Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations

VALERIE CHEN, Carnegie Mellon University, USA
Q. VERA LIAO, Microsoft Research, Canada
JENNIFER WORTMAN VAUGHAN, Microsoft Research, USA
GAGAN BANSAL, Microsoft Research, USA

AI explanations are often mentioned as a way to improve human-AI decision-making, but empirical studies have not found consistent evidence of explanations' effectiveness and, on the contrary, suggest that they can increase overreliance when the AI system is wrong. While many factors may affect reliance on AI support, one important factor is how decision-makers reconcile their own *intuition*—beliefs or heuristics, based on prior knowledge, experience, or pattern recognition, used to make judgments—with the information provided by the AI system to determine when to override AI predictions. We conduct a think-aloud, mixed-methods study with two explanation types (feature- and example-based) for two prediction tasks to explore how decision-makers' intuition affects their use of AI predictions and explanations, and ultimately their choice of when to rely on AI. Our results identify three types of intuition involved in reasoning about AI predictions and explanations: intuition about the task outcome, features, and AI limitations. Building on these, we summarize three observed pathways for decision-makers to apply their own intuition and override AI predictions. We use these pathways to explain why (1) the feature-based explanations we used did not improve participants' decision outcomes and increased their overreliance on AI, and (2) the example-based explanations we used improved decision-makers' performance over feature-based explanations and helped achieve complementary human-AI performance. Overall, our work identifies directions for further development of AI decision-support systems and explanation methods that help decision-makers effectively apply their intuition to achieve appropriate reliance on AI.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Explainable AI, interpretability, human-AI interaction, decision support

ACM Reference Format:
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 (October 2023), 32 pages. https://doi.org/10.1145/3610219

1 INTRODUCTION

Artificial intelligence (AI) systems—often based on machine learning (ML) models—are increasingly used to support decision-makers, even in high-stakes domains like healthcare and finance. Given the complexity of AI systems, it is often suggested that decision-makers could benefit from access to explanations of their predictions. The hope is that such explanations will help decision-makers reason about when and when not to rely on the AI system's predictions, achieving *appropriate reliance* [4]. However, across many domains, empirical studies of explanations have produced mixed

Authors' addresses: Valerie Chen, valeriechen@cmu.edu, Carnegie Mellon University, USA; Q. Vera Liao, veraliao@microsoft.com, Microsoft Research, Canada; Jennifer Wortman Vaughan, jenn@microsoft.com, Microsoft Research, USA; Gagan Bansal, gaganbansal@microsoft.com, Microsoft Research, USA.

2573-0142/2023/10-ART370
https://doi.org/10.1145/3610219

results [4, 11, 49, 53, 58, 68, 82, 87], with some suggesting that explanations increase decision-makers' tendency to rely on AI predictions even when the AI system is wrong [4, 68, 82, 87]—a phenomenon often referred to as *overreliance* [10, 78].

To design explanations that better support appropriate reliance, we must first understand the process through which decision-makers determine whether to rely on AI. Prior studies have typically focused on decision outcomes alone by measuring aggregate performance rather than studying decision-making processes [51], limiting our ability to make generalizable recommendations on how and when explanations can help. One particular blind spot in the existing literature is around the role of a decision-maker's own *intuition* and how it is integrated with AI predictions and explanations during the decision-making process.

While various definitions of "intuition" exist [34, 71, 74], we follow recent work on human-AI decision-making [19, 67] and use the term to broadly refer to beliefs or heuristics that people bring into the decision based on their domain knowledge, experience, instinct, or pattern recognition. For example, a radiologist looking at an X-ray knows where to look and what to look for, and may form a quick opinion about the diagnosis. When given an AI system to support their diagnosis, this intuition may be in agreement or disagreement with the information provided by the AI system. Prior work has suggested that decision-makers' confidence in their own intuition affects reliance on AI [59] and that their intuition about what features matter to a decision can facilitate the detection of model errors from AI explanations [19]. However, little is known about the process through which this happens and how it depends on the type of explanation used.

In this work, we take a bottom-up approach, using a think-aloud protocol to investigate participants' decision-making process with AI predictions and explanations. Our goal is to more holistically understand what types of intuition are involved in engaging with AI decision-support systems and the interplay between these types of intuition and different types of explanations. In light of the well-known pitfall that AI explanations can increase overreliance [4, 68, 82, 87], we take a particular interest in decision-makers' paths to *non-reliance* (i.e., overriding the AI prediction) when the AI system is incorrect. Specifically, we conduct a mixed-methods study (N=26) to investigate how participants make decisions with two common types of explanations: feature- and example-based. Feature-based explanations consist of scores or weights that describe the extent to which each feature of a decision instance contributed to the model's prediction. Example-based explanations support case-based reasoning by providing examples of similar instances and the model's prediction (and often also their ground truth labels [51]). We study these explanations in the context of two decision tasks based on different types of data (tabular and text).

Our exploratory quantitative analysis shows that, consistent with prior work [4, 82, 87], feature-based explanations did not improve participants' decision outcomes compared to their performance without AI support, and led to overreliance on incorrect AI predictions. In contrast, example-based explanations reduced overreliance and helped achieve human-AI *complementary performance*—outperforming human or AI alone. We analyze participants' think-aloud data during the decision tasks to explain these observations. In summary, our work makes the following contributions:

- *Contributing to a fundamental understanding of people's decision-making process with AI predictions and explanations:* We leverage a think-aloud protocol and mixed-methods study to investigate this process. Our analysis highlights three types of intuition that decision-makers apply to override the AI prediction when they believe it is wrong, henceforth referred to as *intuition-driven pathways*: (1) strong intuition about the decision *outcome* that disagrees with the AI prediction; (2) intuition about *features* which they use to reason about explanations and identify evidence that discredits the AI prediction; and (3) intuition about *AI limitations* that they use to infer *prediction unreliability*.

- *Demonstrating and explaining the effectiveness of example-based explanations for decision support:* Our experiment looks beyond the commonly studied feature-based explanations. We find them less effective in supporting decision-making than an example-based explanation that presents the AI prediction and ground-truth labels for two similar training examples. Using the set of intuition-driven pathways, we identify the benefits of example-based explanations, including less disruption of and more compatibility with people's natural intuition formation process, supporting inductive reasoning with additional context to form new intuition about features, and appropriately signaling prediction unreliability.
- *Design implications for AI decision-support systems:* Based on these findings and participants' post-study interviews, we suggest design recommendations for explainable AI methods and AI decision-support tools more broadly that better accommodate these intuition-driven pathways to support appropriate reliance on AI.

2 RELATED WORK AND RESEARCH QUESTIONS

First, we overview related work on explainable AI (XAI), focusing on XAI for ML models—what is sometimes referred to as *interpretability* in the ML literature [27, 57]. We then overview why XAI is believed to be useful for human-AI decision-making and gaps in the community's understanding.

2.1 Overview of XAI

Given the increasing use of AI and ML systems, there is a growing need for people interacting with these systems to understand the underlying models. The technical field of explainable AI (XAI) grew out of these concerns. A diverse set of XAI methods have been proposed that surface technical details of ML models, as surveyed by several authors [1, 2, 18, 35, 38]. XAI methods can be broadly grouped into two categories: inherently interpretable models that are thought to be intuitive to understand (e.g., rule-based models and linear regressions), and post-hoc techniques that generate explanations for complex models like deep neural networks and random forests. Explanations can also be classified as global or local [1, 38]. Global explanations summarize a model's overall behavior, while local explanations shed light on the model's behavior for a particular instance.

Our study focuses on two types of local, post-hoc explanations: feature-based and example-based. Feature-based explanations assign a value to each feature that is generally interpreted as its "contribution" to a given prediction. Popular algorithms to generate feature-based explanations include LIME [70], SHAP [60], GradCAM [73], and Integrated Gradients [76]. Example-based explanations typically select "representative samples" from the training set, with two common approaches to selecting these samples. One is to select prototypes that are representative of a prediction class to help people understand why the model predicts the current instance belongs to that class [11, 13, 41, 49]. The other is to select examples that resemble the current instance and show the AI predictions (and ground truth) on those examples. This approach helps people understand not only how the model makes decisions but also how it might make mistakes [8, 26, 53].

Despite the rapid development of XAI techniques, there remain open questions about what these methods are useful for [21, 57]. Additionally, researchers in the AI, HCI, and CSCW communities have called for more human-centered approaches [29, 55, 81, 83] to investigate what people need and how they interact with AI explanations in specific use cases. Common use cases of AI explanations include supporting model debugging, assisting decision-making, auditing models, and knowledge discovery [2, 19, 56, 77]. Aligning with this line of work, we study how people interact with explanations, focusing on the use case of assisting decision-making.

2.2 Human-AI Decision-Making

Human-AI decision-making, also referred to as AI-assisted decision-making, broadly encompasses set-ups where an ML model is used to help users to make a final judgment or decision [23, 31, 36, 51]—often considered as a form of collaboration between human and AI system. While AI assistance typically provides an ML model's predictions, there is growing work studying whether additional information—performance measures [85], explanations [4, 82], or information about prediction uncertainty [69, 87], for example—can further improve decision outcomes [51]. Among other goals, a common interest is to study what form of AI assistance can help decision-makers outperform both human and AI alone to achieve human-AI *complementary performance* [58, 87]. However, many empirical studies on human-AI decision-making did not observe complementary performance [6, 16, 37, 52, 58, 61, 68, 82, 84, 87]. To achieve complementary performance, it is important to encourage *appropriate reliance* [4, 15, 59, 82, 87]—following the AI system when it is likely to be correct and not following it when it is likely to be wrong.

2.2.1 Can XAI methods improve appropriate reliance? Prior work has asked whether providing decision-makers with explanations of an AI system's predictions can improve appropriate reliance, since decision-makers might be less likely to follow an AI prediction if an explanation suggests flawed model reasoning [5, 12, 17]. As surveyed by Lai et al. [51], the majority of prior empirical studies of AI decision support have focused on feature-based explanations, with a limited set of studies on other explanation types like example-based explanations, rule-based explanations, and counterfactual explanations. These studies have been conducted across various decision tasks with different data types including text data [4, 42, 53, 58], tabular data [68, 82, 87], and image data [11, 49].

Unfortunately, empirical studies have generally failed to confirm that providing explanations can improve appropriate reliance. On the contrary, several studies found that showing feature-based explanations increases people's tendency to *over-rely* on the model when it is wrong, compared to showing only the AI predictions [4, 68, 82, 87]. Wang and Yin [82] found that example-based explanations underperform feature-based explanations for improving appropriate reliance, though other studies suggest example-based explanations are better at helping people detect model errors [11, 13, 49], usually by recognizing the dissimilarity between the instance and selected examples. However, we note that these studies utilized different kinds of example-based explanations and investigated different tasks with different types of data (tabular, text, and image). Our work aims to understand the underlying causes of overreliance with different kinds of explanations. We return to these studies to interpret their mixed results in the context of our findings in Section 5.1.

