- Hypothesis 2 (H2): Anchoring bias has a negative effect on human-AI collaborative decision-making accuracy when AI is incorrect.
- Hypothesis 3 (H3): If the human decision-maker has complementary knowledge then allocating more time can help them sufficiently adjust away from the AI prediction.
- Hypothesis 4 (H4): In our setup, the confidence-based time allocation yields better performance than Human alone and AI alone.
- Hypothesis 5 (H5): In our setup, the confidence-based time allocation yields better human-AI team performance than constant time and random time allocations.

We now describe the different components of Experiment 2 in detail.

Participants. In this study, 176 participants were recruited in the same manner as described in Section 4.1. 35 participants were between ages 18 and 29, 77 between ages 30 and 39, 77 between ages 40 and 49, and 23 over age 50. The average completion time for this user study was 30 minutes, and participants received compensation of $5.125 on average (roughly equals an hourly wage of $10.25). The participants received an average base pay of $4.125 and bonus of $1 (to incentivize accuracy).

Task and AI model. The binary prediction task in this study is the same as the student performance prediction task used before. In this experiment, our goal is to induce optimal human-AI collaboration under the assumptions illustrated in Figure 4. In real-world human-AI collaborations, it is not uncommon for the decision-maker to have some domain expertise or complementary knowledge that the AI does not, especially in fields where there is not enough data such as social policy-making and design. To emulate this situation where the participants have complementary knowledge, we reduced the information available to the AI, given the unavailability of human experts and the limited training time in our experiment. We train the assisting AI model over 7 features, while the participants have access to 3 more features, namely, hours spent studying weekly, hours spent going out with friends weekly, and enrollment in extra educational support. These 3 features were the second to fourth most important ones as deemed by a full model.

To implement the confidence-based time allocation strategy, we had to identify trials belonging to classes $C_L$ and $C_H$. Ideally, for this we require a machine learning algorithm that can calibrate its confidence correctly. As discussed in Section 5.1, we use the AI’s predicted probability $\hat{C}_i$ (termed as AI confidence) and choose the threshold for $C_H$ as $\hat{C}_i \geq 0.75$. This study has 40 questions in the testing section, from which 20 belong to $C_L$ and 20 belong to $C_H$.

Study procedure. As in Experiment 1, this user study has two sections, the training section and the testing section. The training section is exactly the same as before
where the participants are trained over 15 examples selected from the training dataset. To induce anchoring bias, as in Experiment 1, we reinforce that the AI predictions are 85% accurate in the training section.

The testing section has 40 trials, which are sampled randomly from the test set such that the associated predicted probability values (of the predicted class) estimated by the machine learning algorithm are distributed uniformly. While the set of trials in the testing section is fixed for all participants, the order they were presented in was varied randomly.

To test hypotheses H2, H3, H4 and H5, we randomly assigned each participant to one of four groups:

1. **Human only**: In this group, the participants were asked to provide their prediction without the help of the AI prediction. The time allocation for each trial in the testing section is fixed at 25 seconds. This time is judged to be sufficient for humans to make a prediction on their own, based on the results of Experiment 1 (for example the time usage distributions in Figure 3).

2. **Constant time**: In this group, the participants were asked to provide their prediction with the help of the AI prediction. The time allocation for each trial in the testing section is fixed as \( t = \frac{t_L + t_H}{2} = 17.5 \) seconds. We rounded this to 18 seconds when reporting it to the participants.

3. **Random time**: This group has all factors the same as the constant time group except for the time allocation. For each participant, the time allocation for each trial is chosen uniformly at random from the set \{10, 25\} such that the average time allocated per trial is 17.5 seconds.

4. **Confidence-based time**: This is our treatment group, where we assign time according to confidence-based time allocation, \( t_L = 25 \) seconds and \( t_H = 10 \) seconds, described in Section 5.1.

Out of the 176 participants, 41 were in "Human only", 44 were in "Constant time", 44 were in "Random time", and 47 were in "Confidence-based time". In the groups where participants switch between 10-second and 25-second conditions, that is, "Random time" and "Confidence-based time", the accuracy would likely be affected by rapid switching between the two time conditions. Hence, we created blocks of 5 trials with the same time allocation for both groups. The complete testing section contained 8 such blocks. This concludes the description of Experiment 2.

