Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills

ZANA BUÇINCA, Harvard University, USA

SIDDHARTH SWAROOP, Harvard University, USA

AMANDA E. PALUCH, University of Massachusetts Amherst, USA

FINALE DOSHI-VELEZ, Harvard University, USA

KRZYSZTOF Z. GAJOS, Harvard University, USA

People's decision-making abilities often fail to improve or may even erode when they rely on AI for decision-support, even when the AI provides informative explanations. We argue this is partly because people intuitively seek contrastive explanations, which clarify the difference between the AI's decision and their own reasoning, while most AI systems offer "unilateral" explanations that justify the AI's decision but do not account for users' thinking. To align human-AI knowledge on decision tasks, we introduce a framework for generating human-centered contrastive explanations which explain the difference between AI's choice and a predicted, likely human choice about the same task. Results from a large-scale experiment (N = 628) demonstrate that contrastive explanations significantly enhance users' independent decision-making skills compared to unilateral explanations, without sacrificing decision accuracy. Amid rising deskilling concerns, our research demonstrates that incorporating human reasoning into AI design can foster human skill development.

Additional Key Words and Phrases: AI-assisted decision-making, human-AI interaction, explainable AI, human skills, contrastive explanations

1 INTRODUCTION

Imagine if AI decision-support tools not only improved the quality of our decisions but also enhanced our decision-making skills in the process. Competence, mastery, and skill growth are fundamental drivers of motivation in the workplace and beyond [25, 26]. Individuals are inherently driven to refine their abilities in the tasks with which they engage, whether it's making more informed treatment decisions for patients, sharpening writing skills, or mastering a new programming language. The ongoing process of self-improvement not only leads to better outcomes — decisions, papers, or code — it also provides intrinsic satisfaction by fulfilling people's fundamental need for competence [25]. As AI systems become more integrated into our decision-making tasks, a critical question arises: How will this assistance affect our skill growth and competence in decision-making? Specifically, as AI systems increasingly offer ready-made
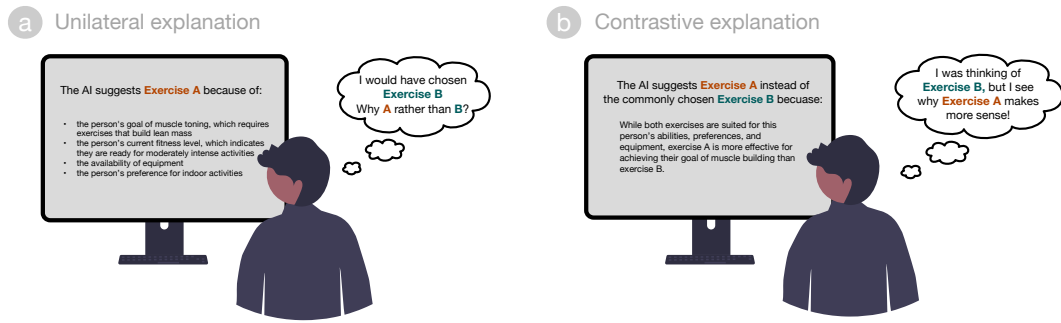
Fig. 1. A simplified illustration of (a) unilateral explanations, which list all the features contributing to the AI's decision, and (b) contrastive explanations, which highlight the differences between the AI's choice and a likely human response for an exercise recommendation task.

"solutions", do individuals develop and improve the underlying skills needed to evaluate and generate high-quality decisions independently, or do they risk becoming overly reliant on AI recommendations?

Fueling broader concerns about deskilling [54, 90], emerging empirical evidence [31, 73] suggests that automation and the design of many current AI decision-support systems might not only fail to nurture our skills but could actively degrade them. For instance, Rinta-Kahila et al. [73] found that people's accounting skill degradation became apparent once a system for fixed assets management was discontinued after years of use. And Gajos and Mamykina [31] demonstrated that providing AI-generated recommendations and explanations did not improve people's decision-making abilities, even when those explanations contained facts from which individuals could learn and improve their decision-making abilities.

Some have argued that when people are provided with AI recommendations they may overrely on the recommendation and only superficially process the explanation [9, 10], thus not improve their learning [31]. Building on this, we posit that even when people attend to the explanation, the reason AI systems fail to improve people's skills, can be partially attributed to the nature of the explanations provided, which often fail to address the specific knowledge gaps that users seek to fill. Typically, these systems offer, what we call, *unilateral* explanations — justifications that focus on why a particular AI recommendation was made, often by detailing the relevant features [72], highlighting regions [79], or presenting the general reasoning in support of the decision. For example, an AI system might recommend treatment P as the best option for a patient because it addresses symptoms X, Y, and Z, without contrasting it to alternative treatments that may address some of those symptoms as well. Yet, research in social science and cognitive psychology has put forth that people naturally seek explanations that are *contrastive* rather than *unilateral* [41, 56, 64]. When people ask why a certain event occurred, or a certain choice was made — "Why P?"— they are often implicitly asking "Why P (referred to as *fact*) rather than Q (referred to as *foil*)?" — where foil is a plausible alternative that was considered but not chosen. Jacobs et al. [43] found that clinicians would prefer contrastive explanations from AI decision support systems as well, which, for example compare and contrast the AI suggestions with existing standards of care. Such explanations are intuitive and engaging because they focus only on the knowledge gap, addressing the specific points of divergence that are of interest or confusion to the explainee.

While numerous computational approaches have been introduced for generating contrastive explanations [1, 3, 45, 86], their focus typically lies on computing the contrast between the fact and the foil rather than on determining a high quality foil. Some approaches consider the foil to be the closest class to the fact in the model [86], which we argue

results in a model-centric foil that does not necessarily align with human reasoning about the task. Others assume the foil is provided or explicitly inputted by the user [45]. Such additional step of making initial decisions has been shown to affect both the acceptance of systems and subjective experience in AI-assisted decision-making [10, 30]. Thus, predicting a human-centered foil without asking for explicit user input remains a challenge for generating useful contrastive explanations.

Building on the existing insights into AI-assisted decision-making, in this paper, we propose a novel approach to enhance AI-powered decision-support systems by generating human-centric contrastive explanations. Our method leverages a "mental model" of humans to provide an explanation in the form of "Why P rather than Q?" — where P represents the AI's recommendation and Q is a plausible alternative response from a human perspective. Unlike unilateral explanations, which justify a recommendation by listing all dimensions that contributed to a decision, our contrastive explanations focus on the distinctions between the AI's suggestion and a likely human response, while highlighting only the dimensions in which the two choices differ. We hypothesize that such contrastive explanations, which highlight knowledge gaps between AI and predicted human responses, will foster greater cognitive engagement and enhance task learning compared to unilateral explanations, while maintaining similar decision accuracy. Additionally, we explore how the quality of the foil (predicted vs. random) and timing (before the person makes a decision vs. after an initial decision) of the contrastive explanation affect their effectiveness, hypothesizing that high-quality foils will maximize learning outcomes and pre-decision timing will maximize acceptance.

To generate contrastive explanations with which to test the hypotheses, we introduce a human-centric framework, which we instantiated for an exercise recommendation decision-making task. Our framework consists of four modules: (1) an AI task model that predicts a response to a decision task (fact), (2) a human model that predicts an average human's response for the same task (foil), (3) a contrast module that identifies the relevant dimensions where the fact and foil differ, and (4) an LLM-powered presentation module that formats these differences into an explanation and adds common sense knowledge (within the constraints provided by the other modules) that bridges the knowledge gap between the AI's recommendation and the human response. To test our hypotheses, we conducted an online between-subjects experiment (N=628) comparing five conditions: no AI, unilateral explanations, contrastive explanations with a predicted foil, contrastive explanations with a random foil, and contrastive explanations provided after an initial decision was made (inputted foil). Our results demonstrated that contrastive explanations with a predicted foil resulted in similar decision accuracy, but significantly enhanced human skill on the task (i.e., human learning) compared to unilateral explanations. Within contrastive conditions, we found that timing of contrastive explanations affected subjective experience but not objective outcomes. Participants in the contrastive explanations with predicted vs. inputted foil did not differ significantly in terms of decision accuracy or human learning but contrastive explanations with predicted foils resulted in significantly higher subjective perceptions of competence, autonomy, and relatedness to the AI than contrastive explanations with inputted foils. Further, we found that the quality of the foil matters: although we used a single model to predict human responses, participants interacting with contrastive explanations featuring a predicted foil improved their learning more than those with a random foil, though the difference was only marginally significant. This result suggests that personalized models, fine-tuned for each individual, may offer additional benefits.

In summary, this paper makes the following main contributions:

- We introduced a contrastive explanations framework for generating human-centered contrastive explanations which compare AI's decision choice to a predicted human response for the same task.

- Our results demonstrated that such human-centered contrastive explanations significantly enhance decision-making skills without sacrificing decision accuracy compared to unilateral explanations, a default explanation design in AI-powered decision support.
- We further presented evidence about which design aspects of contrastive explanations affect objective outcomes and people's intrinsic motivation to engage with the decision task.
- Our work is the first to demonstrate that the *content* of explanations significantly impacts the improvement of human skills, opening up new opportunities for developing more effective explanation designs.
- Our research suggests that decision support tools that consider the decision-makers' knowledge and mental model of the task can enhance people's understanding and proficiency in the task more effectively than current designs of decision-support which provide AI-centric unilateral explanations. With the growing adoption of AI-powered support across tasks and settings, we believe that our findings offer a path forward toward AI systems that upskill, augment, and improve human capabilities.

2 RELATED WORK

2.1 Contrastive Explanations

The field of Explainable AI (XAI) has developed a wide range of methods aimed at making AI systems more understandable and useful to people [36]. Seminal approaches include feature-based explanations like LIME [72] and SHAP [59], which demonstrate how individual features influence an AI decision, as well as saliency maps [79], which highlight image regions that contributed to the outcome. These methods, which we refer to as *unilateral* approaches, focus on explaining why the AI made a specific decision but do so in isolation, without explicitly comparing it to other plausible alternatives.

Meanwhile, Miller [64]'s extensive review of social science research has underscored the significance of contrastive explanations, sparking a new line of inquiry in ML and HCI research. Miller's review highlights that, according to social science literature, explanations people seek and provide are predominantly contrastive [53, 56, 64]. Rather than simply asking "Why P?" to receive a list of features or a sequence of causal events, people often want to know "Why P instead of Q?" — seeking an explanation that clarifies the difference between the actual outcome and an (often implicit) alternative they expected. Lipton [56] refers to "P", the actual event, as the *fact*, and the alternative "Q" as the *foil*.

Social science experts emphasize the value of contrastive explanations for two main reasons [65]. Firstly, they arise from a person's surprise over an unexpected event, revealing their preconceived expectations — essentially offering insight into the individual's mental model and highlighting their knowledge gaps [53, 63]. Secondly, providing and asking for contrastive explanations is less complex and cognitively demanding, making the process more efficient for both the inquirer and the respondent [53, 56, 92]. In AI-assisted decision-making, we further hypothesize that because contrastive explanations highlight (1) the knowledge gap of the inquirer and (2) are shorter, and thus easier to parse, they will result in improved knowledge acquisition from the decision-maker compared to explanations that highlight all the decision factors.

In recent years, machine learning scholars have introduced various computational approaches for generating contrastive explanations, such as pairwise class comparisons [1, 3], tree-based methods [81, 86], or identifying pertinent positives and negatives [27]. Unlike in our work in which the foil seeks to convey explainee's thinking and is generated by a separate model, in the existing techniques, the foil is commonly determined as the closest alternative outcome that would alter the model's decision. For example, in tree-based approaches, foils are selected as the closest non-matching

class leaf, while in counterfactual reasoning Hendricks et al. [39], foils are chosen based on their proximity to the input data but belonging to a different class. It is not clear, however, if these methods correctly anticipate how people and the models disagree. We argue that contrastive explanations in which the foil presents the decision-maker's reasoning more accurately reflect the social science understanding of contrastive reasoning, which seek to clarify the gaps in the explainee's reasoning.

One example of contrastive explanations in HCI literature is Zhang et al. [93]'s framework for generating contrastive explanations in vocal emotion recognition. Like other machine learning techniques, this framework highlights the differences between two similar instances with different class labels, using high-level, human-interpretable concepts rather than granular features. Other related systems produce counterfactual explanations which compare decision instances or hypothetical input space changes [48, 89], rather than outcome differences. It is important to note that contrastive explanations are often conflated with counterfactual explanations, which explore how minimal input changes could lead to different outcomes, while contrastive explanations clarify differences between two outcomes (e.g., "Why treatment P rather than Q?"). In multi-class settings, these explanations (contrastive and counterfactual) are distinct, but for binary classifications tasks, the distinction blurs.

2.2 AI-Assisted Decision-Making

2.2.1 Why optimizing human decision-making skills in AI-assisted decision-making matters? A growing concern, especially with the recent developments in generative AI [54, 90], deskilling refers to the process by which workers lose skills or their proficiency in tasks due to a reduced need to actively engage in those tasks [8]. This often occurs when technology, such as AI and automation [83], takes over some or all responsibilities that were previously performed by humans. As individuals become more reliant on these systems to handle complex or repetitive tasks, they may stop developing or maintaining the expertise required to perform those tasks independently [4]. For example, in AI-assisted decision-making, workers might depend on AI to make recommendations or decisions, which can diminish their critical thinking, problem-solving abilities, and overall competence in that domain over time. Indeed, recent empirical evidence shows that the current designs of decision-support tools that provide AI recommendations and explanations do not seem to support people's growth of decision-making skills [31] and evidence from expert-based systems shows that long-term dependence on such systems does lead to deskilling in those very tasks [73].