Initial efforts to explain why AI explanations increase overreliance have focused on the role of cognitive engagement. The hypothesis is that participants (often recruited on crowdsourcing platforms) over-rely on the AI system because they do not deeply engage with the explanations. The dual processing model has been cited as a useful framework [11, 32, 46, 54, 55]: instead of engaging in analytical reasoning with the explanations (system 2 thinking), people may invoke heuristics to make a quick judgment (system 1 thinking), including the judgment of "just deferring to AI."[1] A recent study [28] also suggested that people often invoke positive heuristics that superficially associate AI being explainable with it being trustworthy, which can lead to overreliance. Buçinca et al. [10] showed that incorporating cognitive forcing functions, which aim at deepening cognitive engagement (e.g., slowing people down), can improve appropriate reliance with explanations,

[1]Note the two systems reflect the depth of reasoning instead of content. Heuristics can also be used as part of system 2 thinking [20, 66]. While our investigation of intuition includes heuristics based on domain knowledge (which differs from a superficial heuristic to defer to AI), we do not claim whether participants engaged *only* in system 1 or system 2 thinking. However, the think-aloud setup might have forced participants to be relatively more engaged.

but at the cost of worsened subjective experience from more effortful interactions. A recent work by Vasconcelos et al. [78] further elucidates this lack of cognitive engagement through a cost-benefit framework, which shows that people strategically choose to engage cognitively with explanations by weighing the costs of engaging against simply deferring to the AI system. This work suggests that a fundamental issue of current XAI techniques is that they are too effortful to verify, thus discouraging cognitive engagement.

2.2.2 What is the role of human intuition on reliance? Even if a decision-maker is motivated to engage with explanations, *how* do they decide to rely or not rely on the AI system? This is the overarching question that motivates our study. Most relevant to our study is a theoretical work by Chen et al. [19], which highlights the role of *human intuition* in reasoning about explanations. They propose a conceptual framework that distinguishes between model decision boundaries (how the model makes decisions using features), and task decision boundaries (how the decision *should* be made). Their framework points out that popular feature-based explanations only facilitate understanding of the former. Yet, in order to detect model errors, decision-makers must apply their (correct) intuition about the task decision boundaries and contrast them with the model decision boundaries revealed by the explanation. The authors suggest two types of human intuition useful for detecting model errors—intuition about feature *relevance* and intuition about feature *mechanism* (e.g., weight), which align with prior findings about how experts critique model explanations in interactive ML settings [33, 75]. To our knowledge, beyond this theoretical work, there have not been empirical investigations into what types of human intuition are involved in human-AI decision-making.

While we follow Chen et al. [19] and use the term "intuition" to broadly refer to beliefs and heuristics people bring to the decision, we note that many definitions of intuition exist in different disciplines [34, 71]. One hallmark of intuition is the reliance on knowledge and beliefs stored inside one's cognition, rather than applying formal reasoning to a complete set of information. Intuition has therefore been widely studied for expert decision-making [72] and decision-making under uncertainty [40]. While intuition is often used interchangeably with terms like heuristics, insights, and instinct, literature surveys on this topic acknowledge that there are no agreed-upon sources of intuition [72, 74]. Indeed, sources span domain expertise, experience, associative memory, pattern recognition, emotional and affective awareness, and more. In this work, we focus on investigating *what* types of intuition participants bring to make decisions, without drawing conclusions on *how* they are generated (which is not permitted by our think-aloud method).

A source of human intuition that has been studied in the context of human-AI decision-making is one's domain expertise about the decision task. One line of work explored whether and how decision-makers with more domain expertise may benefit from AI decision support differently. For example, people with higher overall performance on their own (a proxy for more domain knowledge) tend to achieve higher performance when AI is introduced [44, 58]. Higher confidence in one's own decisions compared to novices could also impact interactions with AI support [59]. For example, Cheng and Chouldechova [22] found that less experienced child welfare caseworkers tend to make decisions more closely following algorithmic risk scores, while senior workers tend to engage in further screening on their own.

2.3 Research Questions

In summary, our work contributes to the growing area of research on XAI for decision support by studying the role of human intuition when decision-makers interact with such systems. To understand why prior studies have found that XAI support is ineffective, and even increases overreliance, we move beyond quantitatively studying decision outcomes to investigating the

decision-making process. Different from qualitative studies that aim to understand the experience of decision-makers retrospectively [14, 45, 47, 65, 80], we adopt a think-aloud protocol to investigate the real-time process as participants engage with the AI output [43, 86], and use the data to understand why two types of explanations—feature-based and example-based—improve or inhibit appropriate reliance. We focus on the following research questions:

- **RQ1:** What types of human intuition are involved in engaging with AI predictions and explanations, and how do they affect reliance on AI?
- **RQ2:** Does human intuition come into play differently with feature- and example-based explanations, and do these differences explain their different effects (if any) on decision accuracy and appropriate reliance?

In addition, we explore participants' subjective experiences with the two types of AI explanations. Our aim is to inform the future development of XAI techniques and broader approaches to provide better AI support for decision-making.

3 METHODS

We describe the set-up of our study and analysis, the two prediction tasks that we asked participants to engage with, and the types of explanations they were shown. We then overview our experimental design and study procedure and discuss the approaches used to analyze the data collected during our study. We note that participation was voluntary and the study was IRB approved.

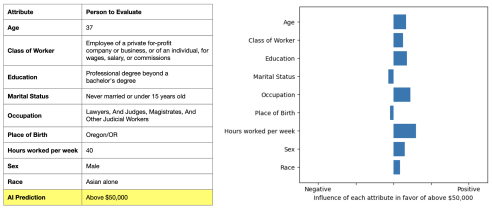
3.1 Participants

Since our research questions do not target any specific population, we chose to start with a convenience sampling strategy and then diversify our selection from a large pool of sign-ups based on their background. Specifically, we advertised our study on Twitter and internal message boards within a large, international technology company to target a variety of communities including researchers, ML engineers, and cross-functional teams. We limited participation to people located in the US since one of our tasks required participants to make judgments about US salaries. In the sign-up form, we inquired about education level, job role, and self-reported knowledge about ML and XAI. 237 people signed up, and we selected 26 participants by diversifying on ML and XAI backgrounds.

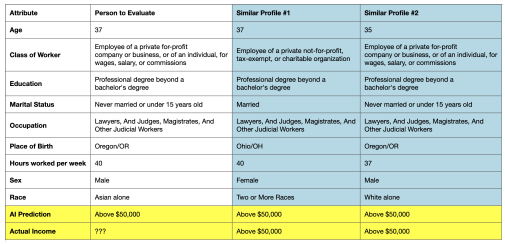
Our participants were from both industry and academia. 73% had graduate degrees, while the remaining 27% had college degrees. The common roles participants held were software engineer, data scientist, research engineer/scientist, and Ph.D. student. In terms of self-reported experience with ML or AI tools, 4% had no experience, 23% had limited experience, 31% had used them often, and 42% consider themselves experts. In terms of self-reported experience with XAI, 15% had no experience, 15% had limited experience, 54% had used them often, and 16% considered themselves experts. Appendix A contains more details on participants. Despite our effort to diversify, our participants were skewed toward an ML-experienced and highly educated population. We acknowledge that this is a potential limitation of our study. However, we note that our primary focus is a qualitative understanding of the role of intuition, rather than quantifying its effect or distribution. Furthermore, during exploratory analysis, we tested the effect of participants' background attributes gathered in the sign-up form and did not find any significant effect. Thus, we proceeded with analyzing all participants' data together.

3.2 Prediction Tasks

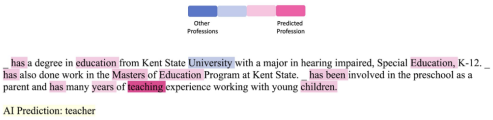
To select two prediction tasks for the user study, we considered the following four criteria: (1) the tasks should not require specialized expertise, but rather the general population should be able to



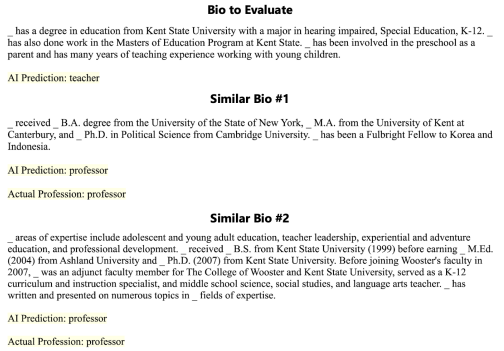
(a) Feature-based explanation for income prediction.



(b) Example-based explanation for income prediction.



(c) Feature-based explanation for biography classification.



(d) Example-based explanation for biography classification.

Fig. 1. Examples of the feature- and example-based explanations (columns) for each of the two prediction tasks of income prediction and biography classification (rows).

reason about and perform the tasks; (2) the tasks should be difficult enough that participants cannot obtain high accuracy without AI assistance; (3) each individual decision should take no more than a couple of minutes to make; and (4) the selected tasks should have different data modalities (e.g., tabular and text) so that we can better understand the generalizability of our results. Using these criteria, we selected two tasks: income prediction and biography classification. Figure 1 shows examples of both tasks.

Income Prediction. In this task, the participant judges whether an individual's income is more than $50,000 USD given a profile that includes the individual's age, class of work (e.g., employed at a private, for-profit company), education level, marital status, occupation, place of birth, number of hours worked per week, sex, and race. The cut-off of $50,000 is close to the median income in the US at the time the data was generated. The AI system used in this task is a random forest model that we trained using data from the Folktables dataset [25] of individuals sampled from the 2018 US Census data. This model achieved an accuracy of 80.8% on a held-out test set. As shown in Figure 1 (a), participants were given an individual's profile displayed in a table with the AI prediction (when applicable) highlighted at the bottom.

Biography classification. In this task, the participant is asked to guess an individual's profession based on the individual's online biography. The AI system used in this task was trained on the BIOS dataset [24], which contains hundreds of thousands of online biographies from the CommonCrawl corpus. While the original dataset consisted of individuals with 29 professions, we simplified the task by narrowing it down to five professions, following Liu et al. [58]: psychologist, physician,

surgeon, teacher, and professor. We embedded each biography using a bag-of-words approach and trained a random forest model on this embedding. This model achieved an accuracy of 75.4% on a held-out test set. As shown in Figure 1 (c), participants were given the full text of the individual's biography with the AI prediction (when applicable) displayed below the text.

3.3 Explanation Methods

Participants were shown two types of explanations: feature-based and example-based.

Feature-based explanations. We generated feature-based explanations using LIME [70], a post-hoc explanation technique.[2] LIME works by computing a simple linear approximation of the model's decision boundary in the local region of the instance to be explained and using the coefficients as a measure of each feature's contribution to the model's prediction. We chose LIME because it is widely used and can be readily applied to both tasks (with tabular and text data). For the income prediction task, we presented the coefficients in a bar chart as in Figure 1 (a). Following Hadash et al. [39], we fix the direction of feature contribution to mean "increasing income level" so features with positive coefficients contributing to a prediction of "above $50,000" have bars on the right and features with negative coefficients to such a prediction have bars to the left. The length of the bar represents the magnitude of the contribution. For the biography classification task, we used a standard approach [62] of color-highlighting the words (features), particularly those with the top 10 highest-magnitude coefficients, as shown in Figure 1 (c). Words with positive coefficients (indicating that they support the model's prediction) are highlighted in red, while words with negative coefficients (indicating that they support other professions) are highlighted in blue, with the shade of the highlight representing the magnitude of the contribution. For very small coefficients, the highlight may not be visible.

