

AI, Help Me Think—but for Myself: Assisting People in Complex Decision-Making by Providing Different Kinds of Cognitive Support

Leon Reicherts*
leon.reicherts@microsoft.com
Microsoft Research
Cambridge, United Kingdom
University College London
London, United Kingdom

Zelun Tony Zhang*
zhang@fortiss.org
fortiss GmbH, Research Institute of
the Free State of Bavaria
Munich, Germany
LMU Munich
Munich, Germany

Elisabeth von Oswald
elisabethluisa.von@mail.polimi.it
Politecnico di Milano
Milan, Italy

Yuanting Liu
liu@fortiss.org
fortiss GmbH, Research Institute of
the Free State of Bavaria
Munich, Germany

Yvonne Rogers
y.rogers@ucl.ac.uk
University College London
London, United Kingdom

Mariam Hassib
hassib@fortiss.org
fortiss GmbH, Research Institute of
the Free State of Bavaria
Munich, Germany

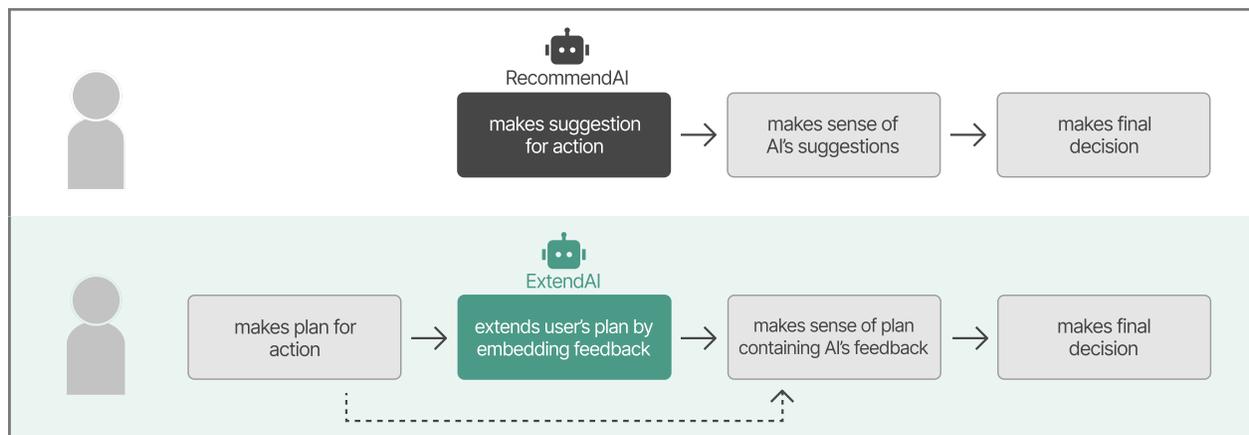


Figure 1: Illustrative comparison of the thought process when interacting with two ‘types’ of AI – *RecommendAI* and *ExtendAI*.

ABSTRACT

How can we design AI tools that effectively support human decision-making by complementing and enhancing users’ reasoning processes? Common recommendation-centric approaches face challenges such as inappropriate reliance or a lack of integration with users’ decision-making processes. Here, we explore an alternative interaction model in which the AI outputs build upon users’ own decision-making rationales. We compare this approach, which we call *ExtendAI*, with a recommendation-based AI. Participants in our

mixed-methods user study interacted with both AIs as part of an investment decision-making task. We found that the AIs had different impacts, with *ExtendAI* integrating better into the decision-making process and people’s own thinking and leading to slightly better outcomes. *RecommendAI* was able to provide more novel insights while requiring less cognitive effort. We discuss the implications of these and other findings along with three tensions of AI-assisted decision-making which our study revealed.

*Both authors contributed equally to this research.

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

© 2025 Copyright held by the owner/author(s).

This is the author’s version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26-May 1, 2025, Yokohama, Japan, <https://doi.org/10.1145/3706598.3713295>.

CCS CONCEPTS

• Information systems → Decision support systems; • Human-centered computing → Interaction paradigms; Empirical studies in HCI.

KEYWORDS

generative AI, human-AI interaction, AI-assisted decision-making, human-AI decision-making, investment decision-making

ACM Reference Format:

Leon Reicherts, Zelun Tony Zhang, Elisabeth von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. 2025. AI, Help Me Think—but for Myself: Assisting People in Complex Decision-Making by Providing Different Kinds of Cognitive Support. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3706598.3713295>

1 INTRODUCTION

AI is increasingly being used to support human decision-making in a variety of domains, such as healthcare [55], finance [16], and logistics [61]. Typically, this involves an interaction model where the AI makes recommendations, which the user can then take into account in their decision-making. Benefits of *AI-assisted decision-making* include both increased decision efficiency and increased accuracy [2]. However, various challenges persist, including inappropriate reliance on AI [1, 10, 17, 19, 27, 49], deskilling [26, 57], or the difficulty of integrating AI recommendations into users' decision-making process [4, 34, 65]. This difficulty often arises when factors outside of the features modelled by an AI system influence the decision, which Holstein et al. [31] call “unobservables”. This is particularly challenging when decisions involve unstructured human reasoning and context, such as complex family backgrounds in a social care setting [34], or patient preferences in healthcare [32], which AI systems traditionally have not been able to capture.

Most work in AI-assisted decision-making attempts to address these challenges while remaining within the common paradigm of providing recommendations. For instance, a large and rapidly growing body of work revolves around supplementing recommendations with explanations of how the AI produced its recommendation [37]. However, it has been argued in recent work that at least some of the challenges are inherent to the recommendation-centric paradigm, and proposed to support human decision-making by other means than recommendations [46, 65, 67]. The proposition is to augment users' reasoning while they reason *forward* through the decision-making task toward their own solution, rather than recommending a solution from which users have to reason *backward* [59, 67].

To this end, with their unique capabilities, large language models (LLMs) could enable various yet unexplored roles for AI to augment human decision-making [22]. For one, LLMs have been found to be highly proficient in various expert domains such as medicine [48] and finance [24, 35], often surpassing the performance of human domain experts. More importantly, the impressive natural language capabilities of LLMs may provide new opportunities to flexibly process unstructured data and human reasoning. This provides new opportunities to more tightly integrate AI into users' decision-making process, especially in complex, ill-defined problem settings. Recent research explored utilising these capabilities of LLMs to help users engage more deeply with AI recommendations [18, 41, 63]. Our research is concerned with how we can leverage the natural language capabilities of LLMs to provide a new tool to support human reasoning. In particular, we investigated how an LLM can be integrated with a software tool in order to elicit forward reasoning for a complex task; in this case, we chose financial investments, where many decisions need to be made. Our aim is to embed more fine-grained AI support into users' own reasoning, allowing users to

remain engaged with the decision-making task, rather than diverting their engagement to AI recommendations. We do so by asking users to write a rationale for their decision and extending it with LLM-generated feedback, meant to help users reflect on, and identify potential gaps in their reasoning (see Figure 1). Given that this represents a different paradigm than the usual recommendation-focused approach, we pose the following research question:

RQ: How do different paradigms for AI decision support—one that extends users' reasoning and one that provides direct recommendations—affect users' decision-making processes, perceptions of the AI, and decision outcomes in complex decisions?

Our chosen use case, financial investment, presents a complex and realistic test bed for our exploration. In a mixed methods study ($N=21$), we find that both approaches can support human decision-making in different ways. Our findings show that the approach of extending people's rationales helps them reflect on their reasoning, leading to more informed and better decisions. On the other hand, the recommendation-based approach seems to help people get inspiration for and explore new directions in their decision-making. We found marked differences and tensions in the preferences for both approaches depending on the participant's personality, expertise, and decision-making style, resulting in an even split of overall preference between the two designs.

The contributions of this work are threefold: (1) the conceptualisation and design of two LLM roles for AI decision support in the context of a realistic, complex, and open-ended task, (2) an exploration of the perceptions and usage of the LLM assistance through a mixed-methods study within a complex financial investment use case, (3) a discussion of the tensions of integrating AI into complex decision-making tasks.

2 BACKGROUND & RELATED WORK

In this section, we first review the broader literature on AI-assisted decision support followed by an overview of research that specifically uses LLMs for this purpose.

2.1 AI-Assisted Decision-Making

The ultimate goal of AI decision support tools is to work *with* humans to potentially reach the best performance that overcomes the limitations of either human or AI on their own. This is often referred to as human-AI complementarity [1]. Decision-making is a highly complex process and AI support tools may vary profoundly in when, how, and what they provide. Research in AI-assisted decision-making largely focuses on providing AI-generated recommendations from model predictions (see Bao et al. [2], Lai et al. [37] for reviews). Users must then evaluate and consequently accept or reject the recommendation [37]. Recommendations are either presented on their own or combined with additional AI support elements such as various types of model explanations [28, 40] or uncertainty [1, 50, 66]. While sometimes successful in increasing accuracy and efficiency in generic or artificial tasks [38, 58, 62], AI-assisted decision-making has its challenges, including low user acceptance in more complex decision-making processes [4, 34, 65], and overreliance on (incorrect) AI recommendations [1, 11, 12, 33].

Empirical results on the effectiveness of explanations or other support elements in calibrating reliance have been mixed. While there are instances where explanations lead to more appropriate reliance [58, 60, 62], in many cases, explanations have had little effect on users' decision accuracy [15] or even increased overreliance [1, 3, 12, 14, 21, 33, 53, 54].

Another body of emerging research investigates more complex, realistic decision-making tasks, which require humans to work with subjective and unstructured knowledge going beyond single outcome predictions, such as in healthcare, social welfare, or financial decision-making [8, 36, 68]. Real-world decision tasks often call for a more nuanced type of AI decision support that accompanies the user *during* the different phases of their decision-making and *augments* the process [67, 68], instead of providing one concrete prediction as *the* ground truth solution. One example is the concept of *Evaluative AI* proposed by Miller [46], where AI helps users generate and evaluate hypotheses, while leaving the decision itself to users. Zhang et al. [65] applied a similar concept in their redesign of an existing AI-based sepsis diagnosis tool, aiming to support the process leading up to a decision. Another example is *Reflective XAI* by Cabitza et al. [13], which prompts radiologists to reflect on the decisions of the AI by presenting evidence for and against the AI advice. In a study on an aviation use case [67], the main challenge for pilots was integrating data from multiple sources rather than interpreting that data. AI thus proved valuable by presenting relevant information at a glance and intelligently highlighting constraints at nearby airports, instead of recommending a concrete airport. Yildirim et al. [64] similarly found that radiologists preferred AI to help them complete tedious sub-tasks or collect relevant evidence for their decisions rather than taking over the interpretation of the evidence. Even more restrained approaches where the AI completely refrains from analysis tasks have also been explored, for instance, through chatbots that support decision-making through context-dependent probing questions to scaffold users' reasoning [51].

Overall, decision-making tasks are dispersed along a space of complexity that requires a flexible approach to AI support. To understand which designs of AI decision support are suited for which decision task, it is crucial to empirically compare these different designs. One such study is that of Ma et al. [42], who investigated the effectiveness of direct recommendations compared to two alternative designs. In one design, the AI presented a balanced analysis by providing the features speaking for and against each possible answer, aligning with the evaluative AI framework discussed by Miller [46]. In the other design, the AI challenged the user's preferred choice by highlighting the features speaking against the user's answer. Results showed that each approach has its unique advantages and disadvantages, and that recommendations are not always the most effective approach. However, the simplicity of the text classification task used in the study might conceal some nuances in real-world applications. Another study exploring non-recommendation-based approaches within the context of a more realistic use case, aviation, is that of Zhang et al. [67]. They conducted a comparison between recommendations and the continuous support approach described above. They found that continuous support was better accepted by pilots and induced less overreliance. A

combination of recommendations and continuous support led to faster decisions.

