Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations

Conference Paper · March 2023

DOI: 10.1145/3581641.3584080

4 authors, including:

Federico Maria Cau
Università degli studi di Cagliari
9 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Lucio Davide Spano
Università degli studi di Cagliari
97 PUBLICATIONS **1,109** CITATIONS
SEE PROFILE

Nava Tintarev
Delft University of Technology
111 PUBLICATIONS **2,814** CITATIONS
SEE PROFILE

Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations

FEDERICO MARIA CAU, University of Cagliari, Italy
HANNA HAUPTMANN, Utrecht University, the Netherlands
LUCIO DAVIDE SPANO, University of Cagliari, Italy
NAVA TINTAREV, Maastricht University, the Netherlands

A common criteria for Explainable AI (XAI) is to support users in establishing appropriate trust in the AI – rejecting advice when it is incorrect, and accepting advice when it is correct. Previous findings suggest that explanations can cause an over-reliance on AI (overly accepting advice). Explanations that evoke appropriate trust are even more challenging for decision-making tasks that are *difficult for humans and AI*. For this reason, we study decision-making by non-experts in the high-uncertainty domain of stock trading. We compare the effectiveness of three different explanation styles (influenced by inductive, abductive, and deductive reasoning) and the role of AI confidence in terms of a) the users' *reliance* on the XAI interface elements (charts with indicators, AI prediction, explanation), b) the correctness of the decision (*task performance*), and c) the *agreement* with the AI's prediction. In contrast to previous work, we look at interactions between different aspects of decision-making, including AI correctness, and the combined effects of AI confidence and explanations styles. Our results show that specific explanation styles (abductive and deductive) *improve the user's* ***task performance*** *in the case of high AI confidence* compared to inductive explanations. In other words, these styles of explanations were able to invoke correct decisions (for both positive and negative decisions) when the system was certain. In such a condition, the ***agreement*** between the user's decision and the AI prediction confirms this finding, highlighting a significant agreement increase when the AI is correct. This suggests that both explanation styles are suitable for evoking appropriate trust in a confident AI.

Our findings further indicate a need to consider AI confidence as a criterion for including or excluding explanations from AI interfaces. In addition, this paper highlights the importance of carefully selecting an explanation style according to the characteristics of the task and data.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Uncertainty quantification.*

Additional Key Words and Phrases: XAI, AI confidence, Logical reasoning, Inductive, Deductive, Abductive, Random forest, Stock market prediction

ACM Reference Format:
Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3581641.3584080

1 INTRODUCTION

The spread of innovative Artificial Intelligence (AI) algorithms assists many individuals in their daily life decision-making tasks but also in sensitive domains such as disease diagnosis [4], and credit risk [54]. However, a great majority of these algorithms are of a black-box nature, bringing the need to make them more transparent and interpretable

Manuscript submitted to ACM

along with the establishment of guidelines to help users manage these systems [3, 47]. The eXplainable Artificial Intelligence (XAI) research field tries to achieve these goals by providing tools for supporting users in AI-assisted decision-making and uncovering the AI's error boundaries [8]. The XAI community investigated numerous factors influencing subjective [59, 64] and objective [38, 63] metrics in the user-AI team, such as the effects of presenting AI-related information and explanations to users. There have been contrasting effects of presenting explanations observed in different work in the literature. On the one hand, previous research demonstrated that explanations might cause users to follow the AI's advice more often, even when it is wrong [56], or lead users to create an incorrect mental model of the AI system [16]. On the other hand, if we consider studies focusing on people having low domain expertise, we have results indicating overconfidence [51, 65] (users that rely mainly on their ability to make a decision) but also overreliance [12] (users relying primarily on the automatic support). Such results may derive from the different settings in these studies, indicating the need for further research identifying the factors causing the different user behaviour.

In recent studies, a factor that gained attention is AI confidence in predictions, which quantifies how likely the AI will correctly classify an individual prediction. Some results show an influence of a confident AI on users' trust and agreement with the AI's predictions, even if its suggestion is wrong [48, 55]. However, there is also evidence that such confidence does not improve the task performance (i.e., the ability to make a correct decision) of the AI-user team [61].

Other studies focused on identifying factors improving task performance. Lai and Tan [39] found that showing the AI's prediction increases task performance. Other research [13, 35] shows that the correctness of the AI predictions strongly influences the user's decision. Hence, AI-related information like confidence and correctness may play a fundamental role in users' decision-making processes, but their effect also depends on experimental settings.