Example-based explanations. As discussed in Section 2.1, example-based explanations may include either representative prototypes of the class that the AI system predicts for the given instance or examples that are similar to the given instance along with their AI predictions and ground truth labels. We focus on the latter as they are thought to better help people understand how the model makes mistakes when its prediction differs from the ground truth [51]. While there exist more sophisticated approaches to select examples (e.g., specifically looking for examples with different ground truth labels [53, 82] or using influence functions [50]), we opt for a simple method of selecting the two nearest neighbors from the training set. In the income prediction task, we display the two individuals from the training set with the closest Euclidean distance to the current individual, with categorical features one-hot encoded, as shown in Figure 1 (b). In the biography classification task, we display the two nearest neighbors with the smallest Euclidean distance to the given biography using the bag-of-words embedding (i.e., the same feature space used by the underlying model), as shown in Figure 1 (d). If participants used the nearest neighbor's ground truth labels to predict the outcome for the current instance, they would achieve 63% accuracy on the income prediction task and 60% accuracy on the biography classification task. This means that to get a performance boost from example-based explanations, participants would need to perform non-trivial reasoning, beyond simply relying on the ground truth of the nearest neighbors.

3.4 Experimental Design

We used a within-subjects design to study the effect of explanation type. Each participant was randomly assigned to complete either the income prediction task or the biography classification task, but engaged with both explanation types. This allowed us to directly ask participants to

[2]Specifically, we used the open-source implementation of LIME: https://github.com/marcotcr/lime

compare the two types of explanations in the post-study interview. To account for ordering effects, we randomized which explanation type was shown to each participant first. We ensured that half of the participants were assigned to each task and approximately half of those participants saw each explanation type first. Note that we do not intend to make direct comparisons between participants' behavior in the two tasks since the tasks differ along many dimensions (e.g., data modality, domain, ML model accuracy, user interface). However, including both tasks helps us to understand the generalizability of our observations.

Prior works that involved controlled experiments on human-AI decision-making have typically adopted one of two workflows [51]: The participant either saw a prediction from an AI system and then made their own decision, or the participant made their own decision first and then was given an opportunity to update their decision once they were shown the AI prediction. One advantage of the latter workflow is that it allows for studying how a participant's behavior differs for instances where they initially agree or disagree with the AI prediction. It also allows a baseline measure of participants' performance without AI support. The drawback is that it may not resemble common real-world use cases of AI decision-support, and asking participants to state their own decision immediately before seeing the AI system's may inhibit their tendency to follow the AI prediction. To balance these benefits and drawbacks, we first asked participants to complete all instances of the decision task on their own without AI support and then showed participants the same set of instances with AI support (in random order).

Each participant was asked to make predictions on 16 instances selected from a large test bank with stratified sampling—10 from those instances where the model made correct predictions and 6 from those where it made incorrect predictions. This means that while the accuracy of the models on held-out data was higher (80.8% for income prediction and 75.4% for biography classification, as described above), participants' "experienced accuracy" was only 62.5%. We over-sampled instances on which the AI prediction was wrong so that we would be better able to explore whether and how participants would over-rely on the AI system. We chose this sampling strategy rather than training a worse-performing model to begin with because the explanations and errors made by a worse-performing model would not be realistic for an acceptable decision-support AI system.

Our above experimental design was informed and refined based on a pilot study with 5 participants. The goal of the pilot was to understand (1) how many decisions participants could make within 30 minutes and if they experienced decision fatigue after the allotted time, (2) whether the study workflow felt unnatural to participants (e.g., if they were bothered by the repeated tasks), and (3) whether the instructions provided about the task and explanations were easy to follow. Participants in the pilot study (and in the final study) often did not recognize that instances were repeated or could not immediately recall what decisions they made without the AI system, perhaps due to the complexity and quantity of instances they saw, which we believe mitigates the concern about the unnaturalness of our workflow with repeated tasks.

3.5 Procedure

The study was conducted online. Prior to participating, each participant filled out a consent form in which they were asked to consent to their camera and screen being recorded during the study. At the start of the study, participants were asked to enter a video conferencing platform, check their microphone, speaker, and camera, and share their screen with the moderator (the first author). The remainder of the study was divided into four phases, as described below. The study took an average of 45 minutes to finish and each participant was compensated with $35.

Phase 1: Practice with the task. In order to familiarize participants with their decision task (either income prediction or biography classification) and give them a reasonable sense of the task domain,

participants were first asked to make decisions on 5 randomly sampled instances of their task. After each decision, they were provided with feedback on whether they were correct, exposing them to ground truth labels for the task.

Phase 2: Human-alone decision-making. Participants were then asked to make decisions on an additional 16 instances without receiving feedback. As described above, this allowed us to determine which instances the participant initially agreed or disagreed with the AI prediction on. It also let us measure the participant's baseline accuracy on the task with no AI assistance.

Phase 3: AI-supported decision making. In this phase, participants engaged with the two types of explanations (feature-based and example-based) in random order. For each explanation type, the participant was first introduced to the type of explanation with an illustrated example, as shown in Appendix C, and then asked to make 8 decisions while seeing the AI system's prediction and explanation, as shown in Figure 1. The 16 instances shown in this phase were the same 16 instances used in phase 2 (order randomized). As discussed, for each explanation type, we sampled 5 instances on which the AI prediction was correct and 3 instances on which it was incorrect.

During this phase, we asked participants to think aloud by vocalizing their thought process as they looked at each instance, engaged with the provided information, and made their decision. The moderator prompted participants who were less vocal to share their thought processes. (We did not require participants to think aloud in earlier phases of the study to save time and avoid fatigue.)

Phase 4: Post-task interview. We closed the study with a brief interview to understand the participant's experience making decisions with both types of explanations and how they thought the system could be improved. The full set of interview questions is available in Appendix B.

3.6 Analysis Approach

We used a mix of quantitative and qualitative methods to analyze the study data. On the quantitative side, we performed an exploratory analysis to study the effect of explanation type on participants' accuracy (RQ2). We describe the specific analyses we ran when we discuss the results in Section 4.2.

We collected two types of data for qualitative analysis: participants' think-aloud data and their responses from the post-study interviews. We analyzed the think-aloud data in order to answer both RQs. The first author and second author first performed open coding informed by Grounded Theory research [79] on a common set of four participants' data. They iteratively discussed and developed a set of common codes and themes (discussed in Section 4.1), and then each coded the think-aloud data from half of the remaining participants. Specifically, for each decision made by each participant, they coded what happened during the decision process, and recorded the codes in one row in a spreadsheet. In this step, the coders were blind to whether the decision was correct. After all decision tasks were coded in this manner, the authors filled in columns about the correctness of participants' original and final decisions, and the correctness of the AI's predictions. We then conducted a *comparative analysis* by separating cases in which we observed appropriate or inappropriate reliance on the AI system, and contrasted codes for participants using different explanations. This comparative analysis allowed qualitative insights into how these explanations affect reliance differently (RQ2), as discussed in Section 4.2.

We also analyzed the interview data to understand participants' subjective perceptions. The first author followed the interview structure and extracted themes around participants' perception of the two types of explanations and how they wish to improve them, as discussed in Section 4.3.

4 RESULTS

First, we overview themes from the think-aloud data that reflect common elements in participants' decision-making processes with AI support, focusing on the types of intuition participants brought into the process (RQ1). Next, we quantitatively analyze the effect of explanation type on decision accuracy and reliance and use think-aloud data to explore the reasons why the two types of explanations had different effects (RQ2). Lastly, we discuss participants' subjective perceptions of the explanations and the improvements they suggested during the post-study interviews.

4.1 Common Types of Intuition Applied in the Decision-Making Process (RQ1)

We present common themes identified from participants' think-aloud data to answer RQ1: What types of human intuition are involved in engaging with AI predictions and explanations and how do they affect reliance on AI? For each theme, we first give an overview, and then briefly discuss how the two types of explanations differ around the theme (RQ2). We then answer the second part of RQ1 by discussing how these themes suggest what we refer to as *intuition-driven pathways* for decision-makers to override the AI prediction when they believe it is wrong. We further delve into how these pathways explain the different effects of the two types of explanations in Section 4.2.

Intuition about the outcome. Participants commented on their own intuition or "gut feeling" about what the decision outcome should be—whether the person more likely made above or below $50,000 in income prediction task, and which profession the person more likely had in biography classification task—sometimes *before even considering the AI prediction and explanation.* While our experiment setup did not explicitly isolate participants' own intuition (in phase 3) from the impact of AI output, we observed that some individuals intentionally chose to make their own judgment first before attending to AI outputs.[3] Some participants explicitly commented on their "*first reaction*," such as "*it is medicine [related]...it doesn't really sound like a professor [as AI predicts], so I'll read down through [the examples]*" (P31).

Participants' comments also reflected the *strength* of their intuition about the outcome (or confidence [59]), which impacted how they engaged with the AI system. When their intuition about the outcome was strong, participants discounted the AI system's output. For example, seeing an individual with the occupation of a truck driver, P11 said "*I actually feel like this is a case where I have a little background and that really contradicts with the AI [prediction]. If truck drivers are doing private goods hauling, they are actually paid quite a lot because they have a pretty tough job... So I'm actually going to ignore the AI prediction.*" When their strong intuition agreed with the AI system's prediction, participants more readily and quickly made a decision, sometimes even without checking the explanations.

In contrast, participants who acknowledged their intuition about the outcome was weak tended to examine the explanations more closely, hoping to find additional information to help them judge the correctness of AI prediction. If they found no evidence that the AI prediction might be wrong or they simply could not reason about the explanation meaningfully, they tended to defer to the AI prediction. For example, P23 spent a long time looking at the feature-based explanations and said "*Doesn't seem unbelievable to me... I'm a little less familiar with this occupation... and I have a lot more uncertainty around what would be a good [indicator] here. And so I think I'm more comfortable relying on the AI prediction.*"

As illustrated in P11's quote above, participants may form their intuition about the outcome by retrieving prototypes or similar examples from previous experiences, which were often prompted by one or a subset of features that caught their attention, such as a profession they know about. In

[3]All participants of course made their own judgments first during phase 2, but as mentioned above, we did not collect think-aloud data during this phase in order to avoid participant fatigue.

some cases, participants anchored their judgment on exceptional or rare feature values. When P28 noticed a given individual worked 55 hours per week, they immediately said "*I'll go with above [$50,000] just based on the hours worked per week.*" Occasionally, participants were influenced by a similar example that they saw earlier in the practice phase.