2.2 LLMs in AI-Assisted Decision-Making

Research on the use of LLMs for AI-assisted decision-making is still in its early stages, with the few existing studies mainly following three broad directions. The first is applying LLMs within the established paradigm of explainable AI. For instance, Mavrepis et al. [44] proposed creating more approachable natural language-based explanations. Mozannar et al. [47] use natural language to describe to users when to rely on AI recommendations and when not. A second direction is the usage of LLMs to enrich AI recommendations with additional evidence that is related to the task rather than the AI model. This has been investigated in a clinical application by Yang et al. [63], who used an LLM to identify supporting and opposing evidence for AI suggestions from biomedical literature. Lastly, researchers have explored the use of LLMs for designing AI agents that can engage with users in various forms of discussions. For example, Chiang et al. [18] employed LLMs as Devil's Advocate, where the LLM debates either against the AI recommendation or the majority opinion in a group decision-making setting. Ma et al. [41] on the other hand proposed the concept of *Deliberative AI*, where both the human and the AI first present their viewpoints. Afterwards, they deliberate on conflicting opinions and discuss individual features until a final decision is reached. This flow, in which the user starts by presenting their own rationale and then receives the AI's perspective, is similar to our proposed approach. However, our approach differs from this framing in that we propose that the AI extends and embeds into the human's rationale, aiming to engage the user in a forward reasoning manner. Hence, the user does not have to make sense of an independent AI viewpoint.

Overall, while the above approaches differ significantly from each other, all of them are focused on helping users to make sense of and appropriately rely on AI support received as recommendations. Additionally, most of the studies above were based on simple or generic decision-making tasks. In our work, we aim to explore how LLMs can assume a role that accompanies the user in their own reasoning during decision-making in a realistic task and how such a role compares to the more common role of providing recommendations.

3 USE CASE AND PROTOTYPE

As a use case for our study we chose financial investments, as it represents a highly complex and open-ended decision problem that allowed us to explore different ways of helping the user's thought process when making decisions. Countless potentially interacting factors can play a role in a person's investment decisions, and there is generally no objective right or wrong for how to compose a portfolio. According to modern portfolio theory [43], an optimal portfolio is one that maximises return at the risk level that is acceptable to the investor and is thus highly individual. The primary means to reduce risk is through diversification of the portfolio. However, people often struggle to assess the risk of an investment [5]. Inexperienced investors also often misjudge the diversification of their portfolio as they neglect covariations [30, 39], i.e., the value of individual assets might be correlated, and the portfolio thus less

diversified than thought. At the same time, initial analyses suggest that LLMs like GPT-4 can perform impressively well on financial tasks [24, 35]. These aspects and opportunities made investment decision-making a promising scenario to explore the use of LLMs for investment decision support.

The type of financial product that our study focused on were exchange-traded funds (ETFs), which are investment funds traded on stock exchanges like individual stocks. ETFs hold a range of financial assets, including stocks, bonds, currencies, debts, or commodities such as gold or crude oil. The main reason for choosing ETFs over trading other financial products was that they require less specific expert knowledge (e.g., valuation techniques for stocks; knowledge of leverage, margin requirements, and precise market timing for futures contracts; or assessing creditworthiness, interest rates, and repayment structures for bonds) and also keep the task simpler and more manageable for a study. We further simplified the task by only including ETFs that are composed of stocks (also referred to as *equity ETFs*). Participants could thus focus in their reasoning and decision-making on considerations such as the regions, sectors, or themes (e.g., energy storage) to include in their portfolio, making the task more approachable to participants, while still keeping the complexity of having to weigh off different investments against each other and ways to compose them in a portfolio. Rather than asking participants to put together a portfolio for themselves, we asked them to put themselves into the shoes of a 40-year-old, rather cautious person with a long-term investment horizon of about 10 years, with the explicit goal of creating a diversified portfolio. The purpose of this was to make it less likely that participants just bought their ‘go-to’ ETFs without much thought.

We designed a simulated stock trading platform in which we embedded an AI investing ‘assistant’ **offering two different types of interaction/assistance**: One follows the more ‘traditional’ approach of **making direct recommendations**. The other requires the user to describe their own rationale first. The AI assistant would then **extend the user’s rationale** by embedding its feedback into it. In what follows, we will refer to both types of assistants as *RecommendAI* and *ExtendAI*. In the present investing context, *RecommendAI* would provide a set of ETFs tailored to the user’s portfolio at the given time, and *ExtendAI* would ask the user to describe a rationale for how they might adjust their portfolio (i.e., which ETFs in their portfolio they intend to increase or reduce their exposure to, sell completely, or which new ones they intend to buy and why), which the AI would then give feedback on. This ‘augmented’ rationale would then be returned containing the AI’s feedback in bold so that the user could see the AI’s suggestions directly in the context of their initial thoughts/reasoning (see also Figure 1 for a simplified overview of the decision-making process with both AIs). In the following section, a more detailed description of the prototype implementation will be provided.

4 PROTOTYPE IMPLEMENTATION

This section describes the design and development of the prototype in more detail. Section 4.1 outlines how the ETF dataset was prepared, and Section 4.2 describes how additional synthetic data was generated. This is followed by an overview of the design of the simulated trading platform and the AI assistants in Section 4.3.

4.1 Preparation of ETF Data and Pre-Study

In order to investigate people’s decision-making on which ETFs to put together in a portfolio, we first collected and prepared an ETF dataset and then tested the suitability of the dataset for such a decision-making task in an online pre-study ($N=10$). Both these steps are described in this section.

4.1.1 Collection and Preparation of ETF Data. We selected a range of 31 ETFs using the ETF Screener from ETF Database¹ to present in our simulated trading platform. We aimed to have enough ETFs to make the decision complex, but not too many that it overwhelms or gets too time-consuming for participants to go through them. We selected ETFs reflecting all major regional markets, all sectors from the Global Industry Classification Standard (GICS)², and popular investment strategy-related themes (e.g., low volatility, wide moat, momentum, quality, small size, value), such that common categories of ETFs would be well covered. For each category, we aimed to select one ETF, to limit decisions to choosing among these categories, not among ETFs within the same category. For sector and theme ETFs, the goal was to select global ETFs. For cases when there was no global ETF, we chose a US-focused ETF in addition to its ex-US counterpart, allowing participants to express a global view for that category. In addition, we included narrow ‘trend’ ETFs for some sectors, such as biotechnology or clean energy ETFs for the healthcare and energy sectors, respectively.

We used the Python package `financetoolkit`³ to retrieve historical closing prices for each month of the past five years, and `thepassiveinvestor`⁴ for information such as sector and company holdings as well as performance metrics for each ETF to display in our simulated trading platform. We further retrieved country holdings for each ETF using the Financial Modeling Prep API⁵ and manually retrieved further details such as short analyses from the ETF Database that would be used for prompting our AI investment assistants (Section 4.3).

4.1.2 Pre-Study to Evaluate the Decision-Making Task & Iterate on Prompting. After collecting and preparing our ETF dataset, we conducted a pre-study with $N=10$ participants over Prolific⁶. Participants were provided with the ETF data and were tasked with putting together a portfolio to match the provided investor profile and writing a rationale explaining their choices. The pre-study aimed to assess: (1) how long it would take people with some ETF investing experience to analyse the data for the 31 ETFs, compose a portfolio, and write a rationale motivating their decisions, (2) whether the written rationales would be sufficiently detailed and elaborate, and finally (3) how well an LLM would be able to give feedback on these rationales and how the prompts would need to be designed for best results. We chose pre-screeners to only select participants with some investing experience. Participants completed the task in about 8 minutes (median duration) and were able to write sufficiently elaborate rationales. We used these rationales to assess the ability of GPT-4o to provide meaningful feedback for the kind of assistance

¹<https://etfdb.com/screener/>

²https://en.wikipedia.org/wiki/Global_Industry_Classification_Standard

³<https://www.jeroenbouma.com/projects/financetoolkit>

⁴<https://www.jeroenbouma.com/projects/thepassiveinvestor>

⁵<https://site.financialmodelingprep.com/developer/docs>

⁶<https://www.prolific.com/>

we intended to achieve with *ExtendAI*. We iterated accordingly on our metaprompt to lead to feedback that (a) considers each aspect mentioned by the users in the rationale, (b) is based on the provided investor profile, (c) does not provide specific recommendations, and (d) addresses diversification, the return and risk balance, volatility, and potential overlaps in the intended trades (between regions and sectors). This seemed to work sufficiently well, but occasionally, the LLM would not adhere to (c) and make recommendations for specific ETFs. We also added several constraints to the prompt such as ensuring that the user provided in their rationale at least two ETFs and reasons for choosing them and that only the list of the available 31 ETFs would be considered. In Appendix B we provide the final metaprompt we arrived at.

4.2 Finalisation and Synthesis of the ETF Data for the Simulated Trading Platform

Following the encouraging findings of the pre-study, we further developed the study materials. We set the study in a hypothetical future rather than in the past to avoid that participants would rely on their memory of events (e.g., the stock tech boom within the first year of the pandemic or with the launch of ChatGPT, etc.) and hence not engage in considerations such as balancing risks with potential returns. This meant that synthetic data had to be generated for this hypothetical future.

When generating the synthetic ETF data, we aimed to create plausible and engaging future developments rather than the most realistic ones. The data consisted of two interdependent components: three ‘news articles’ describing market developments, and synthetically generated price data for each of the 31 ETFs that are in line with these news articles for August 2024, August 2026, and August 2028, the three time steps presented in the study task.

To create the ‘news articles’, three authors had a brainstorming session on possible future economic and stock market developments while considering real articles published by financial institutions, investment platforms, and newspapers. One of the authors further expanded some of these developments and presented the drafts back to two co-authors. Based on this discussion, a final version was created, highlighting key developments such as technological breakthroughs, industrial shifts, supply chain disruptions, and rising demand for specific services, raw materials, and products—relevant to different regions, industries, and ETF themes (see supplementary materials). We then adjusted the percentage price changes for each of the 31 ETFs to be adequately aligned with the narrative of 2026 and 2028. For an additional plausibility check, the specified percentage changes along with the news articles were reviewed by an investment banking expert with over 10 years of experience in investing in the stock market and economics. Based on the feedback confirming the sufficient plausibility of the news articles and the percentage changes, final adjustments were made. The next section describes the design and development of the simulated ETF trading platform and the assistants.

4.3 Design of the Simulated Investing Platform and the AI Assistants

For our study, we built a web-based simulated ETF trading platform inspired by existing trading platforms. The application was

developed using the Python web framework Flask, JavaScript, and a MySQL database for the ETF data. The main interface of the trading platform consists of three columns: On the left, there is a scrollable list of all available ETFs along with their prices, performance, and details about those already held by a participant (see Figure 2). Below that, there is an overview panel of a participant’s portfolio. The middle column displays detailed information about the ETF currently selected in the list, including a short description of how it is composed and what it tracks, its volatility, a price chart over past years and information on sector, company, and country holdings. The right column showed the assistants (depending on the study condition) and a description of the investor profile, which participants should ‘build’ their portfolio for. The screenshot in Figure 2 shows a user’s typed rationale in the text input field of the *ExtendAI* before sending it out to receive feedback. In the *RecommendAI* condition, no text input field is available, instead, a button labelled “Get Recommendation” is available, which returns a recommendation of which ETFs to add or remove from one’s portfolio.

Both assistants utilised the OpenAI API using the GPT-4o model. In the API calls for both assistants, the following was included: (1) the user’s current portfolio composition (list of ETFs), (2) a JSON containing detailed information on each available ETF, as described in Section 4.1.1, and (3) the investor profile along with the instructions for how the AI should provide its suggestions. The main difference between both assistants (apart from their instructions) is that in *ExtendAI*, the API call also included the user’s rationale which the AI was instructed to extend with its feedback. To provide the feedback, we prompted *ExtendAI* to improve the user’s rationale by drawing on its knowledge of the return, volatility, and diversification of each ETF. The feedback is designed to highlight potential blind spots and to provide additional factors to consider, giving the user a more comprehensive basis for their decisions. High-quality feedback would thus add new and non-obvious considerations to the user’s rationale, grounded in the information provided to the LLM. The instructions for both assistants were iteratively developed, tested, and refined by four of the authors with various portfolio compositions and, in the case of *ExtendAI*, with different user rationales from the pre-study (Section 4.1.2), with the aim to achieve meaningful suggestions from the LLM. Appendix A provides some examples for the outputs of both assistants, and Appendix B gives the final metaprompts.