While powerful, AI systems are not infallible. They make errors due to biases in the data, limitations in the model, or unforeseen circumstances and they even hallucinate. In the short term, when humans have strong decision-making skills, they are better equipped to recognize and override AI mistakes, can critically assess the AI's recommendations, apply domain expertise, and contribute meaningfully to the decision-making process, resulting in more accurate and nuanced outcomes. In the long term, nurturing independent and strong decision-making skills is essential for humans to retain autonomy in decision-making, transfer their expertise to new situations, and adapt to evolving technologies. Such independent decision-making protects both accountability and human agency as AI becomes more integrated into workflows.

Our work adds to the nascent body of research in AI-assisted decision-making, which is concerned with improving human decision-making skills in addition to accuracy of the decisions [11, 31].

2.2.2 Eliciting cognitive engagement to calibrate reliance on AI. Early optimism that AI decision-support tools will inevitably enhance human decision quality [61] has dwindled in light of accruing empirical evidence that paints a more complex picture [5, 10, 33, 75, 88]. Intuitive designs that rely on simple XAI approaches, such as providing AI

recommendations alongside (unilateral) explanations, have been shown to lead to overreliance — where users follow incorrect AI advice — across diverse tasks, settings, and explanation styles [5, 10, 32, 44, 75]. This empirical evidence has prompted extensive research in designing interventions beyond explanation content that encourage appropriate human reliance on AI. Some endeavours to addressing this challenge focus on training or onboarding sessions aimed at helping individuals develop a mental model of the AI [18, 19, 49, 67, 68, 71], providing meta-information about the AI's behavior and limitations [15, 67], helping individuals calibrate their own self-confidence about the task [38, 60], or enhance user agency by giving them control over input feature selection and algorithmic assistance [21, 51].

By prompting users to reflect on two choices, contrastive explanations fall within one such growing category of interventions designed to compel deeper cognitive engagement with AI support. Scholars have suggested that overreliance on AI often stems from people's superficial engagement with AI recommendations and explanations [10, 31]. In response, various interventions have been developed to enhance cognitive engagement, including cognitive forcing [10], evaluative AI [66], explanations provided without decision recommendations [31], explanations framed as questions [24], or offering more than one decision suggestion [22, 58].

While many of these interventions have shown promise in human-AI decision-making quality, they often introduce trade-offs, such as reducing subjective experience [10] or requiring more time [20, 84], compared to simply providing AI recommendations with explanations. For instance, cognitive forcing functions [10] compel deeper cognitive engagement by requiring people to make decisions before receiving AI support. While these interventions significantly reduce overreliance compared to presenting AI recommendations and explanations upfront, they also lead to significantly lower subjective experience. Evaluative AI also proposes a paradigm that involves decision-makers making provisional decisions, before seeing an AI critique of their choice [66]. Although no empirical studies have yet operationalized evaluative AI, evidence from Buçinca et al. [10] and Fogliato et al. [30] suggests that people generally tend to dislike receiving AI support *after* having made a decision. Building on this evidence, we hypothesize that providing contrastive explanations *after* prompting a person to make an initial decision may similarly have a negative effect on subjective experience, thus hindering the uptake of systems that provide such support in real-life scenarios. However, there is a trade-off here because providing a contrastive explanation after a person has revealed their initial decision has the obvious advantage of revealing the actual foil to the system. This, in turn, can make the explanations more useful for human decision-making and learning compared to settings where the foils are imperfectly predicted.

Offering more than one decision suggestion or source of advice has also been explored as a mechanism to enhance engagement and calibrate reliance on AI support [5, 22, 58]. For example, Bansal et al. [5] show that providing top two AI predictions and Lu et al. [58] show that offering a "second opinion" in addition to the main AI support, either from another AI or peers, can reduce overreliance on AI recommendations in certain situations. Contrastive explanations also make *two* options salient to the decision-maker — the fact and foil — *along* with reasoning that supports one over the other. It is unclear whether the presence of the explicit comparison in contrastive explanations might dilute the "second opinion" effect that previous studies have shown reduces overreliance.

Finally, the studies mentioned above treat cognitive engagement as a mechanism for fostering appropriate reliance on AI, mostly focusing on optimizing human-AI decision accuracy by encouraging deeper thought about AI recommendations. Building on the work of Gajos and Mamykina [31], our study instead examines cognitive engagement as a means of enhancing human learning about the task.

2.2.3 Assisting decision-making with LLM-generated explanations. The emergence of Large Language Models (LLMs) has sparked interest in their potential to generate explanations that enhance decision-making. In the domain of programming

assistants, Yan et al. [91] introduced an LLM-powered system that generates natural language explanations to clarify the functionality of each code suggestion. For data annotation tasks, Wang et al. [87] leveraged an LLM to predict annotation labels and provide explanations for its choices. In recommendation systems, Silva et al. [78] used LLMs both as the recommendation engine and as a generator of personalized explanations to improve user experience.

In contrast to such approaches, which use LLMs for both task-centric predictions (e.g., code suggestions, recommendations) *and* explanation, our work separates these functions. Like Slack et al. [80], who introduce a chat-based interface to query a predictive machine learning model, we exploit LLMs for their language generation capabilities, and in addition for their common sense reasoning. We rely on a trusted predictive model for generating task-related predictions and explanation dimensions, while the LLM is solely responsible for turning a scaffold produced by the predictive model into natural language, and filling in small common-sense knowledge gaps needed to interpret the model's predictions. This separation preserves the accuracy of predictions and explanations while benefiting from the interpretability and coherence of LLM-generated rationalizations, thereby minimizing the risk of the notorious hallucinations for which LLMs are known [47].

2.3   Human Intrinsic Motivation and AI Assistance

With AI systems redefining workflows and the way tasks are carried out, questions surrounding their effect on people's motivation about the tasks for which they receive assistance are becoming more pressing [11]. According to the seminal Self-Determination Theory (SDT), individuals feel intrinsically motivated when three psychological needs—competence, autonomy, and relatedness—are met during an activity [25]. Competence refers to the need to feel skilled and effective in the activity, autonomy reflects the need to have control over how the activity is carried out, and relatedness involves the need to feel connected to others and to experience a sense of belonging while engaging in the activity. These three needs are fundamental for fostering intrinsic motivation, which leads to greater engagement, performance, and overall satisfaction with the task [25]. The introduction of AI assistance into decision-making processes can affect these psychological needs in multiple ways. For example, while AI might enhance short-term feelings of competence by providing support in the moment of decision-making [29], it may simultaneously undermine long-term mastery, as current designs do not always facilitate skill development [31]. Similarly, AI can diminish a user's sense of autonomy if they feel overly dependent on the system, reducing their ownership of task outcomes.

We hypothesize that both the outcomes of the interaction and the design of the AI system influence perceptions of competence and autonomy. On the outcome side, we expect that AI systems that actively support skill development will enhance feelings of competence. In terms of design, approaches where the AI critiques each decision after it is made (e.g., contrastive after) may undermine users' feelings of competence and autonomy, as they could perceive the AI more as a micromanager than a supportive tool, constantly pointing out flaws and dictating its preferred way of doing things. Additionally, even when AI assistance is provided before a decision, designs that emphasize only one option (e.g., unilateral conditions) can still reduce the sense of autonomy compared to those that present multiple options, broadening the decision-maker's scope of consideration (e.g., contrastive before conditions).

In SDT, relatedness refers to the connection an individual feels toward colleagues or collaborators, typically measured through questions about trust, similarity in reasoning, and willingness to engage in future interactions. We adapt these constructs to assess relatedness to AI, hypothesizing that designs fostering competence and autonomy will similarly enhance relatedness to AI systems.
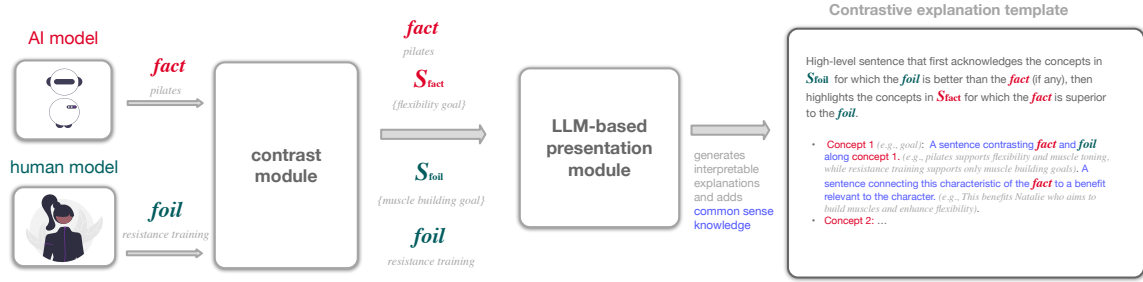
Fig. 2. The Contrastive Explanation Framework: The AI task model predicts the AI's response for a given decision task (fact), while the human model predicts the user's response for the same task (foil). The contrastive module then analyzes the differences between the AI's and the human's responses, generating task dimensions where the fact is superior to the foil (S_{fact}) and, if any, where the foil is superior to the fact (S_{foil}). Finally, the presentation module, powered by a large language model (LLM), formats the information into an interpretable explanation, filling in small common-sense knowledge gaps within the constraints of the predictive models. The example generation outlined in the figure is relates to the character vignette in Figure 3

.

3   THE CONTRASTIVE EXPLANATION FRAMEWORK & HYPOTHESES

Imagine a clinician reviewing an AI-powered decision-support system's recommendation for a patient's treatment plan. The AI suggests Medication A, but the clinician had Medication B in mind based on their experience with this condition. Existing AI systems would simply explain why Medication A is recommended. However, this leaves the clinician wondering why Medication B, which they deemed suitable, is not the better choice. A contrastive explanation may elucidate this knowledge gap as follows: *"While Medication B is a common and a viable choice for most patients because of its short treatment duration, Medication A is recommended due to its lower risk of drug interactions with this patient's current medications."*

We propose the Contrastive Explanation Framework to address the limitations of current AI-powered decision-support systems by providing contrastive explanations that acknowledge human's alternative considerations when suggesting a decision. This framework is composed of four main components: (1) an AI task model, (2) a model of how humans are likely to reason about this task, (3) a contrastive module, and (4) a presentation module. The AI task model is the standard AI system that predicts the AI's response for a given decision task (fact), while the human model predicts an average user's response — a plausible alternative (foil) — for the same task based on a model trained on previous human decisions. Based on AI model (e.g., weights), the contrastive module then analyzes the differences between the AI's and the likely human's responses, generating task concepts in which the fact is superior to the foil (e.g., lower risk of drug interaction) and task concepts, if any, in which the foil is superior to the fact (e.g., shorter treatment duration). Finally, the presentation module, powered by a large language model, formats these dimensions and fills in the common sense knowledge that focuses on the knowledge gap that may lead someone to pick foil as opposed to fact.

To evaluate the effectiveness of contrastive explanations in improving human learning and accuracy in AI-assisted decision-making, we instantiated this framework with an exercise recommendation task, and conducted an experiment in which people were asked to complete a sequence of decisions and were randomized in one of the 5 different conditions:

- **No AI (Baseline).** Participants in the No AI condition completed the study without any AI support.

- **Unilateral** In this condition, participants interacted with the typical AI recommendation and explanation paradigm. The AI suggested a choice and provided reasoning to justify why that choice was the best one. The explanation was unilateral, emphasizing all the concepts and evidence supporting the AI's suggestion.
- **Contrastive predicted (with predicted foil).** The *contrastive predicted* condition was designed to provide participants with a contrastive explanation that compares the AI recommendation (fact) with the alternative (foil) that a human may likely consider, as predicted by the human model. In the interface, we presented the foil as a choice that "many people" would likely make in a similar situation. The explanation highlighted only the concept(s) in which the two choices differ, emphasizing why the AI's recommendation is superior to the foil.
- **Contrastive random (with random foil).** Presentation-wise, the *contrastive random* condition was identical to the contrastive condition. However, in this case, the foil was selected randomly from the six possible choices rather than being predicted by the human model.
- **Contrastive after (with inputted foil).** In the *contrastive after* condition, participants first made their own decision before receiving the AI's recommendation and the contrastive explanation, in which participant's decision was used as the foil. In situations when the inputted foil was the same as the AI suggestion, participants were presented with a unilateral explanation supporting their choice.

3.1 Hypotheses and Research Questions

In our hypotheses, we sometimes refer jointly to *contrastive predicted* and *contrastive after* conditions, in which the foil is not random, as *contrastive with a sensible foil.*

Our main hypotheses are that contrastive explanations with a sensible foil will improve participants' decision-making skills [1] more effectively and result in accuracy that is equal to or better than unilateral explanations. Furthermore, within the contrastive conditions, we hypothesize that *contrastive explanations with a predicted foil* will result in greater human learning than those with a random foil, and offer a superior subjective experience compared to contrastive explanations with an inputted foil.

We categorize these main and other hypotheses and research questions by interaction outcomes — human learning, accuracy, and subjective experience — and elaborate them below. To enhance readability, we abbreviate learning-focused hypotheses and research questions as H-L and RQ-L and accuracy-focused ones as H-A and RQ-A, respectively. For hypotheses related to subjective measures, we use the -S suffix (e.g., H-S1).

