From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis

Zhuoyan Li
li4178@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Hangxiao Zhu
hangxiao@tamu.edu
Texas A&M University
College Station, Texas, USA

Zhuoran Lu
lu800@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Ziang Xiao
ziang.xiao@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

Ming Yin
mingyin@purdue.edu
Purdue University
West Lafayette, Indiana, USA

ABSTRACT

AI-assisted decision making becomes increasingly prevalent, yet individuals often fail to utilize AI-based decision aids appropriately especially when the AI explanations are absent, potentially as they do not reflect on AI's decision recommendations critically. Large language models (LLMs), with their exceptional conversational and analytical capabilities, present great opportunities to enhance AI-assisted decision making in the absence of AI explanations by providing natural-language-based analysis of AI's decision recommendation, e.g., how each feature of a decision making task might contribute to the AI recommendation. In this paper, via a randomized experiment, we first show that presenting LLM-powered analysis of each task feature, either sequentially or concurrently, does not significantly improve people's AI-assisted decision performance. To enable decision makers to better leverage LLM-powered analysis, we then propose an algorithmic framework to characterize the effects of LLM-powered analysis on human decisions and dynamically decide which analysis to present. Our evaluation with human subjects shows that this approach effectively improves decision makers' appropriate reliance on AI in AI-assisted decision making.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

AI-assisted decision making, Explainable AI, Large language model

ACM Reference Format:
Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, Ziang Xiao, and Ming Yin. 2025. From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis . In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642625

CHI '25, April 26–May 1, 2025, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/3613904.3642625

1 INTRODUCTION

AI systems have been significantly integrated into the human decision making process in various domains, such as criminal justice [20, 93] and financial investment [3, 33], thereby creating a new paradigm of human-AI collaboration [94]. In this paradigm, AI models provide recommendations or analysis to assist humans in making decisions, while human decision makers are ultimately responsible for the final decisions [30, 71].

However, many empirical studies evaluating the effectiveness of current AI-assisted decision making systems [41] have demonstrated that when people collaborate with AI in decision making tasks, they rarely engage in analytical thinking to combine their own insights with the AI model's recommendations intelligently [6, 26, 68]. Instead, they often rely on AI inappropriately—accepting an AI model's recommendations when they are incorrect or mistakenly ignoring AI's correct recommendations—leading to either overreliance or underreliance on AI [55, 77]. To address this problem, previous research [2, 78, 79] proposed to display explanations generated by post-hoc explainable AI (XAI) methods [29, 56, 72] along with the AI model's decision recommendations to assist people in evaluating AI's reliability and identifying optimal strategies for relying on AI. However, the computation of AI explanation often requires access to the AI model's internal parameters and structures, while many evaluation studies have revealed that it is challenging for humans to understand and utilize such explanation without substantial effort to teach them how to interpret the explanation [96]. Consequently, even with the presence of AI explanations, decision makers often still exhibit a low level of appropriate reliance on AI, let alone the case when the AI explanations are not available.

As such, one would naturally wonder if it is still feasible to guide decision makers to critically and systematically reflect on AI's decision recommendations and appropriately utilize them *without easily accessible or available AI explanations.* Interestingly, in real-world decision making across various domains like healthcare and finance, when decision makers find the initial recommendations lack transparency and clarity, they often seek additional insights or interpretations from secondary sources. To this end, the exceptional conversational and analytical capabilities exhibited by the latest

Income Prediction	
Education Level	*Value:* 10 years *Concept:* Higher education is commonly linked to higher earning potential. *In this case:* With 10 years of education, this might be slightly below the threshold for high-earning positions, which **decreases** the likelihood of making over $50000 per year.
Recidivism Prediction	
Charge Degree	*Value:* Felony *Concept:* The severity of the charge can predict recidivism, with felonies often leading to harsher predictions than misdemeanors. *In this case:* Facing a felony charge, which **increases** the likelihood of recidivating because felonies are associated with more severe criminal behavior.

Table 1: Examples of analysis generated by the LLM for (top) how the education level of a person might have affected AI's prediction on this person's income level; and (bottom) how the charge degree of a defendant might have affected AI's prediction on this defendant's recidivism status.

state-of-the-art large language models (LLMs) could offer strong promise [13, 31, 45, 59, 88, 102]. For example, LLMs can analyze a decision making task and AI's decision recommendation on it, and then provide potential reasons for why the AI recommends such a decision in a natural language format, which is straightforward for humans to process. While the AI model serves as the primary advisor for human decision makers, when LLM-powered analyses are used to augment the AI model's recommendations, the LLM effectively serves as a secondary advisor to provide additional perspectives and justifications through its analysis. These analyses may help the human decision maker better interpret the recommendation of the primary advisor. They may also offer a starting point for decision makers to organize their thoughts and reflect on both the AI model's decision recommendation and their own judgment, which may help them calibrate their trust in the AI model.

Therefore, in this paper, we start by conducting a randomized human-subject experiment to examine whether incorporating LLM-powered analyses in AI-assisted decision making can improve the performance of human-AI teams and promote more appropriate reliance on AI models in the absence of actual explanations of the AI models. Given a decision making task as well as an AI model's decision recommendation on it, we first prompted OpenAI's GPT-4 model [67] to generate analyses for how each feature of the task might have led to the AI model's decision recommendation on the task (see Table 1 for examples). Depending on whether and how to present these LLM-powered analyses, we created three treatments in our experiment—Control (where participants would not receive any analysis from the LLM), Seq (where participants receive the analysis about each feature sequentially), and All (where participants receive analyses about all features at once). Our experimental results show that presenting LLM-powered analysis either sequentially or concurrently to human decision makers does not significantly improve their decision accuracy compared to those decision makers who did not receive any LLM-powered analysis. This suggests that more intelligent ways should be used to present LLM-powered analyses to people to facilitate their utilization of this information and promote their effective decision making.

In light of this, we propose an algorithmic framework to adaptively present LLM-powered analyses to decision makers—based on the historical data on how human decision makers react to different LLM-powered analyses, our algorithmic framework learns to present LLM-powered analysis selectively and progressively to maximize the chance for the decision maker to rely on the AI model's decision recommendations appropriately and make the correct decisions. To do so, we first learn a human behavior model that characterizes the effects of LLM-powered analysis on human decisions. We then dynamically decide which analysis to present (among the LLM-powered analyses for all features of the decision making task) by comparing the expected maximum utility of presenting each analysis. To evaluate the effectiveness of this algorithmic approach in selecting the best set of LLM-powered analyses to help improve decision makers' appropriate reliance and decision accuracy in AI-assisted decision making, we conducted another randomized human-subject experiment. We find that compared to other baseline approaches for presenting LLM-powered analysis, when the LLM-powered analyses are selected using our algorithmic approach, human decision makers can achieve significantly higher accuracy in their final decisions and reduce overreliance on the AI model across different types of decision making tasks. Additional analysis suggests that our algorithmic approach selects fewer but more informative LLM-powered analysis to show to decision makers compared to baseline approaches.

Together, our study provides important experimental evidence regarding the effectiveness of incorporating LLMs in AI-assisted decision making, and how to design intelligent interactions between humans and LLMs to promote better human-AI collaboration in decision making. We conclude by discussing the implications and limitations of our study.

2 RELATED WORK

2.1 AI-Assisted Decision Making

The increasing prevalence of AI-assisted decision making has led to a growing line of research to investigate how people engage with, trust in, and rely on AI models in this new collaboration paradigm [11, 41, 54]. Early studies focus on empirically identifying factors that influence AI-assisted decision making, including the AI model's performance [70], the explanation of the model recommendation [74, 78, 79], the decision making workflow [12, 69], and the influence of task complexity on human-AI interactions [75].

While it is expected that the complementarity between AI models and humans could enable the human-AI team to outperform either party alone, in practice, the collaboration between humans and AI in decision making is widely observed to be suboptimal [77]. It is observed that people usually exhibit inappropriate reliance

on AI models [85]. For instance, the design of conversational interfaces can influence users' trust, sometimes causing overreliance on AI recommendations [23]. In addition, people may also blindly rely on AI in time-pressured environments, where the presence of AI suggestions may speed up decision making at the cost of accuracy [89]. In contrast, people could also reject the AI model recommendation even when it is correct, noted as underreliance on AI [24, 61, 66]. Recent research has also discussed how misaligned AI outputs can contribute to people's underreliance on AI systems despite their accuracy [22]. To help decision makers interact with and rely on the AI model more appropriately, a wide range of approaches was recently developed [7, 9, 25, 48, 49, 52, 53]. For instance, the cognitive forcing function encourages people to engage with AI more cognitively, thus potentially reducing people's overreliance on the AI model [6, 16, 41, 76]. Ma et al. [58] explored the calibration of user trust in AI-assisted decision making by inferring the correctness likelihood of both human and AI on a decision case, which informs the adaptive presentations of the AI model's decision recommendations.

In addition, providing AI explanations generated by various post-hoc explainable AI (XAI) methods [57, 73] that reveal the decision rationale of AI models is another popular approach used, aiming to improve humans' understanding of the AI model's behavior and enable humans to calibrate their trust in AI. However, many empirical studies have observed that people often struggle to process and comprehend these explanations [44, 50, 91, 96], letting alone utilize the insights revealed from these explanations to trust AI more appropriately. To realize the positive utility of explanations in AI-assisted decision making, recent research highlights the need to provide explanations selectively or progressively to aid human comprehension [17, 43, 50, 83, 84] . For instance, Lai et al. [43] demonstrated that selectively highlighting AI explanations, which align with the user's own decision rationale, can increase agreement between human decisions and AI model predictions and reduce human overreliance on AI recommendations. Springer and Whittaker [83] showed that users may benefit from initially simplified feedback that hides potential AI system errors and assists users in building working heuristics about how the AI system operates progressively. In this work, we make an initial attempt to explore that *in the absence of AI explanations*, whether the incorporation of the natural-language-based, LLM-powered analysis of the AI recommendations on decision making tasks can promote more appropriate reliance behavior of humans on AI models in decision making, and how to present such analysis in the most effective way.

2.2 Human-LLM Interaction

Recently, large language models (LLMs) have demonstrated their exceptional capabilities across various applications to assist humans, including creative writing [47, 95, 99], software engineering [64, 65], and generative design [28], which has sparked significant interest within the HCI community to investigate the interaction between humans and LLMs [12, 18, 19, 35, 46]. On the one hand, LLMs are increasingly utilized to directly create content or solve problems, which is shown to match or even surpass humans' performance. For example, Mirowski et al. [63] presented the framework leveraging LLMs to create coherent scripts and screenplays with humans in the loop. In other cases, LLM-based services provide foundational support for human creation, such as generating coding schemes for qualitative analysis [10]. In these human-LLM collaboration scenarios, a key challenge is that laypeople often lack the skill to effectively prompt LLMs to generate the outputs that they desire [100]. To address this challenge, novel approaches like AI Chains [97], automatic prompting methods [80], and interactive interfaces [51, 92] are developed to enhance the effectiveness of human-LLM interaction, either by improving LLMs' usability [27, 98] or by guiding humans' engagement with LLMs.