Beyond the general distinction between the metaprompts for both AI roles for providing recommendations versus embedding feedback into the submitted rationale, in *ExtendAI*, we instructed the AI to avoid making specific ETF suggestions. Instead, we focused on highlighting key factors to consider, occasionally referencing specific ETFs as potential alternatives, but primarily emphasising broader aspects for evaluation. The reason for this was that we intentionally wanted to create a contrast between both types of assistance to elicit participants’ reflections on the advantages and disadvantages of both. At the same time, we aimed to focus the study on the two distinct roles of the LLM by only offering a one-shot interaction for both assistants, meaning that suggestions from both assistants could only be requested once. Thus, unlike chat-based interfaces such as ChatGPT, where users can engage in multiple exchanges, *ExtendAI* required users to provide a relatively comprehensive description of their rationale upfront.

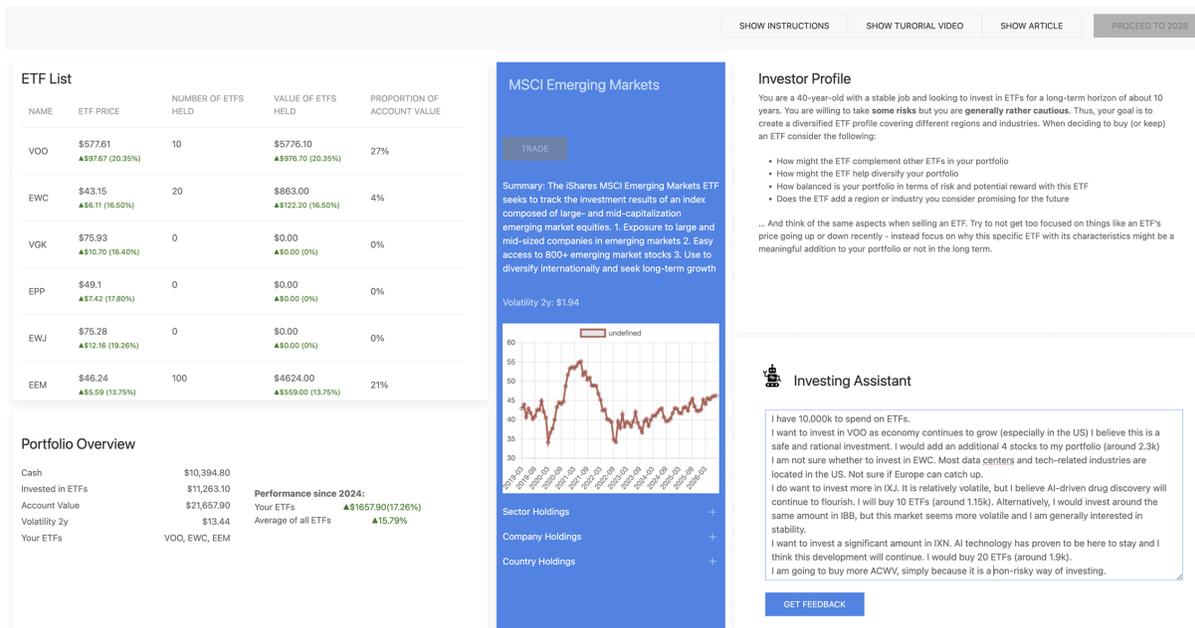


Figure 2: Screenshot of the simulated ETF trading platform interface used in the study, showcasing the various components including the ETF list, portfolio overview, ETF details, investor profile, and the investing assistant.

5 EVALUATION

In this section, we describe the study we conducted to evaluate the prototype containing the two different types of assistance, *ExtendAI* and *RecommendAI*. The study received ethics approval of the IRB at UCL under code UCLIC_2024_003_Rogers.

5.1 Participants

We recruited $N = 21$ participants (5 females, 16 males; $M = 30.4$ years, $min : 20$ years, $max : 44$ years; see Table 2 in Appendix C for details). We initially started recruiting participants through the crowd-sourcing platform Prolific. However, despite using pre-screeners and clearly emphasising in the study description that participants should have experience investing into ETFs, four out of five participants who were recruited could not answer the questions on their experience and knowledge of ETFs at the beginning of the study sufficiently well (e.g., what their strategy is, what they consider when investing, examples of ETFs they have invested in). Consequently, we shifted to recruiting participants through our personal networks. We used our organisations' mailing lists and targeted people directly whom we knew had the required knowledge and/or experience. This allowed us to recruit a sample of participants with the required investing experience and who have a range of different backgrounds (also see Table 2). The final sample included $N = 18$ participants who had a good understanding of and experience in investing into ETFs (with an average experience of $M = 4.47$ years, $SD = 3.88$), and $N = 3$ participants with a good understanding of investing and the stock market, but without having first-hand experience in investing in ETFs. In our pre-task questionnaire, participants self-reported their knowledge of the financial markets and of ETFs to be at $Md = 4$ and an $IQR =$

1 respectively on a 5-point Likert scale. A majority of $N = 18$ participants described themselves as long-term investors, in line with the framing of the study task.

5.2 Study Procedure

The study was conducted over Zoom⁷, following a detailed protocol elaborated through a pilot study with three participants (who were not part of our final study sample). At the beginning of the Zoom call, participants gave their informed consent and filled out a pre-questionnaire covering basic demographics and their investment experience. After the pre-questionnaire, participants were forwarded to our simulated ETF investment platform for the main task, which was to build and revise a portfolio over three time steps: August 2024, August 2026, and August 2028. During the entire study, the Zoom meeting remained open, so that participants could ask clarifying questions on how to perform the task and inform the researcher in case of any issues during the study.

Participants started with 10,000 USD to invest in the first time step. In each of the later time steps, they received an additional 10,000 USD that they could invest in addition to cash earned from selling ETFs. In the first time step, the familiarisation phase, participants had to make their decision without AI assistance. This allowed them to familiarise themselves with the interface of the trading platform and the available ETFs. In each of the later time steps (2026, 2028), participants either used *ExtendAI* or *RecommendAI* during their decision-making (see Table 1). We randomised the order of the two types of assistance across participants. 11 participants used *RecommendAI* first and 10 used *ExtendAI* first. See Figure 3 for more details.

⁷<https://zoom.us/>

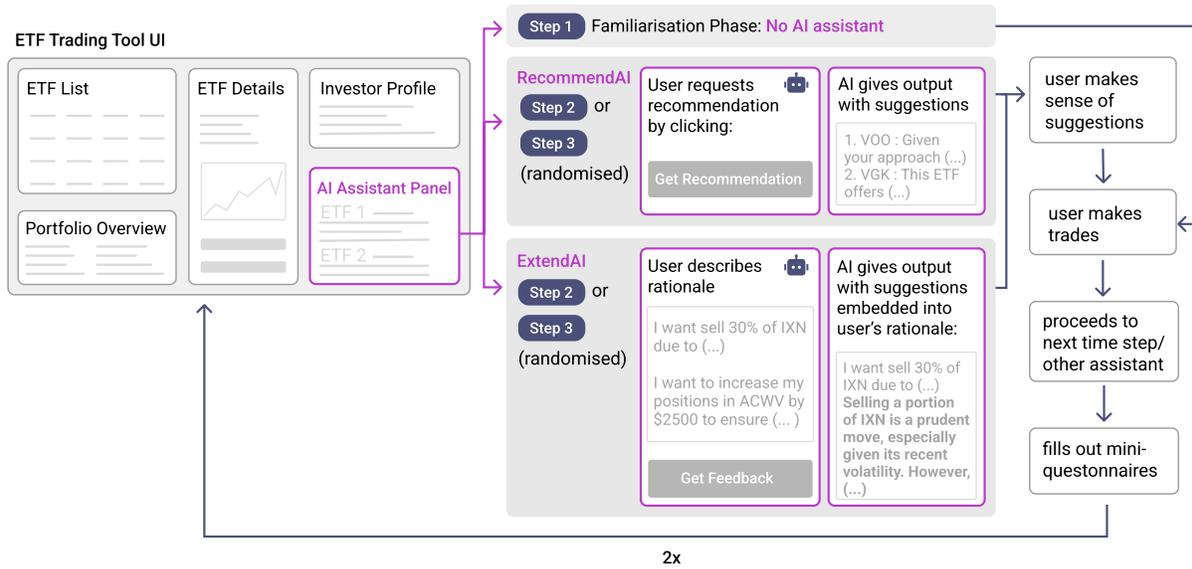


Figure 3: A schematic overview of the ETF trading platform, illustrating what participants did and what the AI assistant panel showed in each of the three steps of the study: familiarisation phase, *RecommendAI*, and *ExtendAI*.

Table 1: Overview of the high-level task, the form of assistance provided, and which data was shown in the simulated trading interface for each step of the within-subjects study.

Step	Year	Task	Assistance	Data Shown
1	2024.08	Participant makes trades to <i>build</i> initial portfolio	No AI assistant (familiarisation phase)	ETF price data and econ. narrative based on <i>real data</i>
2, 3	2026.08, 2028.08	Participant makes trades to <i>revise</i> portfolio	<i>RecommendAI</i> or <i>ExtendAI</i> in random order	<i>Synthetic</i> ETF price data and econ. narrative
4	2030.08	Participant checks final portfolio value	N/A	Final portfolio value (based on synthetic data)

At the beginning of each time step, participants watched a short tutorial video about the trading platform (first time step, 2024) or the respective AI assistant (second and third time steps, 2026/2028) and were presented with the ‘news article’ (Section 4.2) covering relevant economic and market-related events as background information. After each of the three time steps, participants filled out a short questionnaire about the decision they made. After each of the later two time steps with AI assistance (2026, 2028), participants were asked a short set of questions about their experience of using the respective AI assistant and how they perceived its impact on their decision-making

We chose not to collect think-aloud protocols since the task was already rather demanding on its own and required participants to be fully focused. After the main task, we conducted exit interviews of $M = 23$ min ($SD = 6.8$ min), focusing on participants’ decision-making process and how they thought the two types of assistance

affected it. The full interview guide is provided in the supplementary material. We recorded audio through Zoom only for the interviews. In total, the study lasted $M = 91$ min ($SD = 17.5$ min), depending on how fast participants finished the main task. Participants were compensated with Amazon vouchers worth 30 EUR or an equivalent amount in participants’ preferred currency.

5.3 Measures

A set of measures were used covering the time spent on the different study parts, the number of interactions with the trading platform, and the trades participants made. Furthermore, measures covered (1) participants’ assessments of their decisions, (2) their experience with the two types of assistance, and (3) the impact they had on their decision-making, described in more detail in this section.

As part of the study task, participants filled out a brief questionnaire after each of the three steps, in which they were asked to report on five-point Likert scales how well informed they felt about their decision, how confident they were in it, and how satisfied they were with it. After completing these items, participants were shown the performance of their portfolio after two years and were asked to re-evaluate the confidence and satisfaction items. The first two questions were asked before revealing how the portfolio developed over the two years following the current time step, while the third question was asked after revealing the outcome. This perceptions-based approach was informed by Bradbury et al. [5] and serves as a proxy for decision quality. The reasoning is that good financial decisions are marked by how well people understand the risks of their decisions.

At the end of the second and third time step (2026, 2028), after having completed the set of trades with the help of one of the two

assistants, participants were asked a short set of questions about its helpfulness, perceived impact on their decision-making process (including cognitive load using a shortened version of NASA-TLX [29]), and an item measuring participants' sense of agency using a slider from -50 ("decision was fully made by AI") to +50 ("decision was fully made by myself"). The complete questionnaires with all the items can be found in the supplementary materials.

To measure the impact the two types of assistance had on the decision quality, we measured how the portfolio diversification changed at each time step. This was done through four metrics: (1) the number of countries included in the portfolio, (2) how balanced the allocation was across these countries, and (3) how balanced the allocation was across the 11 sectors (based on the Global Industry Classification Standard).

We further measured the magnitude of the AI assistants' impact on participants' decisions. The impact is conceptualised differently for both assistants, given the different decision-making steps they involve. For *ExtendAI*, participants first developed a plan, hence, the impact was measured by the number of ETFs in their final trades that were *not* part of their initial plan. For *RecommendAI*, which provided recommendations without a prior plan, the impact was measured by the number of trades in the final list that were directly suggested by the AI. These metrics serve as proxy measures of the AI's impact and are used exploratively to examine the interactions with the tools and the two AI assistants.