![Figure 5: Average accuracy and agreement ratio of participants in Experiment 2 across the four different conditions, marked on the y-axis. We note that the error bars in (a) for 'All' trials are smaller than the marker size.](image)

### 5.3 Results

Figure 5 shows that our effort to create a scenario where the AI knowledge is complementary to human knowledge is successful because the AI only and "Human only" conditions have similar overall accuracy (around 60%, black diamonds), and yet humans only agreed with the AI in 61.3% of the trials. Moreover, on trials where AI is incorrect, "Human only" has accuracy of 50% on trials in \( C_L \), and 33.7% on trials in \( C_H \). Thus, the participants showed more complementary knowledge in trials in \( C_L \) compared to \( C_H \).

Given this successful setup of complementary knowledge between humans and AI, there is good potential for the human-AI partnership groups, especially the "Confidence-based time" group, to outperform the AI only or "Human only" groups (H4). Based on the overall accuracy shown in Figure 5(a), however, this did not seem to occur. The mean accuracy of the human-AI team in "Confidence-based time" is 61.2% while the accuracy of "Human only" is 61.5% and the accuracy of the AI model is 60%. Thus, regarding H4, the results suggest that the accuracy in "Confidence-based time" is greater than AI alone \((p = 0.097, t(94) = 1.30)\), whereas they do not provide sufficient evidence for "Confidence-based time" being better than "Human only" \((p = 0.58, t(92) = -0.21)\).

However, we did see that the anchoring bias affected
team performance negatively when the AI is incorrect (H2). Figure 5b) shows evidence of anchoring because the agreement percentage in the "Human only" group is much lower than those in the collaborative conditions ($p < 0.001, t(184) = 6.73$). When the AI was incorrect (red triangles and green circles), this anchoring bias clearly reduced team accuracy when compared to the "Human only" accuracy ($p < 0.001, t(370) = −6.68$). Although, it is important to note that the "Human only" group received longer time (25s) than the collaborative conditions on average. Nevertheless, if we just compare "Human only" and "Confidence-based time" within the low confidence trials (red triangles), where both were assigned the same amount of time (25s), we observe similar disparity in agreement percentages ($p < 0.001, t(92) = 4.97$) and accuracy ($p < 0.001, t(92) = −4.74$). Hence, the results are consistent with H2.

Next, we examine the differences between the three collaborative groups. Figure 5a shows that their average accuracy over all trials (black diamonds) are again very close, with "Confidence-based time" at 61.2%, "Random time" at 60.4% and "Constant time" at 62%. Thus, regarding H5, we do not see sufficient evidence for "Confidence-based time" being better than "Constant time" ($p = 0.65, t(91) = −0.38$) and "Random time" ($p = 0.26, t(93) = 0.63$).

However, judging by the agreement percentage, we see that the participants are overall less anchored in "Confidence-based time" than in the other two collaborative conditions, and particularly in the low confidence trials when AI is incorrect (red triangles) ($p = 0.16, t(138) = −1.00$), which suggests that giving people more time indeed helped them adjust away from the anchor sufficiently, in these trials (H3). This de-anchoring also led to slightly higher accuracy in these trials (red triangles) for "Confidence-based time" (40.4%) when compared to the other two collaborative conditions ($p = 0.10, t(138) = 1.24$) with "Random time" at 34.2% and "Constant time" at 36.8%.

The reason that the overall accuracy of "Confidence-based time" is not significantly better than the other two collaborative conditions is likely because of the relatively low accuracy and low agreement percentage in trials in $C_H$ (green circles, purple pentagons). Based on the results of Experiment 1, we expected that the agreement percentage for the 10-second trials would be high and since these align with the high AI confidence trials for "Confidence-based time", we expected these trials to have a high agreement percentage and hence high accuracy. Instead, we observed that "Confidence-based time" has low agreement percentage (84%) in $C_H$, even lower than that of "Random time" (87.9%), with an average time allocation of 17.5 seconds, and "Constant time" (88.1%), with a constant time allocation of 17.5 seconds. This lower agreement percentage translates into lower accuracy (86%) when AI is correct (purple pentagons). In the next section, we discuss how this points to possible distrust of AI in these high confidence trials and its implications.