Researchers have also explored the potential of LLMs in AI-assisted decision making. For example, LLMs could directly provide decision recommendations. However, it was found that the overconfident and seemingly convincing LLM outputs can mislead people to believe them to be correct [87] and result in people's overreliance on LLM [14, 36]. Recently, Slack et al. [82] developed an interactive dialogue system that allows users to inquire about the reasons behind the AI model's predictions. This system leverages a LLM to parse user intent and match it with pre-specified, handcrafted answers, demonstrating significant potential to enhance user understanding and decision performance through conversational interactions with the AI model. Different from the previous work, in this paper, we explore how to utilize LLMs to analyze an AI model's decision recommendations and augment them, and how to build an algorithmic framework to dynamically decide what information to present to humans from the rich information generated by LLMs.

3 EMPIRICAL EXAMINATIONS OF THE IMPACTS OF LLM-POWERED ANALYSIS IN AI-ASSISTED DECISION MAKING

We start by investigating whether the incorporation of LLM-powered analysis can enhance human decision makers' decision performance and promote their more appropriate reliance on AI models in AI-assisted decision making. To do so, we conducted a randomized human-subject experiment on Prolific.

3.1 Decision Making Task

In our experiment, we considered two types of decision making tasks that have widely been used as the testbeds in AI-assisted decision making research [12, 58, 101]:

- **Income Prediction** [38]: Human decision makers were asked to determine whether a person's annual income level is higher or lower than $50k with the assistance of an AI model. Specifically, in each task, we presented the participant with a person's profile containing 7 features, which include the person's gender, age, education level, marital status, occupation, work type, and working hours per week. We trained a random forest model to provide decision recommendations, and the accuracy of the model was 76%.
- **Recidivism Prediction** [15]: Human decision makers needed to predict whether a defendant would reoffend within two years. Each task presented a defendant profile with 8 features, including basic demographics (e.g., gender, age, race), criminal history (e.g., the count of prior non-juvenile crimes, juvenile misdemeanor crimes, and juvenile felony crimes

committed), and information related to their current charge (e.g., charge issue, charge degree). We again trained a random forest model to provide decision recommendations, and the accuracy of the model was 62%[1].

3.2 Generation of LLM-powered Analysis

We used LLMs to generate an analysis for each AI-assisted decision making task. Specifically, we prompted GPT-4 to analyze the decision making task and the AI model's decision recommendation. The prompts for GPT-4 to generate the analysis for both the income prediction and recidivism prediction tasks consist of three parts:

- **Introduction Prompt**: `Please take on the role of a data analyst and prepare to analyze the provided task instances. Your task is to explain how the features in the presented task instances contribute to the AI model predictions provided. Each profile includes various features that you will need to consider in your analysis, [INTRODUCE THE FEATURE NAMES AND DESCRIPTIONS].`
- **Instruction Prompt**: `For each presented task, assess how each feature might contribute to [AI MODEL PREDICTION]. For each task, your analysis should contain 1 identifier (index), [NUMBER OF FEATURES] concepts (explanations of how the features could support the model prediction), and [NUMBER OF FEATURES] case descriptions (specific explanations of how the feature values in the current profile support the model prediction). [AN EXAMPLE OF THE OUTPUT ANALYSIS].`
- **Emotional Stimuli Prompt**: `This is an academic study aimed at enhancing human trust in AI system advice through reasonable explanations. The knowledge gained will help improve human-AI collaboration. This mission is critical to the whole human society. Please analyze the task instance thoroughly and provide diverse insights.`

The LLM examined the task features and determined how each feature may have contributed to the AI model's prediction. Consequently, the LLM generated a set of analyses for each task instance, associating each task feature with one analysis to indicate its possible contribution to the AI recommendation. Table 1 shows several examples of analyses generated by GPT-4. This set of analyses serves as the LLM-powered analysis to be incorporated into AI-assisted decision making in our study (see the supplemental materials for more examples of the analyses). While such LLM-powered analysis can be readily applied to decision making tasks with tabular data where the task-related information is presented in a structured manner as a collection of features and their values, similar analysis can also be conducted on decision tasks involving other types of data (e.g., images, texts) after transforming the unstructured data into structured formats (see more discussions on this in Section 6.5).

Note that we do not consider the analysis generated by the LLM as necessarily reflecting the AI model's true decision rationale. Instead, it is only the LLM's *interpretation/justification* of the AI recommendation, and is used to augment the AI recommendation in the absence of explanations to the AI model. Alternatively, since we prompted the LLM to justify a specific decision (i.e., the decision that is consistent with the AI model's recommendation), one can also view the LLM-powered analysis as the LLM's own explanations to the specified decision.

3.3 Experimental Treatments

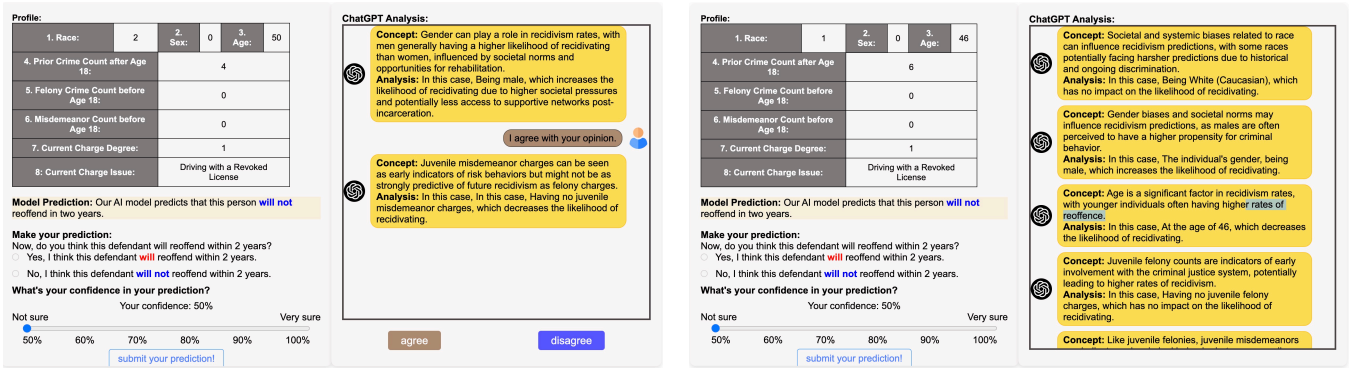
In our experiment, participants were asked to complete a series of decision making tasks. For each task, they were provided with the AI model's prediction along with the task instance, and they needed to make the final decision. We created 3 experimental treatments by varying whether and how LLM-powered analysis was introduced into the AI-assisted decision making process. Specifically:

- Control: In this treatment, we did not incorporate LLM-powered analysis into the AI-assisted decision making process. Participants assigned to this treatment were asked to make decisions with the assistance provided by the AI model alone, without any additional analysis from the LLM.
- Sequential (Seq): In this treatment, participants started working on the decision making task seeing only the task instance and the AI model's recommendation without receiving any LLM-generated analyses. Then, we told participants that an LLM had analyzed how different features of the task instance may contribute to the AI model's recommendation on this task. Participants were required to interact with the LLM through a designated interface where, in each turn, the LLM's analysis about one task feature's contribution to the AI recommendation would be *randomly* sampled from the generated set and presented to the participant. The participant could respond to the analysis by indicating whether they agreed or disagreed with it. The participant must interact with the LLM for *at least X* rounds where X is randomly sampled between 1 and 3 for each task. After meeting the minimum interaction requirement, participants could continue to review the LLM-powered analysis on more features, or they could stop the interaction and make their final decisions at any point that they wish. Figure 1a shows an example of the task interface used in this treatment.
- All: In this treatment, we presented all the LLM-powered analyses for each one of the task features to participants at once, along with the task instance and the AI model's decision recommendation. After reviewing all this information, participants made their final decisions. Figure 1b shows an example of the task interface used in this treatment.

3.4 Experimental Procedure

Our experiment was conducted on Prolific. Upon the arrival of a participant, we randomly assigned them to one of the two types

[1]For the random forest models used in both the income prediction and recidivism prediction tasks, we used grid search to fine-tune the model parameters such as the depth of the tree and the number of trees. The relatively low level of prediction accuracy of the random forest model was primarily due to the inherent difficulty and uncertainty of the task. We also experimented with using zero-shot and few-shot approaches to prompt the state-of-the-art LLM, GPT-4, to directly provide binary recommendations on these tasks. When evaluating on the same test dataset, we found that the accuracy of GPT-4 on income prediction tasks and recidivism prediction tasks were 59% and 56%, respectively, which were lower than the random forest models.



(a) Seq treatment **(b) All treatment**

Figure 1: The example interfaces used in the Seq and All treatments of our experiment for the recidivism prediction task.

of tasks and one of the three treatments. In the experiment, participants were asked to first fill out an initial survey to report their demographic information and knowledge of AI models. Then, they started the formal experiment by completing a tutorial that described the decision making task they needed to work on. To help participants get familiar with the decision making task, we set up a training stage in which participants completed five decision making tasks independently without seeing the AI model's recommendation or any LLM-powered analyses. During these training tasks, we immediately provided participants with the correct answer at the end of each task. After completing the training tasks, participants moved on to the real tasks. In the real tasks, each participant was asked to complete a total of 15 decision making tasks in the assigned treatment. We offered a base payment of $1.20 and a raffle with $1 bonus if the participant's accuracy was above 85%. The experiment was open to U.S.-based workers only, and each worker could only complete the experiment once. Additionally, we included two attention check questions in the experiment where participants were required to select a pre-specified option, and only the data of those subjects who passed both attention checks was considered valid. After filtering out the inattentive participants, for the income prediction task, we obtained data from 134 participants (Control: 41, Seq: 45, All: 48), while for the recidivism prediction task, we obtained data from 150 participants (Control: 49, Seq: 40, All: 61). The median working time for participants was about 8 minutes, which translates to a median hourly payment of $8.9 per hour. For more details of the experiment and participant demographics, please see the supplemental material.

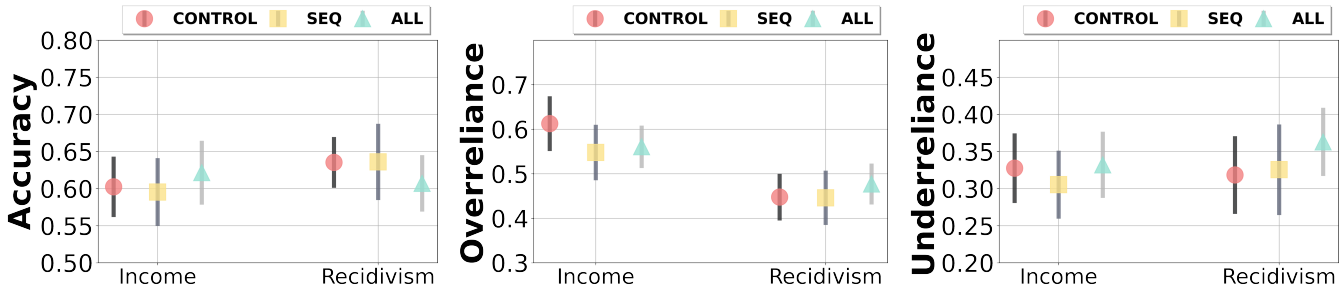
3.5 Experimental Results

Following previous work [42, 101], we used participants' *decision accuracy* to measure the human-AI team performance in decision making, while *underreliance* and *overreliance* were used to quantify the degree to which participants' reliance on the AI model is appropriate. Overreliance refers to the fraction of tasks in which the participant's decision was the *same* as the AI model's decision among all tasks where the AI model's decision was incorrect. Underreliance refers to the fraction of tasks in which the participant's decision was *different* from the AI model's decision among all tasks where the AI model's decision was correct. *Lower* overreliance and underreliance indicate that participants' reliance on AI is more appropriate.