How did intuition about the outcome differ across the two types of explanations? Interestingly, we observed that participants were *more likely to acknowledge that they had weak intuition about the outcome with feature-based explanations.* One reason may be that feature-based explanations are more prominently displayed, particularly in the biography classification task where the explanation is visually overlaid on top of the text as highlights, which is typical for explaining text data. For example, P31 complained that the text highlights made them "*think less*" even when "*non-informative*" words were highlighted. P22 said "*my approach with it in this is to scan the pink, scan the blue [instead of] read more. That's what this interface is leading me to do.*" In contrast, example-based explanations allowed participants to easily focus on the current instance first before attending to the explanation, and even to ignore the explanation altogether when they had a strong intuition about the outcome.

Applying intuition about features to reason about explanations. While participants paid attention to prominent features to arrive at their own intuition about the outcome, as described above, a significant portion of the think-aloud data included participants' comments on their intuition about features when *engaging with explanations.* This suggests that showing explanations created an additional step in the decision-making process, i.e., to judge the impact of features that had not been previously considered or to reason more precisely about certain features. For example, even though P9 came to the same prediction as the AI system, seeing the explanation prompted an additional comment "*[the explanation] is really picking up on age a lot. I agree that it is important, but I'm surprised by the magnitude.*" When reasoning about explanations, participants would look for evidence that indicated an incorrect model prediction, but sometimes ended up updating their own intuition about features when doing so.

We first summarize the types of intuition that participants applied when reasoning about explanations, then elaborate on how they were used differently for the two types of explanations.

- *Feature relevance and weights.* Participants most frequently commented on the weight of a feature presented in the explanation such as "*it is saying occupation is a big deal, and I agree...engineering would be a positive*" (P32). With text data, they tended to talk about binary "relevance." Explanations also prompted some participants to comment on the relative weights of multiple features. For example, P1 and P23 believed age should have a higher weight than education in determining income level.
- *Interaction between features.* Participants also reacted to features in relation to each other, bringing in nuanced domain knowledge, such as "*self-employed, not sure if that would necessarily make the most difference because this is a trade occupation*" (P23), or "*adjunct clinical lecturer—the model did not pick [adjunct] up but 'university.' That indicates professor is not the actual profession.*" (P31).
- *Infer additional features.* For income prediction, we observed that participants inferred additional features by combining existing features, such as inferring an individual's "*career stage*" (P26) based on the combination of education, age, and profession, then using that inference to challenge the AI explanation that under-weighted age. This pattern was even more frequent in biography classification, where participants not only utilized combinations of words and phrases, but also higher-level features such as "*this type of art journals*," or "*academic associations*" (P30), as well as meta-features like the writing style and format.

- *Assign granular or different meanings to a feature and judge weights accordingly.* Interestingly, we observe some participants assigned a granular or different meaning to a feature based on their pre-existing beliefs or prototypical cases they could recall. For example, when seeing the occupation "engineers," multiple participants attempted to assign a specific type of engineer to reason about the feature weight, as "*there [are] maybe trigger words in my brain that make me to believe that it's a certain job*" (P32).
- *Assign granular meaning to a label and update feature weights accordingly.* For example, after seeing the AI system predict professor and also highlight medical-related keywords, P31 realized that "*it could be medical professor*" and accepted the AI explanation and prediction.

How did intuition about features differ across the two types of explanations? With feature-based explanations, participants applied intuition about features to determine if they agreed or disagreed with the model's reasoning. *Feature information they disagreed with is considered evidence that discredits the model's prediction.*

When given example-based explanations, intuition about features was most commonly used in two ways: First, intuition about features, including all the types discussed above, guided evaluations of whether a given instance was indeed similar to the provided examples. For example, P4 judged an example to be similar based on features that they believe should carry more weight: "*The [given individual] works a similar number of hours to [similar] profile 1 and has a similar education background.*" P32 judged an example to be dissimilar based on the interaction of features: "*at their age, an associates degree versus bachelors degree can matter.*" Second, intuition about features affected reasoning about the impact of feature values that differed between the instance and the example to infer the likely outcome. For example, P1 looked at the examples and said "*similar profile and also this person is hard work with more hours. So I chose [to go with more than $50,000].*" Both judgments can help identify evidence to discredit the AI prediction, *either by confirming that the instance is similar (dissimilar) to an example that the model predicted incorrectly (correctly) or by identifying features with different values that indicate a different outcome than the prediction.* Additionally, participants were *more likely to identify and infer the impact of new features* with example-based explanations through reasoning about the similarity of examples. For example, P22 noticed an organization shared by one example and learned that this is "*some teaching thing...book club*" and then switched from their initial intuition of "psychologist" to "teacher." This suggests that example-based explanations can provide additional context and support inductive reasoning [11] to help people form additional intuition about features.

Intuition about AI limitations and prediction unreliability. Lastly, we observed comments on the limitations of the AI system, particularly when participants recognized signals revealed by the explanations that indicated the *unreliability* of the model's prediction on the current instance. While previous work suggests that a desideratum of XAI methods should be uncertainty awareness [18, 82], we avoid using the term "uncertainty" as participants' comments reflect their subjective perception of reliability rather than a measure of the actual model uncertainty. Recognizing that a prediction was unreliable could boost a participant's outcome intuition if they disagreed with the AI prediction, or prompt self-doubt and further deliberation if they initially agreed.

Participants, primarily those with ML experience, also mentioned other general limitations of AI to justify cases where they discounted the AI prediction. One common intuition was regarding biases that can be embedded in ML models, which led participants to discount predictions in cases where the feature-based explanation gave a high weight to gender or in cases where the example-based explanation showed different predictions for examples with different genders. Another common intuition is AI's limitation in considering contexts or multiple features. For example, P9 commented that "*once it's like in the context of this occupation, the contribution [of other features should] change.*

I don't know if that happened in this case for the model." Lastly, participants commented that AI might not be good at predicting rare instances, such as biographies with an uncommon format.

How did intuition about AI limitations differ across the two types of explanations? We found participants utilized different signals of prediction unreliability with different frequencies for the two types of explanations. With the example-based explanation, most participants identified prediction unreliability in instances where the AI prediction was incorrect on similar examples, such as when "*the [AI] predictions [of the examples] are completely off from the actual income*" (P32). With the feature-based explanation, (only) a few participants noted a pattern of unreliability when the weights did not trend strongly in either direction. For example, in the tabular setting, participants noted when there were fairly equal numbers of features providing evidence in both directions, or when the scores were "*pretty uniform*" [P17].

Summary: Three Intuition-Driven Pathways to Non-reliance on AI. In light of prior findings showing explanations lead to overreliance on AI predictions, we summarize three pathways that participants used in our study to apply different types of intuition to override the AI prediction.

- **Pathway 1:** Form a strong intuition about the outcome that disagrees with the AI prediction.
- **Pathway 2:** Apply intuition about features to reason about the AI explanations and identify evidence that discredits the AI prediction.
- **Pathway 3:** Recognize AI limitations, especially signals of prediction unreliability.

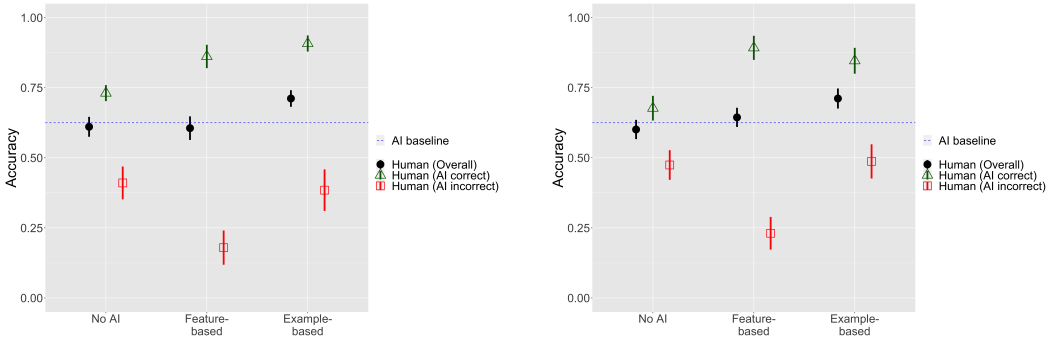
We note that these pathways are identified from reasoning that was frequently mentioned in participants' think-aloud data, which does not allow isolating or quantifying their effect. Therefore, these pathways should not be taken as mutually exclusive or having different levels of impact. Reasoning along these pathways also does not imply that non-reliance is appropriate.

4.2 The Effect of Explanation Type on Accuracy and Reliance (RQ2)

We now evaluate whether the choice of feature- or example-based explanations leads to differences in decision accuracy and appropriate reliance, and if so, whether this can be explained by differences in how intuition comes into play (RQ2). To do so, we first quantitatively analyze the decision outcomes from our study. We then explain the observed effects using insights from qualitative analysis, through the lens of the three intuition-driven pathways identified above. Given the small sample size and that the focus of our study is not on hypothesis testing, we consider the quantitative analysis exploratory. We encourage readers to use caution when interpreting the p-values and focus instead on the trends.

4.2.1 Exploratory quantitative analysis. To begin, we compare participants' decision accuracy across explanation types, where accuracy is the percentage of instances for which a participant's decision matches the ground truth label. The decisions we look at are the participant's final decisions in phase 3 of the study (8 instances for each explanation type). As a baseline, we also compute the accuracy of participants' decisions in phase 2 (16 instances), which we refer to as the "No AI" condition. Figure 2 shows the mean and standard deviation of accuracy for each of these conditions, as well as the accuracy for cases in which the AI prediction was correct or incorrect, respectively, as we discuss more below.

The average overall accuracy and standard error of participants with no AI, AI support with feature-based explanations, and AI support with example-based explanations were 61.1% ± 2.9%, 60.6% ± 4.2%, and 71.1% ± 2.9% respectively in the income prediction task, and 60.0% ± 4.4%, 64.4% ± 4.3%, and 71.2% ± 4.6% respectively in the biography classification task (Figure 2, in black). Since the accuracy of the models we trained on the particular instances presented in the study is



(a) Income Prediction (b) Biography Classification

Fig. 2. Participant's accuracy in different conditions—no AI (from phase 2), AI support with example-based explanations (from phase 3), and AI support with feature-based explanations (from phase 3)—compared to the accuracy of the AI baseline on the two prediction tasks. Across tasks, example-based explanations achieved complementary decision performance whereas feature-based explanations did not.

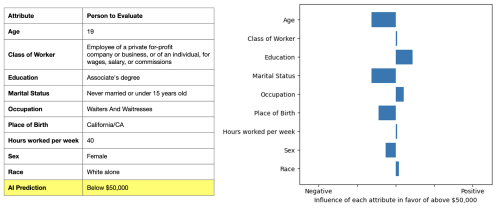
62.5%, this means that the example-based explanations led to *complementary performance*—accuracy higher than human or AI alone—on both tasks.

We ran a mixed-effect regression on the decision accuracy, using participants as a random-effects variable, and explanation type as a fixed-effects variable, where No AI is used as the reference level. We find that feature-based explanations did not have a significant effect on decision accuracy over No AI for either task ($SE = 0.00, p = 0.88$ for income prediction and $SE = 0.04, p = 0.51$ for biography classification), whereas the effect of example-based explanations is marginally significant ($SE = 0.10, p = 0.07$) for income prediction and significant ($SE = 0.11, p = 0.04$) for biography classification.