6 Discussion

Lessons learned. We now discuss some of the lessons learned from the results obtained in Experiment 2. As noted in Section 5.3 we see that "Confidence-based time" has a low agreement rate on trials in $C_H$ where the time allocated is 10 seconds and the AI prediction is 70% accurate. Moreover, we see that the agreement rate is lower than "Human only" and "Constant time" on trials in $C_H$, where the AI prediction is correct as well as those where the AI prediction is incorrect. This behavior suggests that the participants in "Confidence-based time" may have grown to distrust the AI, as they disagreed more with the AI on average and spent more time on the trials where they disagreed. The distrust may be due to "Confidence-based time" assigning longer times (25s) only to low-AI-confidence trials, perhaps giving the impression that the AI is worse than it really is. Our observations highlight the importance of accounting for human behaviour in such collaborative decision-making tasks. We envisage that giving the decision-maker the reasoning behind the time allocation, that is, informing them about AI prediction confidence and then allocating time according to AI confidence, would help increase the agreement rate and accuracy in line with the two other collaborative conditions. This in turn may lead to overall accuracy that is better than both "Human only" and AI only.

This conjecture is also supported by findings in Cham-bon et al. (2020). In this work, the authors observe that the human brain is primed to learn with a bias that is pegged to our freely chosen actions. Choice tips the balance of learning: for the same action and outcome, the brain learns differently and more quickly from free choices than forced ones. Thus, providing valid reasons for time allocation would help the decision-maker make an active choice and hence, learn to collaborate with AI better than in Experiment 2 where the time allocation motivation is unclear to the participants.

Another insight gained from Experiment 2 is that the model should take into account the sequentiality of decision-making where the decision-maker continues to learn and build their perception of the AI as the task progresses, based on their interaction with the AI. Dynamic Markov models have been studied previously in the context of human decision-making (Busemeyer et al. 2020 Lieder et al. 2018). We believe that studying dynamic
cognitive models that are cognizant of the changing interaction between the human and the AI model would be helpful in creating more informed policies for human-AI collaboration.

Conclusions. In this work, we foreground the role of cognitive biases in the human-AI collaborative decision-making setting. Through literature in cognitive science and psychology, we explore several biases and present mathematical models of their effect on collaborative decision-making. We focus on anchoring bias and the associated anchoring-and-adjustment heuristic that is important towards optimizing team performance. We validate the use of time as an effective strategy for mitigating anchoring bias through a user study. Furthermore, through a time-based resource allocation formulation, we provide an optimal allocation strategy that attempts to achieve the "best of both worlds" by capitalizing on the complementary knowledge presented by the decision maker and the AI model. Using this strategy, we obtain human-AI team performance that is better than the AI alone, as well as better than having only the human decide in cases where the AI predicts correctly. When the AI is incorrect, the information it provides the human distracts them from the correct decision, thus reducing their performance. Giving them a choice as detailed in the lessons learned above might alleviate some of these issues and bring us closer to the ideal Human-AI team performance shown in Figure 4.

Future work. Our work shows that a simple time-based strategy, built on the cognitive tendencies of the decision-maker in a collaborative setting, can help decision-makers adjust their decisions correctly. More generally, our work showcases the importance of accounting for cognitive biases in decision-making, where in the future we would want to study other important biases such as confirmation bias or weak evidence effect. This paper opens up several directions for future work where explanation strategies in this collaborative setting are also studied and designed based on the cognitive biases of the human decision-maker. Another interesting direction is to utilize the resource allocation framework for other cognitive biases based on their debiasing strategies. It is also of interest to study the effectiveness of combining our time-based strategy with an explanation strategy to improve team performance.

Acknowledgements. The work of Charvi Rastogi was supported in part by NSF CIF 1763734.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., et al. (2019). Factsheets: Increasing trust in ai services through supplier’s declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1.

Arnott, D. (2006). Cognitive biases and decision support systems development: a design science approach. *Information Systems Journal*, 16(1):55–78.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019a). Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. (2019b). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437.

Bansal, G., Wu, T., Zhu, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. S. (2020). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *arXiv preprint arXiv:2006.14779*.

Barnes JR., J. H. (1984). Cognitive biases and their impact on strategic planning. *Strategic Management Journal*, 5(2):129–137.

Baudel, T., Verbockhaven, M., Roy, G., Cousergue, V., and Laarach, R. (2020). Addressing cognitive biases in augmented business decision systems. *arXiv preprint arXiv:2009.08127*.

Bromme, R., Hesse, F. W., and Spada, H. (2010). *Barriers and Biases in Computer-Mediated Knowledge Communication: And How They May Be Overcome*. Springer Publishing Company, Incorporated, 1st edition.

Busemeyer, J. R., Kvam, P. D., and Pleskac, T. J. (2020). *Comparison of markov versus quantum dynamical models of human decision making*. *Wiley Interdisciplinary Reviews: Cognitive Science*.

Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., and Palminteri, S. (2020). Information
about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behavior.*

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*(7):287 – 291. Special issue: Probabilistic models of cognition.