Finally, human-centered aspects such as presenting and selecting the appropriate explanation technique are usually overlooked in the literature. The focus is usually on algorithms or comparing the presence and the absence of explanations. However, even the same technique may have different effects if presented through different visualizations. Recent studies started covering these aspects, for instance, by contextualizing explanations [11] or comparing visual, textual or hybrid explanations [61]. We focus here on the reasoning triggered by explanations, which results in an effective or ineffective understanding of the AI suggestions if not carefully selected according to the presented data. Previous literature in this field is sparse but includes attempts to classify the techniques into inductive, abductive, and deductive styles according to Pierce's theory [15], and highlighting different effects between inductive and deductive styles in the image classification domain [13].

In this paper, we investigate the interactions between different aspects of decision-making, focusing in particular on the combined effects of AI confidence and the explanation reasoning style (inductive, abductive, and deductive). We hypothesise that a confident AI creates consistent explanations, which users can effectively use for accepting or rejecting the AI suggestion only if they trigger the appropriate reasoning type. To demonstrate this, we set up a user study controlling AI-related information in an XAI interface, including **a)** the correctness of the AI suggestion, **b)** AI confidence and **c)** explanations presented with logical reasoning styles (i.e., inductive, abductive, and deductive) [15]. We analyse these factors on **i)** users' reliance on the XAI interface elements (stock charts, AI prediction, and explanation), **ii)** users' task performance, and **iii)** agreement with the AI. We do this in the stock market domain, allowing us to study decision-making in a high-uncertainty domain like a stock trading task, and which factors evoke appropriate trust in decisions that are difficult for both humans and AI.

We evaluate the effectiveness of the considered factors in an online study with 184 participants, where users interacted with an AI-assisted trading platform simulator to buy or sell four different stocks. Our results show that AI confidence impacts the relative ranks between the use of the different information types presented in the XAI interface – users

rely more on the instance data in case of a low-confidence prediction while rely equally on the AI prediction and the instance data in case of high confidence. In addition, we registered a positive significant effect of the abductive and deductive explanation styles on task performance when the AI confidence is high. The same configurations (high AI confidence plus abductive or deductive style explanations) lead to an increased agreement with the prediction when the AI is correct. To summarize, this paper makes the following contributions:

- We show that AI-related factors such as confidence and correctness interact with human-centered properties of an XAI interface, particularly the explanation reasoning style, in the decision-making process. On the one hand, the quality of the explanation depends on the confidence of the AI prediction. On the other hand, users receive such information only if its presentation triggers an effective reasoning style.
- We provide a set of guidelines for effectively including (or excluding) explanations in XAI interfaces when non-expert users decide in a high-uncertainty domain, validating the effects on different aspects of the decision-making process (reliance, agreement and task performance) of the AI confidence, AI correctness, and different explanation styles.

The paper has the following organisation. Section 2 introduces the related work. Section 3 describes the method, hypotheses and settings of the user study, while Section 4 discusses the results. Section 5 proposes a discussion of the results by highlighting their implications and limitations. We conclude the article and describe our plans for future work in Section 6.

2 RELATED WORK

This section covers the research we used for i) contextualizing our study in the stock trading domain and ii) motivating the questions investigated in this paper. We start from summarizing the frequently used XAI techniques concerning financial forecasting, for providing an overview the available options for the classification model and the explanations. Then, we cover the state of the art in the estimation of the AI confidence on machine learning models. Next, we briefly describe humans' logical reasoning styles and motivate why they may improve XAI explanations. Finally we frame the state of the art in XAI system evaluations, focusing on identifying methods and metrics employed.

2.1 XAI on Stock Market Prediction

The ever-growing field of machine learning applications has led to considerable advancements in many domains, including financial forecasting. Nevertheless, most of these techniques are black-box, needing to explain why a model reached a specific output(s). A performant and widely used class of models for predicting stock market trends are Tree-based, like Random Forest (RF) [2, 9, 33, 36, 37, 50], which are recommended for financial forecasting and suitable for both classification/regression tasks. However, if we consider a classification task like predicting future stock market trends, these models' performances are naturally bounded to the selected stocks and trading window [9, 50]. Thus, we may expect a great accuracy performance (> 90%) in predicting stock trends, for example, 30 days ahead, but the accuracy decreases (< 75%) when predicting price trends that are 5 or 7 days ahead. Since we are interested in predicting stock market trends on a short-term window (7 days) while having a reasonable model accuracy (at least > 70%), we chose a Random Forest for classification tasks in our experiments. The stock market prediction task enables us to set up a user evaluation where the AI's predicted trend can actually support novice traders in buying or selling stocks since we expect that novice users are unfamiliar with stock trading. However, additional information is needed for users to understand better AI's decisions, which can be delivered using eXplainable Artificial Intelligence (XAI) techniques.