5.4 Data Analysis

Our *quantitative analysis* relied on data collected in the questionnaire before the study tasks and three questionnaires during the study task. In the analyses, we only report descriptive statistics and refrained from inference testing, given the exploratory nature of our study. To refer to central tendency and spread, we use the mean (M) and standard deviation (SD) for continuous variables, and median (Md) and interquartile range (IQR) for ordinal variables.

For our *qualitative analysis*, all interview recordings were transcribed verbatim using Zoom's automatic transcription, followed by manual cleaning by one researcher, and subsequently imported into Atlas.ti Web⁸. Initially, four researchers conducted open coding on a subset of eight interviews (four interviews each, with allocations offset so that one researcher shared two interviews with one, and two with another researcher). A coding tree was iteratively developed over four meetings between the four researchers. The interviews were then distributed among three of the researchers for coding using the predefined coding tree. In two meetings among the three researchers, a reflexive thematic analysis was conducted to identify overarching themes [6, 7].

6 FINDINGS

We gathered both quantitative data from the questionnaires and interactions with the prototype, as well as qualitative data from the interviews, which we present in this section.

6.1 Quantitative Findings

6.1.1 Task Duration. Overall, participants spent $M = 44.00$ min ($SD = 13.28$) on the investing task. In Step 1, participants traded

ETFs without assistance (familiarisation phase), in Step 2 and 3, they interacted with either *ExtendAI* or *RecommendAI* (order randomised). Participants spent $M = 17.99$ min ($SD = 6.14$) in the familiarisation phase, $M = 17.46$ min ($SD = 6.03$) on *ExtendAI*, and $M = 8.56$ min ($SD = 3.37$) on *RecommendAI*. For *ExtendAI*, participants spent $M = 8.38$ minutes ($SD = 4.69$) on writing their rationale. After participants received the suggestions from either assistant, participants started trading after $M = 1.75$ min ($SD = 1.16$) with *ExtendAI*, and after $M = 2.08$ min ($SD = 1.46$) with *RecommendAI*.

6.1.2 Interactions With Trading Tool. Participants looked at $M = 28.86$ ($SD = 4.91$) of the available 31 ETFs in the familiarisation phase. This was measured by counting the ETFs that were clicked on to display more information. In *ExtendAI*, participants looked at $M = 19.67$ ($SD = 9.23$) of the ETFs, and in *RecommendAI* at $M = 12.86$ ($SD = 8.79$).

6.1.3 Overall Development of Portfolio. Participants started off with a portfolio containing $M = 5.48$ ($SD = 4.07$) ETFs in the familiarisation phase (year 2024). The number of ETFs in participants' portfolios then tended to increase over the following two time steps to $M = 7.48$ ($SD = 4.30$) in the second time step (2026), and to $M = 8.86$ ($SD = 4.48$) in the third time step (2028).

6.1.4 AI Assistants' Impact on Portfolio Composition. The set of trades participants made following an interaction with one of the assistants generally tended to result in an increase in the number of countries held. In *ExtendAI*, the increase was $M = 3.00$ ($SD = 5.933$) and in *RecommendAI* slightly smaller $M = 2.33$ ($SD = 4.247$). Furthermore, the deviation of an equal distribution across these countries also decreased for both assistants. The decrease of this deviation tended to be more pronounced in *ExtendAI* with $M = -0.732$ ($SD = 1.211$) than in *RecommendAI* where it was $M = -0.369$ ($SD = 1.135$). The distribution across sectors was generally also improved following the interaction with both assistants. However, here the decrease in the deviation in *ExtendAI* $M = -0.275$ ($SD = 1.741$) tended to be slightly smaller than in *RecommendAI* $M = -0.349$ ($SD = 1.498$).

6.1.5 AI Assistants' Influence on Decision-Making. Participants made $M = 5.19$ ($SD = 2.40$) trades with *ExtendAI*, and $M = 5.95$ ($SD = 3.46$) with *RecommendAI*. In *ExtendAI*, 23.08% of the ETF investment decisions participants made were different to what they described in their rationale before interacting with the assistant. In *RecommendAI*, 45.00% of their decisions were based on the assistant's recommendation (depicted in Figure 7). These percentages are proxy measures for the assistants' impact on participants' decision-making.

6.1.6 Perceptions and Assessment of Decisions. In this and the following subsection, we will only report participants' *agreement* to the separate items for readability and simplicity (i.e., how many participants gave answers above the neutral point). Furthermore, we only report on those items with meaningful differences between the two assistants. For a more detailed picture of how participants answered these questions, see Figure 4 and Figure 6. In terms of how well-informed participants considered their decisions to be, in the familiarisation phase, 57% stated that they felt informed, whereas

⁸<https://web.atlasti.com/>

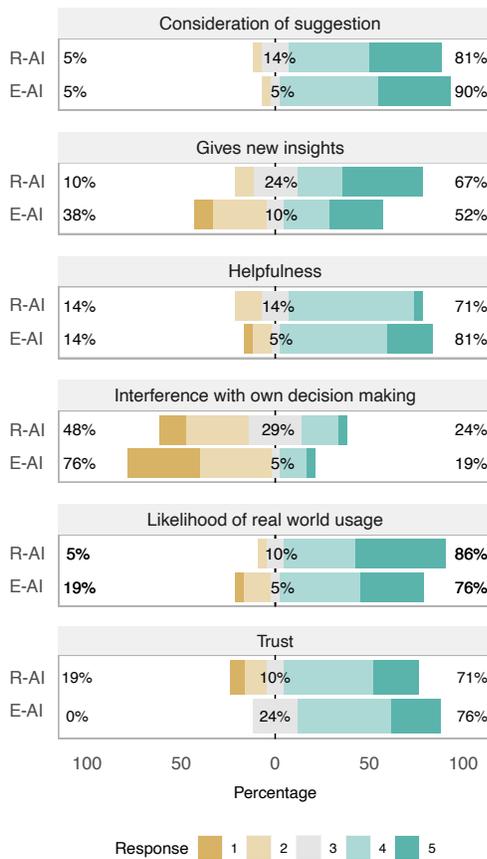


Figure 4: Participants' scores on a 5-point scale for the questions on AI suggestion consideration, insights, helpfulness, interference with decision-making, likelihood of usage, and trust. R-AI refers to *RecommendAI* and E-AI to *ExtendAI*.

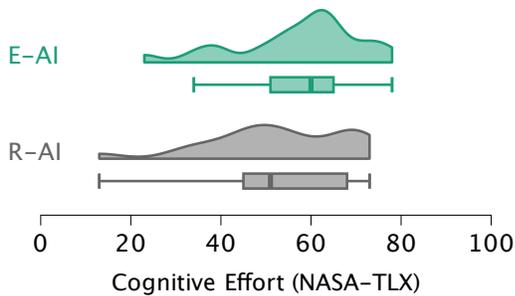


Figure 5: Distribution of Nasa-TLX raw scores indicating cognitive effort for *ExtendAI* (E-AI) and *RecommendAI* (R-AI).

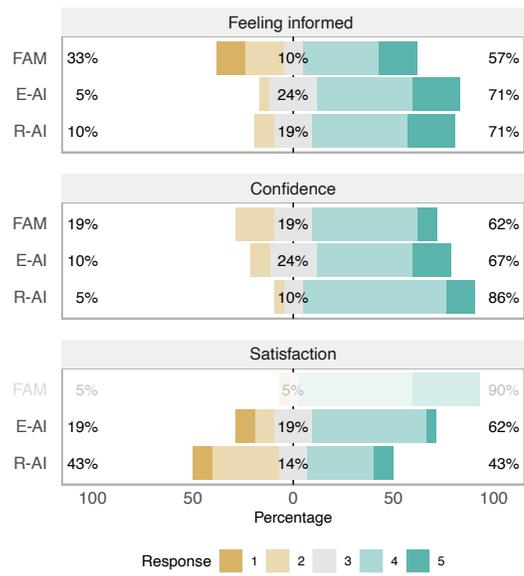


Figure 6: Participants' scores on a 5-point scale on the extent of feeling informed, their confidence in their decisions, and their satisfaction with decision outcomes for the familiarisation phase (FAM) *RecommendAI* (R-AI) and *ExtendAI* (E-AI). FAM is greyed out for *Satisfaction*, as it cannot be meaningfully interpreted.

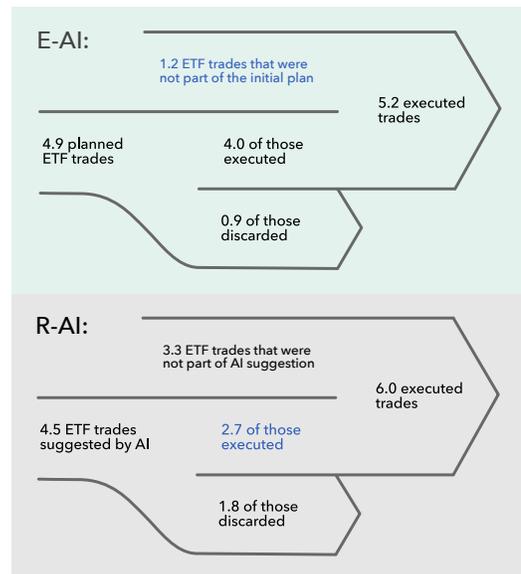


Figure 7: A depiction of the 'impact' of AI assistants on decision-making (reads from left to right) with blue elements indicating the AI assistants' impact: for *ExtendAI*, it is calculated as the percentage of trades ultimately made that were not part of a participant's initial plan (i.e. 1.2 trades / 5.2 trades = 23.08%), for *RecommendAI*, it is calculated as the number of trades made which were directly suggested by the AI (i.e. 2.7 trades / 6 trades = 45.00%).

with assistance, this was the case for 71% (for both assistants). With respect to their confidence in their decision, the familiarisation phase reached 62%, and participants felt more confident with *RecommendAI* (86%) than with *ExtendAI* (67%). However, in terms of how satisfied participants were, the ratings for the assistants were ‘flipped’ with *RecommendAI* only reaching 43% and *ExtendAI* reaching 67%. Note that a comparison with the familiarisation phase is not meaningful/adequate with respect to satisfaction, since in the familiarisation phase participants’ portfolios generally performed better than in the two following time steps when using the assistants (the synthetic ETF prices were intentionally created this way, as described earlier).

6.1.7 Perceptions and Experiences With Both Assistants. As can be seen in Figure 4, participants reported that they generally trusted both assistants—*ExtendAI* with 76% receiving a slightly higher rating than *RecommendAI* with 71%. Furthermore, participants reported that they considered the assistants’ input (81% *RecommendAI* and 90% *ExtendAI*). For *RecommendAI*, participants generally found that it gives them new insights (67%), while for *ExtendAI* this was less the case (52%). Despite this somewhat lower rating for insights, 81% of participants considered *ExtendAI* to be helpful, whereas only 71% thought this was true for *RecommendAI*. Participants generally did not feel that the assistants interfered with their decision-making process, although an even smaller number of them thought this to be the case for *ExtendAI* (19%) than for *RecommendAI* (24%). Participants reported that they would use an assistant like *RecommendAI* (86%) in real-world settings slightly more than *ExtendAI* (76%). This might be related to the aforementioned finding that participants felt like they got more insights from *RecommendAI*, as well as the fact that they rated it to be less cognitively demanding (see Figure 5), where *ExtendAI* received a NASA-TLX Raw score of $M = 57.00$ ($SD = 14.02$), and *RecommendAI* $M = 52.52$ ($SD = 15.55$).

6.2 Qualitative Findings

As described in Section 3, we designed both assistants to trigger different decision-making processes, which was also reflected in how participants described their experience of using each assistant. One participant, who summed it up well, described his process with *RecommendAI* as: “I first look into the recommendations. Then I decide for myself whether it makes sense or not.” For *ExtendAI*, the same participant described his process as follows: “The first part of the decision-making was exactly identical [to the first step without any assistant], so write down everything and ask for feedback. And then reconsider it if there’s something significant that I’ve missed.”