3.1.1 Human Learning.

H-L1: Contrastive explanations with sensible foil — predicted **(H-L1a)** or inputted **(H-L1b)** — will lead to more learning than providing people with no AI support.

H-L2: Contrastive explanations with sensible foil — predicted **(H-L2a)** or inputted **(H-L2b)** — will lead to more learning than unilateral explanations.

H-L3: Contrastive explanations with predicted foil will lead to more learning than contrastive explanations with a random foil.

RQ-L1: Will contrastive explanations with predicted foil (provided at the decision-making time) lead to different learning than contrastive explanations after the decision is made (contrastive after)?

3.1.2 Accuracy & Overreliance.

---

[1]In this paper, we use the terms "improving human learning" and "improving decision-making skills" interchangeably.

(a) Sample task with contrastive explanation

(b) Unilateral explanation

(c) Contrastive explanation after

Fig. 3. Illustration of the exercise recommendation decision-making task featuring different explanation designs. 3a shows a sample of the task with *contrastive* explanation, whereas 3b and 3c depict only the explanations for the respective conditions. In the *contrastive random* condition, the presentation was identical to the contrastive condition, but with the alternative (foil) selected randomly. In the *no-AI* condition (not illustrated), participants made decisions without any AI assistance.

H-A1: Contrastive explanations with sensible foil — predicted **(H-A1a)** or inputted **(H-A1b)** — will lead to equal or better decision accuracy compared to unilateral explanations.

RQ-A1: Will contrastive explanations — predicted or random — which present two choices, reduce overreliance on AI, compared to unilateral explanations?

3.1.3 Subjective Experience.

H-S1: Contrastive explanations with predicted foil will lead to higher perceived competence, autonomy, and relatedness to AI than unilateral explanations.

H-S2: Contrastive explanations with predicted foil will lead to higher perceived competence, autonomy, and relatedness to AI than contrastive explanations with inputted foil.

In the following sections, we describe an exercise recommendation task and the instantiation and implementation of the contrastive explanations framework for the exercise recommendation task.

4 EXERCISE RECOMMENDATION TASK DESIGN

To create a decision-making task accessible to laypeople on crowd-sourcing platforms while presenting cognitive challenges similar to high-stakes decisions (e.g., treatment selection), we collaborated with a kinesiology expert, a co-author of this paper. We designed scenarios for an exercise recommendation task, as shown in Figure 3. Participants are tasked to choose the best exercise from a list of options based on a fictional character's description, goals, and preferences. This

task is designed to be easy to understand yet complex enough to mimic clinical treatment decisions. Clinicians consider many (sometimes competing) factors when selecting treatments, such as patient condition, treatment preferences, side-effect tolerance, and constraints. Similarly, selecting the right exercise involves weighing the individual's goals, preferences, and capabilities, requiring analogous cognitive steps.

4.1 Generating the fictitious characters

We generated vignettes of fictitious people by randomly sampling their demographics from probabilities obtained from the US Census[2], Centers for Disease Control and Prevention[3], and the US Bureau of Labor statistics[4] (name, age, gender, BMI, physical activity level, occupation). According to the sampled fictitious character, we manipulated or randomly sampled the following factors which were deemed important for exercise prescription by the expert: (1) their fitness level and maximal intensity (based on demographics), (2) their exercise goal (*e.g.*, building muscles, weight loss, flexibility), and (3) their exercise preference (*e.g.*, indoor/outdoor, group/individual). We implemented these steps as fictitious character generation process that allowed us to generate different characters.

4.2 Curating the exercises

To build an exercise repository for recommending activities to fictional individuals, we curated a list of 59 leisure activities from a comprehensive compendium, which included various physical activities, from sports to everyday tasks like housework and occupational activities [2]. In the compendium, each activity was labeled with its MET (metabolic equivalent), which denotes the energy requirement for basal homeostasis (1 MET is roughly the energy required to sleep or watch TV). Moderate activities require 3-6 METs, while vigorous activities require more than 6 METs. We also labeled the exercises based on (i) their goals (cardio, muscle building, flexibility), (ii) whether they are typically performed indoors or outdoors, and (iii) whether they are typically performed individually or in a group. From this list, we selected seven representative exercises for the dropdown menu: aerobics, bicycling, boxing, jog/walk combination, pilates, resistance training, and swimming. See Appendix for a detailed description of the selection process.

4.3 Representing characters and exercises

To prescribe exercises to characters, we first represented them in a joint representation space. Guided by the domain expert, we constructed a relatively simple representation space consisting of three broad concepts: (1) intensity, (2) goal, and (3) preference. Each exercise and generated character was encoded onto these three broad concepts as described below.

Intensity. For exercises, intensity captures the level of exertion or effort the exercise requires, measured in METs. One MET is defined as the oxygen consumption of 3.5 milliliters of oxygen per kilogram of body weight per minute (3.5 ml/kg/min), which is roughly the rate of oxygen consumption at rest.

For characters, intensity captures the level of exertion or effort a character can sustain during physical activity (i.e., their cardiorespiratory fitness). It is quantified by the reserve oxygen uptake (VO_{2_R}), which represents the additional oxygen consumption capacity a person has beyond their resting state. This reserve is determined by subtracting the resting oxygen uptake (3.5 ml/kg/min or 1 MET) from the maximal oxygen uptake ($VO_{2_{max}}$), which is the highest rate at which the body can use oxygen during intense physical activity. Maximal oxygen uptake is assessed in clincial

---

settings using a treadmill test, but various equations have been proposed as useful proxies [70]. Following Jang et al. [46], we calculated the cardiorespiratory fitness of a character as a function of age, sex, BMI, and current physical activity level [5](rating of physical activity on a 7-point scale).

Goal. For exercises, goal captures the type of benefit the exercise has in the body, consisting of three dimensions: cardiovascular improvement, muscle building, or flexibility. For characters, goal reflects what the character aims to achieve through exercise, in terms of the same three dimensions: improving cardiovascular health, building muscle, or enhancing flexibility. Note that additional domain knowledge (e.g., cardio is beneficial for weight loss) is necessary to convert some of the higher level character's exercise goals (e.g., losing weight) to the representation space.

Preference. For exercises, preference indicates whether the exercise is typically performed indoors or outdoors, and whether it is usually done individually or in a group. For characters, preference captures the character's preferred exercise environment (indoor/outdoor) and social setting (individual/group).

4.4 Designing the objective function

Having constructed joint representations for characters and exercises, we now formalize our setting and explain the objective function we designed for recommending exercises to characters.

Let a fictitious character representation be $\mathbf{x} \in \mathbb{R}^D$ and an exercise representation be $\mathbf{y} \in \mathbb{R}^D$, where $D = 6$ and both representations are structured with dimensions representing intensity, goals, and preferences (e.g., $\mathbf{x} = [x_{\text{MET}}, x_{\text{cardio}}, x_{\text{muscle}}, x_{\text{flexibility}}, x_{\text{environment}}, x_{\text{social setting}}]^T$, with $\mathbf{y}$ following a similar structure). Our goal was to create a function that scores the "goodness" of an exercise for the given character. We designed a linear objective function:

$$f(\mathbf{g}(\mathbf{x}, \mathbf{y}), \mathbf{w}) = \mathbf{w}^T \mathbf{g}(\mathbf{x}, \mathbf{y}), \tag{1}$$

where $\mathbf{g}(\mathbf{x}, \mathbf{y})$ is a piece-wise vector-valued function (devised with the expert) that returns a joint representation (vector) of the person and the exercise for each dimension. $\mathbf{g}(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{D+1}$ concerning the following aspects: intensity, goal, and preference.

$$\mathbf{g}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \min(0, x_1 - y_1) \\ \min(0, y_1 - x_1) \\ [\mathbb{1}[x_c > 0]((y_c - x_c) + \mathbb{1}[y_c == x_c])]_{c \in \{2,3,4\}} \\ \mathbb{1}[y_c == x_c]_{c \in \{5,6\}} \end{bmatrix} \tag{2}$$

$\leftarrow$ **Intensity**: Penalize exercises exceeding character's capabilities.

$\leftarrow$ **Intensity**: Penalize exercises underutilizing character's capabilities.

$\leftarrow$ **Goal**: Match each stated subgoal (cardio, muscle building, flexibility).

$\leftarrow$ **Preference**: Match each preference (environment, social setting).

In the equation above, the subscripts refer to dimensions of $\mathbf{x}$ and $\mathbf{y}$, and $\mathbb{1}$ denotes the indicator function, which takes the value 1 if the condition inside the brackets is true and 0 otherwise.

4.5 Learning the expert weights

The parameterized objective function (equation 1) enables learning weights $\mathbf{w}$ from different sources of labels. We aim to learn f_{expert} with weights $\mathbf{w}_e$ based on expert labels, and f_{human} with weights $\mathbf{w}_h$ based on crowdworker labels. The expert model f_{expert}, takes a description of a fictitious character and exercise, and outputs a real-valued score indicating how well the exercise matches the goals, abilities and preferences of the fictitious character. We trained and validated

---

[5]$VO_{2max} = 48.392 - 0.088(age) + 12.335(sex; men = 1, women = 0) - 0.386(BMI) + 0.693(PA)$

this model on expert labels of optimal exercises for a series of characters. Similarly, the human model, which captures how humans reason on average (described in 5.2), was trained on crowdworkers' (i.e., laypeople's) labels.

Generating a series of diverse fictitious characters, we asked the kinesiology expert to select top exercises for them from a list of top 15 exercises (out of 59) selected with a "dummy" scoring function which equally weighted each dimension. For every character, the expert provided a best set of exercises $\mathcal{S}_1$ typically consisting of two or three similar exercises (e.g., pilates, yoga), and a second best set of exercises $\mathcal{S}_2$ that would still be a reasonable choice but not as good as the first set. With 15 exercises in the list, these labels provided multiple pairwise comparisons between the individual exercises. For every exercise $\mathbf{y}^i \in \mathcal{S}_1$, then $\mathbf{y}^i$ is a better choice than $\mathbf{y}^j$ for every other exercise $\mathbf{y}^j \notin \mathcal{S}_1$. Similarly, for every exercise $\mathbf{y}^i \in \mathcal{S}_2$, then $\mathbf{y}^i$ is a better choice than $\mathbf{y}^j$ for every $\mathbf{y}^j \notin \{\mathcal{S}_1 \bigcup \mathcal{S}_2\}$.

With a dataset of rankings, our goal was to learn the expert weights $\mathbf{w}_e$ from equation 1. Ranking problems, particularly with linear ranking functions, can be transformed into classification problems by considering pairwise differences between elements [40]. This approach involves transforming the ranking task of a set of items (e.g., exercises) into several binary classification tasks. For each item pair, a difference vector of their features $(\mathbf{u}^i - \mathbf{u}^j)$ is generated and the label corresponds to their relative order (e.g., the label $v = 1$ if item i is a better choice than item j and -1 otherwise). In our setting, the items correspond to exercises. A binary classifier is then trained on these labeled pairs to predict which of the given two items should be ranked higher. When using a linear binary classifier $v = sign(\mathbf{w}^T(\mathbf{u}^i - \mathbf{u}^j) + b)$, the coefficients of the model represent the weights of the feature differences, thereby indicating the importance of each feature dimension in determining the ranking.

Let exercise $\mathbf{y}^*$ be a better choice than exercise $\mathbf{y}^i$ for a character $\mathbf{x}$. In our setting this looks as follows:

$$[\mathbf{g}(\mathbf{x}, \mathbf{y}^*) - \mathbf{g}(\mathbf{x}, \mathbf{y}^i), 1] \text{ or } [\mathbf{g}(\mathbf{x}, \mathbf{y}^i) - \mathbf{g}(\mathbf{x}, \mathbf{y}^*), -1],$$

where the first element in the square brackets is the input to our classifier model, and the second element is the label.

To avoid biasing the classifier, we randomly assign a pair to either have a positive (1) or a negative (-1) label (i.e., "$\mathbf{y}^*$ is better than $\mathbf{y}^i$" or "$\mathbf{y}^i$ is worse than $\mathbf{y}^*$"). We fit an SVM classifier with a linear kernel to these tuples of data with expert labels, thereby recovering the coefficients as the expert weights $\mathbf{w}_e$ for the scoring function: $f(\mathbf{g}(\mathbf{x}, \mathbf{y}), \mathbf{w}_e) = \mathbf{w}_e^T \mathbf{g}(\mathbf{x}, \mathbf{y})$. (For implementation details and the evaluation of the expert model see Appendix A.1.1.) We followed a similar approach to learn the human model weights from crowd-sourced data, as described in Section 5.2.

5 APPLYING THE CONTRASTIVE EXPLANATION FRAMEWORK TO THE EXERCISE TASK

Our goal is to generate contrastive explanations (using the framework in section 3 for the exercise recommendation task explained in section 4. In this section, we describe this process, and we end up with contrastive explanations like the ones shown in figure 3. To do so, we use a simulated AI model (we control the accuracy of this model), generate foils using the human model weights, generate contrast concepts using our representation $\mathbf{g}(\mathbf{x}, \mathbf{y})$, and generate the explanations using an LLM.

5.1 Simulated AI model: Generating the *fact*

The AI model in our framework represents the common way in which models are trained for specific tasks (e.g., disease diagnosis) by exposing them to vast amounts of data, which allows them to identify patterns and make decisions based on learned statistical relationships. However, because these models operate solely within the confines of the data they

have encountered, they achieve high performance in familiar decision instances, but they also make mistakes when encountering novel or poorly-represented scenarios.