The eXplainable Artificial Intelligence (XAI) [8] research presents many ways to explore the reasons behind predictions, making models more trustworthy and offering investors and traders the tools for making better decisions. The two most common techniques used in the literature for explaining financial forecasting are LIME (Local Interpretable Model-Agnostic Explanations) [49] and SHapley Additive exPlanations (SHAP) [44]. These techniques are generally valid for financial market forecasting because they explain an opaque model's decision locally or globally, giving insights into the features (i.e., technical indicators, stock-related news, buy/trigger signals, etc.) that contributed to the model's outputs. While LIME creates a linear model from the black-box one to interpret its predictions by perturbing the input of data samples, SHAP explains individual predictions by computing the contribution of each feature to the prediction leveraging the coalition game theory. For example, the authors of [6] and [23] created an interactive dashboard for price prediction movements based on time series and integrated it with LIME explanations on the stock-related news to trigger buy or sell signals. Further, Benhamou et al. [10] used SHAP contributions to explain potential stock market crashes at a given date, while Gradojevic et al. [24] used SHAP to get an insight into option pricing before and during the COVID-19 pandemic. Nevertheless, there is still a lack of studies that examine how these XAI techniques impact users in a real stock trading scenario.

2.2 AI Confidence Estimation

As mentioned in the previous sections, we can catalogue the stock market prediction task as a high-uncertainty domain considering both humans and artificial decision-makers. Consequently, this domain needs to include relevant information like AI confidence accompanied by explanations to explain the model's decisions. We, therefore, illustrate the notion of confidence used in this paper concerning previous work. In recent years, many articles focused on computing how likely a single model prediction would be correct, formally called *confidence* or uncertainty. Previous research [31, 34] categorizes uncertainty into two types: epistemic and aleatoric [31]. *Epistemic* uncertainty refers to the uncertainty generated by a lack of knowledge of the model and can be reduced by adding more data. *Aleatoric* uncertainty refers to the notion of variance and randomness that is intrinsic in any process and cannot be reduced with more data. Considering machine learning (ML) models like Decision Trees and Random Forests, uncertainty estimation can be accomplished using approaches based on relative likelihood [52]. Further, a novel method to estimate local confidence in ML models accounting of both epistemic and aleatoric uncertainty based on nearest neighbors is MACEst (Model Agnostic Confidence Estimator) [27], which provides trustworthy and calibrated [28] confidence estimates. We thus use MACEst for extracting AI confidence estimates from the Random Forest model.

2.3 Human Reasoning Styles

Explanations inform users about the AI's decisions and may elicit cognitive patterns aligned with how users think and reason. Thus, human reasoning styles may act as a bridge to improve XAI explanations and mitigate cognitive biases [45, 57]. Previous literature proposed different explanation styles that can be represented via the theory of Pierce [20], which defines three logical reasoning styles: inductive, abductive, and deductive. Inductive reasoning involves drawing a general conclusion from a set of specific observations. Abductive reasoning begins with an incomplete set of observations and proceeds to the likeliest possible explanation. Deductive reasoning starts with general rules and examines the possibilities to reach a specific, logical conclusion. Only a few studies analysed the impact of presenting explanations using logical reasoning styles on users. Buçinca et al. [13] briefly discussed how inductive and deductive reasoning explanations, which were designed via example-based explanations and general rules from the simulated AI respectively, impacted users in a nutrition-related scenario. Another article that studied explanation styles which

falls into Pierce's theory is from Van Der Waa et al. [56], which compared contrastive example-based and rule-based explanations' effects on users. The example-based ones referred to historical situations similar to the current one and resemble inductive reasoning, while the rule-based ones were rendered via *if... then...* statements and elicit deductive style. Consequently, we investigate the impact of presenting explanations via logical reasoning styles (i.e., inductive, abductive, and deductive) considering metrics like users' reliance [18, 30, 43], task performance [13] and decision agreement with the AI [61].