Figure 2 shows the comparisons of participants' decision accuracy, overreliance, and underreliance on the AI model across the three treatments for both the income prediction and the recidivism prediction tasks. We found that compared to the Control treatment where participants did not receive any LLM-powered analysis, incorporating LLM-powered analyses in AI-assisted decision making does not appear to significantly change participants' decision accuracy or reliance on AI, no matter how they are presented (i.e., sequentially or concurrently). Our one-way ANOVA test results further confirmed that the differences in accuracy, overreliance, and underreliance across the three treatments are not significant at the level of $p = 0.05$ for both types of decision making tasks. In other words, the ways that human decision makers interact with the LLM-powered analysis in both the All and Seq treatments may still be not effective for them to critically reflect on the task and calibrate their reliance on the AI recommendation. For example, in the All treatment, the sheer volume of information that people need to process may cause a significant cognitive burden, and make it challenging for people to grasp the essential insights from all the information. Meanwhile, in the Seq treatment, although the LLM-powered analysis is presented sequentially to enable decision makers to digest and reflect on each analysis, the random order in which the analysis is presented may imply a miss of opportunity to help decision makers prioritize the most crucial information needed for correct decisions.

4 ALGORITHMIC SELECTION OF LLM-POWERED ANALYSIS IN AI-ASSISTED DECISION MAKING

Results of our experimental study suggest that in AI-assisted decision making, although LLMs possess the analytical ability to generate additional information to assist humans, the current ways that humans interact with them are not yet optimal. This suboptimal interaction makes it difficult for humans to effectively utilize the information provided by the LLM, hindering their ability to identify essential insights and make informed decisions. Given these



(a) Accuracy (b) Overreliance (c) Underreliance

Figure 2: Comparing the average decision accuracy, overreliance, and underreliance on the AI model for participants across the Control, Seq, and All treatments, for both the income prediction and the recidivism prediction tasks. Error bars represent the 95% confidence intervals of the mean values.

challenges, a natural question arises: *How can we enhance the interaction between humans and the LLM to help humans better utilize the analysis provided by the LLM in AI-assisted decision making?* To answer this question, we propose an algorithmic framework that dynamically and strategically selects the most useful LLM-powered analysis to present to human decision-makers, aiming to help them rely on the AI model more appropriately and make correct decisions.

4.1 Modeling the Effects of LLM-powered Analysis on Human Decision

To enable the optimal selection of the LLM-powered analysis, we start by quantitatively characterizing how the presentation of LLM-powered analysis impacts humans' decision making in an AI-assisted task. Specifically, consider a human who needs to complete a decision making task with the aid of an AI model. The human is initially provided with the task $\boldsymbol{x} \in \mathcal{X}$ and the AI model's decision recommendation $y^m \in \mathcal{Y}$. Subsequently, the human interacts with the LLM over several rounds to obtain analyses of the task features. In each interaction round t ($1 \le t \le T$)[2], the human receives a LLM-powered analysis $\boldsymbol{r}^t \in \mathcal{R}^t = \mathcal{R} \setminus \{\boldsymbol{r}^1, \ldots, \boldsymbol{r}^{t-1}\}$ from the LLM, where $\mathcal{R}$ is the entire set of analyses generated for the task $\boldsymbol{x}$ across all task features. The human can then reflect on this analysis $\boldsymbol{r}^t$ and indicate their attitude towards it by selecting an option $a^t \in \{\text{agree}, \text{disagree}\}$. After T rounds of interaction with the LLM, the human makes a final decision $y^h \in \mathcal{Y}$. To quantitatively model the effects of these LLM analyses on the human's final decision, we begin by expressing the probability of the final decision given the sequence of interactions as:

$$\mathcal{P}(y^h|\boldsymbol{x}, y^m, \boldsymbol{r}^1, a^1, ..., \boldsymbol{r}^T, a^T) = \int \mathcal{P}(y^h|\boldsymbol{h}^T)\mathcal{P}(\boldsymbol{h}^T|\boldsymbol{x}, y^m, \boldsymbol{r}^1, a^1, ..., \boldsymbol{r}^T, a^T)d\boldsymbol{h}^T \quad (1)$$

where $\boldsymbol{h}^T$ reflects the human's *hidden state* at interaction round T. Without loss of generality, we assume that human's hidden state in any round t (i.e., $\boldsymbol{h}^t$) is only dependent on the previous hidden state $\boldsymbol{h}^{t-1}$, the LLM-powered analysis presented in the current round (i.e., $\boldsymbol{r}^t$), and the human's reaction to this analysis (i.e., a^t). Thus, we can decompose the above probability as follows:

$$\mathcal{P}(y^h|\boldsymbol{x}, y^m, \boldsymbol{r}^1, a^1, \ldots, \boldsymbol{r}^T, a^T) = \int \underbrace{\mathcal{P}(\boldsymbol{h}^0|\boldsymbol{x}, y^m)}_{\text{Initial State Mapping}} \underbrace{\left(\prod_{t=1}^{T} \mathcal{P}(\boldsymbol{h}^t|\boldsymbol{h}^{t-1}, \boldsymbol{r}^t, a^t)\right)}_{\text{Hidden State Updating}} \underbrace{\mathcal{P}(y^h|\boldsymbol{h}^T)}_{\text{Human Final Decision}} d\boldsymbol{h}^0 \cdots d\boldsymbol{h}^T \quad (2)$$

Based on this decomposition, our behavior model characterizing how the human's decision is influenced by the LLM-powered analysis consists of three components (see Figure 3 for a graphical illustration):

1. Initial State Mapping: This component captures the human decision maker's initial hidden state $\boldsymbol{h}^0$, before they receive any analysis from the LLM. As shown in Figure 3A, we assume that the initial hidden state $\boldsymbol{h}^0$ is only influenced by the task instance $\boldsymbol{x}$ and the AI model's recommendation y^m, and a model parameterized by $\boldsymbol{\theta}_{\text{init}}$ can be learned to characterize the conditional probability distribution of $\boldsymbol{h}^0$:

$$\boldsymbol{h}^0 \sim \mathcal{P}(\boldsymbol{h}_0|\boldsymbol{x}, y^m; \boldsymbol{\theta}_{\text{init}}) \quad (3)$$

2. Hidden State Updating: This component characterizes how the human decision maker's hidden state evolves over time as they interact with the LLM, i.e., seeing the LLM-powered analysis in each interaction round, for which they may or may not agree with. As shown in Figure 3B, the hidden state $\boldsymbol{h}^t$ in the t-th round is decided by the previous hidden state $\boldsymbol{h}^{t-1}$, the LLM-powered analysis presented in the current round $\boldsymbol{r}^t$, and the human's reaction to it a^t. A model parameterized by $\boldsymbol{\theta}_{\text{update}}$ can be learned to characterize the conditional probability distribution of $\boldsymbol{h}^t$:

$$\boldsymbol{h}^t \sim \mathcal{P}(\boldsymbol{h}^t|\boldsymbol{h}^{t-1}, \boldsymbol{r}^t, a^t; \boldsymbol{\theta}_{\text{update}}) \quad (4)$$

Note that the current hidden state $\boldsymbol{h}^t$ encapsulates the cumulative information gathered through all previous human interactions with the LLM. It achieves this by iteratively encoding the LLM's analysis $\boldsymbol{r}$ and the human reasoning processes (as indicated by human reactions to LLM's analysis $\boldsymbol{a}$) to update the hidden state. Each iteration integrates new insights from the latest LLM analysis and

[2] T is the maximum rounds of interaction occurred, which varies with the specific decision task and may vary across decision makers.

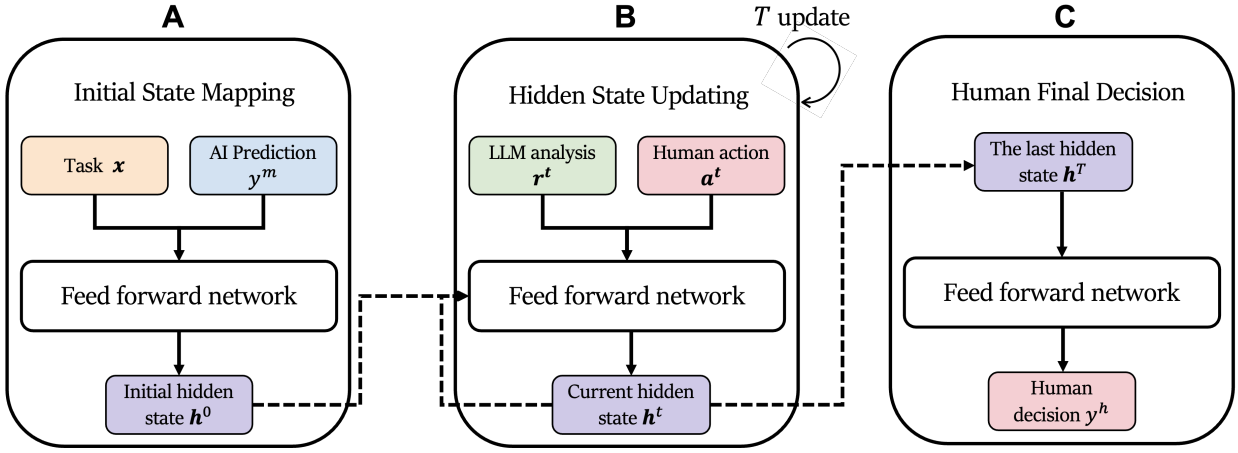


Figure 3: Our human behavior model comprises three components: A) Initial State Mapping: this component encodes the decision making task and AI recommendation into the human's initial hidden state, which serves as the foundational setup to integrate the task details and initial AI insights into human decision making process. B) Hidden State Updating: This component characterizes how the human's hidden state evolves based on the presented LLM-powered analysis and the human's reactions (i.e., whether they agree or disagree with the LLM's analysis). Each update is dependent on the previous hidden state, reflecting the iterative incorporation of new information and human reasoning process into the decision making process. C) Final Decision: This component maps the human's latest hidden state to the actual decision made on the task. It translates the cumulative understanding and reasoning process through the hidden states into the human actual decision outcome.

human responses to reflect humans' evolving understanding of the decision making task.