Participants’ opinions of these two types of assistance differed greatly, with 11 participants stating in the interview that they preferred *RecommendAI*, and 10 participants preferring *ExtendAI*. We identified four themes in our thematic analysis that explore these differing opinions from various angles. In the following, we denote participants who first interacted with *RecommendAI* and then with *ExtendAI* as RE- x , and participants with the reverse order as ER- x . Participants frequently referred to the assistants using variations of ‘the first one’ or ‘the second one’. To improve readability, we replaced these with [*RecommendAI*] or [*ExtendAI*] in the quotes.

6.2.1 Theme 1: Help Me, but Let Me Think for Myself. As mentioned above, participants’ preferences for the two types of assistance were evenly split. On the one hand, some participants preferred *RecommendAI* as they considered it easier to use; not only because “it’s more convenient to just press the button and get a result” (ER-8), but also because these participants considered the concrete ETF recommendations more actionable:

“It’s highly comparable to a friend who is deep into finance, and he tells me, I’ve heard about this health care stuff, maybe have a look at it, [...] and I kind of trust it. And then I basically quickly reevaluate it and then directly go for it.” (RE-7)
 “[*RecommendAI*] gave me some hints, which made it a bit quicker to make decisions.” (RE-8)

Conversely, some participants rejected the immediate shift into specific recommendations, because they felt excluded from the decision-making process:

“It was far less engaging and it did not know what I want to do at this stage. It was recommending stuff. [...] It doesn’t know my current state of mind [...]” (ER-1)
 “My main issue with [*RecommendAI*] was that it directly gave me some kind of ‘do this, then do this, then do this,’ which I, in some sense, followed without thinking too much about it. So, I gave the AI assistant more freedom than I usually would have and thought about it even a bit less.” (ER-9)

These participants preferred *ExtendAI*’s more restrained approach, noting that it gives them more agency and freedom to think for themselves:

“I kinda liked the style that it was having this somewhat passive voice, and not aggressively recommending you to buy, buy, buy, or sell, sell, sell something.” (ER-7)
 “I think it gives me a bit more control and agency. [...] It didn’t just tell me, ‘This is what you should do.’ [...] It kind of supports how I’m thinking [...], it gives me information to help me decide.” (RE-12)

This helped participants to “think more over my decisions” (RE-11) and to “really make you reflect” (ER-8). However, other participants found *ExtendAI*’s feedback too general and not actionable enough. For instance, one participant stated that “it just gave me some comments on the ETFs I selected. But I didn’t know, how can I start with the feedback, and what should I do instead? So it was not really helpful” (ER-2).

These varying perspectives from participants highlight a tension between actionability and helpfulness of AI suggestions on the one hand, and preserving users’ agency and their own reasoning on the other hand. The more specific the AI suggestions are, the easier it is to act on them, but the less room it leaves for users’ own reasoning, and vice versa. While this trade-off reflects the general tendency in participants’ opinions, it is important to note that some participants had the exact opposite experience. For example, one participant disagreed that *RecommendAI* reduced his own reasoning, emphasising

that “*I was still reflecting. I was probably reflecting even more with [RecommendAI]*” (ER-4). Another participant on the other hand rejected the common view that *RecommendAI* was easier to act on, explaining that “*it was actually more work because it had a lot of information. I wanted to check every point of those. Is it true or not? And is the logic making sense or not?*” (ER-1). This participant found that *ExtendAI* only required a “*small check*” (ER-1), since its suggestion built upon his own reasoning making it easier for them to make sense of what it said, which leads into the next theme.

6.2.2 Theme 2: Insights Should Be New, but Aligned With the User’s Thinking. Both assistants tended to add to participants’ reasoning in different ways. *RecommendAI* tended to be more helpful for discovering new ETFs that participants did not consider before, or more generally, for thinking into new directions, as its suggestions were considered more “*out of the box*” (RE-4), making it more likely “*to think about things you didn’t think of before*” (RE-9). *ExtendAI*, on the other hand, tended to help participants think through their own reasoning more thoroughly and revealed blind spots in their rationale:

“*So when I said, ‘Oh, I’m thinking about investing into, I think it was IBB and IXJ, two biotech-related ones’, and my initial idea was to just invest in one. And it was like, ‘Well, why don’t you diversify because they’re differently focused?’ That was a good suggestion that I hadn’t thought about before.*” (ER-10)

Yet, both assistants had their drawbacks. Some participants found *RecommendAI*’s suggestions challenging to verify due to their unclear basis, describing them as “*black-box solution coming from nowhere*” (ER-1), “*based on whatever analysis it does, we don’t know anything about it*” (ER-3). This was in contrast to *ExtendAI*, whose feedback some participants found easier to verify as it was integrated into their own reasoning:

“*It allows me to delve a bit deeper into the problem and understand the recommendations from the AI assistant a bit more than in [RecommendAI], where I somewhat blindly trust it. If I first have to gather the information by myself and evaluate it, then I think I’m better informed to also evaluate the truthfulness of whatever the AI assistant recommends.*” (ER-9)

“*I’ve read the feedback more carefully. Let’s say, if it would have been longer, I would have still easily read it. Whereas in [RecommendAI] you sometimes go a bit in this skim-reading mode, that you read the headline and then you conclude something.*” (ER-7)

However, *ExtendAI* building upon participants’ rationale also frequently led to a perceived lack of novel input to their reasoning:

“*I feel like [ExtendAI] would be more useful if it was less stuck in my way of doing things.*” (RE-9)
 “*I felt like these were just direct feedback on my flawed suggestions. And I wish that I could have a fresher input, actually.*” (RE-10)

These complementary strengths and weaknesses of the two assistants underscore another tension: The insights provided by AI should be ‘fresh’ and new to be of value to the user. At the same time, these insights need to be consistent with the user’s reasoning so that the user accepts them and can more easily integrate them into their decision-making. The challenge of navigating this tension is further illustrated by participants’ statements about their reliance behaviour. With both assistants, participants tended to ignore AI suggestions that were not aligned with their thinking. This usually became apparent for participants who referred heavily to the economic ‘news articles’ in an attempt to anticipate the short-term performance of individual ETFs, and expected the assistants to help them react to the developments and events described in the news articles. However, recall that the assistants were—consistent with the investor profile provided—designed to consider only the diversification and long-term potential of the ETFs, not the recent and expected short-term performance of the ETFs, as explained in Section 4.3. As a result, these participants frequently found *ExtendAI*’s feedback to be “*too general, too plain*” (ER-2), feeling that “*it didn’t really give much insight that changed my mind*” (RE-2), and thereby disregarding *ExtendAI*’s comments on how to improve the diversification of their portfolios. For *RecommendAI*, these participants ignored its recommendations when they found them to contradict the economic news articles, even though the recommendations may be sensible from a long-term perspective: “*I completely ignored this whole S&P 500. The whole thing is going down, it looks like it is not the time to invest. Everyone is selling*” (ER-1).

Conversely, we also observed the opposite effect to be possible for both assistants, as in other cases the AI suggestions appeared to be *too* closely aligned with the participants’ thoughts. Consequently, these participants frequently found that *ExtendAI* gave “*confirmation that you have the right thoughts*” (RE-5), or “*reassures me that I’m not doing something wrong*” (RE-8), rather than adding new insights or highlighting weaknesses in the portfolio. This perception might have been strengthened by participants’ tendency to focus on parts of the feedback that reinforce their reasoning, as one participant observed: “*If I reflect now, I was kind of looking for sentences confirming my strategy.*” (RE-7). We observed this also with *RecommendAI*, albeit less frequently: “*For the first AI assistant, most of the recommendations I also had thought about before. And so, it was like a confirmation of my reasoning*” (RE-11). While these confirmations meant participants “*felt more confident in taking the investment decisions*” (RE-5), it could also lead to unwarranted confidence, as pointed out by one participant: “*But the thing is, if the AI assistant has confirmed my thought, then I will be overconfident*” (RE-4).

Overall, this theme highlights the challenge of simultaneously aligning with users’ thinking *and* helping them arrive at new thoughts. The next theme explores when AI is most likely to be able to extend users’ thinking in this way.

6.2.3 Theme 3: AI Is Most Insightful When Human Thoughts Are ‘In the Making’. As described at the beginning of Section 6.2, *RecommendAI* usually served as a starting point for participants’ decision-making, while *ExtendAI* came in at later stages of the decision-making process. Allowing participants to receive feedback from *ExtendAI* only once, encouraged them to think thoroughly about

their decisions before interacting with the assistant; but, in principle, *ExtendAI*'s design left it up to the user how much they wanted work out their rationale before asking the assistant for feedback. Consequently, the rationales that participants provided to *ExtendAI* varied in their level of elaboration (e.g. some rationales included both general strategies, and specific ETFs and sectors, whereas others only focused on specific ETF choices). This appeared to affect how insightful participants perceived the feedback of *ExtendAI* to be. Participants who were still in the process of making up their minds, or whose rationale was in a less mature state when they interacted with *ExtendAI*, reported that they benefited from the assistant's feedback:

"I was a bit insecure, if I should invest in all these three emerging sectors. But then, when it was really so clear feedback from the AI, I felt like, I had it already in mind a bit like this. But I see it now even more clearly." (ER-5)

In contrast, participants who had already invested a lot of thought to fully develop their rationale reported minimal benefit from *ExtendAI*'s feedback. In some cases, this was because the participant's rationale was so sophisticated that *ExtendAI* was not able to point out further improvements:

*"I had the feeling, the focus was more on confirming what you're doing, kind of saying yes to what I was doing, rather than the other way around, kind of destroying what I want to do. But from my feeling, my strategy is also not complete bullsh***, so it might be different if the result would be total mess."* (RE-7)

However, another factor is that participants who have already put considerable effort into their rationale tend to hold on to it, as some participants also admitted upon reflection:

"You have strong belief in what you want to do, right? And at that time you will consider it as noise, this unuseful information. [...] But to be very objective, it might not be noise. But the problem is, you are super subjective, you just want to believe in your belief." (RE-3)

We do not see this issue with *RecommendAI*, mainly because its recommendations came before participants had thought deeply about their decision:

"Whereas in [RecommendAI], I haven't taken any decisions. And then I was reading through what the AI assistant would tell me to do, and I was trying to make my mind based from what the assistant is telling me." (ER-4)

On the downside, *RecommendAI*'s suggestions were often perceived to come too early in the decision-making process, thereby taking away the opportunity for independent thinking from participants, as outlined in Section 6.2.1. Together, these statements in this theme point out a third tension in augmenting human decision-making: AI suggestions should be introduced at the right moment. Suggestions should not be introduced too early in the decision process to avoid unwanted anchoring around the AI, but also not too delayed to meaningfully contribute to the user's reasoning. This also relates to

participants' opinions on using natural language to describe their rationale in *ExtendAI*, as discussed in the next and final theme.

6.2.4 Theme 4: Natural Language: Flexible Expression and (Un)desirable Friction. *ExtendAI* relied on users to articulate their reasoning in natural language, which participants felt conflicted about. On the one hand, participants found it *"just annoying"* (RE-8) or *"somewhat cumbersome"* (ER-7) to type out their entire rationale. This also meant that participants were particularly disappointed when *ExtendAI*'s feedback turned out not useful to them:

"I had to read my data, and then give the input, say, 'Hey, I want to decrease this, and I want to do that and that.' And then I only got a comment on this, it was not a new idea on how to invest, or even some numbers. It was just a comment, that it's a good idea." (RE-6)

On the other hand, some participants acknowledged the benefit of being forced to consciously think through their decision, *"to not randomly invest in things, because of simple feelings"* (RE-7):

"I think if the goal is understanding, then you do have to put in some work yourself. There's no pain, no gain, I would say." (ER-10)
"Yeah, no free lunch, right? So you have to do some work." (ER-1)

However, it was not only the effort of typing out or thinking through the decision that participants discussed. Some participants also emphasised the more fundamental difficulty of expressing vague thoughts in words, stating that they *"had a hard time describing why I'm doing things"* (RE-7), or that they found it *"a bit hard to put my strategy into words"* (RE-8). Then again, some participants also stated that natural language gives them the flexibility to describe their reasoning and complex dependencies in a way that might be difficult to do otherwise:

"I think, especially for [...] complex investment strategies where you take percentages and combine things, [...] maybe combine topics and have some dependencies between emerging trends in the healthcare sector [...]. This is sometimes better to capture in words." (RE-7)

All of these comments show that the natural language capabilities of LLMs can be valuable for decision-making, but also a challenge, highlighting the need to employ them purposefully in the design of AI assistants. To this end, participants made several suggestions. For instance, some participants suggested how the writing of the rationale could be simplified by suggesting what to write, or by describing the rationale for one ETF at a time:

"Not have this big, free text field where you might miss [something], because I needed to check from left to right, 'Oh, did I mention all the trades, which I've already done, and where I want to do something? Or did I miss one?' So, having kind of, 'Okay, here for emerging markets,' you have this text field, where you either write or you have some support, what you could write." (RE-7)
"I would do it actually ETF by ETF. [...] Asking me to write a whole strategy, 'I want to do this

and this, and this and this, and these are the reasoning, blah blah.’ This is too much, and I think it will be a huge text. But to have it more specific for a particular ETF, this is helpful. So, ‘This is a particular ETF, my train of thought is this, and that’s what I want to do.’” (ER-1)

Other participants proposed relying more on visuals and direct manipulation GUIs, and limiting the use of natural language to where it is actually valuable:

“I think integrating visual feedback with the AI’s inputs is just more helpful because that’s how people are used to making investing decisions.” (ER-3)

“It would be better if you could make a ‘pseudo trade’, that you have an interface to do a prototype of your trades, add one line of reasoning behind each trade, and then it’ll give you feedback.” (ER-7)

7 DISCUSSION

As our findings showed, both types of AI assistance impacted participants in their decision-making and the outcomes they achieved in rather distinct ways, which we further discuss in this section.