2.4 Evaluating Explainable AI Systems

The widespread usage of complex AI systems supporting users during decision-making in diverse applications led researchers to find more rigorous ways to evaluate explainable AI systems [17, 41, 46, 62]. We built our user evaluation based on the taxonomies described below. Doshi-Velez and Kim [17] suggested a taxonomy for evaluating XAI systems approaches on interpretability, categorized into i) application-grounded evaluation, which involves domain experts evaluated in actual tasks, ii) human-grounded evaluation, which considers novice users evaluated in simplified tasks, and iii) functionally-grounded evaluation, which requires no user human experiments and a formal definition of interpretability serves as a proxy for explanation quality. Mohseni et a. [46] presented a survey and a framework for a multidisciplinary approach to XAI interfaces focused on design goals for different XAI user groups and the corresponding evaluation measures. In a more recent article, Zhou et al. [62] propose a taxonomy for XAI evaluation methods, further distinguishing two types of metrics: subjective and objective. Subjective metrics consider users' experience on the AI-assisted decision task such as trust and satisfaction, while objective ones involve measuring information like users' task performance or task completion time.

In our study, we went for a human-grounded evaluation or real task [13], where users completed actual decision-making tasks on stock trading assisted by an AI. For the evaluation metrics, we decided to use the reliance [18, 30, 43] subjective measure, which is frequently used in the literature to collect information on users' trust in the ML model. We decided to measure users' reliance in ranking the available elements of the XAI interface (charts with indicators, AI prediction with confidence, and explanation) with different levels of AI confidence. For the objective measures, we measured users' task performance [13], and agreement [61] with AI decisions. For all three measures we are considering different levels of AI confidence, logic-style explanations, and AI correctness on predictions.

3 STUDY DESIGN

We carried out a user study based on a stock market trading task to assess the effect of AI confidence, AI correctness and reasoning style explanations on reliance, task performance, and agreement. We asked participants to give their decision on buying/selling a stock providing them with an instance (stock chart with indicators), the AI prediction, and prediction confidence (i.e., AI uncertainty expressed as confidence percentage), and one among the four explanation styles considered (NO EXPLANATION, INDUCTIVE, ABDUCTIVE, AND DEDUCTIVE).

Below we list the levels for each of our assessed independent variables

- The **explanation type**, has four levels: "no explanation", inductive, abductive, and deductive.
- The **AI confidence**, which has two levels: low and high.
- The **AI correctness**, which has two levels: wrong and correct.

We measured their effect on three dependent variables:

- The users' **reliance** on the different types of information provided to the user, including the stock chart with indicators, the AI information (prediction and confidence), and the explanation, measured as a ranking.
- The **task performance**, which is whether the action (buy/sell) the user decides to assign to the current stock is correct or not (i.e., it brings financial benefit).
- The **agreement** with the AI, which is whether the user confirms the AI prediction with his/her decision.

3.1 Materials

Datasets. For defining the stock market trading tasks, we used daily data about four different stocks available at Yahoo Finance[1], considering the time between May 2017 and August 2022: Cipla Limited (CIPLA.NS), United States Steel Corporation (X), Redington (India) Limited (REDINGTON.NS) and Kohl's Corporation (KSS). We chose these stocks randomly from a pool of about 500 since they were the ones for which our model performed best considering metrics such as accuracy (above 70%) followed by precision, recall and F1 score. Although most of these stocks have a history of 20 years and more, we decided to set the historical data to train the model to five years for the following reasons. First, we decided to train our model for mid-term forecasting, and a five-year range was suitable for developing performant models considering the abovementioned metrics. Second, we believe that users with almost no experience with trading cannot conduct any technical analysis on the historical chart price, and presenting them with 20 years or more of price history could be overwhelming and misleading for the scope of the task. Instead, providing them with short time spans (e.g., one year) may not provide enough context. Lastly, we would like to guide users to think more about the stock in the short-mid term, letting them focus on the technical indicators' meaning and guidance together with the AI suggestion, confidence, and explanation. Similarly to other work on predicting stock market prices using times-series data [9], we performed exponential smoothing ($\alpha = 0.65$) as a good practice to remove random variation in the data and improve the model training process. Afterwards, we computed several well-known technical indicators we will use as features in our models, which we explain and motivate below.

Classification problem. For providing its advice to the user, the AI has to predict the price trend of the considered stock for the next week (7 days ahead), as the difference between the closing price of the next week and the closing price of today. If this value is positive, the stock will increase in price, and the AI should recommend buying the stock in the *buy* task and to not selling the stock in the *sell* task. If the price trend is negative, the stock will decrease in price, and the AI will recommend selling the stock in the *sell* task and to not buying the stock in the *buy* task. Hence, the decision variable will have two values, resulting in a binary classification:

$$Price\_Trend = \begin{cases} Increase & if \quad P_{7D} \geq P_{today} \\ Decrease & if \quad P_{7D} < P_{today} \end{cases}$$

Classification models. To solve this classification problem, we used a Random Forest (RF) model, a popular and performant approach suitable for this type of task [2, 9, 50]. We initially trained one RF model for each of the four stocks, using different technical indicators as features. We considered the Relative Strength Indicator (RSI) [60], the Stochastic Oscillator on K days (STOCH %K) [40], the Advance-Decline Line (ADX) [60], the Moving Average Cross-over Divergence (MACD) [5], the Price Rate of Change (PROC) [1], the On Balance Volume (OBV) [25], the Accumulation

[1]https://finance.yahoo.com/

Distribution Line (ADL), the Momentum (MOM) [42], the Average True Range (ATR) [60], the Daily News Sentiment Index [53], the Ease of Movement (EMV) [32], and the 200-day moving average.