To better understand the effect of explanation type on *appropriate reliance*, we separately analyze accuracy for instances on which the AI system was correct and instances on which it was incorrect. For cases where the AI system made correct predictions, consistent with prior work [82], we find that both types of explanations led to increased accuracy on both tasks (Figure 2, in green). Running an analogous regression analysis to the one for overall accuracy above, we find that, for these AI-correct cases, feature-based explanations had a marginally significant effect on accuracy ($SE = 0.13, p = 0.06$) for income prediction and a significant effect on accuracy ($SE = 0.22, p = 0.01$) for biography classification. Example-based explanations had a significant effect ($SE = 0.18, p = 0.01$) for income prediction and a marginally significant effect ($SE = 0.17, p = 0.06$) for biography classification.

For cases where the AI system was incorrect, consistent with prior work finding that feature-based explanations could reduce decision accuracy by increasing people's overreliance [4, 68, 87], we find that feature-based explanations led to decreased accuracy compared with the No AI condition. This effect is marginally significant ($SE = -0.23, p = 0.09$) for income prediction and significant ($SE = -0.24, p < 0.01$) for biography classification. In contrast, we find no such effects for example-based explanation ($SE = -0.03, p = 0.80$ for income prediction and $SE = 0.01, p = 0.43$ for biography classification); participants were able to maintain a similar accuracy as they had without AI support in these cases. These results are illustrated in Figure 2 in red.

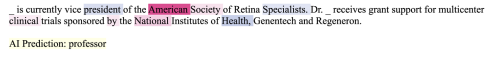
In summary, our analyses echo prior work and suggest that **feature-based explanations may not improve decision accuracy because they increase participants' reliance on AI** regardless



(a) Example of a *Low Diff* feature-based explanation for income prediction.

Attribute	Person to Evaluate	Similar Profile #1	Similar Profile #2
Age	42	42	43
Class of Worker	Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions	Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions	Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions
Education	Some college, but less than 1 year	1 or more years of college credit, no degree	Some college, but less than 1 year
Marital Status	Married	Married	Married
Occupation	Laborers And Freight, Stock, And Material Movers, Hand	Laborers And Freight, Stock, And Material Movers, Hand	Laborers And Freight, Stock, And Material Movers, Hand
Place of Birth	California/CA	California/CA	California/CA
Hours worked per week	40	40	40
Sex	Male	Male	Male
Race	White alone	White alone	White alone
AI Prediction	Above $50,000	Above $50,000	Above $50,000
Actual Income	???	Below $50,000	Below $50,000

(b) Example of a *Both Wrong* example-based explanation for income prediction.



(c) Example of a *Low Diff* feature-based explanation for biography classification.

Bio to Evaluate

_ believes sensory-friendly performances can particularly help families who face social stigmas. "It's a great opportunity for them to actually do something 'normal' with their children where they're not getting looked at by other adults that typically judge them," _ said. "A lot of times they think their children are out of control and they're not controlling their children."

AI Prediction: teacher

Similar Bio #1

_'s friendly, open, and, according to all reports, the kind of doctor who makes _ patients feel cared about. How do the nice doctors like _ stay that way? "The nice ones were born that way," _ replies. "Why people who are not nice are in medicine is the question--why they're allowed to go into it, why they're encouraged by their teachers.

AI Prediction: teacher

Actual Profession: surgeon

Similar Bio #2

For over fifteen years _ has used humor and compassion to help children learn to manage and control their emotions while figuring out who they are and what they want in this crazy, fast-paced world. _ has developed successful strategies aimed at helping parents to feel more in control as they learn to empower their kids to do the same.

AI Prediction: teacher

Actual Profession: psychologist

(d) Example of a *Both Wrong* example-based explanation for biography classification.

Fig. 3. Examples of feature- and example-based explanations that signal prediction unreliability.

of whether the AI system is correct or incorrect. In contrast, **example-based explanations helped achieve complementary human-AI performance** in our study by increasing appropriate reliance when the AI system was correct while helping participants maintain their accuracy when the AI system was incorrect. Next, we further delve into the reasons for these observations.

4.2.2 Why did example-based explanations better support decision-making than feature-based explanations? To answer this question, we investigate participants' reliance on the AI system through the lens of the three intuition-driven pathways identified in Section 4.1. To do this, we further break down participants' decisions into four different scenarios based on whether the participant's initial decision in phase 2 was correct (denoted P2 ✓) or incorrect (P2 ✗), and whether the AI prediction was correct (denoted AI ✓) or incorrect (AI ✗). Table 1 shows participants' phase 3 accuracy in each of these scenarios.

Quantifying the existence of unreliability signals in explanations. To explore the effect of pathway 3—how explanations help participants recognize prediction unreliability—we define a way of measuring whether a particular explanation signals unreliability. For feature-based explanations, as discussed in Section 4.1, participants mentioned using the fact that the total feature weights did not trend strongly in either direction as a signal of an unreliable prediction. To capture this, we heuristically define the following proxy measure, which we refer to as *Low Diff*. First, we calculate the absolute difference between the sum of positive weights and the sum of negative weights for each instance used in the study. We use the value of the 25th percentile of the absolute differences as a threshold, and consider any instance where this difference is lower than the threshold as signaling unreliability. We refer to instances where the difference is above the threshold as *High*

Diff. Examples of *Low Diff* explanations for the two tasks are shown in Figure 3 (a) and (c), while the explanations in Figure 1 (a) and (c) are *High Diff*.

Table 1. For each task (income prediction, top and biography classification, bottom) and each type of explanation (feature-based and example-based), we break down participants' accuracy by four scenarios that capture different forms of (non-)reliance based on whether their initial decision in phase 2 was correct (denoted by a ✓in the P2 column) or incorrect (✗) and whether the AI prediction was correct (denoted by a ✓in the AI column) or incorrect (✗). For each scenario and each phase 3 decision (P3 column) we additionally report the number of times that the provided explanations did or did not exhibit signals of unreliability.

Income Prediction Task								
		Feature-Based			**Example-Based**			
(AI, P2)	P3	Accuracy	High Diff	Low Diff	Accuracy	Both Right	Mixed	Both Wrong
(✗, ✓)	✓	42.9%	3	3	50.0%	2	1	6
	✗		6	2		4	4	1
(✗, ✗)	✓	4.0%	0	1	28.6%	1	0	5
	✗		10	14		10	4	1
(✓, ✓)	✓	91.7%	32	12	100.0%	37	10	0
	✗		4	0		0	0	0
(✓, ✗)	✓	70.6%	11	1	66.7%	12	0	0
	✗		5	0		0	6	0

Biography Classification Task								
		Feature-Based			**Example-Based**			
(AI, P2)	P3	Accuracy	High Diff	Low Diff	Accuracy	Both Right	Mixed	Both Wrong
(✗, ✓)	✓	41.2%	4	3	75.0%	5	3	7
	✗		4	6		2	2	1
(✗, ✗)	✓	9.1%	0	2	21.1%	1	1	2
	✗		10	10		8	7	0
(✓, ✓)	✓	95.7%	42	3	100.0%	29	12	0
	✗		2	0		0	0	0
(✓, ✗)	✓	72.2%	8	5	58.3%	9	5	0
	✗		3	2		6	4	0

For example-based explanations, participants tended to notice whether the AI system made mistakes on the two examples provided. We define two heuristic measures of whether an explanation signals unreliability based on whether the system made incorrect predictions on both (*Both Wrong*) or one (*Mixed*). We refer to instances where the AI system made only correct predictions as *Both Right*. Examples of *Both Wrong* explanations are shown in Figure 3 (b) and (d), while Figure 1 (b) and (d) are examples of *Both Right*.

Note that these measures only reflect the theoretical presence of unreliability signals and do not imply that all participants recognized these signals. As discussed, there were significantly fewer participants who commented on such signals for feature-based explanations.

What happened in these four scenarios? We now discuss observations about each of the four scenarios, referencing the results summarized in Table 1 combined with observations from the

comparative analysis of the think-aloud data, as described in Section 3.6. For each scenario, we further break down cases by whether the participants' final decision was correct (denoted by a ✓ in the P3 column) or incorrect (✗ for P3). For each scenario, we include columns that show the number of cases in which the explanation exhibited signals of (un-)reliability as described above.

AI ✗, P2 ✓. Participants' accuracy is higher in this scenario (row 1 in Table 1) than it is when both the AI prediction and their own decisions were wrong (row 2) across tasks and explanation types. This supports the idea that *participants' intuition about the outcome (pathway 1) played a role in their correct non-reliance on AI.* However, in both tasks, participants had higher accuracy with example-based explanations compared with feature-based explanations. Across both tasks, for cases where participants correctly overrode the AI prediction (P3 ✓, first sub-row), there is a high percentage of instances where the AI prediction of the two similar examples were *Both Wrong*. This suggests that two wrongly predicted examples form a *strong unreliability signal (pathway 3) that played a role in participants' correct non-reliance.* However, having mixed examples did not appear to be a strong enough signal to push participants to not rely on the AI prediction.

While these statistics only suggest correlations, the think-aloud data support that most participants in this scenario were eager to discredit the AI prediction after noticing it was wrong on both examples. In almost all cases, participants felt more confident about their own intuition and presented their own reasoning to explain the AI system's mistakes. When only one example was wrong, participants more carefully reasoned about the similarity of examples and the impact of different feature values. They had correct non-reliance when they found the instance to be similar to the example, which had different ground truth (pathway 2). Interestingly, in 4 instances of biography classification where participants had correct non-reliance, they commented about *new features* they learned from the examples to help them judge the similarity. For example, P14 correctly disagreed with the AI prediction (teacher) because they noticed that the current profile resembled the two examples (both professors) by all mentioning "*sophisticated*" art venues.

In contrast, for feature-based explanations, there is not a strong correlation between the *Low Diff* proxy measure and participants' decision correctness. Moreover, there were only two instances where participants noticed this signal of unreliability. When they had correct reliance, they primarily relied on their own intuition (pathway 1) or picked up on feature weights they disagreed with from the explanations (pathway 2). Strikingly, in 5 out of the 8 cases where participants incorrectly relied on the AI prediction for income prediction, they acknowledged they had weak intuition about the outcome and deferred to AI predictions. The remaining 3 went with the AI prediction because they found the explanation agreeable, but only focused on a subset of features. Out of the 10 such cases for biography classification, 3 acknowledged weak intuition, 3 agreed with the explanation, and 4 disagreed with the explanation (e.g., finding highlighted keywords meaningless) but still went with the AI prediction. All these observations support the claim that *feature-based explanations disrupt people's natural intuition* (despite making correct decisions in phase 2), especially when participants think the AI explanation makes sense or are unable to reason about the explanation.