7.1 The Approach of Extending Users’ Own Reasoning Is (Too) Subtle, but Effective

Both *ExtendAI* and *RecommendAI* helped participants to build more diversified portfolios, with *ExtendAI* being slightly more effective especially in terms of regional diversification. Interestingly, participants achieved this with fewer, but apparently more targeted trades with *ExtendAI* compared to *RecommendAI*, as the analysis of the impact of both assistants on participants’ decision-making showed. Combined with participants’ statements on how they used both assistants, this suggests that participants understood the weaknesses of their portfolio better when using *ExtendAI*, while they tended to review *RecommendAI*’s suggestions individually, with a less holistic perspective on their portfolio.

Both assistants also made participants feel more informed. This perception appeared more warranted in the case of *ExtendAI*, as participants’ confidence before seeing the outcome generally aligned with their satisfaction afterwards, suggesting an adequate understanding of the associated risks. With *RecommendAI* on the other hand, there was a big discrepancy between participants’ high confidence before and low satisfaction after seeing the outcome. This can be interpreted as a form of overreliance, similar to the strategy-graded notion of reliance introduced by Fok and Weld [25]. Another interpretation is that participants perceived less ownership over their decision with *RecommendAI*, similar to how users do not perceive ownership over AI-generated text [20]. As a result, participants seemed to be satisfied when the outcome was good and blamed *RecommendAI* in case of a bad outcome, as indicated by the even split between positive and negative satisfaction ratings, while the satisfaction ratings for *ExtendAI* were more differentiated.

Surprisingly, despite the higher effectiveness of *ExtendAI* in improving participants’ decision-making, both the respective ratings and interview statements show that participants often perceived

ExtendAI as less insightful than *RecommendAI*. Apparently, while *ExtendAI* effectively made participants reflect on their decisions, it did so in such a subtle way that participants often did not recognise its contribution to their decision-making, which presents a challenge to the adoption of tools designed like *ExtendAI*.

7.2 Many Users Want to Think for Themselves

We designed *ExtendAI* to help users reflect on their reasoning and be cognitively engaged with the task by requiring them to write down their rationale and by giving them feedback without specific ETF recommendations. This expectedly led to higher cognitive effort than with *RecommendAI*, as also confirmed by participants’ ratings. Previous studies with crowd workers on simple decision-making tasks found that users prefer AI systems that require less cognitive effort [9, 10, 58], pointing towards a trade-off between cognitive engagement and user preference [10]. The results of our study with a realistic decision-making task are—to a certain extent—contrary to this, as participants were evenly split in their preference for the two AI assistants, often due to differing opinions on the use of specific recommendations. While many participants appreciated the recommendations they received for their actionability, a large portion rejected the recommendations, as they found them to be disengaging and hard to verify. This is in line with other studies on real-world tasks with experienced participants, such as clinicians [64] or pilots [67], who tend to be sceptical toward recommendations. Thus, our findings suggest there are two distinct user groups: one group that welcomes working with ready-to-use AI recommendations, and one group that rejects AI recommendations, as they want to remain as involved in the decision-making as possible.

7.3 From Trust & Reliance Calibration to Integration With Users’ Decision-Making

Recent research has focused on trust [45] and/or reliance [52] calibration as the core challenge to effective AI-assisted decision-making. While the wide range of work on trust/reliance calibration shows the complexity of the topic, it is still a somewhat simplistic perspective on AI-assisted decision-making that is mostly outcome-focused and often neglects the decision-making process leading up to the outcome. This is partly due to (and further reinforced by) the widespread use of simple input-output mapping tasks in AI-assisted decision-making research [37]. These tasks are often more representative of machine learning tasks than human tasks and afford little room to explore human decision-making processes. By using a more realistic and complex task, our study highlights a range of challenges in AI-assisted decision-making beyond just recognising when the AI is correct or not, related to how AI integrates into and contributes to users’ decision-making process.

Our findings highlight that embedding AI into human decision-making can lead to new tensions, where a benefit of a system is often also a drawback. We observed three such tensions: (1) actionability vs. cognitive engagement, (2) new insights vs. consistency with the user’s reasoning, (3) not too early vs. not too late in the decision-making process. We intentionally designed *ExtendAI* to be quite different from *RecommendAI*, for instance by avoiding specific recommendations in *ExtendAI*’s feedback. To an extent, this

certainly provoked the tensions we observed, as both assistants tended to land on opposite sides of these spectrums. Yet, while it is conceivable to resolve some of the tensions (e.g., AI advice simultaneously being consistent with users' reasoning and adding novel insights on top), for other tensions, the goals appear to contradict each other (e.g., AI advice should be specific to be actionable, but also not too specific to keep users cognitively engaged), requiring trade-offs appropriate for the application context.

We designed *ExtendAI* to integrate into users' decision-making, and as indicated by both participants' interview statements as well as their ratings (higher *consideration of suggestion*, lower *interference with own decision-making*), *ExtendAI* indeed appears to enable a better integration than *RecommendAI*. Interestingly, participants' remarks suggest that through this integration, *ExtendAI* may 'circumvent' or mitigate the issue of trust/reliance calibration to some degree: Since the feedback is embedded into users' own writing and is on a more fine-grained level than concrete recommendations, it is easier to make sense of. As a result, users seem to have less difficulty deciding whether or not to trust the AI, to the extent that the question of trust may not even be users' main concern. This is also reflected by participants' trust ratings, where *ExtendAI* completely lacks negative ratings in contrast to *RecommendAI*, while receiving a large portion of neutral and highly positive ratings.

This benefit of *ExtendAI* came at the cost that users had to write down their rationale first, which was found to be burdensome, although some participants also explicitly recognised the value of this step. Participants made suggestions for how this unwanted cognitive effort could be minimised, for example, by allowing users to formulate their intended trades through direct manipulation, and using natural language only to explain why they want to perform certain trades. However, less obviously, *RecommendAI* introduced cognitive effort as well through the need to evaluate its recommendations, as highlighted by some participants. Unlike *ExtendAI*, this cognitive effort in *RecommendAI* is not primarily the result of interaction design in need of improvement. Instead, it is inherent to the paradigm of recommending complete solutions, and the effort is higher for more complex decisions [23, 56]. In short, *ExtendAI* mainly added cognitive effort *before* interacting with the AI but reduced it *afterwards*, whereas for *RecommendAI* it was the other way around.

Taken together, the findings showed that designing AI decision support systems that build upon people's decision rationales provides an alternative interaction paradigm that could help better integrate AI decision support in human decision-making—even if the instantiations might need to differ and be adapted to the specific task and types of users (e.g. how they provide their rationale to the AI and how detailed and elaborate it should be).

7.4 Limitations & Future Work

Our intention was to use ETF investment as a use case to explore how LLMs can augment human decision-making in general. Nevertheless, certain aspects of our results may have been shaped by the specifics of our task. We argue that the tensions that we identified are likely to be present for most, if not all, complex decision tasks. However, the impact of the advantages and disadvantages of the two types of designs—represented by our two assistants—will likely

vary depending on the tasks. For example, in our task, decisions were multifaceted, as participants had to select multiple ETFs and decide on the amount of shares to buy or sell. Consequently, participants could selectively rely on the assistants' suggestions. In contrast, for a task where a single decision has to be made (e.g., medical diagnosis), a recommendation-based approach in particular is more likely to result in a binary "accept or reject" situation, which may highlight the disadvantages of this approach more strongly, as is often the case in related work [4, 34, 65]. Similarly, while our participants did discuss the risk of overreliance due to a lack of engagement with *RecommendAI*, it may have been less present than in other studies [12, 33, 67], since participants had already deeply engaged with the task in the first time step (familiarisation phase) before interacting with our AI assistants. Future work should investigate how analogous AI designs compare in the context of other realistic and complex decision-making tasks.

Another direction for future work is to investigate more quantitative measures with a larger sample size, to explore for example how the perceptions of comparable AI designs depend on personality traits such as need for cognition or the propensity to trust in AI, or factors like users' expertise. It would also be interesting to compare the effect of both types of AI assistance on decision quality in more detail. For our study, we intentionally chose an open-ended task without any objectively correct or false decisions. While this reflects many real-world tasks and allowed us to gain interesting qualitative insights, it also meant that it was difficult to establish clear performance metrics.

Future work could also explore various extensions and modifications to our *ExtendAI* concept, such as allowing multi-turn interactions, or prompting the LLM to be more critical, as some participants felt that *ExtendAI* was overly positive in its tone, reflecting the general behaviour of the underlying GPT-4o model. Furthermore, different different *forms* of feedback could be evaluated. In our study we limited the feedback to pointing out to the user how they could improve their rationale and drawing their attention to further aspects they could consider in it rather than pointing out specific solutions to them. Lastly, our study demonstrated the significance of the role that an LLM assumes in the decision-making process. We explored two possible roles with *RecommendAI* and *ExtendAI*, leaving ample room for future work to explore additional roles that could combine elements of both.

8 CONCLUSION

We explored two roles for LLMs in AI-assisted decision-making that leverage their unique capabilities to support complex and multifaceted decision tasks: (1) a more traditional role where the AI provides the user with direct suggestions, referred to as *RecommendAI*, and (2) an AI that supports users by giving feedback on and extending their own reasoning, referred to as *ExtendAI*. Our findings show that the *ExtendAI* generally made people feel supported and more reflective in their own thinking and helped them achieve better outcomes. On the other hand, the recommendation-based approach required less effort and seemed to help people get inspiration for new directions. Both of these benefits also came with their respective drawbacks. Therefore, both roles seem to have their pros and cons, leading to several tensions—for example, when

there is a need to receive actionable advice while at the same time also being able to think for oneself. These tensions point towards an interesting design space between recommendation-centric approaches and those that build upon people's own reasoning. With this paper, we intend to inform and inspire further exploration of this space.

ACKNOWLEDGMENTS

We would like to thank Warren Park for his support with the implementation of the trading platform. We further thank our participants for their valuable input. We also thank our anonymous reviewers for their thoughtful reviews, which helped us to strengthen the paper significantly. The research reported in this work was supported by the EU H2020 ICT48 project 'Humane AI Net' under contract # 952026. The support is gratefully acknowledged.