3. Final Decision: This component maps the human decision maker's last hidden state at the end of the interaction to the final decision they make on the task. As shown in Figure 3C, the final decision y^h is only decided by the last hidden state $\boldsymbol{h}^T$, and a model parameterized by $\boldsymbol{\theta}_{\text{decision}}$ is used to characterize the conditional probability distribution of y^h:

$$y^h \sim \mathcal{P}(y^h|\boldsymbol{h}^T; \boldsymbol{\theta}_{\text{decision}}) \tag{5}$$

With a set of human behavior data indicating how humans react to LLM-powered analysis and then make decisions, i.e., $\mathcal{D} = \{\boldsymbol{x}_i, y_i^m, \{\boldsymbol{r}_i^t, a_i^t\}_{t=1}^T, y_i^h\}_{i=1}^N$, we can use maximum likelihood estimation to learn the behavior model parameters $\Theta = \{\boldsymbol{\theta}_{\text{init}}, \boldsymbol{\theta}_{\text{update}}, \boldsymbol{\theta}_{\text{decision}}\}$.

4.2 Selecting the LLM-powered Analysis

Given a learned model Θ that characterizes the impacts of LLM-powered analyses on humans' decisions, we next explore how to dynamically select the optimal analysis $\boldsymbol{r}^t$ from the set of candidate analysis (i.e., $\mathcal{R}^t$) to maximize human's appropriate reliance on AI models in AI-assisted decision making. To achieve this, we first need to measure the reliability of the AI model's prediction y^m on each task instance $\boldsymbol{x}$. Recent work [32, 86] has proposed methods to leverage the complementary strengths of humans and AI in decision making tasks by combining the human's independent decisions and an AI model's decision recommendations intelligently (e.g., using a Bayesian modeling framework), which often yields more accurate decisions than those made by either the human or the AI model alone. Specifically, given the human's independent decision $y^h_{\text{independent}}$, the AI model's recommendation y^m, and the task instance $\boldsymbol{x}$, these methods learn models to combine $y^h_{\text{independent}}$ and y^m to produce a combined result:

$$y_{\text{combine}} = \text{CombineModel}(y^h_{\text{independent}}, y^m, \boldsymbol{x}) \tag{6}$$

In this study, we adopted the human-AI combination method proposed by Kerrigan et al. [32] to obtain y_{combine}. Since the accuracy of y_{combine} was shown to be higher than either $y^h_{\text{independent}}$ and y^m, we treated y_{combine} as the "target" decision and we selected the LLM-powered analysis $\boldsymbol{r}^t$ in a way to nudge humans into making this target decision[3]. In other words, when $y_{\text{combine}} = y^m$, we selected a LLM-powered analysis to nudge humans towards relying on the AI recommendation; otherwise, we nudged humans towards not relying on the AI recommendation.

To effectively nudge the human towards making the target decision y_{combine}, we first define an *immediate utility function* for evaluating the selection of an analysis $\boldsymbol{r}$ in round t given the human's hidden state by the end of the previous round is $\boldsymbol{h}^{t-1}$. Since our goal is to maximize the probability that the final decision made by the human aligns with the target decision, the utility function $U(\cdot)$ is defined as:

$$U(y_{\text{combine}}|\boldsymbol{r}, \boldsymbol{h}^{t-1}) = \log \frac{\mathbb{E}_{a^t \sim \text{Bern}(0.5)}[\int \mathcal{P}(\boldsymbol{h}^t|\boldsymbol{h}^{t-1}, \boldsymbol{r}, a^t; \boldsymbol{\theta}_{\text{update}}) \mathcal{P}(y^h = y_{\text{combine}}|\boldsymbol{h}^t; \boldsymbol{\theta}_{\text{decision}}) d\boldsymbol{h}^t]}{\mathcal{P}(y^h = y_{\text{combine}}|\boldsymbol{h}^{t-1}; \boldsymbol{\theta}_{\text{decision}})} \tag{7}$$

Here, inside the log term, the numerator represents the probability for the human to make the target decision y_{combine} if they see the analysis $\boldsymbol{r}$ in round t, and are asked to *immediately* make a decision by the end of round t. Since the human's reaction a^t to the analysis $\boldsymbol{r}$ is unknown at this point, when computing this probability, we assume that the human is equally likely to agree or disagree with the

[3]We evaluated various human-AI combination models, and our results showed that the method proposed by Kerrigan et al. [32] generally resulted in combined decisions that outperform AI solo and independent human decisions, as well as other combination methods. The evaluation details can be found in the supplementary material.

analysis, i.e., $a^t \sim \text{Bern}(0.5)$. Moreover, the denominator represents the probability that the human would have made the target decision at the end of the round $t-1$. Intuitively, $U(y_{\text{combine}}|\boldsymbol{r}, \boldsymbol{h}^{t-1})$ reflects the immediate *probability gain* for the human to select the target decision after they are presented with analysis $\boldsymbol{r}$ in round t—when $U(y_{\text{combine}}|\boldsymbol{r}, \boldsymbol{h}^{t-1}) > 0$, it means that presenting $\boldsymbol{r}$ to the human in round t *increases* their chance of selecting the target decision compared to that at the end of previous round; otherwise, the human's probability of selecting the target decision would decrease or remain the same.

Note that when we need to select the analysis to be presented in round t, instead of knowing the human's precise hidden state $\boldsymbol{h}^{t-1}$ at the end of the previous round, we can only recursively estimate a *distribution* of this hidden state using the learned model Θ and the history of past interactions $\{\boldsymbol{x}, y^m, \{\boldsymbol{r}^k, a^k\}_{k=1}^{t-1}\}$. We denote this distribution of the human's hidden state prior to round t as the "*state belief*" $\mathcal{B}(t)$:

$$\mathcal{B}(t) \propto \mathbb{E}_{\boldsymbol{h}^{t-2} \sim \mathcal{B}(t-1)} \left[\mathcal{P}(\boldsymbol{h}^{t-1} | \boldsymbol{h}^{t-2}, \boldsymbol{r}^{t-1}, a^{t-1}; \boldsymbol{\theta}_{\text{update}}) \right] \quad \forall t \geq 2 \tag{8}$$

and $\mathcal{B}(1) = \mathcal{P}(\boldsymbol{h}^0 | \boldsymbol{x}, y^m; \boldsymbol{\theta}_{\text{init}})$. Thus, given a state belief $\mathcal{B}(t)$, the *expected immediate utility* for selecting the analysis $\boldsymbol{r}$ in round t is defined as $\rho(\mathcal{B}(t), y_{\text{combine}}, \boldsymbol{r}) = \mathbb{E}_{\boldsymbol{h}^{t-1} \sim \mathcal{B}(t)}[U(y_{\text{combine}}|\boldsymbol{r}, \boldsymbol{h}^{t-1})]$, which represents the *expected probability gain* for the human to select the target decision after they are presented with the analysis $\boldsymbol{r}$ in round t and are asked to immediately make a decision by the end of round t.

However, note that the human does not have to immediately make a decision by the end of round t—instead, we could choose to present more LLM-powered analyses to the human if they can help further increase the human's probability of selecting the target decision y_{combine}. Therefore, to determine the optimal analysis that maximizes the ultimate probability for humans to select y_{combine}, we define a value function V to represent the *maximum expected overall utility* that is achievable from the current state belief $B(t)$ given the set of remaining analyses $\mathcal{R}^t$:

$$V(\mathcal{B}(t), \mathcal{R}^t, y_{\text{combine}}) = \max_{\boldsymbol{r} \in \mathcal{R}^t} g(\mathcal{B}(t), \boldsymbol{r}, y_{\text{combine}}) \tag{9}$$

$$g(\mathcal{B}(t), \boldsymbol{r}, y_{\text{combine}}) = \underbrace{\rho(\mathcal{B}(t), y_{\text{combine}}, \boldsymbol{r})}_{\text{expected immediate utility}} + \underbrace{V(\mathbb{E}_{a^t \sim \text{Bern}(0.5)}[\mathcal{B}(t+1)], \mathcal{R}^t \setminus \{\boldsymbol{r}\}, y_{\text{combine}})}_{\text{maximum expected future utility}}$$

In this definition, $g(\mathcal{B}(t), \boldsymbol{r}, y_{\text{combine}})$ represents the *expected overall utility* that is achievable from round t onward when the state belief prior to round t is $\mathcal{B}(t)$ and the analysis $\boldsymbol{r}$ is presented to the human in round t. It is composed of two parts. The first part is the expected immediate utility $\rho(\mathcal{B}(t), y_{\text{combine}}, \boldsymbol{r})$, which represents the *immediate probability gain* for the human to select the target decision in the t-th round after $\boldsymbol{r}$ is presented. The second part is the maximum expected future utility $V(\mathbb{E}_{a^t \sim \text{Bern}(0.5)}[\mathcal{B}(t+1)], \mathcal{R}^t \setminus \{\boldsymbol{r}\}, y_{\text{combine}})$, which represents the *maximum future probability gain* for the human to select the target decision in the $(t+1)$-th round and beyond if we continue to present the human with the optimal LLM-powered analyses selected from the set $\mathcal{R}^t \setminus \{\boldsymbol{r}\}$, while our state belief prior to the $(t+1)$-th round is $\mathcal{B}(t+1)$. $\mathcal{B}(t+1)$ is updated from $\mathcal{B}(t)$ according to Equation 8, assuming that the human is equally likely to agree or disagree with the analysis $\boldsymbol{r}$ that is presented in the t-th round. Finally, the optimal LLM-powered analysis $\boldsymbol{r}^t \in \mathcal{R}^t$ for round t is selected to maximize $g(\mathcal{B}(t), \boldsymbol{r}, y_{\text{combine}})$, and the expected overall utility associated with this optimal choice of analysis is denoted as $V(\mathcal{B}(t), \mathcal{R}^t, y_{\text{combine}}) = g(\mathcal{B}(t), \boldsymbol{r}^t, y_{\text{combine}})$.

We can iteratively update the value function $V(\mathcal{B}(t), \mathcal{R}^t, y_{\text{combine}})$ until convergence, which yields the optimal policy $\pi(\mathcal{B}(t), \mathcal{R}^t, y_{\text{combine}})$ for selecting the optimal analysis $\boldsymbol{r}^t$ to present in the t-th round :

$$\boldsymbol{r}^t = \pi(\mathcal{B}(t), \mathcal{R}^t, y_{\text{combine}}) = \begin{cases} \text{Not presenting and stop interaction} & \text{if } V(\mathcal{B}(t), \mathcal{R}^t, y_{\text{combine}}) \leq 0 \\ \arg\max_{\boldsymbol{r} \in \mathcal{R}^t} g(\mathcal{B}(t), \boldsymbol{r}, y_{\text{combine}}) & \text{otherwise} \end{cases} \tag{10}$$

If the value function $V(\mathcal{B}(t), \mathcal{R}^t, y_{\text{combine}})$ is less than or equal to zero, it indicates that further interaction with the LLM is not expected to increase the chance for the human to make the target decision. Therefore, we stop presenting LLM analyses and let the human make the final decision. Otherwise, we will present the analysis that maximizes the expected overall utility.