To emulate real-world situations, we designed a *simulated* AI model such that it performs better than the average human but that it also occasionally makes mistakes. We chose to simulate the AI model because we wanted to have control over the types of mistakes the AI makes. Our formative studies indicated that unassisted people achieve on average 30% accuracy on selecting the top exercise out of 7 choices, and we designed the AI model to have an accuracy 71.4%. We used the expert model weights to decide the top exercise recommendation when the AI made a correct decision. For a given decision instance (i.e., character), the top AI suggestion is the exercise with the highest score under the expert model weights $\mathbf{y}_{\text{fact}} = \text{argmax}_{\mathbf{y}^i}(f(\mathbf{g}(\mathbf{x}, \mathbf{y}^i), \mathbf{w}_e))$. In the contrastive explanation framework, we refer to the AI generated exercise as the *fact* (even when it is a wrong suggestion).

To make the AI err, we chose to select a reasonable alternative from among the exercises rather than a random one, as the latter would make AI errors too obvious to participants. Therefore, the AI suggestion in such instances was the foil — the top exercise selected by the human model (as described in Section 5.2: this is always different to the expert model's top exercise.).

5.2 Human model: Generating the *foil*

As motivated in previous sections, we believe that contrastive explanations are most effective when the foil represents a likely human answer. For instance, in contexts with established guidelines, such as medical decision-making, the foil could be the guideline-recommended action [43]. In situations without established guidelines, the foil can be inferred from prior human decisions. In our implementation of the contrastive explanation framework for the exercise recommendation task, we chose to implement the *foil* as the likely human response to a given question. Specifically, we build a human model that predicts the exercise laypeople would select for previously unseen fictitious characters by training on unassisted human responses. We implemented a generic model to represent human decision-making, which was sufficient for our simple task. However, depending on the context, personalized models that adapt and update as they learn more about individual users could be more appropriate.

We generated a series of fictitious characters and ran an online study on Prolific to collect responses from crowd workers who served as non-domain experts. See Appendix A.2.1 for details of the data collection study and the evaluation of the human model. To learn the human model weights, we followed the same procedure as we did for the expert model weights, and as described in section 4.5.

Given a character and two exercises, our learnt linear SVM classifier predicted which exercise is more likely to be selected by the human non-expert. The coefficients of the classifier with which this decision was achieved yielded the human model weights for each concept (i.e., goal, intensity, preference). Therefore, we constructed a scoring function based on human model weights as well: $f(\mathbf{g}(\mathbf{x}, \mathbf{y}), \mathbf{w}_h) = \mathbf{w}_h^T \mathbf{g}(\mathbf{x}, \mathbf{y})$.

In our implementation, we selected the foil as the exercise with the highest score under the human weights that was not the same as the expert choice: $\mathbf{y}_{\text{foil}} = \text{argmax}_{\mathbf{y}^i}(f(\mathbf{g}(\mathbf{x}, \mathbf{y}^i), \mathbf{w}_h))$, where $\mathbf{y}^i \neq \mathbf{y}_{\text{fact}}$. This approach selects the most likely *incorrect* human answer. When the simulated AI was to provide a wrong suggestion (i.e., the *fact* was suboptimal), the output of this human model was presented as the *fact*, and the new *foil* was the second most likely incorrect human model answer: this is still an incorrect choice, but less likely to be selected by people than the first one.

5.3 Contrast Module: Generating the contrast concepts

The goal of the contrast module is to generate the dimensions or features in which the fact and the foil differ. Specifically, what aspects render the fact superior to the foil, and in what aspects (if in any) is the foil superior to the fact.

In our setting, these dimensions indicate the three main concepts of the task: *intensity*, *goal*, and *preference*. To generate these dimensions we employed the following approach. Let $\mathbf{y}_{\text{fact}}$ and $\mathbf{y}_{\text{foil}}$ be the two exercises generated by the AI and the human model for character $\mathbf{x}$, respectively. Our goal is to identify the dimensions in which these two exercises differ based on the expert model's weights. For each exercise, we computed the element-wise product of the expert model weights with the joint character-exercise representation $\mathbf{g}(\mathbf{x}, \mathbf{y})$, resulting in the weighted vectors $\mathbf{w}_e \circ \mathbf{g}(\mathbf{x}, \mathbf{y}_{\text{fact}})$ and $\mathbf{w}_e \circ \mathbf{g}(\mathbf{x}, \mathbf{y}_{\text{foil}})$ for the AI-generated exercise and the human model-generated exercise, respectively.

Next, we calculated the difference between these two weighted vectors to determine the dimensions along which the exercises differ according to the expert model's weighting scheme. This difference vector, $\Delta \mathbf{g}_{AI}$, is given by:

$$\Delta \mathbf{g}_{AI} = \mathbf{w}_e \circ \mathbf{g}(\mathbf{x}, \mathbf{y}_{\text{fact}}) - \mathbf{w}_e \circ \mathbf{g}(\mathbf{x}, \mathbf{y}_{\text{foil}}) \tag{3}$$

Non-zero dimensions of $\Delta \mathbf{g}_{AI}$ indicate where the two exercises differ. A positive value indicates that the fact is superior to the foil in that dimension, while a negative value indicates that the foil is superior to the fact. Therefore, the contrastive module generates two sets of dimensions, dimensions for which the fact is superior to the foil: $\mathcal{S}_{\text{fact}} = \{c \mid \Delta \mathbf{g}_{AI}[c] > 0\}$ and those for which the foil is superior to the fact $\mathcal{S}_{\text{foil}} = \{c \mid \Delta \mathbf{g}_{AI}[c] < 0\}$, where c denotes the dimension. Because the foil may not be superior to the fact in any dimension, $\mathcal{S}_{\text{foil}}$ can be an empty set. However, by definition $\mathcal{S}_{\text{fact}} \neq \emptyset$.

5.4 Presentation Module: Generating interpretable explanations

Once the *fact*, *foil*, and the dimensions where they differ are generated, the presentation module's purpose is to convert this information into a format that is easily understood by humans. We chose to implement an LLM-powered presentation module which is guided by our trusted predictive model, allowing little room for hallucinations. Given $\mathbf{y}_{\text{fact}}$, $\mathbf{y}_{\text{foil}}$, and the sets for which each are superior ($\mathcal{S}_{\text{fact}}$, $\mathcal{S}_{\text{foil}}$), the LLM-powered presentation module adds common sense knowledge and turns the explanations into prose.

Specifically, the LLM adds knowledge to create the mapping from the the representation space (i.e., concepts) in which the predictive model operates to the input (i.e., vignette) and output spaces (i.e., exercises). For example, let $\mathbf{x}$ be a fictitious character whose goal is to lose weight. Let $\mathbf{y}_{\text{fact}}$ correspond to the representation of activity *running* and $\mathbf{y}_{\text{foil}}$ correspond to the representation of activity *pilates*. Further, let $\mathcal{S}_{\text{fact}}$ include {*goal_cardio*}. In other words, *running* is superior to *pilates* because it supports *cardio goals*. The remaining domain knowledge required to fully understand the explanations are the following: 'cardio benefits weight loss', 'running is a cardio exercise' and 'pilates is not a cardio exercise'.

Therefore, highlighting cardio as a differing dimension may not be enough without explaining those domain facts. The LLM is prompted to fill in these knowledge gaps, given the information '*running* is superior to *pilates* in supporting *cardio* goals'. Note that there is little room for the LLM to hallucinate facts, because we are constraining the generation process with the fact, foil, and concepts ($\mathcal{S}_{\text{fact}}$, $\mathcal{S}_{\text{foil}}$) that are generated by the predictive models.

The LLM was always shown the character's vignette, and told the representation space dimensions that we identified as important (from section 4.4).

As shown in Fig 2, for contrastive explanations, the LLM was given $\mathbf{y}_{fact}$, $\mathbf{y}_{foil}$, $\mathcal{S}_{fact}$, $\mathcal{S}_{foil}$. For unilateral explanations, only $\mathbf{y}_{fact}$ was provided to the LLM. Templates which guided the LLM to generate the explanations are provided in the Appendix A.3. We used the OpenAI API [69] and chose GPT-4 to generate the explanations [6].

6 EXPERIMENT

6.1 Task description

Participants were shown vignettes of fictitious characters and were asked to select the optimal exercise for the character in question based on their goals, capabilities, and preferences. They had to make a selection of the top exercise among 7 exercises, which were fixed choices across vignettes and alphabetically ordered in the drop-down list: aerobics, bicycling, boxing, jog/walk combination, pilates, resistance training, and swimming.

6.2 Conditions

Participants were randomized into one of the five conditions:*no AI*, *unilateral*, *contrastive predicted*, *contrastive after*, and *contrastive random*, as described in Section 3. Figure 3 provides a sample of a decision task with illustrations of the key conditions.

6.3 Procedure

Participants accessed the study online through Prolific, where they first provided informed consent. They then completed pre-task questionnaires, including a brief demographic survey, a six-item Need for Cognition (NFC) Scale [55], and a seven-item Actively Open-minded Thinking (AOT) Scale [37]. The study consisted of three blocks: pre-test and post-test blocks, each with 5 exercise prescription tasks without AI support which served for measuring human learning, and an intervention block with 14 tasks where participants interacted with one of the AI interaction designs (or no AI, depending on their randomization). After completing the tasks, participants filled out a shortened version of the Intrinsic Motivation Inventory (IMI) [62, 74], a self-reported instrument intended to measure participants' subjective experience with the task, which assessed their perceived autonomy, competence, relatedness to AI, and interest/enjoyment, using 4 questions for each construct (except for relatedness, for which 3 questions where used). An additional question was included to assess mental demand.

6.4 Participants

We conducted a power analysis using G*Power [28] to determine the required sample size for detecting a small effect size in our study with 5 conditions. With a small effect size, an α error probability of 0.05, and a desired power of 0.80, the analysis indicated that a total of 548 participants would be needed to achieve sufficient power to detect the effect. To account for filtering of spammers, a total of 800 participants were recruited to complete the task via Prolific. Participation was limited to US adults fluent in English. Recruited in batches, participants received an average compensation of $2.70 (USD) per task. To ensure a compensation rate of $12 per hour, we adjusted the payment from $2.40 in the initial small batches to $2.75 in later batches, reflecting the median time participants spent on the study. The average age of participants was $M = 35.76$ ($SD = 11.71$) and their education distribution was 0.5% pre-high school, 19.4% high school, 75.8% college, 5.7% post-graduate degree, and 4.6% did not disclose their education.

---

[6]The first author manually reviewed the generated explanations to verify whether the LLM introduced any hallucinations; we elaborate on this process in the limitations section.

6.4.1 Exclusion criteria. We retained 628 participants for analyses. To ensure meaningful engagement, participants with a median response time under 4 seconds were excluded, as this suggested insufficient consideration of the tasks, which required reading vignettes and selecting exercises. Those with any response time exceeding 2.5 minutes (90th percentile) were also removed to avoid data distortion from distractions. Additionally, participants in AI-assisted conditions who performed near random (below 20% accuracy) or selected the same exercise for more than half of the study were excluded for potential misunderstanding. For subjective experience analyses, 6 participants were removed due to technical issues they encountered during the post-study questionnaire.

6.5 Approval

This study received approval from our institution's IRB under protocol number [anonymized for review].

6.6 Design & Analysis

This study followed a between-subjects design, with the condition as the factor. Each participant interacted with one of the five conditions.

We collected the following indicators of performance and learning:

- **Accuracy:** Percentage of correct answers provided by participants in the intervention block, where a correct answer is one that matches the ground truth.
- **Overreliance:** Percentage of answers that matched the AI's suggestions in questions for which participants received AI support and the AI's suggestion was incorrect.
- **Learning:** Percentage of correct answers on *post*-intervention questions (controlled by participant's performance on *pre*-intervention questions).

For accuracy, learning, and overreliance in text and in figures we report the *marginal means* produced by the regression models that included performance on pre-intervention questions as a covariate.

To assess the subjective experience, we collected the following measures assessed on a 5-point Likert scale, unless stated differently (See Appendix A.4 for the questionnaire):

- **Perceived Competence:** Four questions adapted from the Intrinsic Motivation Inventory (IMI) to measure participants' feelings of effectiveness and competence in the task.
- **Perceived Autonomy (Choice):** Four questions adapted from the IMI capturing the degree of autonomy and freedom participants felt in their decision-making.
- **Relatedness to AI:** Three questions adapted from the IMI measured on a Likert scale, to evaluate participants' sense of connection and trust in the AI.
- **Interest/Enjoyment:** Four questions adapted from the IMI to assess participants' interest and enjoyment during the study.
- **Mental Demand:** A single question, measuring the cognitive effort required by participants.