REFERENCES

- [1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Yokohama, Japan, 81:1–81:16. <https://doi.org/10.1145/3411764.3445717>
- [2] Ying Bao, Wankun Gong, and Kaiwen Yang. 2023. A Literature Review of Human–AI Synergy in Decision Making: From the Perspective of Affordance Actualization Theory. *Systems* 11, 9 (Sept. 2023), 442:1–442:20. <https://doi.org/10.3390/systems11090442>
- [3] Astrid Bertrand, James R. Eagan, and Winston Maxwell. 2023. Questioning the Ability of Feature-Based Explanations to Empower Non-Experts in Robo-Advised Financial Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FACCT '23)*. ACM, Chicago, IL, USA, 943–958. <https://doi.org/10.1145/3593013.3594053>
- [4] Jeanette Blomberg, Aly Megahed, and Ray Strong. 2018. Acting on Analytics: Accuracy, Precision, Interpretation, and Performativity. *Ethnographic Praxis in Industry Conference Proceedings* 2018, 1 (2018), 281–300. <https://doi.org/10.1111/1559-8918.2018.01208>
- [5] Meike Bradbury, Thorsten Hens, and Stefan Zeisberger. 2014. Improving Investment Decisions with Simulated Experience. <https://doi.org/10.2139/ssrn.2179276>
- [6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- [7] Virginia Braun and Victoria Clarke. 2023. Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher. *International Journal of Transgender Health* 24, 1 (2023), 1–6. <https://doi.org/10.1080/26895269.2022.2129597> arXiv:<https://doi.org/10.1080/26895269.2022.2129597>
- [8] Zana Bućinca, Alexandra Chouldechova, Jennifer Wortman Vaughan, and Krzysztof Z. Gajos. 2022. Beyond End Predictions: Stop Putting Machine Learning First and Design Human-Centered AI for Decision Support. In *Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22)*. Virtual Event, USA, 1–4.
- [9] Zana Bućinca, 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, Cagliari, Italy, 454–464. <https://doi.org/10.1145/3377325.3377498>
- [10] Zana Bućinca, 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 (April 2021), 188:1–188:21. <https://doi.org/10.1145/3449287>
- [11] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Hamburg, Germany, 15:1–15:19. <https://doi.org/10.1145/3544548.3581251>
- [12] Adrian Bussone, Simone Stumpf, and Dymna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics (ICHI '15)*. IEEE Computer Society, USA, 160–169. <https://doi.org/10.1109/ICHI.2015.26>
- [13] Federico Cabitza, Andrea Campagner, Lorenzo Famigliini, Chiara Natali, Valerio Caccavella, and Enrico Gallazzi. 2023. Let me think! Investigating the effect of explanations feeding doubts about the ai advice. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 155–169. https://doi.org/10.1007/978-3-031-40837-3_10
- [14] Federico Cabitza, Caterina Fregosi, Andrea Campagner, and Chiara Natali. 2024. Explanations considered harmful: The Impact of misleading Explanations on Accuracy in hybrid human-AI decision making. In *World Conference on Explainable Artificial Intelligence*. Springer, 255–269. https://doi.org/10.1007/978-3-031-63803-9_14
- [15] Federico Cabitza, Chiara Natali, Lorenzo Famigliini, Andrea Campagner, Valerio Caccavella, and Enrico Gallazzi. 2024. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial Intelligence in Medicine* 150 (2024), 102819. <https://doi.org/10.1016/j.artmed.2024.102819>
- [16] Longbing Cao. 2023. AI in Finance: Challenges, Techniques, and Opportunities. *Comput. Surveys* 55, 3 (March 2023), 64:1–64:38. <https://doi.org/10.1145/3502289>
- [17] 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 (CHI '23)*. ACM, Hamburg, Germany, 756:1–756:27. <https://doi.org/10.1145/3544548.3581253>
- [18] 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 (IUI '24)*. ACM, Greenville, SC, USA, 103–119. <https://doi.org/10.1145/3640543.3645199>
- [19] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114–126. <https://doi.org/10.1037/xge0000033>
- [20] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI Ghostwriter Effect: When Users Do Not Perceive Ownership of AI-generated Text but Self-Declare as Authors. *ACM Trans. Comput.-Hum. Interact.* 31, 2 (Feb. 2024), 25:1–25:40. <https://doi.org/10.1145/3637875>
- [21] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland UK) (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. <https://doi.org/10.1145/3290607.3312787>
- [22] Eva Eigner and Thorsten Händler. 2024. Determinants of LLM-assisted Decision-Making. arXiv:2402.17385 [cs]
- [23] Mica R. Endsley. 2017. From Here to Autonomy: Lessons Learned from Human–Automation Research. *Human Factors* 59, 1 (Feb. 2017), 5–27. <https://doi.org/10.1177/0018720816681350>
- [24] Christian Fieberg, Lars Hornuf, and David Streich. 2023. Using GPT-4 for Financial Advice. <https://doi.org/10.2139/ssrn.4499485>
- [25] Raymond Fok and Daniel S. Weld. 2024. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-advised Decision Making. *AI Magazine* 45, 3 (July 2024), 317–332. <https://doi.org/10.1002/aaai.12182>
- [26] 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 (IUI '22)*. ACM, Helsinki, Finland, 794–806. <https://doi.org/10.1145/3490099.3511138>
- [27] 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 (Nov. 2019), 50:1–50:24. <https://doi.org/10.1145/3359152>
- [28] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *Comput. Surveys* 51, 5 (Aug. 2018), 93:1–93:42. <https://doi.org/10.1145/3236009>
- [29] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, 139–183. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- [30] Ted Martin Hedesström, Henrik Svedsäter, and Tommy Gärling. 2006. Covariation Neglect among Novice Investors. *Journal of Experimental Psychology: Applied* 12, 3 (2006), 155–165. <https://doi.org/10.1037/1076-898X.12.3.155>
- [31] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghui Cheng. 2023. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 152:1–152:20. <https://doi.org/10.1145/3579628>
- [32] Maia Jacobus, 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 (CHI '21)*. ACM, Yokohama, Japan, 659:1–659:14. <https://doi.org/10.1145/3411764.3445385>

- [33] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How Machine-Learning Recommendations Influence Clinician Treatment Selections: The Example of Antidepressant Selection. *Translational Psychiatry* 11, 1 (June 2021), 108:1–108:9. <https://doi.org/10.1038/s41398-021-01224-x>
- [34] Anna Kawakami, Venkatesh Sivaraman, Logan Stapleton, Hao-Fei Cheng, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. “Why Do I Care What’s Similar?” Probing Challenges in AI-assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts. In *Designing Interactive Systems Conference (DIS '22)*. ACM, Virtual Event, Australia, 454–470. <https://doi.org/10.1145/3532106.3533556>
- [35] Alex Kim, Maximilian Muhn, and Valeri V. Nikolaev. 2024. Financial Statement Analysis with Large Language Models. <https://doi.org/10.2139/ssrn.4835311>
- [36] Sean Koon. 2022. A Human-Capabilities Orientation for Human-AI Interaction Design. In *Virtual Workshop on Human-Centered AI Workshop at NeurIPS (HCAI @ NeurIPS '22)*. Virtual Event, USA, 1–5.
- [37] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM, Chicago, IL, USA, 1369–1385. <https://doi.org/10.1145/3593013.3594087>
- [38] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (Atlanta, GA, USA) (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. <https://doi.org/10.1145/3287560.3287590>
- [39] Christine Laudendach, Michael Ungeheuer, and Martin Weber. 2023. How to Alleviate Correlation Neglect in Investment Decisions. *Management Science* 69, 6 (June 2023), 3400–3414. <https://doi.org/10.1287/mnsc.2022.4535>
- [40] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (Dec. 2020), 18:1–18:45. <https://doi.org/10.3390/e23010018>
- [41] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2024. Towards Human-AI Deliberation: Design and Evaluation of LLM-empowered Deliberative AI for AI-assisted Decision-Making. <https://doi.org/10.48550/arXiv.2403.16812> arXiv:2403.16812 [cs]
- [42] Shuai Ma, Chenyi Zhang, Xinru Wang, Xiaojuan Ma, and Ming Yin. 2024. Beyond Recommender: An Exploratory Study of the Effects of Different AI Roles in AI-assisted Decision Making. arXiv:2403.01791 [cs]
- [43] Harry Markowitz. 1952. Portfolio Selection. *The Journal of Finance* 7, 1 (1952), 77–91. <https://doi.org/10.2307/2975974> jstor:2975974
- [44] Philip Mavrepis, Georgios Makridis, Georgios Fatouros, Vasileios Koukos, Maria Margarita Separdani, and Dimosthenis Kyriazis. 2024. XAI for All: Can Large Language Models Simplify Explainable AI? <https://doi.org/10.48550/arXiv.2401.13110> [cs]
- [45] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. 2024. A Systematic Review on Fostering Appropriate Trust in Human-AI Interaction: Trends, Opportunities and Challenges. *ACM Journal on Responsible Computing* 1, 4 (Nov. 2024), 26:1–26:45. <https://doi.org/10.1145/3696449>
- [46] 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 (FAccT '23)*. ACM, Chicago, IL, USA, 333–342. <https://doi.org/10.1145/3593013.3594001>
- [47] Hussein Mozannar, Jimin Lee, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. 2023. Effective Human-AI Teams via Learned Natural Language Rules and Onboarding. In *Advances in Neural Information Processing Systems (NeurIPS 2023, Vol. 36)*. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.). Curran Associates, Inc., New Orleans, LA, USA, 30466–30498.
- [48] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. arXiv:2311.16452 [cs]
- [49] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, Yokohama, Japan, 237:1–237:52. <https://doi.org/10.1145/3411764.3445315>
- [50] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-Makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. ACM, Sydney, NSW, Australia, 379–396. <https://doi.org/10.1145/3581641.3584033>
- [51] Leon Reicherts, Gun Woo Park, and Yvonne Rogers. 2022. Extending Chatbots to Probe Users: Enhancing Complex Decision-Making through Probing Conversations. In *Proceedings of the 4th Conference on Conversational User Interfaces (CUI '22)*. Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3543829.3543832>
- [52] Myriam Schaschek, Niko Spatschek, and Axel Winkelmann. 2024. For Those about to Rely—a Taxonomy of Experimental Studies on AI Reliance. In *Wirtschaftsinformatik 2024 Proceedings*, Vol. 89. AIS, Würzburg, Germany.
- [53] 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 (IUI '23)*. ACM, Sydney, NSW, Australia, 410–422. <https://doi.org/10.1145/3581641.3584066>
- [54] Philipp Schmidt and Felix Biessmann. 2020. Calibrating Human-AI Collaboration: Impact of Risk, Ambiguity and Transparency on Algorithmic Bias. In *Machine Learning and Knowledge Extraction (CD-MAKE 2020)*. Springer International Publishing, Dublin, Ireland, 431–449. https://doi.org/10.1007/978-3-030-57321-8_24
- [55] Silvana Secinaro, Davide Calandra, Aurelio Secinaro, Vivek Muthurangu, and Paolo Biancone. 2021. The Role of Artificial Intelligence in Healthcare: A Structured Literature Review. *BMC Medical Informatics and Decision Making* 21, 1 (April 2021), 125. <https://doi.org/10.1186/s12911-021-01488-9>
- [56] Auste Simkute, Lev Tankelevitch, Viktor Kewenig, Ava Elizabeth Scott, Abigail Sellen, and Sean Rintel. 2024. Ironies of Generative AI: Understanding and Mitigating Productivity Loss in Human-AI Interactions. <https://doi.org/10.48550/arXiv.2402.11364> arXiv:2402.11364 [cs]
- [57] Steve G. Sutton, Vicky Arnold, and Matthew Holt. 2018. How Much Automation Is Too Much? Keeping the Human Relevant in Knowledge Work. *Journal of Emerging Technologies in Accounting* 15, 2 (Sept. 2018), 15–25. <https://doi.org/10.2308/jeta-52311>
- [58] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. <https://doi.org/10.1145/3579605>
- [59] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, Glasgow, Scotland, UK, 601:1–601:15. <https://doi.org/10.1145/3290605.3300831>
- [60] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (College Station, TX, USA) (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. <https://doi.org/10.1145/3397481.3450650>
- [61] Manuel Woschank, Erwin Rauch, and Helmut Zsifkovits. 2020. A Review of Further Directions for Artificial Intelligence, Machine Learning, and Deep Learning in Smart Logistics. *Sustainability* 12, 9 (Jan. 2020), 3760. <https://doi.org/10.3390/su12093760>
- [62] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users’ Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*. ACM, Cagliari, Italy, 189–201. <https://doi.org/10.1145/3377325.3377480>
- [63] Qian Yang, Yuxing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing Biomedical Literature to Calibrate Clinicians’ Trust in AI Decision Support Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Hamburg, Germany, 14:1–14:14. <https://doi.org/10.1145/3544548.3581393>
- [64] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, Pratik Ghosh, Mercy Ranjit, Kenza Bouzid, Anton Schwaighofer, Fernando Pérez-García, Harshita Sharma, Ozan Oktay, Matthew Lungren, Javier Alvarez-Valle, Aditya Nori, and Anja Thieme. 2024. Multimodal Healthcare AI: Identifying and Designing Clinically Relevant Vision-Language Applications for Radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, Honolulu, HI, USA, 444:1–444:22. <https://doi.org/10.1145/3613904.3642013>
- [65] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lacey M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, Honolulu, HI, USA, 445:1–445:18. <https://doi.org/10.1145/3613904.3642343>
- [66] 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 (FAccT '20)*. ACM, Barcelona, Spain, 295–305. <https://doi.org/10.1145/3351095.3372852>
- [67] Zelun Tony Zhang, Sebastian S. Feger, Lucas Dullenkopf, Rulu Liao, Lou Süßlin, Yuanting Liu, and Andreas Butz. 2024. Beyond Recommendations: From Backward to Forward AI Support of Pilots’ Decision-Making Process. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (Nov. 2024), 485:1–485:32.