AI ✗, P2 ✗. Interestingly, in this scenario (row 2), while there were very few cases where participants arrived at the right decision with feature-based explanations, example-based explanations surprisingly supported correct decisions for a sizable number of cases. In the majority of these cases, the example-based explanations were *Both Wrong* or *Mixed* (pathway 3). In these cases, we observed that all participants carefully examined the examples to confirm the similarity between the instance and examples with different ground truths (pathway 2), which prompted them to not only override the AI prediction but also their own intuition. In two cases, participants also learned about new features from the examples (pathway 2). In the only three instances where

participants were correct with feature-based explanations, they made the correct decision because they disagreed with the feature weights (pathway 2).

AI ✓, P2 ✓. Surprisingly, a few participants made incorrect decisions in this scenario (row 3) when using feature-based explanations. This was mainly because they picked up on information they disagreed with in the explanation, e.g., the precise weight of a feature (pathway 2, but incorrect non-reliance). In one case, the participant picked up on the unreliability signal, commenting that there were factors pushing in both directions (pathway 3, but incorrect non-reliance). Even though the AI prediction was correct, there were still a sizable number of cases for which feature-based explanations exhibited *Low Diff*; this is not surprising since instances near the decision boundary can still be predicted correctly. Meanwhile, example-based explanations had *more informative unreliability signals*—zero cases had both examples predicted wrong. As a result, participants did not have any pathway for non-reliance and reached 100% accuracy for both tasks.

AI ✓, P2 ✗. Most participants mentioned their intuition disagreed with the AI prediction (pathway 1, but incorrect non-reliance) in this scenario (row 4), contributing to lower accuracy compared to row 3. This is the only scenario in which feature-based explanations led to higher accuracy than example-based explanations. In all correct reliance cases, participants either found nothing they disagreed with in the feature-based explanation (no pathway 2) or acknowledged they had weak intuition about the decision outcome (no pathway 1). For example-based explanations, none of the cases where participants had correct reliance were cases of *Both Wrong* (no pathway 3), boosting their confidence in the AI prediction. Considering cases with *Mixed* examples, in the biography classification task, participants changed their decision about half the time, either because they found the example with consistent ground truth to be more similar or they learned new features from the examples. (There were no *Mixed* cases where participants changed decisions in the income prediction task.)

These observations about each pathway allow us to summarize why example-based explanations better supported appropriate reliance:

- For pathway 1 (intuition about the outcome): Example-based explanations led to less disruption of people's natural intuition about the outcome. One reason for this is that they are less visually overwhelming and allow people to focus on the instance first. Additionally, they do not force people to attend to features they would otherwise not pay attention to or quantify the impact, as feature-based explanations often do, which may disrupt or prompt self-doubt about one's own intuition.
- For pathway 2 (intuition about features to reason about explanations): Example-based explanations better supported recognizing or even learning new features because the similar examples and their ground truth labels brought in additional context that could promote inductive reasoning. Reasoning about example-based explanations was also more aligned with people's natural intuition, since it does not require precise quantification of the impact of features. In contrast, when faced with quantified feature importance, participants either found it challenging to judge or difficult to look past information they disagree with, even if it is a relatively trivial disagreement—e.g., "*the weight for occupation is not high enough*" (P26). This opened the door to making a wrong decision, even when both the AI prediction and their initial intuition about the outcome were correct.
- For pathway 3 (intuition about AI limitations): Example-based explanations provided strong signals of unreliability, particularly when the AI system predicts both examples incorrectly. These signals were not only accurate (highly correlated with AI incorrectness) but were also

easily noticeable. These signals helped boost people's outcome intuition when they disagree with the AI prediction and dampen it when they agree with the AI prediction, increasing the chance that people make correct decisions even when both the AI prediction and their own initial intuition are wrong. In contrast, pathway 3 was generally weak for feature-based explanations, at least considering the *Low Diff* proxy measure.

- Differences between the two tasks (tabular vs. text data): When provided with example-based explanations, participants were more likely to learn about new features from the examples in the biography classification task than in the income prediction task, which we conjecture is due to the biography classification task using text data. This made them more likely to override incorrect intuition even when they were shown examples with mixed ground truth labels (row 4). Also, when given feature-based explanations, pathway 2 may have had a slightly weaker impact on the biography classification task, where there were more cases where participants disagreed with the explanation (e.g., picking up on meaningless keywords in the text) but still chose to rely on the AI system.

4.3 Post-Study Interview Responses

We analyzed responses to post-study interview questions to understand, subjectively, how participants perceived the two types of explanations and their preferences, and also their suggestions on how to improve both types of explanations in future systems.

4.3.1 Subjective preferences of explanation types. While the empirical findings above show that example-based explanations are better at supporting decision-making, subjectively, participants showed rather mixed preferences: about half favored feature-based explanations (N=11) and the other half preferred example-based explanations (N=10). The remaining participants (N=5) were impartial, noting the potential benefits and drawbacks of each explanation type.

Participants' subjective responses favoring example-based explanations further supported our conclusions in Section 4.2. Multiple participants commented that the *unreliability signals or the lack thereof* (i.e., correct predictions on both similar examples) *played a key role* in their confidence in the AI system. For example, P17 said "*when the data was super consistent, I was gonna go with it*" (P32). P1 noted that "*I trusted [ground-truth of the similar examples] more than the predictions of the AI itself.*" Others liked that the similar examples provide them with *more context to form intuition about features*, such as to "*compare and contrast [the current input] with the neighbors. For example, I can narrow down what features should I look at*" (P27), or "*in breaking ties*" (P30).

Feature-based explanations were often preferred because subjectively they were *easier to consume*. Some participants found feature-based explanations more intuitive and useful to understand how the AI system made the prediction: "*seeing if the factors that have contributed most to the decision were aligned with what my thought process was*" (P9) and the highlights helped form "*a picture in my mind for how the AI model might have been acting*" (P31). Interestingly, while some participants found the overlay presentation of feature-based explanations distracting, others found it appealing to "*guide my attention*" (P18) or "*help speed up my scanning process*" (P8) without realizing the risk of feature-based explanation disrupting them from forming their own intuition.

We find a disparity between what participants *thought* was helpful and what was actually helpful for their decisions: 7 out of 11 participants who preferred feature-based explanations performed better with example-based explanations. A recent work by Buçinca et al. [10] called out that when interacting with AI decision support, there is tension between cognitive engagement, which can lead to better decision outcomes, and subjective user experience, which may be influenced by ease of use and efficiency. Participants' comments suggest that there may exist such tension for the explanation types we studied. While example-based explanations encouraged more independent

and deeper reasoning than feature-based explanations, these explanations could also be subjectively perceived as more cognitively demanding and time-consuming.

4.3.2 Participant suggestions for improvement. We asked how the decision support could have been improved. Here we summarize suggestions from the 20 participants who offered them.

Better support reasoning about example-based explanations. Following the above observations that example-based explanations were sometimes perceived to be less intuitive to use, participants suggested ways to improve them. As discussed, in pathway 2 reasoning, participants looked for evidence that the AI prediction might be wrong in the explanation based on an instance's similarity with examples or the impact of different feature values. To better support this, multiple participants suggested interactive example-based explanations that allow users "*to interactively change some of the factors in the person to be judged to see if that changed the prediction of the AI*" (P24) or more sophisticated sampling methods that either select contrasting examples with different ground truth labels or show examples that vary in different features to understand the model's "*decision boundaries*" (P23). Inspired by feature-based explanations, some participants suggested combining the two explanation types by "*highlighting the keywords in examples*" (P14), "*highlighting the different and similar features*" (P20) between the instance and examples, or explicitly showing the impact of different feature values.

Better support intuition about AI limitations. Participants suggested several approaches that relate to improving their intuition about the AI system's limitations and boosting pathway 3 reasoning. In addition to suggesting that the AI system support include confidence or uncertainty quantification for each prediction, participants asked for support to have a global understanding of how the model assigns weights to different features in general (i.e., a global explanation [35, 38]) "*to show the influence of the attributes on the entire data set*" (P20) or on what kind of cases the model is likely to be unreliable. Some noted that this kind of global model intuition can also help them better reason about local explanations. For example, P20 commented that if they were aware that the model is generally biased, they would discount the impact of demographic features when looking at the explanations. For users with more ML knowledge, P27 suggested providing more technical information about the underlying algorithm, training data, and how the explanations are generated: "*Are we using a deep learning model or are we using something more classic? Overall, where does the training data come from? ... What is the highlighting [method]? Is it SHAP? Is it LIME?*"

Better align explanations with human reasoning. Participants further pointed out that the current explanations do not fully align with human reasoning. For example, applying explanations to embeddings at the token level (i.e., individual words) does not align with features that people use to judge a biography. As discussed in Section 4.2, people tended to reason about features that were either "*two words together or phrases*" (P25) or higher-level semantic features, linguistic markers, or writing styles. We note that recent XAI research has begun to explore how to "translate" raw features used by a model into higher-level concepts that are familiar to people [30, 48]. Other suggestions relate to presenting the explanations as a human would, i.e., in a more interactive, conversational, and narrative manner rather than using bar charts. P2 said "*I saw the AI as more of a person or a friend explaining it to me. If the information was presented in a story format or a paragraph format, that would also help weigh my decision.*"

5 DISCUSSION

We investigated the types of human intuition present in human-AI decision-making with explanations. We identified three intuition-driven pathways to override AI predictions: (1) using strong outcome intuition to disagree with the AI prediction; (2) applying intuition about features to reason

about explanations to discredit the AI prediction; and (3) recognizing AI limitations through signals of prediction unreliability. In this section, we first discuss how these pathways can support a more generalizable understanding of when and what explanations can help decision-making. Then we make design recommendations for AI support that helps decision-makers appropriately apply their own intuition.

5.1 Understanding Overreliance in the Human-AI Decision-Making Process

To facilitate the responsible design of AI systems by preventing harmful overreliance [54], a fundamental understanding of people's decision-making process with AI is key. We believe the set of intuition-driven pathways we identified can be a useful tool to help understand why—and even anticipate when—inappropriate reliance may happen. We illustrate this with two use cases below.

These pathways can help explain the effect of different explanations. First, we can use the pathways to interpret the mixed results of prior empirical studies on different forms of example-based explanations. Contrary to our results, a previous study by Wang and Yin [82] found no evidence that the example-based explanations they used can improve decisions over feature-based explanations. However, they adopted a different strategy to select examples, focusing only on examples where the model predictions are *correct*, and always displaying one example for which the AI system predicts the same class and one example for which it predicts a different class. This approach may not provide easily recognizable prediction unreliability signals (pathway 3). We also suspect that this selection method may not result in examples that are consistently similar, then participants are less likely to utilize pathway 2 to find evidence that a prediction might be wrong. In another study, Kim et al. [49] tested example-based explanations that showed representative prototypes for each class, rather than nearest neighbors of the given instance, and found less overreliance on incorrect AI predictions compared to feature-based explanations. We note that prototype-based explanations do not directly support pathway 3, but rather support pathway 2 when users recognize dissimilarity between the prototypes and the instance, which Kim et al. [49] also explicitly asked participants to rate.