5 EVALUATION OF ALGORITHMIC FRAMEWORK

In this section, we explore whether and how our proposed framework, which adaptively presents LLM-powered analysis by estimating the human's hidden state and the effects of LLM-powered analysis on human decisions, can enhance human's decision performance in AI-assisted decision making and calibrate human trust in AI models.

5.1 Operationalizing the Algorithmic Framework

We operationalized our proposed algorithmic framework in Section 4 in the context of AI-assisted income prediction and recidivism prediction tasks. Specifically, we utilized the data collected in Section 3 under the SEQ treatment to learn parameters Θ of the human behavior models for both types of decision making tasks. The behavior models are optimized using Adam [37] with an initial learning rate of $1e-4$ and a batch size of each training iteration of 128. The number of training epochs is set as 15. The 5-fold cross validation on the collected data shows that the average accuracy of the learned models in predicting humans' decisions under the assistance of AI recommendations and LLM-powered analysis is 0.74 for income prediction and 0.71 for recidivism prediction, respectively. To enable the use of the human-AI combination method [32] to infer the target decision for each decision making task, we also conducted a pilot study collecting humans' independent judgments on various income prediction and recidivism prediction tasks. Using this pilot data, we trained two models of humans' independent decision making, which achieved an average accuracy of 0.81 and 0.84 for predicting humans' independent judgment in income prediction and recidivism prediction, respectively. Finally, we utilized these learned human behavior models and human independent decision making models to dynamically select the LLM-powered analysis for humans in the following study. For more details related to the algorithm setting, please refer to the supplementary material.

5.2 Experimental Treatments

In addition to the three baseline treatments discussed in Section 3.3 (i.e., Control, Seq, All), we introduced two additional experimental treatments for this phase of evaluation:

- Algorithmic (Alg): In this treatment, participants started working on the decision making task seeing only the task instance and the AI model's recommendation without receiving any LLM-generated analysis. Then participants were required to interact with the LLM, where in each turn, the LLM-powered analysis to be presented was selected based on Equation 10 to nudge the participant towards relying on the AI model's recommendation appropriately.
- Rank: This treatment followed the same experimental procedure as the Seq treatment regarding participants' interaction and decision making processes. However, the Rank treatment differed in how the LLM-powered analysis to be presented was selected: We first used the post-hoc XAI method LIME [73] to generate feature importance scores for each task instance and then ranked all task features based on the absolute values of their importance scores. We then selected the LLM-powered analysis to present according to a decreasing order of the absolute importance score of the corresponding feature (instead of in a random order as done in the Seq treatment). This treatment is designed to examine whether selecting LLM analysis based on our proposed algorithm—which takes into account potential human reactions to such analyses—can enhance human decision making accuracy compared to selection of LLM analysis that is based solely on heuristic feature importance, should it be available.

Finally, as a reference, we also included a Human-Solo treatment where participants completed the decision making tasks on their own *without* receiving either the AI model's recommendation or any LLM-powered analysis.

5.3 Data Collection

Following the experimental procedure described in Section 3.4, we again recruited participants from Prolific to complete AI-assisted income prediction and recidivism prediction tasks in the six treatments. For each participant in the income prediction task, we randomly sampled 15 different tasks from a pool of about 500 task instances, which were different from the instances used in either the Section 3 study or our pilot study (i.e., these task instances have not be used previously for learning human behavior models or human independent decision models). Similarly, in the recidivism task, we also randomly sampled 15 different tasks from a pool of about 200 task instances which were different from the task pool used in the Section 3 study and our pilot study. We offered a base payment of $1.20 and a potential bonus of $1.00 if the participant's decision accuracy was above 85%. We also excluded participants who had previously participated in our study in Section 3 or our pilot study from taking this study. After filtering out inattentive participants, for the income prediction task, we obtained data from 447 participants, while for the recidivism prediction task, we obtained data from 397 participants. The median working time of the participants was 9.3 minutes, which translates to a median hourly pay of $8.3 per hour. For more details of the collected data, see the supplementary material.

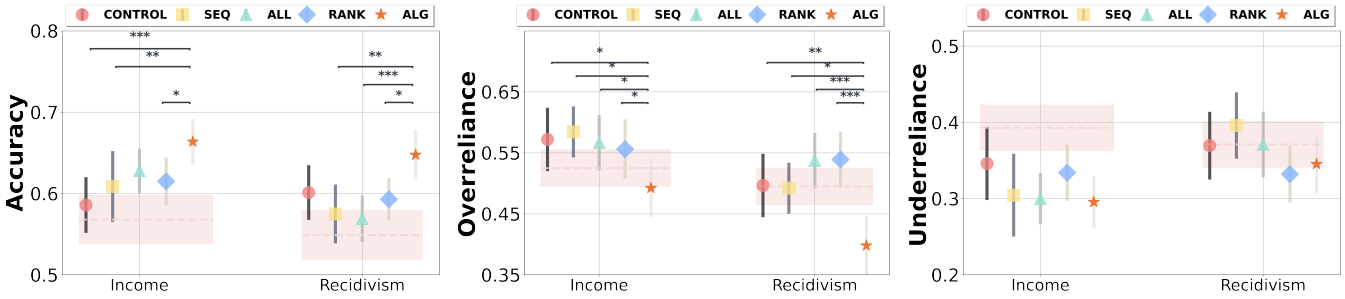
5.4 Experimental Results

Below, we analyze whether our proposed algorithmic framework can help decision makers make more accurate decisions, rely on the AI model's decision recommendation more appropriately, and interact with LLM in an efficient manner.

5.4.1 Comparisons of Decision Accuracy. Figure 4a compares the average decision accuracy of our participants across treatments. Visually, it appears that participants in the Alg treatment achieve the highest decision accuracy among participants in all treatments for both types of tasks.

To examine whether these differences are statistically significant, we conducted regression analyses. Specifically, the primary independent variable of the regressions was the treatment participants were assigned to. The dependent variable was the participants' decision accuracy. To minimize the impact of potential confounding variables, we included a set of covariates in the regression models, such as participants' demographic background (e.g., age, gender, race, and education level), their knowledge of AI models, and the accuracy of the AI recommendation they received in the tasks. Our regression results indicate that our proposed algorithmic framework can significantly improve humans' decision making accuracy in both the income prediction and recidivism prediction tasks. Specifically, in the income prediction task, participants in the Alg treatment achieved significantly higher accuracy compared to participants in the Control ($p < 0.001$), Seq ($p = 0.007$), Rank ($p = 0.041$) and Human-Solo ($p < 0.001$) treatments. Similarly, in the recidivism prediction task, participants in the Alg treatment achieved significantly higher accuracy compared to participants in the Seq ($p = 0.006$), All ($p < 0.001$), Rank ($p = 0.047$) and Human-Solo ($p < 0.001$) treatments.

5.4.2 Comparisons of Appropriate Reliance on AI. Figures 4b and 4c compare participants' overreliance and underreliance on AI across treatments, respectively. For participants in the Human-Solo treatment, despite they did not see the AI model's decision recommendations, we still computed their hypothetical overreliance and underreliance (i.e., computed as if the participant was presented with the AI recommendation on each task) to reflect the natural tendency for participants' independent judgment to agree with an incorrect AI recommendation (Figure 4b) or disagree with a correct AI recommendation (Figure 4c). Here, we again see that participants in the Alg treatment almost always achieve the lowest level of overreliance and underreliance on AI among participants in all treatments. Our regression analyses suggest that for participants in the Alg treatment, the decrease in their overreliance on AI is statistically significant compared to participants in other treatments. For example, in the income prediction task, our proposed framework led to participants' significantly decreased overreliance on AI compared to that of participants in the Control ($p = 0.028$), Seq ($p = 0.011$), All ($p = 0.014$), and Rank ($p = 0.034$) treatments. Similarly, in the recidivism prediction task, our proposed framework significantly decreased overreliance compared to Control ($p = 0.002$), Seq ($p = 0.013$), All ($p < 0.001$), and Rank ($p < 0.001$)



(a) Accuracy (b) Overreliance (c) Underreliance

Figure 4: Comparing the participants' average decision accuracy, overreliance, and underreliance on AI in different treatments for income prediction and recidivism prediction tasks. The pink dashed lines show that for participants in the Human-Solo treatment, (a) the accuracy of their decisions, (b) the frequencies at which their decisions align with AI recommendations (despite not seeing them) when AI recommendations are wrong, and (c) the frequencies at which their decisions differ from AI recommendations (despite not seeing them) when AI recommendations are correct. Error bars (shade) represent the 95% confidence intervals of the mean values. *, **, and *** denote significance levels of 0.05, 0.01, and 0.001, respectively.

Treatment	Income prediction	Recidivism prediction
Seq	4.88 ± 1.79	4.71 ± 1.89
Rank	4.87 ± 1.03	3.72 ± 1.33
Alg	**2.99 ± 1.51**	**2.50 ± 1.19**

Table 2: The mean and standard deviation in the round of interactions between participants and the LLM in the Seq, Rank, and Alg treatments in a single decision making task. According to the results of the ANOVA test, followed by Tukey's HSD test, the number of interaction rounds in the Alg treatment is significantly lower than the number in the other treatments.

treatments, and it even made participants agree with the wrong AI recommendations less than the natural degree of agreement exhibited by participants in the Human-Solo treatment ($p < 0.001$). On the other hand, our regression results suggest that the decrease in participants' underreliance on AI in the Alg treatment was not statistically significant compared to other treatments; the only exception was that on the income prediction task, participants in the Alg treatment disagreed with correct AI recommendations significantly less than the natural degree of disagreement exhibited by participants in the Human-Solo treatment ($p < 0.001$).

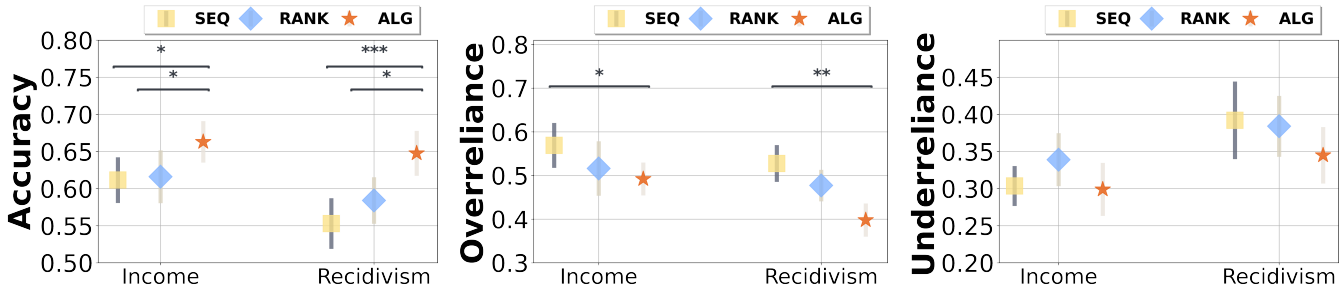
5.4.3 Comparisons of Efficiency of Interactions. Lastly, we looked into whether the proposed algorithm helps human decision makers process the most informative information from the LLM-powered analysis in an efficient manner. First, Table 2 compares the average number of interaction rounds between participants and the LLM in the Seq, Rank, and Alg treatments in a single decision making task. Results of ANOVA tests indicate that the number of interaction rounds is significantly different across treatments for both types of decision making tasks ($p < 0.001$). We then proceed with post-hoc pairwise comparisons using Tukey's HSD tests. We found that, for both the income prediction task and the recidivism prediction task, our proposed approach led to significantly fewer rounds of interactions between participants and the LLM compared to the Rank ($p < 0.001$ for both tasks) and Seq ($p < 0.001$ for both tasks) treatments. This suggests that our proposed algorithm potentially decreased decision makers' cognitive load and helped them make decisions in a time-efficient manner.