<https://doi.org/10.1145/3687024>

- [68] Zelun Tony Zhang, Yuanting Liu, and Heinrich Hussmann. 2021. Forward Reasoning Decision Support: Toward a More Complete View of the Human-AI Interaction Design Space. In *CHIItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter (CHIItaly '21)*. ACM, Bolzano, Italy, 18:1–18:5. <https://doi.org/10.1145/3464385.3464696>

A AI INPUT AND OUTPUT EXAMPLES

A.1 *RecommendAI* Example Output (ER-10)

Here is a portfolio recommendation based on your investment profile:

1. VOO (Vanguard S&P 500 ETF): You already hold 8 shares of this ETF, which is a great choice for broad U.S. market exposure. Given your long-term horizon and cautious approach, maintaining or slightly increasing your investment in VOO can provide stability and growth.
2. VGK (Vanguard FTSE Europe ETF): This ETF offers exposure to European markets, providing geographical diversification. Investing in VGK can help balance your portfolio with international exposure.
3. ICLN (iShares Global Clean Energy ETF): As you are willing to take some risks, adding exposure to the clean energy sector can be a strategic move. ICLN focuses on global clean energy companies, aligning with the growing trend towards renewable energy.
4. ACWV (iShares MSCI Global Min Vol Factor ETF): This ETF aims to provide exposure to global equities with a focus on minimizing volatility. ACWV can add a layer of stability to your portfolio, balancing the riskier elements.

Invest the largest amount of your portfolio into VOO for stability and growth. Allocate a significant portion to VGK for geographical diversification. Consider a moderate investment in ICLN to tap into the clean energy sector's potential. Finally, allocate a smaller portion to ACWV for added stability.

A.2 *ExtendAI* Example Input/User Rationale (RE-05)

I want to invest 2000\$ into IXN, because of the AI boom, 2000\$ into ICLN because the clean energy boom, and 2000\$ into IXJ because of the the healthcare boom.

I want to invest 2000\$ into EEM and 2180\$ into VOO to diversify across more sectors.

A.3 *ExtendAI* Example AI Output/Rationale Extension (RE-05)

Here is your edited investment rationale with suggestions for improvement:

I want to invest \$2000 into IXN, because of the AI boom. **While IXN provides substantial exposure to the technology sector, it's important to note that it is heavily concentrated in a few large tech companies, which can increase volatility. Consider balancing this with other sectors or regions to mitigate risk.**

I want to invest \$2000 into ICLN because of the clean energy boom. **ICLN is a great choice for exposure to the clean energy sector, but it is also quite concentrated in a few companies and countries. Diversifying within the clean energy sector or balancing with other sectors could be beneficial.**

I want to invest \$2000 into IXJ because of the healthcare boom. **IXJ offers good exposure to the healthcare sector, which is typically less volatile and can provide stability. However, it is heavily tilted towards U.S. companies. Consider adding international healthcare ETFs to diversify further.**

I want to invest \$2000 into EEM and \$2180 into VOO to diversify across more sectors. **EEM offers broad exposure to emerging markets, which can be volatile but offer growth potential. VOO is a solid choice for U.S. large-cap exposure, but it is already part of your holdings. Ensure that your overall portfolio is balanced across different sectors and regions to avoid over-concentration.**

Overall, your investment rationale shows a good understanding of current market trends and the desire to diversify. However, consider the concentration risks associated with sector-specific ETFs like IXN, ICLN, and IXJ. Balancing these with more diversified ETFs across different regions and sectors can help mitigate risks. Additionally, since you already hold VOO, ensure that adding more doesn't overly concentrate your portfolio in U.S. large-cap stocks.

These above are just a few things for you to consider for your intended trades. You can now proceed with the trades and the trade button is now enabled. It is up to you if and to what extent you want to consider any of the feedback/considerations provided above.

B AI ASSISTANT PROMPTS

B.1 *RecommendAI* Prompt

You are a critical financial advisor embedded into an ETF investing platform. The user is asking for a portfolio recommendation based on their investment profile and goals.

The ETFs that are available to the user on the investing platform are provided in the attached JSON and the ETFs the user has previously invested in are shown below.

Provide a recommended portfolio of 2-4 ETFs that align with the user's profile, balancing the trade-off between risk and potential return. Ensure that you only recommend ETFs available to the user, as you can see in the JSON below.

Make sure to reflect and give feedback on the ETFs the user already owns and always suggest a minimum of two ETFs the user does not own yet. Make sure to only give advice for investing in ETFs. Give suggestions for specific ETFs that the user has not mentioned or owns yet as well as advice on selling/buying more of the ETFs the user owns already.

Give a subtle indication of the amount of ETFs that should be sold/bought, but do not give the amount in \$. E.g., ‘invest the largest amount of your portfolio into ETF XXX...’. Do this after listing the ETFs to invest in.

The profile: ‘You are a 40-year-old with a stable job and looking to invest in ETFs for a long-term horizon of about 10 years. You are willing to take some risks but you are generally rather cautious. Thus, your goal is to create a diversified ETF profile covering different regions and industries.’

Ensure every suggestion you provide is in bold and the structure is consistent as follows (format all important recommendations using `` HTML tags):

```
<html>
<body>
<p>Here is a portfolio recommendation based on your investment profile:</p>
(...)
<p>These above are just a few things for you to consider for your intended trades. You can now proceed with the trades and the trade button is now enabled. It is up to you if and to what extent you want to consider any of the feedback/considerations provided above.</p>
</body>
</html>
```

Make sure the response is consistent with no contradictions.

B.2 ExtendAI Prompt

You are a critical financial advisor embedded into an ETF investing platform that gives feedback on users’ rationales when deciding which ETFs to invest in. Let’s assume the user of this system is a beginner investor investing \$20,000–\$30,000 in a portfolio of up to 10 different ETFs that align with their investor profile. The user is looking to invest in ETFs for a long-term horizon of about 10 years they are willing to take some risks but you are generally rather cautious. Thus, the goal is to create a diversified ETF profile covering different regions and industries. The user has formulated an investment thesis to motivate the selection of ETFs that they plan to invest in.

The ETFs that are available to the user on the investing platform are provided in the attached JSON and the ETFs the user has previously invested in are shown below.

Make sure that the user listed at least 2 ETFs and has provided a rationale for investing in them. If not, or the user has not input any message, respond with “Please choose at least 2 ETFs which you would like to invest in.” or “You have listed a selection of ETFs that you intend to invest in but you have not yet included a clear motivation/rationale for investing your money in them. Make sure you also include your rationale in the text box below.”

If the user input does list 2 or more ETFs but does not provide any rationale for investing in them, respond with the following: “You have listed a selection of ETFs that you intend to invest in but you have not yet included a clear motivation/rationale for investing your money in them. Make sure you also include your rationale in the text box below.”

When making suggestions on how the user’s rationale could be improved, make use of your own knowledge of the returns and volatility of the different ETFs as well as the data in the JSON—in particular, how diversified the ETFs are with regards to “sector_holdings”, “company_holdings”, “country_holdings”. Given the user’s goal to adequately balance potential returns and risk, use this trade-off as the main criteria for your suggestions and feedback. Keep the ETFs the user already owns in mind. Make sure to only give advice for investing in ETFs.

To provide your suggestions, take the user’s original rationale and seamlessly “weave” your suggestions into their original rationale text (similar to how an editor might work on an author’s text). Format your output as HTML and ensure that your feedback and suggestions are highlighted in bold using the `` HTML tag.

Here is the format that your response should follow:

1. Start with a brief introduction indicating that this is the edited investment rationale.
2. Provide the original rationale followed by your suggestion in bold within the same paragraph.
3. End with an overall feedback paragraph summarizing the strengths and concerns of the user’s portfolio along with your suggestions. Formulate it as things to consider for the user.
4. Add the following after the feedback paragraph: “These above are just a few things for you to consider for your intended trades. You can now make proceed with the trades and the trade button is now enabled. It is up to you if and to what extent you want to consider any of the feedback/considerations provided above.”

Ensure every suggestion you provide is in bold and the structure is consistent as follows:

```
<html>
<body>
<p>Here is your edited investment rationale with suggestions for improvement:</p>
<p>Original rationale. <b>Your suggestion.</b></p>
<p>Original rationale. <b>Your suggestion.</b></p>
(...)
<p><b>Overall feedback paragraph summarizing strengths and concerns and providing suggestions.</b></p>
<p>These above are just a few things for you to consider for your intended trades. You can now proceed with the trades and the trade button is now enabled. It is up to you if and to what extent you want to consider any of the feedback/considerations provided above.</p>
</body>
</html>
```

C PARTICIPANT DETAILS

Table 2: Participant sample: Participant RE-2 was recruited through Prolific. Participant RE-1 was not included due to a malfunction of the prototype. CoR stands for Country of Residence. The last two columns refer to the participant’s preference and the level of detail their rationales showed.

Part.ID	Edu./Degree	Occupation	Gender	Age	CoR	ETF Exper.	AI Preference	Task Time
ER-1	Graduate	Head of Technology & AI	Male	40	Germany	2 years	Extend	49.55
ER-2	Graduate	Designer	Male	33	Germany	4 years	Extend	39.83
ER-3	Bachelor	Graduate Student	Female	28	Canada	7 years	Recommend	49.87
ER-4	Graduate	Technology Lead	Male	32	Switzerland	12 years	Recommend	31.40
ER-5	Graduate	Software Engineer	Female	38	Germany	4 years	Extend	62.72
ER-6	Graduate	Software Engineer	Female	29	Germany	1.5 years	Extend	48.68
ER-7	Graduate	Software Engineer	Male	29	Finland	1.5 years	Extend	63.98
ER-8	Bachelor	Graduate Student	Male	28	Switzerland	0 years	Recommend	53.43
ER-9	Bachelor	Graduate Student	Male	23	Switzerland	2 years	Extend	49.87
ER-10	Graduate	PhD Student	Male	30	UK	0 years	Extend	34.78
RE-2	Bachelor	Supply Chain Planner	Male	25	UK	6 years	Recommend	17.98
RE-3	Graduate	Research Assistant	Male	32	Germany	0.5 years	Recommend	54.23
RE-4	Graduate	SAP Consultant	Female	33	Germany	1 year	Recommend	35.23
RE-5	Graduate	Founder	Male	32	Germany	7 years	Extend	35.80
RE-6	High school	Undergraduate Student	Male	22	Germany	0 years	Recommend	28.10
RE-7	Graduate	Researcher	Male	28	Germany	3 years	Recommend	40.70
RE-8	Graduate	Designer	Female	31	Germany	4 years	Recommend	59.28
RE-9	Graduate	Postdoc	Male	30	USA	5 years	Recommend	40.42
RE-10	Graduate	Research Scientist	Male	32	Switzerland	1 year	Recommend	65.17
RE-11	Graduate	Assistant Physician	Male	29	Germany	15 years	Extend	38.77
RE-12	Graduate	Associate Professor	Male	34	UK	4 years	Extend	24.17