These pathways can help understand which individuals will benefit more from AI support and why. Understanding how individual differences affect the viability of each pathway can make it possible to anticipate who will or will not benefit from certain AI support. We discuss two concrete examples. The first is individuals with a low level of domain knowledge, where the only viable pathway is pathway 3 because pathways 1 and 2 require substantial domain knowledge. Since example-based explanations better support appropriate non-reliance through pathway 3, we expect it to be particularly helpful for people with low domain knowledge. Indeed, by using participants' decision accuracy over all instances in phase 2 (N=16) as a proxy measure of their domain knowledge, we found that the 5 participants with the lowest domain knowledge (average phase 2 accuracy = 43.8%) improved significantly with example-based explanations (average phase 3 accuracy = 70.0%) but not with feature-based explanations (average phase 3 accuracy = 52.5%).

Second, we can anticipate that people whose own outcome intuition is highly consistent with the AI system will benefit less from AI support because of a weak pathway 1, which has been alluded to by prior works as complementary knowledge [3, 87]. By using agreement between the participants' phase 2 decisions and AI predictions over all instances in phase 2 (N=16) as a proxy measure of *complementary outcome intuition*, we find a moderate negative correlation with their phase 3 accuracy when using feature-based explanations ($r = -0.34$ for income prediction and $r = -0.41$ for biography classification). Interestingly, there is little to no correlation between this proxy measure and accuracy with example-based explanations, possibly because the example-based explanations we used tend to encourage participants to utilize pathway 3, as discussed in Section 4.2.

5.2 Design Recommendations for AI Decision-Support Systems

We make the following recommendations for future work to develop more effective AI support that accounts for different types of human decision intuition as well as individual differences in them.

Accommodate human decision intuition of varied strengths. Our study suggests that even for the same decision task, human decision-makers may have a strong intuition for some instances but not others, presenting different opportunities for AI support. On the one hand, when people do not have a strong intuition, it would be difficult to identify when the AI system is incorrect through pathway 1 (but possible through other pathways). On the other hand, when people have strong intuition, even though they may be more capable of overseeing the AI system, they may also be less motivated to engage with complex information. In this case, AI support should also avoid disrupting people's natural intuition and instead encourage people to examine their own intuition even when they are in agreement with the AI system. We can also envision adaptive or personalized systems that provide different kinds of information support (e.g., when to show a certain type of explanation) depending on the decision-maker's confidence in their own decision.

Make explanations compatible with human decision rationale. We observed that when using feature-based explanations, participants sometimes mistakenly override the AI prediction because they disagree with certain, potentially trivial, aspects of the explanation. Social science literature suggests that natural human explanations are predominantly *qualitative* [63]—while people may have an intuition about which feature is relevant to the decision, or which feature is more relevant than another, they may not be able to articulate a precise quantification. This literature suggests that the common way to present feature-based explanations, which shows a contribution score for each feature, does not align with how people reason. Future work should explore more natural explanations that, for example, start with qualitative narratives and provide more precise information upon request. More broadly, our study identifies types of feature-based intuition that participants applied to reason about explanations (Section 4.1). We hope they inspire future work to develop AI systems that leverage these reasoning patterns to design more compatible explanations.

Design explanations to facilitate people's use of intuition-driven pathways. Previous studies have attributed the cause of overreliance on AI to a lack of cognitive engagement [10, 32, 55]. However, we show that even when people are relatively engaged (by thinking aloud), current XAI techniques do not reduce overreliance. This points to the inherent limitations of feature-based explanations in supporting participants to override incorrect predictions (i.e., the disruption to and incompatibility with natural intuition). To override AI predictions through pathway 2, we find people naturally look for evidence in explanations to discredit AI predictions (e.g., looking for multiple similar examples with a different or incorrect prediction in example-based explanations). We encourage future work to first explore what people consider to be evidence to discredit predictions for different types of explanations, and then to design interfaces to help people identify such evidence.

Develop decision-support features that effectively reflect AI limitations. Lastly, we highlight the need for explanations to help people form intuition about AI limitations (pathway 3). Participants' interview comments suggested that global explanations and documentation of the model's failure cases can inform intuition about model limitations. While prior work has suggested that showing prediction uncertainty or confidence measures may be more effective than explanations for this purpose [87], the former can suffer from miscalibration (i.e., the presented value does not correspond to the actual error probability). We also suggest further work on evaluation metrics that capture an XAI method's ability to *reflect and communicate prediction unreliability*. Similar metrics have been used to evaluate the calibration of uncertainty quantification [7, 9, 64]—the estimated uncertainty

should reflect the observed error rate. Our study suggests that such metrics should also consider how people actually *perceive* the information. For example, we can envision a metric that asks targeted users to rate the reasonableness of explanations and then quantifies the correlation with errors in the test data.

5.3 Limitations

We acknowledge several inherent limitations of using a think-aloud method. First, by asking participants to verbalize their thought processes, our protocol may not completely resemble a realistic decision-making setting. It is possible that when not thinking aloud, especially in a low-stakes setting, people will be less cognitively engaged with AI explanations [32] and some effects found in the study will not be observed. With this limitation, we again encourage readers to interpret the quantitative results with caution and focus on the trends. Second, think-aloud data are observational and do not allow for isolating the effect of different types of intuition or for identifying precise causes of how a participant arrived at a certain decision. Therefore, our results focused on the themes that emerge from the observational data rather than attempting to draw conclusions about their precise relations and impact on decisions.

We also acknowledge the trade-off of a two-phase study design that showed participants the same set of instances in both phases. While this design allowed us to analyze cases where the human-alone decisions agree or disagree with the AI predictions separately, this design may have strengthened, in some cases, participants' prior intuition more than a realistic human-AI decision-making setting. That being said, we do not foresee any *type* of intuition to be contingent on the set-up of our study design.

Our study was also limited by the choice of decision tasks and explanation methods, as well as the relatively small sample size. The two tasks are relatively low stakes and do not require specialized domain knowledge. They are also not representative of all data and feature types. One method in our study is a popular post-hoc feature-based explanation that is known to be not completely faithful to the underlying prediction model. The potential noise that can be introduced into decision-making by unfaithful post-hoc explanations, as opposed to directly interpretable models, is an open question that should be explored in future work. Our participants were not experts on these tasks and, despite our effort to diversify, the sample was biased toward more highly educated and ML-experienced individuals than the general population. Therefore, we acknowledge that the types of intuition identified in this work may not be complete or fully generalizable. We encourage future work to further study the interplay between human intuition and different types of explanations across domains and populations to verify the themes identified in this study.

6 CONCLUSION

We designed a mixed-methods study to understand how human decision-makers reconcile their own intuition with AI predictions and explanations. In light of prior work, which finds that feature-based explanations can increase overreliance on incorrect AI predictions, we study how decision-makers can apply their own intuition to override AI and have appropriate reliance. Our analysis of participants' think-aloud data revealed three intuition-driven pathways to reduce overreliance on AI: (1) using strong outcome intuition to disagree with the AI prediction; (2) applying intuition about features to reason about explanations to discredit the AI prediction; and (3) recognizing AI limitations through signals of prediction unreliability. We use these pathways to explain why the example-based explanations we used helped lead to complementary human-AI performance and better supported appropriate reliance in comparison to feature-based explanations: they were less disruptive of people's natural intuition about outcomes, they better promoted inductive reasoning about features and the decision task generally, and, in particular, they provided strong and accurate

signals of prediction unreliability. These pathways also highlight the limitations of feature-based explanations, providing reasons why they lead to overreliance when the AI system is incorrect: they disrupt outcome intuition, may conflict with intuition about features, and do not present clear ways to reason about prediction unreliability. Our findings provide fundamental knowledge about the human-AI decision-making process that could support a generalizable understanding of when and what explanations can help decision-making, and point to user needs for AI decision-support systems that better accommodate human decision intuition, are more compatible with human intuition, and support a more critical understanding of AI.

ACKNOWLEDGMENTS

This research was conducted at Microsoft Research. We thank our participants for their time and the reviewers for their feedback. We also thank Zana Buçinca, Han Liu, Andreas Madsen, Stephanie Milani, and researchers in the Microsoft Research FATE group for their thoughtful suggestions.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 2–11.

[4] 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.

[5] Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M Mack, George Ruiz, Mark S Smith, and Eric Horvitz. 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PloS one* 9, 10 (2014), e109264.

[6] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[7] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 401–413.

[8] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[9] Glenn W Brier et al. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review* 78, 1 (1950), 1–3.

[10] 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[11] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[12] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[13] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.

[14] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during

medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
[15] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.
[16] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.
[17] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 1721–1730.
[18] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.
[19] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. 2023. Machine Explanations and Human Understanding. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=y4CGF1A8VG
[20] Serena Chen and Shelly Chaiken. 1999. The heuristic-systematic model in its broader context. (1999).
[21] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. 2022. Interpretable machine learning: Moving from mythos to diagnostics. *Queue* 19, 6 (2022), 28–56.
[22] Lingwei Cheng and Alexandra Chouldechova. 2022. Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening. *arXiv preprint arXiv:2204.05478* (2022).
[23] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[24] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.
[25] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems* 34 (2021).
[26] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.
[27] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
[28] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv preprint arXiv:2107.13509* (2021).
[29] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable AI: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.
[30] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 263–274.
[31] 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.
[32] Krzysztof Z Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*. 794–806.
[33] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 235 (jan 2021), 28 pages. https://doi.org/10.1145/3432934
[34] Gerd Gigerenzer. 2007. *Gut feelings: The intelligence of the unconscious*. Penguin.
[35] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.
[36] 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.
[37] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[38] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[39] Sophia Hadash, Martijn C Willemsen, Chris Snijders, and Wijnand A IJsselsteijn. 2022. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *CHI Conference on Human Factors in Computing Systems*. 1–9.

[40] Katherine H Hall. 2002. Reviewing intuitive decision-making and uncertainty: The implications for medical education. *Medical education* 36, 3 (2002), 216–224.

[41] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5540–5552. https://doi.org/10.18653/v1/2020.acl-main.491

[42] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. https://doi.org/10.1145/3544548.3581025

[43] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[44] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, LibuŠe Hannah VepŘek, and Gabrielle Quinn. 2023. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Trans. Comput.-Hum. Interact.* (mar 2023). https://doi.org/10.1145/3534561 Just Accepted.

[45] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: A sociotechnical lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[46] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*.

[47] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[48] 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.

[49] Sunnie SY Kim, Nicole Meister, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. Hive: Evaluating the human interpretability of visual explanations. In *European Conference on Computer Vision*. Springer, 280–298.

[50] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[51] 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).

[52] 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.

[53] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[54] Q Vera Liao and S Shyam Sundar. 2022. Designing for Responsible Trust in AI Systems: A Communication Perspective. *Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency* (2022).

[55] Q Vera Liao and Kush R Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).

[56] Q Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.

[57] Zachary C Lipton. 2018. The Mythos of Model Interpretability. *Commun. ACM* 61, 10 (2018), 36–43.

[58] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[59] 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.

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

[61] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering* 2, 10 (2018), 749–760.

[62] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural NLP: A survey. *Comput. Surveys* 55, 8 (2022), 1–42.

[63] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[64] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 29.

[65] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 2021. Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[66] Richard E Petty and Pablo Briñol. 2011. The elaboration likelihood model. *Handbook of theories of social psychology* 1 (2011), 224–245.