In addition, as our proposed algorithm resulted in the highest decision accuracy among all treatments, it is natural to ask if this increase in accuracy was caused by the decreased number of LLM analysis shown to participants, or by the nature of the LLM analysis selected. To gain a deeper understanding on this, we conducted another human-subject experiment with three treatments—Seq, Rank, and Alg—and we controlled the number of interaction rounds in a decision making task in the Seq or Rank treatments to match that experienced by participants in the Alg treatment[4]. For each type of decision making task, we recruited 50 participants for each treatment. Figure 5 compares participants' decision accuracy, overreliance, and underreliance on AI across the three treatments. Again, we found that participants in the Alg treatment achieved significantly higher accuracy compared to participants in the Seq (income prediction: $p = 0.044$, recidivism prediction: $p < 0.001$) and Rank (income prediction: $p = 0.032$, recidivism prediction: $p = 0.012$) treatments. Moreover, we observed that participants in the Alg treatment significantly decreased their overreliance on AI compared to those in the Seq treatment for both the income prediction task ($p = 0.042$) and the recidivism prediction task ($p = 0.004$). This means that the proposed algorithm improved the accuracy of participants' decisions and promoted their appropriate reliance on the AI recommendation primarily as it selected the most informative LLM-powered analysis to be presented to people.

5.5 Exploratory Analyses

Finally, to gain deeper insights into why the proposed algorithm effectively nudged decision makers towards making more accurate decisions and relying on AI recommendations more appropriately,

[4]Based on our results in Table 2, for the income prediction task, we set the number of interaction rounds in a task to be 3. For the recidivism prediction task, we set the number of interaction rounds in a task to be 2 or 3 uniformly randomly.



(a) Accuracy (b) Overreliance (c) Underreliance

Figure 5: Comparing the participants' average decision accuracy, overreliance, and underreliance on AI in different treatments for income prediction and recidivism prediction tasks, when fixing the number of interaction rounds at the same level. Error bars represent the 95% confidence intervals of the mean values. *, **, and *** denote significance levels of 0.05, 0.01, and 0.001, respectively.

Task	Alignment Rate (%)			
	Initial Analysis		All Analyses	
	AI Correct	AI Incorrect	AI Correct	AI Incorrect
Income Prediction	59	25	51	21
Recidivism Prediction	51	38	50	37

Table 3: The alignment rate between the LLM analyses and the AI model recommendations when the AI model's decision recommendation is correct or incorrect for the income prediction and recidivism prediction tasks.

Task	Alignment Rate (%)			
	Initial Analysis		All Analyses	
	AI = Target	AI ≠ Target	AI = Target	AI ≠ Target
Income Prediction	66	5	64	11
Recidivism Prediction	64	23	69	15

Table 4: The alignment rate between the LLM analyses and the AI model recommendation when the AI model's decision recommendation matches or does not match the target decision for the income prediction and recidivism prediction tasks.

we conducted exploratory analyses to understand the nature of the LLM-powered analysis selected by the algorithm.

As the example analysis shown in Table 1, given a decision task, the LLM typically provides its interpretation on how the value of a task feature influences the prediction—the value of a feature could increase, decrease, or has no influence on the likelihood of a certain prediction. The direction of this suggested influence can either align or not align with the AI model's actual recommendation. For example, on an income prediction task, if the LLM suggests the education level of the person in the task decreases the likelihood of them making over $50k per year, and the AI model's prediction on the task is indeed "below $50k", then this analysis "*aligns*" with the AI prediction (i.e., on this task, the person's education level provides *supporting evidence* to the AI model's recommendation). However, if the AI model's prediction on the task is "above $50k", then this analysis does not align with the AI prediction, and the value of the person's education level provides a *contradictory evidence* to the AI model's recommendation.

For all participants in the Alg treatment of our experiment, we analyzed whether each LLM-powered analysis presented to them on a decision task aligned with the AI model's recommendation on that task. In Table 3, we compared the fraction of selected LLM analysis that aligned with AI recommendation for tasks in which the AI recommendation was correct and tasks in which the AI recommendation was wrong, and such comparison was conducted when considering only the first analysis selected by the LLM on each task (see the "Initial analysis" column), or considering all analyses selected by the LLM on each task (see the "All analyses" column). Clearly, for both income prediction and recidivism prediction tasks, we found that LLM analyses that aligned with the AI recommendation were significantly more likely to be selected on tasks where the AI recommendation was correct than on tasks where the AI recommendation was wrong (proportion tests suggest that $p < 0.001$). In other words, when the AI recommendation was correct, our algorithm was more likely to select analysis that provides "supporting evidence" to the AI recommendation, while analysis that provides "contradictory evidence" was more likely to be selected when the AI recommendation was wrong.

Table 4 shows an even larger discrepancy in the alignment rate of the LLM-powered analysis selected when focusing on the comparison between tasks where the target decision was the same as the AI recommendation (hence the algorithm aimed to nudge the

participant towards relying on AI), versus tasks where the target decision was different from the AI recommendation (hence the algorithm aimed to nudge the participant towards not relying on AI). This means that the algorithm primarily presents supporting evidence to humans to nudge them to rely on AI, while primarily presents contradictory evidence to humans to nudge them towards not relying on AI. As a qualitative example, in an income prediction task, suppose the AI model predicts the person's income would be above $50k. If the algorithm aims to increase participants' reliance on this prediction, the top three LLM analysis selected by the algorithm across all decision making tasks are "*With an occupation of professional specialty, this might increase the likelihood of making over $50k per year*", "*With the sex being male, this might increase the likelihood of making over $50k per year*", and "*With the work type being in the private sector, this might increase the likelihood of making over $50k per year*". In contrast, if the algorithm aims to decrease participants' reliance on this prediction, the top three LLM analysis selected by the algorithm across all decision making tasks are "*With an age of X [X is a value that is below median], this might decrease the likelihood of making over $50k per year*", "*With a marital status of divorced, this might decrease the likelihood of making over $50k per year*", and "*With X [X is a value that is below median] years of education, this might decrease the likelihood of making over $50k per year*".

6 DISCUSSIONS

In this paper, via two phases of study, we explore how to effectively incorporate the analytical capabilities of LLMs in AI-assisted decision making to improve human-AI team performance in the absence of AI explanations. Based on our findings, we discuss the potential societal impacts, design implications, and limitations of our study.

6.1 Algorithmic Selection of LLM-powered Analysis Could Be a Double-Edged Sword

In our study, we seek to enhance human-AI team performance in decision making by selectively and progressively presenting LLM-generated analyses that nudge humans towards making decisions that are considered as optimal by a rational integration of human and machine intelligence. This practice demonstrated potential benefits, such as improving the accuracy of human-AI team's decisions and reducing human overreliance on AI models. As we have shown in our study, the integration of carefully selected LLM-powered analyses in AI-assisted decision making can, under controlled conditions, lead to improved decision making performance by augmenting AI recommendations with detailed task analysis and enabling humans to reflect on the AI recommendations in a structured way.

However, our findings also raise concerns about the susceptibility of human behavior to algorithmic selection of the information that humans receive in their decision making. Despite the apparent benefits, the ease with which human decisions can be influenced by algorithmically selected LLM-powered analysis poses notable risks. Our study reveals that it is relatively straightforward to set a nudge direction that subtly manipulates human decision outcomes. This manipulation, while potentially benign and intended to correct for known biases or decision making flaws, could also be maliciously used by adversarial actors to achieve unethical goals. In the context of recidivism prediction, an example of such misuse could involve an adversarial actor manipulating the human decision making process to be unfairly biased against certain groups [50]. By setting an unethical nudge goal, the LLM-powered analysis can be algorithmically presented in a manner that selectively emphasizes certain aspects over others. This selective presentation might influence human judicial decisions, nudging them towards more punitive measures for targeted populations, which reinforces existing societal biases and compromises the fairness.

To counteract the potential adversarial uses of LLM-powered analysis, it is crucial that further research not only focuses on developing and enhancing the capability of AI models to support human decision making, but also on devising strategies to prevent their misuse in manipulating decisions adversely. For instance, to mitigate the risks of adversarial nudges in AI-assisted decision making, strengthening security measures around AI systems like implementing both physical security measures and cybersecurity protocols designed to guard against unauthorized access, hacking, and manipulation is critical. In addition, in our study, the successful nudging of human decisions to improve the human-AI team performance was based upon the accurate modeling of human behavior. This modeling was fundamentally based on empirical human-AI interaction data. As such, protecting this data from misuse is crucial. Strict controls must be in place to ensure that only authorized and well-intentioned parties have access to sensitive interaction data, to prevent the misuse of algorithmic nudges. Finally, implementing continuous monitoring of decisions when humans interact with AI/LLM-powered systems is necessary to detect any unusual patterns in human behavior that may indicate potentially misleading or biased AI information.

6.2 On Determining Nudging Directions through Combining Human Decisions and AI Recommendations

In our proposed framework, a key step is to determine the trustworthiness of the AI recommendation and decide whether to present LLM analysis to nudge human decision makers towards relying on the AI recommendation or not. We did so by leveraging the "human-AI complementarity"—we inferred a "target decision" on each decision making task using existing methods (e.g., [32]) to combine the predicted human's independent decision on the task and the AI model's recommendation on the task, and nudging human decision makers towards making this target decision. While these combination methods could always be used to generate a target decision, the quality of the target decision—to what extent the target decision is more accurate than both human's independent decision and AI's recommendation and therefore provides useful information on the trustworthiness of the AI recommendation—may vary with many factors. For example, the correlation between human and AI decisions was found to be a significant factor that would limit the human-AI complementarity—the more correlated humans' and AI's decisions are, the less likely the combined decision outperforms both human and AI alone [86]. This implies that if the AI model is trained based on historic decisions made by

humans to mimic human decision making, the algorithmic combination of human and AI decisions may not yield target decisions of significantly higher accuracy. Another key influencing factor is the accuracy differences between humans and AI—the larger the accuracy difference, the less likely the combination of human and AI decisions would outperform the decision of the more accurate party [5, 86]. Different combination methods may also yield target decisions with varying levels of accuracy, as each method has its own assumptions when modeling human decisions and AI decisions, which may or may not be valid for a specific decision making task.