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

Das, T. and Teng, B.-S. (1999). Cognitive biases and strategic decision processes: An integrative perspective. *Journal of Management Studies, 36*(6):757–778.

Dhurandhar, A., Graves, B., Ravi, R. K., Maniachari, G., and Ettl, M. (2015). Big data system for analyzing risky procurement entities. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015*, pages 1741–1750. ACM.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*

Ehrlinger, J., Readinger, W., and Kim, B. (2016). Decision-making and cognitive biases. *Encyclopedia of Mental Health.*

Englich, B., Mussweiler, T., and Strack, F. (2006). Playing dice with criminal sentences: The influence of irrelevant anchors on experts’ judicial decision making. *Personality and Social Psychology Bulletin, 32*(2):188–200.

Epley, N. and Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological science, 12*(5):391–396.

Epley, N. and Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4):311–318. PMID: 16623688.

Fernbach, P., Darlow, A., and Sloman, S. (2011). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition, 119*:459–67.

Furnham, A. and Boo, H. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*:35–42.

Fünkranz, J., Kliegr, T., and Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. *Machine Learning, 109*(4):853–898.

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

Griffiths, T. L., and Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*(9):767–773.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330.

Janssen, J. and Kirschner, P. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: towards a research agenda. *Educational Technology Research and Development, 1*. 1–23.

Klayman, J. (1995). Varieties of confirmation bias. In *Psychology of learning and motivation*, volume 32, pages 385–418. Elsevier.

Kliegr, T., Bahník, Š., and Fünkranz, J. (2018). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *arXiv preprint arXiv:1804.02969.*

Lai, V., Liu, H., and Tan, C. (2020). Why is ‘chicago’ deceptive? towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Lai, V. and 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, FAT' '19*, page 29–38, New York, NY, USA. Association for Computing Machinery.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1):50–80. PMID: 15151155.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society, 5*(1):2053951718756684.

Lieder, F., Griffiths, T. L., Huys, Q. J. M., and Goodman, N. D. (2018). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin and Review, 25*:322–349.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3):31–57.
Matsumori, K., Koike, Y., and Matsumoto, K. (2018). A biased bayesian inference for decision-making and cognitive control. *Frontiers in Neuroscience*, 12:734.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175 – 220.

Okamura, K. and Yamada, S. (2020). Adaptive trust calibration for human-ai collaboration. *PLOS ONE*, 15(2):1–20.

Oswald, M. and Grosjean, S. (2004). Confirmation bias. In R. F. Pohl (Ed.). *Cognitive Illusions. A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*, pages 79–96. Hove and N.Y.: Psychology Press.

Payzan-LeNestour, E. and Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Comput Biol*, 7(1):e1001048.

Payzan-LeNestour, E. and Bossaerts, P. (2012). Do not bet on the unknown versus try to find out more: Estimation uncertainty and “unexpected uncertainty” both modulate exploration. *Frontiers in Neuroscience*, 6:150.

Phillips-Wren, G., Power, D. J., and Mora, M. (2019). Cognitive bias, decision styles, and risk attitudes in decision making and dss. *Journal of Decision Systems*, 28(2):63–66.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. (2018). Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Preece, A. (2018). Asking ‘why’in ai: Explainability of intelligent systems—perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72.

Siou, K. and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31:47–53.

Silverman, B. G. (1992). Human-computer collaboration. *Human–Computer Interaction*, 7(2):165–196.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, 63(2):129.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, 1(1):161–176.

Solomon, J. (2014). Customization bias in decision support systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3065–3074.

Springer, A., Hollis, V., and Whittaker, S. (2018). Dice in the black box: User experiences with an inscrutable algorithm. *arXiv preprint arXiv:1812.03219*.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *Advances in neural information processing systems*, pages 59–68.

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., and Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware ai. *Patterns*, 1(4):100049.

Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, page 1–15, New York, NY, USA. Association for Computing Machinery.

Yeom, S. and Tschantz, M. C. (2018). Discriminative but not discriminatory: A comparison of fairness definitions under different worldviews. *arXiv preprint arXiv:1808.08619*.

Yeung, A., Joshi, S., Williams, J. J., and Rudzicz, F. (2020). Sequential explanations with mental model-based policies. *arXiv preprint arXiv:2007.09028*.

Zhang, Y., Bellamy, R. K., and Kellogg, W. A. (2015). Designing information for remediating cognitive biases in decision-making. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 2211–2220.

Zhang, Y., Liao, Q. V., and 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*, pages 295–305.