To assess the effects of experimental conditions on learning, accuracy, and subjective measures, we employed analysis of covariance (ANCOVA). For human learning, ANCOVA was applied to the average post-intervention correctness per participant, with pre-test performance as a covariate and condition as a fixed factor. A Shapiro-Wilk test was conducted on the residuals to check the normality assumption, which was not violated ($W = .993, p = .137$). Holm-Bonferroni corrections [42, 76] were used to adjust for multiple comparisons across our eight hypotheses and planned analyses related to learning. Adjusted p-values are reported wherever a correction was applied. For accuracy, we again used
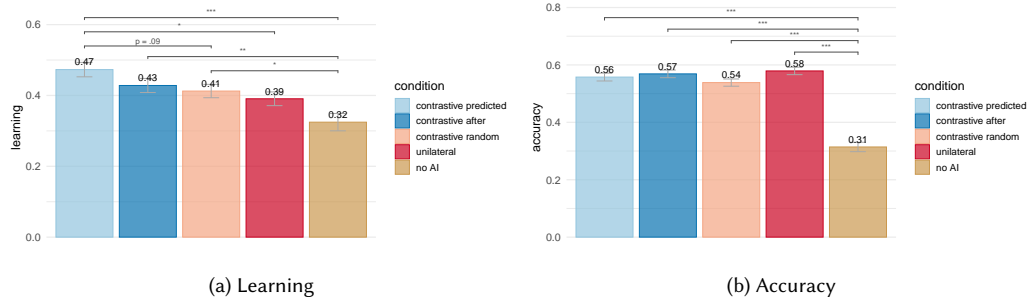
(a) Learning   (b) Accuracy

Fig. 4. Main results. Marginal means of human learning (post-intervention performance, controlled for pre-intervention performance) and accuracy accross different conditions. Error bars represent one standard error. Significance levels after Holm-Bonferroni correction are presented only for significant (or marginally significant) differences, indicated by: * p < 0.05, ** p < 0.01, *** p < 0.001.

ANCOVA, this time on the average correctness during the intervention. Pre-test performance was included as a covariate due to its significant correlation with intervention question performance, while condition was treated as a fixed factor. Subjective measures were analyzed using ANOVA, with condition as the fixed factor. Post-hoc pairwise comparisons between conditions were corrected using Holm-Bonferroni method to account for multiple hypotheses. Throughout the results, we report effect sizes using Cohen's d along with 95% confidence intervals. Effect sizes and accompanying confidence intervals provide valuable information, particularly when interpreting results where we hypothesize no significant differences. When the confidence interval for an effect size includes 0, it suggests that the true effect could be negligible or even nonexistent [23, 52, 85]. For correlations, Pearson's r is provided.

7 RESULTS

7.1 Main results

7.1.1 Human Learning. Main results for learning are depicted in Figure 4a. We report adjusted p-values, corrected with Holm-Bonferroni to account for multiple comparisons. As hypothesized (**H-L1a** & **H-L1b**), participants experienced statistically significantly greater learning in the *contrastive predicted* ($M = 0.47$, $F_{1,209} = 38.62$, $p = 0.00004$, $d = 0.65$ [0.37, 0.94]) and *contrastive after* ($M = 0.43$, $F_{1,216} = 26.68$, $p = 0.006$, $d = 0.47$ [0.19, 0.74]) conditions compared to participants in the no AI condition ($M = 0.32$).

Participants in the *contrastive random* condition also showed significantly higher gains than those in the no AI condition ($M = 0.41$, $F_{1,230} = 23.42$, $p = 0.02$, $d = 0.40$ [0.13, 0.67]). Conversely, participants in the *unilateral* condition did not signficantly improve their learnring compared to *no AI* ($M = 0.39$, $F_{1,222} = 29.66$, $p = n.s.$, $d = 0.30$ [0.03, 0.57]).

Comparing contrastive conditions with sensible foil to unilateral explanations, as hypothesized (**H-L2a**), participants in the *contrastive predicted* condition ($M = 0.47$) learned statistically significantly more than participants in the *unilateral* condition ($M = 0.39$, $F_{1,260} = 40.99$, $p = 0.02$, $d = 0.35$ [0.11, 0.60]). However, the difference between *contrastive after* ($M = 0.43$) and *unilateral* explanations was not significant ($F_{1,267} = 33.05$, $p = n.s.$, $d = 0.16$ [−0.08, 0.40]), not lending support to **H-L2b**.

Within the contrastive conditions, participants in the *contrastive predicted* condition demonstrated greater learning ($M = 0.47$) compared to those in the *contrastive random* condition ($M = 0.41$). However, this difference was only marginally significant ($F_{1,268} = 31.82$, $p = 0.09$, $d = 0.26$ [0.02, 0.50]), offering partial support for hypothesis **H-L3**.
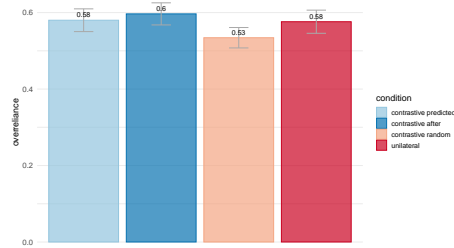
Fig. 5. Overreliance across conditions

Fig. 6. Subjective results. Error bars represent one standard error.

Addressing research question **RQ-L1**, the *contrastive predicted* condition did not result in significantly different learning compared to the *contrastive after* condition ($M = 0.43$, $F_{1,254} = 41.44$, $p = n.s.$, $d = 0.18\ [−0.07, 0.43]$).

7.1.2 Accuracy and Overreliance. Figure 4b summarizes results of human accuracy on the decision task with different conditions. As hypothesized (**H-A1a** & **H-A1b**), the accuracy of participants in the *contrastive predicted* ($M = 0.56$, $F_{1,260} = 20.04$, $p = n.s.$, $d = −0.15\ [−0.39, 0.10]$) and *contrastive after* ($M = 0.57$, $F_{1,267} = 9.13$, $p = n.s.$, $d = −0.08\ [−0.32, 0.16]$) conditions was not significantly different from that of participants in the *unilateral* condition ($M = 0.58$).

While participants improved their performance on the task significantly on average when they received AI support ($M = 0.56$) compared to receiving no AI support ($M = 0.31$, $F_{1,627} = 195.32$, $p << 0.0001$), their performance also significantly degraded when AI suggestions were suboptimal ($M = 0.14$) compared to receiving no support ($M = 0.29$, $F_{1,627} = 36.34$, $p << 0.0001$). Note that the different means for no AI support ($M = 0.31$, $M = 0.29$) in this analysis occur because we split the performance of participants in the no AI condition based on whether AI, if provided, would have been correct or incorrect, to allow a fairer comparison with other conditions that received incorrect suggestions for only a subset of questions.

In situations when AI provided a suboptimal recommendation, participants in the *contrastive predicted* ($M = 0.58$, $F_{1,260} = 6.53$, $p = n.s.$, $d = 0.03\ [−0.22, 0.27]$) and *contrastive random* ($M = 0.54$, $F_{1,281} = 3.83$, $p = n.s.$, $d = −0.12\ [−0.35, 0.12]$) exhibited similar overreliance as those in the unilateral condition ($M = 0.58$), addressing **RQ-A1**.

Similarly, presenting contrastive explanations immediately (*contrastive predicted*), resulted in similar overreliance ($M = 0.58$) compared to presenting contrastive explanations after a decision was made (*contrastive after*) ($M = 0.59$, $F_{1,254} = 10.61$, $p = n.s.$, $d = −0.01\ [−0.25, 0.24]$).
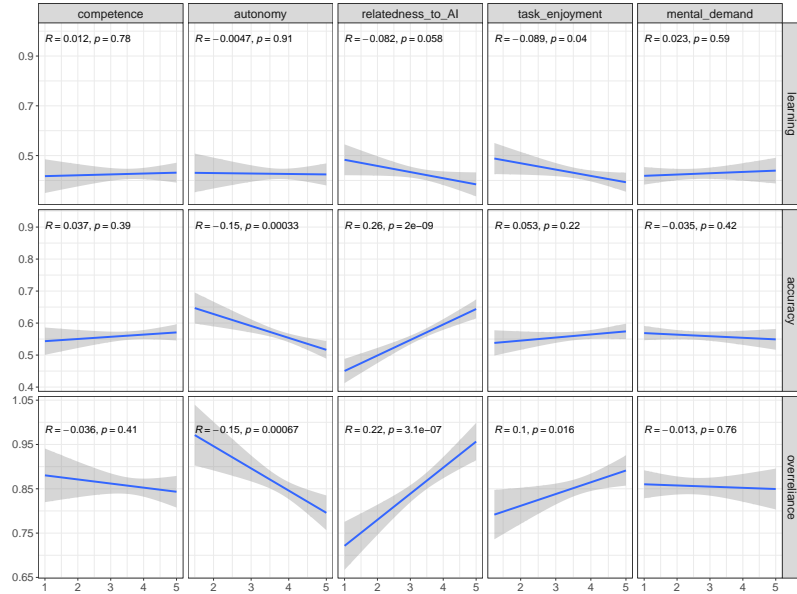
Fig. 7. Relationship between subjective experience vs. objective outcomes. R indicates Pearson's *r*. Only conditions with AI support are included in the analysis.

7.1.3 Subjective Experience. Subjective results are summarized in Figure 6.

Condition was a significant predictor of perceived competence ($F_{4,618} = 6.40$, $p << 0.00005$). A Holm-Bonferroni corrected post-hoc test revealed that participants in the *contrastive predicted*, *contrastive random*, and *unilateral* conditions reported significantly higher competence compared to those in the *contrastive after* and *no AI* conditions.

Perceived autonomy (*i.e.*, choice) was also significantly predicted by condition ($F_{4,618} = 8.85$, $p << 0.00001$). A Holm-Bonferroni corrected post-hoc test revealed that participants in the *contrastive predicted*, *contrastive random*, and *no AI* conditions perceived significantly higher autonomy compared to those in the *unilateral* and *contrastive after* conditions.

Condition was also a significant predictor of relatedness to AI (computed only for conditions involving AI) ($F_{3,533} = 9.02$, $p << 0.00001$), with a Holm-Bonferroni post-hoc test revealing that participants in the *contrastive after* condition felt significantly less related to the AI compared to those in other AI conditions.

Task enjoyment/interest was not significantly predicted by condition ($F_{4,618} = 1.66$, $p = n.s.$) and neither was mental demand ($F_{4,618} = 0.47$, $p = n.s.$).

Across measures, the subjective results support **H-S2**, showing that contrastive explanations with a predicted foil led to significantly higher perceptions of competence, autonomy, and relatedness to the AI compared to the contrastive after condition. Additionally, our findings partially support **H-S1**: while contrastive explanations with a predicted foil significantly increased perceived autonomy compared to unilateral explanations, no significant differences were observed for competence and relatedness to the AI.

7.1.4 Subjective vs. objective measures. Figure 7 shows the relationship between subjective experience and objective outcomes across conditions with AI support (*i.e.* , no AI condition was not included in the analysis). Our analysis

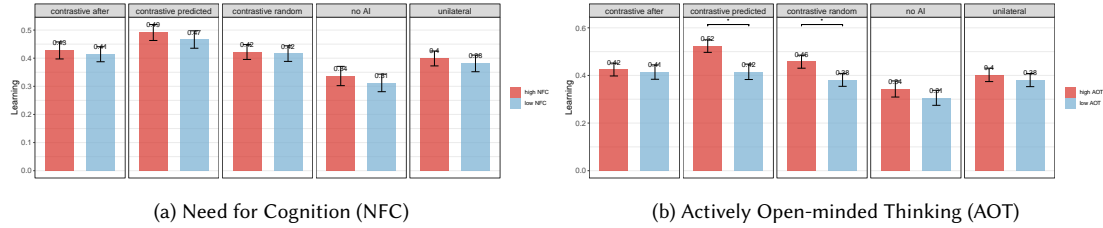(a) Need for Cognition (NFC) (b) Actively Open-minded Thinking (AOT)

Fig. 8. Auditing for intervention generated inequalities: Learning (marginal means) for different individual differences. Error bars represent one standard error. Significance levels after Holm-Bonferroni correction are indicated by: * p < 0.05.

revealed that there was no correlation between actual learning and competence, autonomy, or mental demand and that actual learning was very weakly inversely correlated with relatedness to AI ($r = -0.08, p = 0.06$) and task enjoyment ($r = -0.09, p = 0.04$). Accuracy was significantly positively correlated with relatedness to AI ($r = 0.26, p << 0.0001$), a construct that included questions about trust in AI too, and it was significantly negatively correlated with perceived autonomy ($r = -0.15, p = 0.0003$). Similarly, overreliance was significantly positively correlated with relatedness to AI ($r = 0.22, p << 0.0001$), and significantly negatively correlated with perceived autonomy ($r = -0.15, p = 0.0006$). In addition, overreliance was significantly positively correlated with task enjoyment ($r = 0.1, p = 0.02$).

7.2 Audit for intervention-generated inequalities

Intervention-generated inequalities occur when an intervention, while beneficial on average, disproportionately benefits some groups over others [57]. Disaggregating results by relevant demographics or variables can help uncover these disparities. Informed by prior research in AI-assisted decision-making [10, 31], we conduct a self-audit and examine whether contrastive explanations, introduced as interventions to enhance human decision-making skills, benefit different groups equally.

Previous studies have shown that individual differences in information processing can significantly impact the effectiveness of AI support and interventions, particularly for cognitively demanding outcomes like learning. One individual difference that may affect the effectiveness of our interventions is Need for Cognition (NFC), a stable trait that reflects an individual's motivation to engage in deep thinking and information processing [16]. NFC has been consistently identified as a predictor of performance in cognitive tasks such as problem-solving and decision-making [17]. In the context of AI-assisted decision-making, NFC has been found to influence whether cognitive forcing reduces overreliance on AI [13] and how effectively individuals learn from AI assistance [14, 31].

Another important individual difference that we reasoned would be particularly relevant for interventions that require consideration of multiple viewpoints is Actively Open-Minded Thinking (AOT). People high in AOT are more likely to critically evaluate new evidence, weigh it against their existing beliefs, take sufficient time to solve problems, and carefully consider others' opinions when forming their own [6, 37]. We investigate whether individuals with varying levels of AOT benefit differently from contrastive explanations, which provide alternative "viewpoints" for consideration.