[67] Martin Potančok. 2019. Role of data and intuition in decision making processes. *Journal of Systems Integration* 10, 3 (2019), 31–34.

[68] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[69] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[70] 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.

[71] Allan D Rosenblatt and James T Thickstun. 1994. Intuition and consciousness. *The Psychoanalytic Quarterly* 63, 4 (1994), 696–714.

[72] Eduardo Salas, Michael A Rosen, and Deborah DiazGranados. 2010. Expertise-based intuition and decision making in organizations. *Journal of management* 36, 4 (2010), 941–973.

[73] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[74] Debbie A Shirley and Janice Langan-Fox. 1996. Intuition: A review of the literature. *Psychological reports* 79, 2 (1996), 563–584.

[75] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International journal of human-computer studies* 67, 8 (2009), 639–662.

[76] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.

[77] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[78] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2022. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *arXiv preprint arXiv:2212.06823* (2022).

[79] Diane Walker and Florence Myrick. 2006. Grounded theory: An exploration of process and procedure. *Qualitative health research* 16, 4 (2006), 547–559.

[80] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[81] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

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

[83] Jennifer Wortman Vaughan and Hanna Wallach. 2021. A Human-Centered Agenda for Intelligible Machine Learning. In *Machines We Trust: Perspectives on Dependable AI*, Marcello Pelillo and Teresa Scantamburlo (Eds.). MIT Press.
[84] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.
[85] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.
[86] Wencan Zhang and Brian Y Lim. 2022. Towards Relatable Explainable AI with the Perceptual Process. In *CHI Conference on Human Factors in Computing Systems*. 1–24.
[87] 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.

A PARTICIPANT INFO

Table 2 contains demographic information about the study participants.

Table 2. Table of participant information. For ML and XAI knowledge, participants rated themselves using the following scale: 0 = "None"; 1 = "Limited experience, I know the basic concepts"; 2 = "I have experienced them often or they are part of my day-to-day life"; 3 = "I consider myself an expert on them."

	Education	Role	ML Knowledge	XAI Knowledge	Experimental Condition
P1	Graduate Degree	Software Engineer	2	0	Income Prediction
P2	College	Software Engineer	0	0	Income Prediction
P4	Graduate Degree	Software Engineer	3	2	Income Prediction
P8	College	Data Scientist	2	2	Biography Classification
P9	Graduate Degree	Applied Scientist	3	2	Income Prediction
P11	Graduate Degree	PhD Student	3	2	Income Prediction
P12	College	Engineer	3	3	Income Prediction
P14	Graduate Degree	PhD Student	2	2	Biography Classification
P15	Graduate Degree	PhD Student	3	2	Biography Classification
P16	College	Editor / Data analyst	1	1	Biography Classification
P17	Graduate Degree	PhD Student	2	3	Income Prediction
P18	Graduate Degree	Software Engineer	1	1	Biography Classification
P19	Graduate Degree	Professor	2	2	Biography Classification
P20	Graduate Degree	Research Engineer / Scientist	2	2	Income Prediction
P21	Graduate Degree	Engineer	3	2	Biography Classification
P22	College	Software Engineer	1	0	Biography Classification
P23	Graduate Degree	PhD Student	2	2	Income Prediction
P24	Graduate Degree	PhD Student	3	3	Income Prediction
P25	Graduate Degree	PhD Student	3	3	Biography Classification
P26	Graduate Degree	PhD Student	2	2	Income Prediction
P27	College	PhD Student	3	2	Biography Classification
P28	Graduate Degree	Software Engineer	1	0	Income Prediction
P30	Graduate Degree	Research Engineer / Scientist	3	2	Biography Classification
P31	Graduate Degree	Data Scientist	3	2	Biography Classification
P32	College	PhD Student	1	1	Income Prediction
P33	Graduate Degree	PhD Student	1	1	Biography Classification

B POST-STUDY INTERVIEW QUESTIONS

The following questions were asked in the post-study interviews.

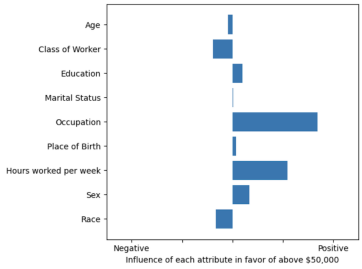
- Can you describe your general strategies to make decisions when AI was not available in phase 2?
- What difference did introducing the AI in phase 3 make?
- How did you reason with AI prediction and explanations?
- What strategies did you take in phase 3 when you had feature contributions versus similar examples? Which explanation type did you prefer?
- In general, do you think you are better at this task than the AI? If yes, would you still want to use the AI if you were asked to perform the task again?
- Is there something else besides feature contribution or similar examples that you wish to know or that would have helped you make better decisions?

C USER STUDY INSTRUCTIONS

Figures 4–7 show the instructions that study participants were given. These variants correspond to cases in which the participant saw example-based explanations before feature-based explanations.

As before, for each individual, you will see the AI system's prediction of whether their annual income is above or below $50,000. However, in this part of Phase 3, you will see a **different type of explanation** showing the extent to which each attribute (Age, Education, etc.) influenced the prediction, using a visualization like the one on the right:

Attribute	Person to Evaluate
Age	37
Class of Worker	Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions
Education	Associate's degree
Marital Status	Married
Occupation	Other Managers
Place of Birth	California/CA
Hours worked per week	48
Sex	Male
Race	White alone
AI Prediction	Above $50,000



In this visualization, bars pointing to the right signify influence in favor of the individual having an annual income above $50,000, whereas bars pointing to the left signify influence in favor an income below $50,000. The longer a bar is, the stronger influence the attribute has. The AI system's prediction combines the influence of each attribute.
For example, for the individual above, the AI system's prediction was that the individual made more than $50,000. This individual's Occupation and Hours Worked Per Week strongly pushed the AI system towards this prediction, while Class of Worker and Race pushed it away.

Please continue thinking aloud when making your guesses.

You will make guesses about 8 individuals in this phase. If you have any questions, please ask now.

Fig. 4. Interface and instructions for how to use feature-based explanations in the income prediction task.

Now, you will make guesses with the help of an Artificial Intelligence (AI) system. For each individual, you will see the **AI system's prediction** of whether their annual income is above or below $50,000. You will also see an **explanation** consisting of two individuals who have similar profiles from the data used to train the system (shown in light blue in the table), the system's predictions for these individuals, and the individuals' actual income group (shown in yellow in the table). The similarities and differences between these individuals and the one you are asked to make a guess about may help you better understand the AI system's prediction and decide whether you want to follow it.

Attribute	Person to Evaluate	Similar Profile #1	Similar Profile #2
Age	25	23	25
Class of Worker	Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions	Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions	Employee of a private for-profit company or business, or of an individual, for wages, salary, or commissions
Education	Regular high school diploma	Regular high school diploma	Some college, but less than 1 year
Marital Status	Never married or under 15 years old	Never married or under 15 years old	Never married or under 15 years old
Occupation	Pharmacy Technicians	Pharmacy Technicians	Surgical Technologists
Place of Birth	California/CA	California/CA	California/CA
Hours worked per week	40	40	40
Sex	Female	Female	Female
Race	White alone	White alone	White alone
AI Prediction	Below $50,000	Below $50,000	Below $50,000
Actual Income	???	Below $50,000	Below $50,000

In Phase 3, **we ask you to think aloud—-that is, try to verbalize your reactions and thoughts while making your guesses and looking at the AI's predictions.** The experimenter may also interrupt and ask follow-up questions. We are especially interested in knowing the reasoning behind your guesses. You are welcome to draw on any background knowledge you have about the task. You are also encouraged to point to specific pieces of information provided.
You will make guesses about 8 individuals in this phase. If you have any questions, please ask now.

Fig. 5. Interface and instructions for how to use example-based explanations in the income prediction task.

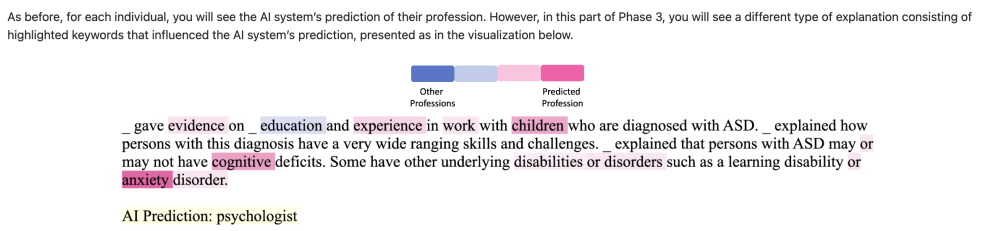


Fig. 6. Interface and instructions for how to use feature-based explanations in the biography classification task.

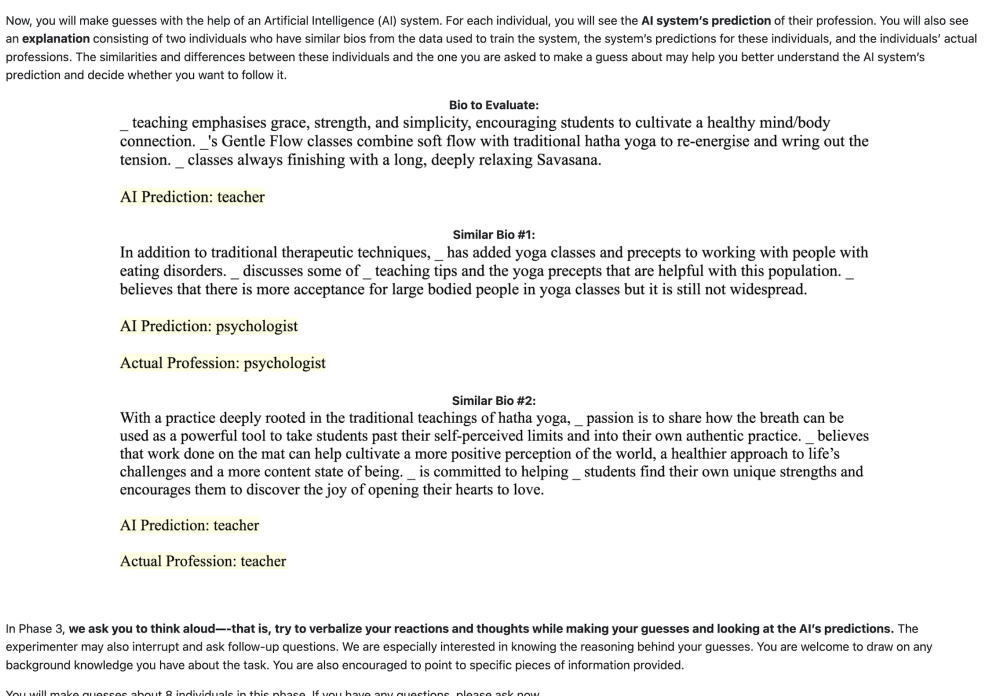


Fig. 7. Interface and instructions for how to use example-based explanations in the biography classification task.

Received January 2023; revised April 2023; accepted May 2023