As the effectiveness of the combination method may vary with many different factors, in practice, given a particular type of decision making task, we recommend first collecting pilot data on human and AI's decisions on this task. This data would enable the comparison of the performance of various combination methods as well as understanding if the combined decisions show true advantages over the independent decisions of either humans' or AI's. If the accuracies of combined decisions are similar to the more accurate party between the human decision maker and the AI model, instead of triggering additional computational cost to compute the combined decisions, one may consider simply nudging the decision maker to always rely on AI (if AI is more accurate than human) or always not rely on AI (if human is more accurate than AI). However, if the combined decisions are more accurate than both humans' and AI's decisions, one should select the combination method that produces the most accurate combined decisions, or even design new combination algorithms that are tailored to the unique characteristics of human and AI decisions in the current decision making task, thereby producing more accurate combined decisions than existing algorithms.

6.3 On the Potential Misalignment between LLM Analysis and True AI Decision Rationales

As discussed earlier, in our framework, the analysis produced by the LLM on a decision making task does not necessarily align with the actual decision rationale of the AI model (e.g., the random forest models used in this study). Since the LLM is not directly informed of the internal workings of the AI model (as we focus on scenarios where internals of AI models are not accessible in this study), its analysis—generated based on general knowledge about the task—may not capture the specific decision boundaries or feature correlation relationships of the AI model. However, we note that in this study, accurately explaining the AI model's decision rationale is not the primary motivation for including the LLM-powered analysis. Instead, LLM-powered analysis is used to provide a subjective interpretation of the AI model's recommendation and prompt decision makers to engage in critical thinking when there is no access to the actual explanations of the AI model. That is, while the AI model serves as the primary advisor for human decision makers and provides them with the decision recommendation, the LLM serves as the secondary advisor supplementing the primary advisor by providing its own justifications to the primary advisor's recommendation, which allows the decision maker to put the recommendation into context.

We argue that the potential lack of alignment between the LLM-powered analysis and the AI model's true decision rationale may not be a concern in many cases. First, the main motivation for including the LLM-powered analysis in our framework is to encourage decision makers' critical reflection of the decision task as well as the AI recommendation. Even if these analyses deviate from the AI model's true decision rationale, it could still effectively draw decision makers' attention to key features related to the decision, thereby guiding decision makers' independent and more thoughtful evaluation of the recommendation, allowing them to act on it cognitively rather than blindly trusting/not trusting it simply due to the lack of transparency. Second, in many scenarios, there may exist multiple reasoning paths to arrive at the same recommendation, making it less practical to align the LLM analysis with the "true" decision rationale of the AI model, which may not even be well-defined. In fact, even when actual AI explanations can be obtained, established explainable AI methods were often found to have limited fidelity [62], and different methods can provide different explanations for the same decision of the same model [39, 40]. Thus, when the actual AI explanations are not accessible, the LLM-powered analysis could just be viewed as one possible reasoning path to arrive at the AI model's recommendation when having the LLM engage in "perspective taking" to rationalize that recommendation, or it could even be viewed as the LLM's independent (and true) reasoning path when it has to arrive at the AI model's recommendation. The degree to which the LLM's reasoning path looks reasonable may provide critical insights into the validity of the AI recommendation, as the perceived reasonableness of the LLM's reasoning may correlate with the plausibility and robustness of the AI recommendation. Finally, when the ultimate goal is to improve the decision maker's appropriate reliance on the AI model and thus increase their decision accuracy, the exact reasoning behind the recommendation of the AI model might not matter as long as the information provided by the secondary advisor (i.e., the LLM) leads to a better-informed decision. In this sense, compared to the precise content of the LLM-powered analysis, the knowledge about to what extent presenting a LLM analysis will nudge decision makers towards making a desirable target decision is more critical for effectively improving humans' decision accuracy. In our algorithmic framework, this knowledge is captured through our human behavior model.

That said, we acknowledge that the when helping decision makers gain accurate understandings of the internal workings of the AI model is a primary end-goal, the LLM-powered analyses may bring about risks as humans may build an inaccurate mental model of the AI's internal workings based on these analyses. In extreme cases, the LLM-powered analyses may even "sugercoat" incorrect AI recommendations or hide ethical issues underneath the AI model, such as model biases [81]. To address this risk, the ultimate solution is to increase the transparency of the AI model to obtain the actual explanations of the model, and the proposed algorithmic framework could still be used to determine how to present these explanations selectively and progressively. However, without access to actual AI explanations, methods should be designed to increase people's awareness of the potential mismatch between the LLM analysis and the true AI decision rationale. Moreover, one may consult multiple

secondary advisors (e.g., multiple LLMs) to analyze the AI recommendation and triangulate the reasoning process; this may help the human decision makers understand the diversity of possible interpretations of the AI recommendation and reduce the likelihood of being misled by the misinterpretation of any single secondary advisor.

6.4 Design Implications for Human-LLM Interaction

Our study demonstrates that while LLMs can generate and provide informative analysis for human decision makers, how to present this information is critical to its effective utilization. The heuristic design of interactions between humans and LLMs, when not carefully curated, often proves inefficient and fails to achieve the intended positive utility of LLM's analytical capabilities. For example, when decision makers are directly supplied with an abundance of LLM-generated information, the information overload would overwhelm users and potentially result in decision fatigue, making it difficult for users to identify relevant information quickly. In addition, the practice of randomly slicing abundant information into pieces or relying solely on standard importance metrics to guide the presentation of data does not adequately consider the cognitive processes of how humans process such information. Such methods may lead to prolonged interactions between humans and LLMs, which may also overwhelm and confuse users, leading to suboptimal engagement and diminished utility of the LLM outputs.

To mitigate these issues and enhance the practical utility of LLM for users, it is essential to integrate considerations of cognitive and contextual factors into the design of interaction paradigms between humans and LLMs to facilitate more effective and efficient interaction. For example, one important consideration in designing these interaction paradigms is to determine the most valuable information to present to users from the large pool of content that LLMs can generate. Given that LLMs are adept at producing vast quantities of information, ranging from seemingly meaningful to less relevant content, it is crucial to implement intelligent selection strategies to group information based on decision making priorities or estimated human cognitive needs. This may allow the LLM to dynamically adjust to the user's immediate needs and contexts by predicting what information is most pertinent based on user behavior and feedback. Additionally, allowing users to customize the presentation and management of information within the interface can be another promising approach to explore in the future. Customization options might include adjusting the volume, complexity, and format of the information to better align with individual processing styles and needs. Finally, incorporating continuous feedback loops within the interface design is crucial for optimizing the interaction between humans and LLMs. These feedback loops enable users to provide input on the usefulness of the information presented, which can inform and refine the algorithms that select and present data, ensuring that the LLM remains dynamically aligned with user needs and preferences.

6.5 Generalization of Methods and Findings

We acknowledge that our study has a few important limitations regarding the generalizability of our methods and findings. First, our study focused on two specific decision making tasks: income prediction and recidivism prediction. These tasks are widely used in previous research on AI-assisted decision making [4, 21, 90, 96, 101] and feature a tabular data format with an explicit structure of features; the property of these tasks allows us to effectively apply LLM in the analysis and estimate how humans might react to these analyses. The success of our proposed framework in these two different tabular-data-based tasks strengthens our confidence in its potential to generalize to other AI-assisted decision making scenarios involving tabular data. However, applying our framework to decision making tasks with different data types, such as vision-based or text-based decision making tasks, presents additional challenges and requires further adaptation. This is because the image or text data do not contain explicit structured information that is amenable to analysis by the LLM in the same manner as tabular data. One potential solution is to convert these unstructured data types into a structured format that fits our proposed framework. For example, in text-based tasks, one could first use an LLM to extract semantically meaningful information from the text (e.g., text sentiment, key subjects in the text). This extracted information can then be treated as features, similar to how features are handled in tabular tasks, and subsequently input into the LLM for generating analyses. Likewise, for vision-based tasks, one could start by segmenting images into superpixels (i.e., groups of pixels representing visually meaningful entities) [1] or identifying relevant concepts in the images [34]. The presence and absence of certain superpixels and concepts would then serve as the features of the image, enabling LLMs to directly analyze them. The LLM analysis obtained could then be integrated into our framework, allowing it to work with a wider range of decision making tasks. When the transformation of unstructured data into structured formats is required before conducting the LLM-powered analysis, the decision on what features to be extracted from the data can be either made automatically (e.g., by the LLM) or manually by the human decision makers. Thus, how to ensure a comprehensive set of features will be extracted from the unstructured data becomes a critical challenge to be addressed.

Secondly, as previously discussed, successfully nudging humans towards relying on AI models more appropriately hinges on the accurate modeling and prediction of how humans will react to different LLM-powered analyses. However, data on human-AI interaction collected in the past to train such behavioral models may not always align perfectly with current human behavior patterns, leading to potential discrepancies in effectiveness when these models are applied. It is thus essential to continually update the human-AI interaction data. This update process ensures that the models can make predictions that align more accurately with current human behavior patterns.

In addition, in our current framework, we model human decision makers' reactions to LLM-powered analysis on a population level without accounting for the unique characteristics of each individual. In other words, our human behavior model characterizes the behavior of an "average" decision maker. In our study, we found that the effects of the proposed algorithmic approach for selecting LLM-powered analysis in improving participants' final decision accuracy and enhancing their appropriate reliance on AI are robust across subpopulations with diverse demographic backgrounds and

varying levels of AI knowledge, suggesting that modeling an average decision-maker is a reasonable modeling choice. That said, we acknowledge that this average modeling approach may neglect crucial individual differences that significantly affect the dynamics of human-AI interaction in AI-assisted decision making, and may indicate missing opportunities to further improve different individuals' decision performance by accounting for their unique characteristics. Future work could integrate various human characteristics (e.g., a person's intuition or prior knowledge about the task [8], need for cognition [6]) into the human behavior models (i.e., one or more of the three model components—the initial state mapping model, the hidden state updating model, and the final decision model) to further accommodate individual preferences and traits. For instance, a person's competence or confidence in a specific decision task could be a critical moderating factor influencing how they would react to the AI recommendation and LLM-powered analyses. As individuals tend to exhibit low reliance on AI when they are more confident [60], explicitly accounting for human confidence in the behavior models may enable more efficient presentation of LLM-powered analyses (e.g., on tasks where humans are highly confident and the target decision suggests AI is not trustworthy, one may need to present fewer LLM analyses to nudge humans towards the target decision). As another example, a person's inherent tendency to trust AI or LLM systems can also be explicitly accounted for in the human behavior models, which may allow the algorithm to dynamically adjust which and how many LLM-powered analyses to be presented to the human decision makers based on their trust inclination.