Figures 8a and 8b depict results disaggregated by NFC and AOT. We did not find any significant differences among the effectiveness of (any) contrastive explanations for people with different levels of NFC (for detailed ananlyses see Appendix, Table 2). However, our findings reveal a notable contrast in the AOT groups: participants with high AOT benefited significantly more from the *contrastive predicted* condition ($M = 0.52$) compared to those with low AOT

($M = 0.42$; $F_{1,122} = 6.67$, $p = 0.01$, $d = 0.47$ [0.11, 0.84]). Similarly, the *contrastive random* condition was more effective for individuals with high AOT ($M = 0.46$) than for those with low AOT ($M = 0.42$; $F_{1,142} = 3.77$, $p = 0.05$, $d = 0.34$ [−0.01, 0.68]), although the difference was only marginally significant (see Table 1 for non-significant conditions). These findings uncover AOT as a relevant individual difference to consider in AI-assisted decision-making and reveal that contrastive explanations may unevenly impact individuals, offering greater advantages to those with higher AOT.

8 DISCUSSION

Recent evidence suggests that while AI decision-support tools often enhance decision accuracy in the moment, they can impede the long-term development of individuals' decision-making skills [14, 31], even when the explanations they provide contain potentially valuable learning opportunities. Human decision-making skills are crucial not only for making informed independent decisions but also for critically evaluating AI-generated outputs. In this work, we investigated whether AI decision support systems that account for the decision-maker's mental model of the task can simultaneously enhance decision accuracy and promote the development of independent decision-making skills.

8.1 On the effectiveness of contrastive explanations in improving human-AI decision-making outcomes

Grounded on social science research [56, 64], we introduced a framework for generating human-centric *contrastive* explanations, which accounts for human reasoning when constructing the explanations. These human-centric contrastive explanations compare AI's choice to a predicted, likely human response for the same decision task, by highlighting only the dimensions where the two differ (if they differ). We hypothesized that such contrastive explanations, in which the foil (contrast case) is the predicted average human answer, will lead to greater human learning than the conventional *unilateral* explanations, which are AI-centric and explain why the AI made a specific choice without considering the decision-maker's point of view.

As expected, our results showed that participants learned significantly more with contrastive explanations with predicted foil compared to unilateral explanations (**H-L2a**) or no AI support (**H-L1a**) (Figure 4a). Moreover, also as hypothesized (**H-A1a**), this improvement in learning was achieved without sacrificing accuracy: participants completing the task with contrastive explanations with predicted foil were as accurate as their counterparts who received unilateral explanations (Figure 4b). Additionally, participants in the contrastive explanations with predicted foil condition reported significantly greater perceived autonomy (but not competence or relatedness to AI) during the task compared to those in the unilateral condition, providing partial support for **HS-1**.

The *contrastive after* condition, where participants received contrastive explanations after making an initial decision (inputted foil), led to significant learning gains compared to receiving no AI support (lending support to **H-L1**) but not significantly different learning compared to unilateral explanations (not supporting **H-L2b**). As expected (**H-A1b**), participants' accuracy in the contrastive after condition was not significantly different from those in the unilateral condition.

Overall, our research provides compelling evidence that contrastive explanations with predicted foils significantly enhance decision-making skills without sacrificing decision accuracy compared to unilateral explanations, which remain the default explanation design in AI-powered decision support. Our study is the first to demonstrate that even when AI offers decision recommendations (rather than explanations alone [14, 31]), users can still cognitively engage with its content and improve their learning about the task when this content is engaging. This finding opens new possibilities for optimizing AI decision-support systems by intervening not only at the *interaction* level, as previous work

suggests [12–14, 94], but also at the *content* level of the explanations themselves to improve human-AI decision-making outcomes.

Lastly, while contrastive explanations with predicted foil improved human decision-making skills, we do not think they are a panacea for human-AI decision-making. For example, our results showed that contrastive explanations (as well as unilateral explanations) still resulted in significant overreliance on AI. Also, they were signficantly more effective for people high in AOT (who are inherently driven to consider multiple viewpoints) compared to those low in AOT. Instead, we believe that contrastive explanations are useful when shown in the right situations, such as when the AI is confident about its decision, and to people who benefit from them (e.g., those high in AOT). As such, these explanations expand the suite of human-AI interaction techniques that can be adaptively selected in appropriate situations to optimize human-AI decision-making outcomes, like decision accuracy and human learning [7, 14, 84].

8.2 What have we learned about the design of contrastive explanations?

Our results provide evidence about which aspects of contrastive explanations matter for objective and subjective outcomes in human-AI decision making.

First, as hypothesized (**H-S2**), our findings show that interaction design matters for subjective experience: contrastive explanations are as effective in objective measures when the foil is predicted as they are when the foil is inputted (the contrastive after condition) — even though the inputted foil is the "perfect" comparison. However, consistent with prior research [10, 30], our results show that providing contrastive explanations *after* people make their own decisions (input their foil) results in significantly lower subjective experience, even if that advice engages with their own input as in contrastive explanations after condition. We found no differences in subjective experience between contrastive explanations with a predicted foil and those with a random foil, suggesting that the contrastive design, applied before a decision is made, is perceived favorably regardless of the foil's quality.

Second, our results show suggestive evidence that quality of the foil matters for improving learning as the objective outcome of the interaction. When contrastive explanations are presented at the decision-making time, high quality foil such as in *contrastive predicted* resulted in greater learning on average compared to a randomly selected foil, albeit the difference was only marginally significant, partially supporting **HL-3**. We believe that one of the reasons why the difference between these two conditions is not more pronounced in our study is that even a "random" foil in our setting is relatively reasonable. A randomly selected exercise from the list still addresses at least part of the needs or preferences of the fictitious character, rendering it a choice worth considering. We believe that in different situations, such as medical treatment decisions, where the choices may consist of a wide variety of treatments for a wide array of diseases, a randomly selected choice would likely be obviously ineffective or harmful, thus a waste of cognitive resources for the clinician to consider. In addition, we believe there is room to further improve the quality of the predicted foil. In our implementation, the foil was generated using a single model that predicted the average human response across all decision-makers. We believe that employing personalized models, which capture each individual's unique mental model of the task, could result in even more accurate foils and, consequently, lead to greater learning gains. Our analysis of the participants' responses used to train the human model revealed high variability in exercise choices across participants (Appendix A.2.2), further supporting the need for personalized models. Future research should investigate how to best fine-tune models to individuals and assess the added value of personalized models compared to average human models for enhancing downstream human-AI decision outcomes.

In this study, we sought to deepen our understanding of the timing of contrastive explanations and the impact of foil quality. We experimented with a simple, intuitive design in which the foil represented the choice of many people,

while the fact reflected the AI's suggestion in the user interface. However, contrastive reasoning can be conveyed in various other forms. For example, two conversational agents—one advocating for the human model's choice and the other for the AI's—could engage in a dialogue, allowing the decision-maker to assess which agent's reasoning is more compelling. Alternatively, designs could focus solely on contrastive dimensions, rather than the fact and foil, by highlighting aspects of the decision that the human decision-maker may be overlooking. This approach could provide insights as intermediate support without offering a direct recommendation (e.g., *cardio supports weight loss goals*). Having demonstrated the effectiveness of one human-AI contrastive design in promoting learning, we believe future research should explore a wider range of design possibilities for representing human-AI misalignment in even more impactful ways.

8.3 What have we learned about the effects of contrastive explanations on overreliance?

Evidence from prior work suggested that presenting more than one AI suggestion (i.e., a "second opinion") to people may reduce their overreliance, as it makes them more likely to consider alternatives [5, 58]. Our results showed that participants in contrastive explanations conditions (with predicted or random foils) exhibited similar rates of overreliance on AI suggestions as those in unilateral condition (**RQ-A1**). We believe that we may not be observing the beneficial effect of second opinion in our study because in situations when the simulated AI provided incorrect recommendations (i.e., when the "fact" was a suboptimal choice), the foil was an even worse choice. Therefore, participants were primed to contrast two suboptimal choices and resorted to the better choice out of the two. Future work should explore whether contrastive explanations would still result in similar overreliance, when the foil is a better alternative than the fact.

Interestingly, we also found that participants' overreliance rate on AI in contrastive after condition was similar to that of the unilateral condition. This contrasts with prior research showing that providing *unilateral* explanations after an initial decision reduces overreliance [13, 34], as people are less likely to follow incorrect AI advice once they have made a decision. In our study, because contrastive explanations directly addressed participants' decision and provided evidence as to why their choice was inferior to the AI's, they seemed more persuasive, potentially diminishing the positive effect of the cognitive forcing.

8.4 What have we learned about intrinsic motivation in AI-assisted decision-making?

We measured participants' perceived competence, autonomy, and relatedness *to AI* as psychological needs underpinning individuals' intrinsic motivation about a task. Our results demonstrate that both interaction and explanation design significantly impact these constructs. First, as hypothesized **H-S2**, we found that the contrastive after condition—in which contrastive explanations "critiqued" individuals' inputted answer and presented evidence that AI's choice was superior—led to significantly lower perceived competence, autonomy and relatedness to AI compared to situations in which contrastive explanations were presented before a decision was made. Second, our results demonstrated that contrastive explanations provided before a decision (whether using a predicted or random foil), which presented two decision choices, led to significantly higher perceived autonomy in task completion—comparable to participants who received no AI support—compared to unilateral explanations that offered only a single option.

Our analysis of intrinsic motivation constructs and objective outcomes revealed that actual learning was not correlated with perceived competence. Increased perceived autonomy was correlated with reduced overreliance but also with lowered accuracy, while stronger perceptions of relatedness to the AI were correlated with greater overreliance on AI and higher accuracy.

These findings suggest that the design of AI support can significantly influence people's intrinsic motivation toward a task, as well as objective outcomes such as accuracy and overreliance. We believe that when developing new AI-assisted decision-making systems, researchers should carefully consider and measure how these designs affect people's intrinsic motivation about the task in addition to the objective outcomes of the interaction.

8.5 Generalizability & Limitations

We conducted a single controlled experiment with a single task and with crowdworkers. While prior prior research on AI-assisted decision-making suggests that experts often exhibit similar behavior to non-experts when relying on AI systems [32], we do not know whether this holds for learning from the AI about a task of their expertise. Jacobs et al. [43] show that clinicians would prefer a system that explains why AI's choice differed from the established clinical guidelines, which suggests they may be open for learning from the AI. Our task choice had inherent learning opportunities (e.g., facts about exercises). Learning may not be as pronounced in some tasks, such as hiring, were opportunities for learning exist but are sparser. Further research is needed to understand how generalizeable our findings are for other tasks, domains, and settings.

We believe that our contrastive explanations framework can be effectively applied to a wide range of tasks and settings. Its modular design allows for flexible adaptation based on specific contexts. For instance, in medical applications, the foil could be derived from treatment guidelines, while in image-based tasks, the contrast module could compute pixel gradients or concept activations [50] that highlight differences between the fact and foil. Similarly, the presentation module can be customized to suit the task, such as employing tailored visualization techniques. However, we believe that contrastive explanations—and by extension, our framework—are most valuable in multiclass classification or ranking scenarios, where the foil is less obvious than in binary decision contexts.

An important consideration about our work is that we chose to implement the presentation module with a large language model (LLM). We used the LLM to turn the scaffold produced by the rest of the modules into a natural language explanation, while providing small gaps in the template for it to fill with domain facts. We believe the approach of constraining the generation of facts within the constraints of more trusted predictive models may be useful for certain settings, such as ours but may not generalize to expert-level domains where the LLM may not have the nuance to fill in the gaps. Moreover, we iteratively arrived at prompts (included in the Appendix) which produced explanations with almost no hallucinations. The first author of the paper reviewed the generations for all the characters included in the experiment, finding the LLMs to only mix the indoor/outdoor preferences of characters at times, but no other major hallucinations. However, such manual review cannot be scaled. We believe that our framework can be extended to include a verification step for the generated explanations. For example, multiagent frameworks can be used with additional agents reviewing the generated explanations.

Another limitation of our study is that, in order to control the AI's mistakes, we chose to simulate the AI. We introduced errors in four randomly selected questions during the intervention phase, where the AI's suggestion was generated by a human model rather than the expert model. This approach may have contributed to overreliance on the AI, as the wrong AI suggestion was a likely human choice.

9 CONCLUSION

In this work, we investigated whether explanation designs that account for human reasoning can improve human decision-making skills in the task in AI-assisted decision-making. We introduced a framework for generating human-centric contrastive explanations by showing the difference between AI's reasoning and a likely human response for

the same task. Our results demonstrated that contrastive explanations significantly enhanced human decision-making skills compared to unilateral explanations, the default method of AI support, without compromising accuracy. Sparking hope about growing deskilling concerns, our work suggests that AI support that accounts for human mental models of the task can be a promising approach toward systems that augment and upskill decision-makers.

ACKNOWLEDGEMENTS

REFERENCES

[1] Carlos Aguilar-Palacios, Sergio Muñoz-Romero, and José luis Rojo-Álvarez. 2020. Cold-Start Promotional Sales Forecasting Through Gradient Boosted-Based Contrastive Explanations. *IEEE Access* 8 (2020), 137574–137586. https://doi.org/10.1109/ACCESS.2020.3012032

[2] Barbara E Ainsworth, William L Haskell, Stephen D Herrmann, Nathanael Meckes, David R Bassett, Catrine Tudor-Locke, Jennifer L Greer, Jesse Vezina, Melicia C Whitt-Glover, and Arthur S Leon. 2011. 2011 Compendium of Physical Activities: a second update of codes and MET values. *Med Sci Sports Exerc* 43, 8 (2011), 1575–1581.

[3] David Alvarez-Melis, Harmanpreet Kaur, Hal Daumé, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. From Human Explanation to Model Interpretability: A Framework Based on Weight of Evidence. https://doi.org/10.48550/ARXIV.2104.13299

[4] Vicky Arnold, Steve G Sutton, et al. 1998. The theory of technology dominance: Understanding the impact of intelligent decision aids on decision maker's judgments. *Advances in accounting behavioral research* 1, 3 (1998), 175–194.

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of CHI '21*.

[6] Jonathan Baron. 1993. Why Teach Thinking?-An Essay. *Applied Psychology* 42, 3 (1993), 191–214.

[7] Umang Bhatt, Valerie Chen, Katherine M Collins, Parameswaran Kamalaruban, Emma Kallina, Adrian Weller, and Ameet Talwalkar. 2023. Learning Personalized Decision Support Policies. *arXiv preprint arXiv:2304.06701* (2023).

[8] Harry Braverman. 1998. *Labor and monopoly capital: The degradation of work in the twentieth century*. nyu Press.

[9] 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 (IUI '20)*. ACM, New York, NY, USA.

[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 (April 2021), 21 pages. https://doi.org/10.1145/3449287

[11] Zana Buçinca. 2024. Optimizing Decision-Maker's Intrinsic Motivation for Effective Human-AI Decision-Making. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–5.

[12] Zana Buçinca, Alexandra Chouldechova, Jennifer Wortman Vaughan, and Krzysztof Z Gajos. [n. d.]. Beyond end predictions: stop putting machine learning first and design human-centered AI for decision support.

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

[14] Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Susan A Murphy, and Krzysztof Z Gajos. 2024. Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning. *arXiv preprint arXiv:2403.05911* (2024).

[15] Ángel Alexander Cabrera, Adam Perer, and Jason I Hong. 2023. Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–21.

[16] John T Cacioppo and Richard E Petty. 1982. The need for cognition. *Journal of personality and social psychology* 42, 1 (1982), 116.

[17] John T. Cacioppo, Richard E. Petty, Jeffrey a. Feinstein, W Blair, G Jarvis, and W. Blair G. Jarvis. 1996. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin* 119, 2 (1996), 197–253. https://doi.org/10.1037/0033-2909.119.2.197

[18] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[19] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[20] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 277 (oct 2023), 26 pages. https://doi.org/10.1145/3610068

[21] Lingwei Cheng and Alexandra Chouldechova. 2023. Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–27.

[22] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 103–119.

[23] Nick Colegrave and Graeme D Ruxton. 2003. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. *Behavioral Ecology* 14, 3 (2003), 446–447.

[24] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[25] Edward L Deci, Anja H Olafsen, and Richard M Ryan. 2017. Self-determination theory in work organizations: The state of a science. *Annual review of organizational psychology and organizational behavior* 4 (2017), 19–43.

[26] Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology* 1, 20 (2012), 416–436.

[27] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*. 592–603.

[28] Edgar Erdfelder, Franz Faul, and Axel Buchner. 1996. GPOWER: A general power analysis program. *Behavior research methods, instruments, & computers* 28 (1996), 1–11.

[29] Matthew Fisher and Daniel M Oppenheimer. 2021. Harder than you think: How outside assistance leads to overconfidence. *Psychological Science* 32, 4 (2021), 598–610.

[30] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1362–1374.

[31] 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* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

[32] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 31.

[33] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.

[34] 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.

[35] Nina Grgić-Hlača, Junaid Ali, Krishna P. Gummadi, and Jennifer Wortman Vaughan. 2024. (De)Noise: Moderating the Inconsistency Between Human Decision-Makers. arXiv:2407.11225 [cs.HC] https://arxiv.org/abs/2407.11225

[36] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI magazine* 40, 2 (2019), 44–58.

[37] Uriel Haran, Ilana Ritov, and Barbara A Mellers. 2013. The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision making* 8, 3 (2013), 188–201.

[38] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[39] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating Counterfactual Explanations with Natural Language. https://doi.org/10.48550/ARXIV.1806.09809

[40] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 2000. Large margin rank boundaries for ordinal regression. (2000).

[41] Denis J Hilton. 1988. *Contemporary science and natural explanation: Commonsense conceptions of causality.* New York University Press.

[42] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[43] 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*. Association for Computing Machinery, New York, NY, USA, Article 659, 14 pages. https://doi.org/10.1145/3411764.3445385

[44] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy Jr, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational Psychiatry* 11 (2021). https://doi.org/10.1038/s41398-021-01224-x

[45] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378* (2021).

[46] Tae-Won Jang, Shin-Goo Park, Hyoung-Ryoul Kim, Jung-Man Kim, Young-Seoub Hong, and Byoung-Gwon Kim. 2012. Estimation of maximal oxygen uptake without exercise testing in Korean healthy adult workers. *The Tohoku journal of experimental medicine* 227, 4 (2012), 313–319.

[47] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 1827–1843.

[48] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. Carepre: An intelligent clinical decision assistance system. *ACM Transactions on Computing for Healthcare* 1, 1 (2020), 1–20.

[49] Anna Kawakami, Luke Guerdan, Yanghuidi Cheng, Matthew Lee, Scott Carter, Nikos Arechiga, Kate Glazko, Haiyi Zhu, and Kenneth Holstein. 2023. Training Towards Critical Use: Learning to Situate AI Predictions Relative to Human Knowledge. *arXiv preprint arXiv:2308.15700* (2023).

[50] 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.

[51] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective explanations: Leveraging human input to align explainable ai. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.

[52] Dong Kyu Lee. 2016. Alternatives to P value: confidence interval and effect size. *Korean journal of anesthesiology* 69, 6 (2016), 555.

[53] David Lewis. 1986. Causal explanation. *Philosophical Papers* 2 (1986), 214–240.

[54] Jie Li, Hancheng Cao, Laura Lin, Youyang Hou, Ruihao Zhu, and Abdallah El Ali. 2024. User experience design professionals' perceptions of generative artificial intelligence. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[55] Gabriel Lins de Holanda Coelho, Paul HP Hanel, and Lukas J. Wolf. 2020. The very efficient assessment of need for cognition: Developing a six-item version. *Assessment* 27, 8 (2020), 1870–1885.

[56] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement* 27 (1990), 247–266.

[57] Theo Lorenc, Mark Petticrew, Vivian Welch, and Peter Tugwell. 2013. What types of interventions generate inequalities? Evidence from systematic reviews. *J Epidemiol Community Health* 67, 2 (2013), 190–193.

[58] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does more advice help? the effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–31.

[59] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 4765–4774.

[60] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[61] Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics* 113 (2021), 103655.

[62] Edward McAuley, Terry Duncan, and Vance V Tammen. 1989. Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research quarterly for exercise and sport* 60, 1 (1989), 48–58.

[63] Tim Miller. 2018. Contrastive explanation: A structural-model approach. *arXiv preprint arXiv:1811.03163* (2018).

[64] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007 arXiv:1706.07269

[65] Tim Miller. 2021. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review* 36 (2021), e14.

[66] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-Driven Decision Support Using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

[67] Hussein Mozannar, Jimin Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. 2024. Effective Human-AI Teams via Learned Natural Language Rules and Onboarding. *Advances in Neural Information Processing Systems* 36 (2024).

[68] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. 2022. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5323–5331.

[69] OpenAI. 2020. OpenAI API. https://openai.com/blog/openai-api Accessed: 2024-08-24.

[70] James E Peterman, Matthew P Harber, Mary T Imboden, Mitchell H Whaley, Bradley S Fleenor, Jonathan Myers, Ross Arena, W Holmes Finch, and Leonard A Kaminsky. 2020. Accuracy of nonexercise prediction equations for assessing longitudinal changes to cardiorespiratory fitness in apparently healthy adults: BALL ST cohort. *Journal of the American Heart Association* 9, 11 (2020), e015117.

[71] Marc Pinski, Martin Adam, and Alexander Benlian. 2023. AI Knowledge: Improving AI Delegation through Human Enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

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

[73] Tapani Rinta-Kahila, Esko Penttinen, Antti Salovaara, and Wael Soliman. 2018. Consequences of discontinuing knowledge work automation-surfacing of deskilling effects and methods of recovery. (2018).

[74] Richard M Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology* 43, 3 (1982), 450.

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

[76] Juliet P. Shaffer. 1995. Multiple Hypothesis-Testing. *Annual Review of Psychology* 46 (1995), 561–584.

[77] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.

[78] Ítallo Silva, Leandro Marinho, Alan Said, and Martijn C Willemsen. 2024. Leveraging ChatGPT for Automated Human-centered Explanations in Recommender Systems. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 597–608.

[79] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

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

[81] Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All The Promise of Interactive Explanations for Machine Learning Transparency. *Kunstliche Intelligenz* (4 Feb. 2020). https://doi.org/10.1007/s13218-020-00637-y

[82] James Surowiecki. 2005. *The wisdom of crowds*. Anchor.

[83] Richard Susskind and Daniel Susskind. 2015. *The Future of the Professions: How Technology Will Transform the Work of Human Experts*. Oxford University Press, Oxford.

[84] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *29th International Conference on Intelligent User Interfaces (IUI '24)*. ACM.

[85] Bruce Thompson. 2007. Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools* 44, 5 (2007), 423–432.

[86] Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. 2018. Contrastive explanations with local foil trees. *arXiv preprint arXiv:1806.07470* (2018).

[87] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.

[88] 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.

[89] Zijie J Wang, Alex Kale, Harsha Nori, Peter Stella, Mark Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. 2021. Gam changer: Editing generalized additive models with interactive visualization. *arXiv preprint arXiv:2112.03245* (2021).

[90] Allison Woodruff, Renee Shelby, Patrick Gage Kelley, Steven Rousso-Schindler, Jamila Smith-Loud, and Lauren Wilcox. 2024. How knowledge workers think generative ai will (not) transform their industries. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.

[91] Litao Yan, Alyssa Hwang, Zhiyuan Wu, and Andrew Head. 2024. Ivie: Lightweight anchored explanations of just-generated code. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[92] Petri Ylikoski. 2007. The idea of contrastive explanandum. In *Rethinking explanation*. Springer, 27–42.

[93] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[94] Zelun Tony Zhang, Sebastian S Feger, Lucas Dullenkopf, Rulu Liao, Lukas Süsslin, Yuanting Liu, and Andreas Butz. 2024. Beyond Recommendations: From Backward to Forward AI Support of Pilots' Decision-Making Process. *arXiv preprint arXiv:2406.08959* (2024).

A APPENDIX

A.1 Task Design: Implementation details

A.1.1 Evaluating the expert model. We developed the objective function and expert model iteratively over multiple discussion sessions with the expert. In each session, we evaluated the critical dimensions for inclusion in the model and assessed its predictions, deciding whether to add or remove dimensions accordingly. Once we had decided the structure of the **g** representation, the expert provided a total of 322 pairwise comparisons among exercises for 12 unique fictitious characters. For learning the expert weights for the objective function **f** in Equation 1, we followed the approach described in Section 4.5. Using the Scikit-learn library in Python, we trained a Support Vector Machine (SVM) model with a linear kernel and a regularization parameter (C) set to 1.0. We evaluated the model's performance using 12-fold cross-validation, where each fold excluded one of the fictitious characters. The model was trained on the remaining characters and tested on its ability to predict pairwise comparisons for the excluded character. The model achieved a mean accuracy of 0.86 with a standard deviation of 0.08 across all folds. Additionally, the mean area under the ROC curve (AUC) was 0.86. It is important to note that these results reflect pairwise comparisons involving the full set of 59 exercises, and not only the subset of 7 exercises with which we populated the drop-down list in the interface. For the final step, we qualitatively assessed the model's choices for new fictitious characters, confirming that the decisions were sound and reasonable. (Providing additional validation that the model effectively captures expert reasoning about the designed task, another self-identified kinesiology expert, who participated in one of our formative studies online, achieved a 96% score in the task—significantly higher than the typical 32% average from crowds.)

A.1.2 Selecting the drop-down exercises. We sought to populate the drop-down list for the interface with a sensible number of exercises that would not overwhelm the participants. To select a representative set of exercises from the larger set of the 59 exercises, we clustered the exercises based on their similarity. We generated a large set of 300 fictitious characters, and scored each of the 59 exercises for each of the characters with the expert scoring function. We then computed the correlations between the scores of the exercises and clustered them based on the similarity of their score profiles using hierarchical clustering (as depicted in Figure 9. This method allowed us to group exercises that received similar scores across the 300 characters into clusters. We applied agglomerative clustering with Ward's method. After generating the dendrogram, we determined an appropriate number of clusters by examining the level at which the clusters remained distinct while minimizing redundancy across exercises.

To select representative exercises from each cluster, we calculated the centroid of each cluster, representing the average score profile across all exercises in that group. From there, we selected the exercise whose score profile was closest to the centroid and that was also a more common or accessible exercise (e.g., aerobics vs. trampoline jumping) , ensuring that the selected exercise would be a good representative of the group as a whole. A set of 7 representative exercises (*aerobics*, *bicycling*, *boxing*, *jog/walk combination*, *pilates*, *resistance training*, *swimming*) was then used to populate the drop-down list in the interface, providing a diverse but manageable selection that reflected the range of exercise options without overwhelming participants with too many choices.