Finally, we note that our study was conducted on Prolific, which primarily involved non-expert users in low-stake decision making scenarios. While this setting provided a suitable testbed for the evaluation of the appropriateness of human trust in AI-assisted decision making, we urge caution should be used when generalizing our conclusions to other populations or decision making scenarios. For example, in high-stake decision making scenarios where decision makers may utilize different cognitive strategies and where the consequences of errors are more significant, it is unclear whether the intelligent interaction paradigms we designed for interactions between humans and LLMs will perform equally well. However, we believe that if sufficient human-AI interaction data can be collected in high-stake scenarios to train highly accurate human decision making models, the potential to successfully nudge human decisions even in these critical environments still persists.

6.6 Other Limitations

Our study has a few additional limitations. For example, our LLM-powered analysis mainly relies on the GPT-4 model. Consequently, our results may not generalize to other fine-tuned LLMs that are specifically designed for decision making support in various specialized fields, such as medical LLMs employed in clinical decision making scenarios. The distinct capabilities and pre-designed functionalities of these specialized models could lead to different outcomes in human-LLM collaboration compared to those observed with GPT-4, which has general-purpose capabilities. Furthermore, in our study, we employed random forest models as the AI assistant to provide decision support. The outcomes observed could vary significantly with the use of different AI models with its own set of processing abilities, training datasets, and optimization goals, all of which could potentially influence the effectiveness and reliability of the decision making support provided.

7 CONCLUSION

In this paper, we present an initial exploration of whether and how incorporating LLM-powered analysis can enhance the performance of human-AI teams in AI-assisted decision making, when explanations of the AI recommendations are not easily accessible or available. Through a randomized experiment, we first show that presenting LLM-powered analysis of each feature in decision making tasks, either sequentially or concurrently, does not significantly improve humans' performance in AI-assisted decision making. We then propose an algorithmic framework to characterize the effects of LLM-powered analysis on human decisions and dynamically decide which analysis to present. Our evaluation with human subjects shows that, by following the proposed approach, humans can achieve higher decision accuracy and exhibit reduced overreliance on AI in AI-assisted decision making. Overall, our study provides important experimental evidence regarding the effectiveness of incorporating LLMs in AI-assisted decision making, and how to design intelligent interaction methods between humans and LLMs to fully unlock the potential of LLMs for promoting better human-AI collaboration in decision making.

ACKNOWLEDGMENTS

We thank the support of the National Science Foundation under grant IIS-2229876 and IIS-2340209 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

REFERENCES

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence* 34, 11 (2012), 2274–2282.

[2] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does explainable artificial intelligence improve human decision-making?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6618–6626.

[3] Matin N Ashtiani and Bijan Raahemi. 2023. News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. *Expert Systems with Applications* 217 (2023), 119509.

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

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

[7] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 471 (nov 2022), 23 pages. https://doi.org/10.1145/3555572

[8] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2022. Machine explanations and human understanding. *arXiv preprint arXiv:2202.04092* (2022).

[9] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article

370 (oct 2023), 32 pages. https://doi.org/10.1145/3610219
[10] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924* (2023).
[11] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
[12] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 103–119.
[13] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061* (2021).
[14] Hyo Jin Do, Rachel Ostrand, Justin D Weisz, Casey Dugan, Prasanna Sattigeri, Dennis Wei, Keerthiram Murugesan, and Werner Geyer. 2024. Facilitating Human-LLM Collaboration through Factuality Scores and Source Attributions. *arXiv preprint arXiv:2405.20434* (2024).
[15] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.
[16] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 43–52.
[17] Shi Feng and Jordan Boyd-Graber. 2022. Learning to Explain Selectively: A Case Study on Question Answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 8372–8382. https://doi.org/10.18653/v1/2022.emnlp-main.573
[18] Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. 2024. A Taxonomy for Human-LLM Interaction Modes: An Initial Exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.
[19] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 11, 29 pages. https://doi.org/10.1145/3613904.3642002
[20] Mehdi Ghasemi, Daniel Anvari, Mahshid Atapour, J Stephen Wormith, Keira C Stockdale, and Raymond J Spiteri. 2021. The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism. *Criminal Justice and Behavior* 48, 4 (2021), 518–538.
[21] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.
[22] Luke M. Guerdan, Kenneth Holstein, Zhiwei Steven, and Steven Wu. 2022. Under-reliance or misalignment? How proxy outcomes limit measurement of appropriate reliance in AI-assisted decision-making. https://api.semanticscholar.org/CorpusID:248913428
[23] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To trust or not to trust: How a conversational interface affects trust in a decision support system. In *Proceedings of the ACM Web Conference 2022*. 3531–3540.
[24] Gaole He, Abri Bharos, and Ujwal Gadiraju. 2024. To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems *(HT '24)*. Association for Computing Machinery, New York, NY, USA, 98–105. https://doi.org/10.1145/3648188.3675130
[25] Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*.
[26] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
[27] Zijin Hong, Zheng Yuan, Qinggang Zhang, Hao Chen, Junnan Dong, Feiran Huang, and Xiao Huang. 2024. Next-Generation Database Interfaces: A Survey of LLM-based Text-to-SQL. *arXiv preprint arXiv:2406.08426* (2024).
[28] Qirui Huang, Min Lu, Joel Lanir, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2024. GraphiMind: LLM-centric Interface for Information Graphics Design. *arXiv preprint arXiv:2401.13245* (2024).
[29] Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207* (2023).
[30] Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business horizons* 61, 4 (2018), 577–586.
[31] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T Hancock. 2023. Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–29.
[32] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. 2021. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems* 34 (2021), 4421–4434.
[33] Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled H Alyoubi, and Ahmed S Alfakeeh. 2022. Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing* (2022), 1–24.
[34] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*. PMLR, 2668–2677.
[35] Callie Y Kim, Christine P Lee, and Bilge Mutlu. 2024. Understanding large-language model (llm)-powered human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 371–380.
[36] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 822–835.
[37] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[38] Ron Kohavi. 1996. Census Income. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5GP7S.
[39] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602* (2022).
[40] Vivian Lai, Jon Z Cai, and Chenhao Tan. 2019. Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. *arXiv preprint arXiv:1910.08534* (2019).
[41] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).
[42] Vivian Lai, Han Liu, and Chenhao Tan. 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*. 1–13.
[43] Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. 7, CSCW2, Article 357 (oct 2023), 35 pages. https://doi.org/10.1145/3610206
[44] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–22.
[45] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *Comput. Surveys* 56, 9 (2024), 1–39.
[46] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. How Does the Disclosure of AI Assistance Affect the Perceptions of Writing?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 4849–4868. https://doi.org/10.18653/v1/2024.emnlp-main.279
[47] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–25.
[48] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2023. Modeling human trust and reliance in ai-assisted decision making: A markovian approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 6056–6064.
[49] Zhuoyan Li, Zhuoran Lu, and Ming Yin. 2024. Decoding AI's Nudge: A Unified Framework to Predict Human Behavior in AI-assisted Decision Making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 10083–10091.
[50] Zhuoyan Li and Ming Yin. 2024. Utilizing Human Behavior Modeling to Manipulate Explanations in AI-Assisted Decision Making: The Good, the Bad, and the Scary. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=7XkwzaPMvX
[51] Zhe Liu, Chunyang Chen, Junjie Wang, Mengzhuo Chen, Boyu Wu, Xing Che, Dandan Wang, and Qing Wang. 2024. Make llm a testing expert: Bringing human-like interaction to mobile gui testing via functionality-aware decisions. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.

[52] Zhuoran Lu, Zhuoyan Li, Chun-Wei Chiang, and Ming Yin. 2023. Strategic Adversarial Attacks in AI-assisted Decision Making to Reduce Human Trust and Reliance.. In *IJCAI*. 3020–3028.

[53] Zhuoran Lu, Syed Hasan Amin Mahmoo, Zhuoyan Li, and Ming Yin. 2024. Mix and Match: Characterizing Heterogeneous Human Behavior in AI-assisted Decision Making. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 12. 95–104.

[54] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does More Advice Help? The Effects of Second Opinions in AI-Assisted Decision Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 217 (April 2024), 31 pages. https://doi.org/10.1145/3653708

[55] Zhuoran Lu and Ming Yin. 2021. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[56] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[57] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[58] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[59] Zilin Ma, Yiyang Mei, Yinru Long, Zhaoyuan Su, and Krzysztof Z. Gajos. 2024. Evaluating the Experience of LGBTQ+ People Using Large Language Model Based Chatbots for Mental Health Support. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 872, 15 pages. https://doi.org/10.1145/3613904.3642482

[60] Syed Hasan Amin Mahmood, Zhuoran Lu, and Ming Yin. 2024. Designing behavior-aware AI to improve the human-AI team performance in AI-assisted decision making. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. 3106–3114.

[61] Hasan Mahmud, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. 2022. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change* 175 (2022), 121390.

[62] Miquel Miró-Nicolau, Antoni Jaume-i Capó, and Gabriel Moyà-Alcover. 2024. Assessing fidelity in xai post-hoc techniques: A comparative study with ground truth explanations datasets. *Artificial Intelligence* 335 (2024), 104179.

[63] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 355, 34 pages. https://doi.org/10.1145/3544548.3581225

[64] Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. 1–13.

[65] Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-tau Yih, Sida Wang, and Xi Victoria Lin. 2023. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*. PMLR, 26106–26128.

[66] Jessica Ochmann, Leonard Michels, Sandra Zilker, Verena Tiefenbeck, and Sven Laumer. 2020. The influence of algorithm aversion and anthropomorphic agent design on the acceptance of AI-based job recommendations.. In *ICIS*.

[67] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[68] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 379–396.

[69] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–22.

[70] Amy Rechkemmer and Ming Yin. 2022. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 chi conference on human factors in computing systems*. 1–14.

[71] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[72] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[73] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[74] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 223–233.

[75] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A missing piece in the puzzle: Considering the role of task complexity in human-ai decision making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 215–227.

[76] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[77] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916* (2022).

[78] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[79] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2024. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[80] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020).

[81] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[82] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence* 5, 8 (2023), 873–883.

[83] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.

[84] Aaron Springer and Steve Whittaker. 2020. Progressive disclosure: When, why, and how do users want algorithmic transparency information? *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–32.

[85] Mark Steyvers and Aakriti Kumar. 2023. Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science* (2023), 17456916231181102.

[86] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences* 119, 11 (2022), e2111547119.

[87] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence* (2025), 1–11.

[88] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. Luminate: Structured Generation and Exploration of Design Space with Large Language Models for Human-AI Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 644, 26 pages. https://doi.org/10.1145/3613904.3642400

[89] Siddharth Swaroop, Zana Buçinca, Krzysztof Z Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 138–154.

[90] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[91] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[92] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 699–714.

[93] Caroline Wang, Bin Han, Bhrij Patel, and Cynthia Rudin. 2023. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology* 39, 2 (2023), 519–581.

[94] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–6.

[95] Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, et al. 2024. Weaver: Foundation Models for Creative Writing. *arXiv preprint arXiv:2401.17268* (2024).

[96] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 318–328.

[97] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.

[98] Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. 2024. Human-centric autonomous systems with llms for user command reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 988–994.

[99] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[100] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[101] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 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*. 295–305.

[102] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).