A.2 The Contrastive Explanation Framework: Implementation Details

A.2.1 Data collection study for training the human model. We conducted an online user study on Prolific for collecting data with which to train the human model. The task and procedure were identical to those used in the main experiment, but participants completed the task without AI assistance, as our goal was to capture the human mental model of the task. In total, 20 participants answered 100 questions, with each participant selecting exercises for 5 characters,
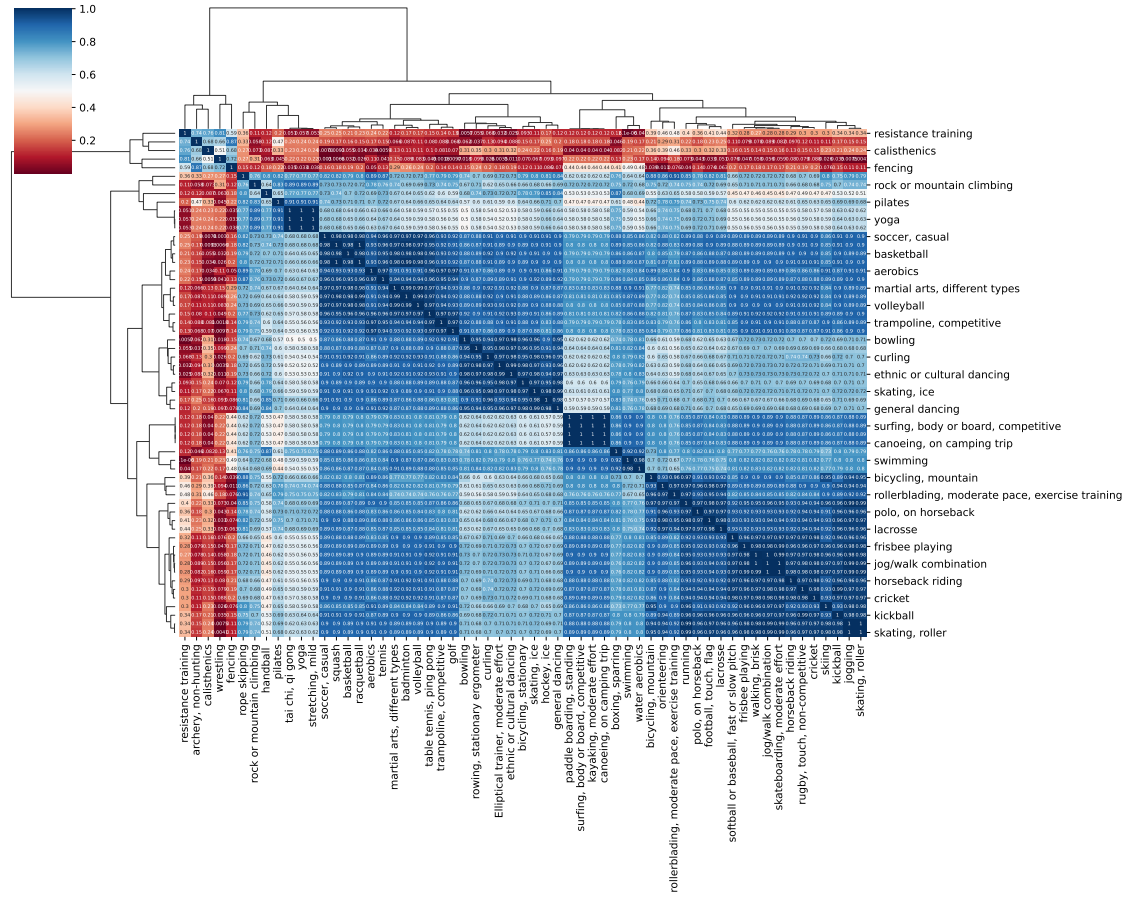
Fig. 9. Correlation heatmaps with hierarchical clustering that served as a basis to select the drop-down exercises for the interface.

randomly sampled from 30 unique characters (distinct from the characters used in the main experiment). Participants achieved a mean accuracy of 30% on the task.

Figure 10 shows the distributions of exercise choices participants selected per fictitious character. To evaluate the variability of participants' responses in selecting exercises for different fictitious characters, we computed normalized entropy [77] per fictitious character. High variability could signal differing decision strategies for the task, while low variability would indicate stronger consensus and shared mental model. We selected *normalized entropy* as it provides a robust measure of uncertainty, independent of the number of available categories, making it ideal for comparing variability across different characters. With a computed mean normalized entropy of $\mu = 0.51$, we found that participants' choices exhibited moderate variability, indicating that while some patterns emerged, responses remained fairly distributed across exercise choices. The standard deviation of $\sigma = 0.35$ further showed notable fluctuation in variability across characters, implying that certain fictitious characters elicited more consistent responses, whereas others triggered more diverse decision-making. This analysis informed our evaluation of the human model, as we expected a moderately, but not highly, accurate model given the variability of participants' responses.
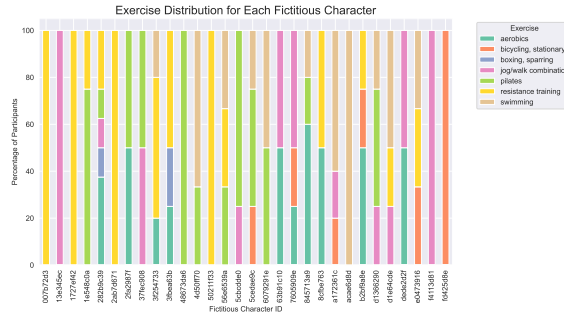
Fig. 10. Distribution of participants' responses for the 30 fictitious character in the data collection study.

A.2.2 Evaluating the human model.

We trained the human model from the responses collected in the data collection study and by following the same approach with which we learned the expert weights. The total 100 choices (each out of 7 exercises) from the data collection study, yielded a total of 600 pairwise exercise comparisons. As with the expert model, we trained a Support Vector Machine (SVM) model with a linear kernel and a regularization parameter (C) set to 1.0. We used a 30-fold cross-validation for evaluating the human model, where in each fold one participant was removed from the training set. The model was then trained on the remaining participants and tested on the excluded one. This process was repeated for all participants, allowing us to assess the model's generalizability and its ability to predict individual behavior across different subsets of the data. As expected from the high variance in participant responses, the model was moderately accurate, with a mean cross-validation accuracy of 0.69 and an AUC of 0.68 for the pairwise comparisons of exercises.

As an additional evaluation, we compared the human model to the expert model. We generated a new set of characters to conduct the evaluation and found that 60% of the unseen 50 characters, both human and expert models produced the same responses. The key differences emerged in specific exercise choices: the human model was less likely to select *boxing* or *aerobics*, which the expert model identified as suitable for some characters. Despite the high variability within and across participants, this demonstrates a "wisdom of the crowd" effect [82], where the average human model captured signal across participants' responses (achieved 60% accuracy on the task, compared to average participant accuracy of 30%), resembling the expert model (an effect also observed in [35]). In the main study, for cases where the human and expert models provided identical responses for the characters, we selected the second-highest-ranked option from the human model as the human response (i.e., the foil).

A.3 LLM Prompts

The variables in double brackets were populated according to the character in question.

A.3.1 Contrastive Explanation Prompt.

[[vignette]]
Here are the aspects that a kinesiology expert considers when making the decision:
(1) Intensity: whether the intensity required to carry out an exercise exceeds the fitness capabilities of the person.
(2) Intensity: whether an exercise matches the intensity the person is capable of exerting.

(3) Goal: whether the exercise matches the person's goals.

(4) Preference: whether the exercise matches the person's preference.

According to the expert's function, `[[fact]]` is better than `[[foil]]` on the following: `[[positive_contributors_fact]]`.

Whereas, `[[foil]]` is better than `[[fact]]` because of: `[[positive_contributors_foil]]`.

Construct an explanation about why `[[fact]]` is better than `[[foil]]` using the following template:

Make it compact. Go into bullet point(s) strictly only for concepts for which the fact is better than the foil according to the expert's function. Do not explicitly say anything about the expert. Acknowledge the benefits of the foil over the fact if any as the first sentence, then highlight the tradeoffs in high-level concepts at the beginning of the explanation.

Use the following structure for each bullet point, one by one, and include only the concepts for which `[[fact]]` is superior to `[[foil]]`:

- Identify the primary characteristic of the superior exercise (e.g., running is a cardio exercise) and contrast this to the other exercise.
- Connect this characteristic to a benefit relevant to the character (e.g., cardio is good for weight loss).

High-level sentence that first acknowledges the concepts for which the foil is better than the fact (if any) or states that the foil is also a good choice, then highlights the concepts for which the fact is superior to the foil. Include only the concepts for which `[[fact]]` is superior to `[[foil]]`.

- Concept 1 (e.g., Goal):
- Concept 2: ...

Format the response as a JSON object:

"high_level_contrastive_explanation": "explanation", "contrast_concepts": [{"Formatted name of concept (e.g., Goal)": "explanation"}].

A.3.2 Unilateral Explanation Prompt.

`[[[vignette]]]`

Here are the aspects that a kinesiology expert considers when making the decision:

(1) Intensity: whether the intensity required to carry out an exercise exceeds the fitness capabilities of the person.

(2) Intensity: whether an exercise matches the intensity the person is capable of exerting.

(3) Goal: whether the exercise matches the person's goals.

(4) Preference: whether the exercise matches the person's preference.

Create a concise explanation for why `[[fact]]` is the best exercise for the specified character, using the following structure in bullet points only for aspects the expert considers:

- Identify the primary characteristic of `[[fact]]` (e.g., running is a cardio exercise).
- Connect this characteristic to a benefit relevant to the character (e.g., cardio is beneficial for weight loss).

Strictly only include aspects recognized by the expert as beneficial for the character, omitting any for which `[[fact]]` may not be optimal or relevant. Do not explicitly say anything about the expert. Use the terms `Goal`, `Intensity`, and `Preference` when describing the relevant 'concept'.

Format the response as a list of JSON records with 'concept' and 'explanation' as the keys for the records.

A.4 Post-study Questionnaire

Perceived Competence (Adapted from the Intrinsic Motivation Inventory (IMI)).

- I think I performed well in making exercise recommendations during this task.
- This was a task that I couldn't do very well. *(reverse Likert)*
- I believe I am skilled at suggesting suitable exercises for different individuals.
- After working at this task for a while, I felt pretty competent.

Perceived Choice (Adapted from IMI).

- I felt like I had a lot of choice in deciding which exercises to recommend.
- I was free to choose the exercises I thought were best suited for the person described.
- I felt like I was strongly influenced by the AI on how to recommend exercises. *(reverse Likert)*
- I recommended exercises in the way I wanted to.

Relatedness to AI (Adapted from IMI).

- I felt I could trust this AI.
- I felt my reasoning on this task was distant from the AI's. *(reverse Likert)*
- I would like a chance to interact with this AI in the future.

Interest/Enjoyment (Adapted from IMI).

- I enjoyed this exercise recommendation task.
- This task did not hold my attention at all. *(reverse Likert)*
- While I was doing this task, I was thinking about how much I enjoyed it.
- I thought this exercise recommendation task was a boring task. *(reverse Likert)*

Mental Demand.

- I found this task mentally demanding.

A.5 Results: Audit for Intervention-Generated Inequalities

Condition	High AOT (SE)	Low AOT (SE)	Significance	Effect Size (d [CI])
contrastive after	0.42 (0.03)	0.41 (0.03)	$F_{1,127} = 0.07, p = n.s.$	0.05 [-0.31, 0.40]
unilateral	0.40 (0.03)	0.38 (0.03)	$F_{1,134} = 0.30, p = n.s.$	0.09 [-0.24, 0.43]
contrastive random	0.46 (0.03)	0.38 (0.03)	$F_{1,142} = 3.77, p = 0.05$	0.34 [-0.01, 0.68]
contrastive predicted	0.52 (0.03)	0.42 (0.03)	$F_{1,122} = 6.67, p = 0.01$	0.47 [0.11, 0.84]
no AI	0.34 (0.03)	0.31 (0.03)	$F_{1,83} = 0.64, p = n.s.$	0.17 [-0.26, 0.61]

Table 1. ANCOVA results by condition for AOT groups, showing marginal means (SE), Significance (F-statistic, p-value), and Effect size (Cohen's d with 95% confidence intervals).

Condition	High NFC (SE)	Low NFC (SE)	Significance	Effect Size (d [CI])
contrastive after	0.43 (0.03)	0.41 (0.03)	$F_{1,127} = 0.11, p = n.s.$	0.06 [-0.29, 0.41]
unilateral	0.40 (0.03)	0.38 (0.03)	$F_{1,134} = 0.20, p = n.s.$	0.08 [-0.26, 0.42]
contrastive random	0.42 (0.03)	0.42 (0.03)	$F_{1,142} = 0.02, p = n.s.$	0.02 [-0.30, 0.35]
contrastive predicted	0.49 (0.03)	0.47 (0.03)	$F_{1,122} = 0.35, p = n.s.$	0.11 [-0.25, 0.46]
no AI	0.34 (0.03)	0.31 (0.03)	$F_{1,83} = 0.28, p = n.s.$	0.11 [-0.32, 0.55]

Table 2. ANCOVA results by condition for NFC groups, showing marginal means (SE), Significance (F-statistic, p-value), and Effect size (Cohen's d with 95% confidence intervals).