To avoid any look-ahead bias, we split each dataset chronologically by picking the first 85% of the instances for the training set and the remaining 15% for the test set.

We trained a model for each stock using 300 estimators (trees) and using six samples as the minimum number of samples required to split an internal node. Next, we used Recursive Feature Elimination (RFE) as a feature selection technique to improve the classification accuracy, reducing the set of features from 12 to 5, which are the same for each stock. The test set accuracy scores after the feature selection procedure are about 71% for each stock, which is reasonable performance compared to other state-of-the-art approaches [2, 9, 50]. The resulting features used to train the model are:

- **MACD:** triggers technical signals when it crosses above (to buy) or below (to sell) the zero line. The further away from zero, the stronger the signal generated.
- **ATR:** measures the volatility of a stock. A stock experiencing a high level of volatility has a higher ATR, and a low-volatility stock has a lower ATR (computed on 14 days).
- **EMV:** fluctuates around the zero line. Positive EMV indicate positive money flow and buying pressure. Negative EMV indicate selling pressure and negative money flow. The further away from zero, the stronger the signal generated (computed on 14 days).
- **RSI:** values range from 0 to 100. When RSI is above 70, the stock is overbought and may be subject to a decrease in price. Instead, when RSI is below 30, the stock is oversold and may be subject to an increase in price (computed on 14 days).
- **News Sentiment:** the Daily News Sentiment Index [53] is a high frequency measure of economic sentiment based on lexical analysis of economics-related news articles. Higher values indicate more positive sentiment, and lower values indicate more negative sentiment (see article [53] for more details).

Instance selection. After deploying the RF model, we proceeded with selecting the instances to include in the user study. We selected 4 instances for each stock with all the combinations of AI confidence and AI correctness (i.e., low-correct, low-wrong, high-correct and high-wrong), 32 in total. For each participant, we randomly assign to an instance a reasoning style (no explanation, inductive, abductive, deductive), ensuring balance across the experimental conditions. After that, we computed the explanation of the AI prediction as described in Section 3.1.1. For selecting the instances, we proceeded as follows. First, we calculated the AI Confidence values on the modified RF models using the Model Agnostic Confidence Estimator (MACEst) [27] algorithm on each of the four stocks, considering only the epistemic uncertainty [31] and converting it into a confidence score ranging from 0 to 100. Then, we computed the quartiles on the confidence scores for each stock and used the second quartile (Q_2) to establish the threshold for high vs low AI confidence. The second quartile (Q_2) threshold value was about 57%, and the confidence score distributions for each stock were very similar. We assigned an instance to a low AI confidence if its value was $\leq Q_2$ and the others to a high AI confidence. Next, we randomly picked 16 low and 16 high AI confidence instances for each stock. Each set contains 8 instances where the AI makes the correct prediction and 8 where the AI is wrong. The final low confidence values we collected ranged from 12% to 55%, and the high confidence values ranged from 75% to 90%.

3.1.1 Generating the explanations. **Inductive explanations**. We use local example-based explanations retrieved by the k-NN algorithm inside MACEst. The example selection technique (k-NN) for generating the explanations has no

binding with the prediction model. Nevertheless, the technique is widely used in the XAI literature [14, 55, 56, 58]. Given a test instance, we visualize the three nearest neighbours in the training set through a table showing i) the date of the neighbour example, ii) the price of the stock, iii) the values of the indicators and iv) the AI prediction on the price increase/decrease of the neighbour sample (see Fig. 1 D, inductive explanation).

Abductive explanations. We use local explanations based on the SHapley Additive exPlanations (SHAP) framework [44], which provides a set of techniques to generate explanations for individual predictions by computing the contribution of each feature in favour or against the final prediction. Given the stock price values and indicators represented as a table row, we map the weight of each feature in the prediction obtained by Shapley values to the cell background colour. We used red to represent contributions to a price increase outcome and blue for a price decrease. The opacity indicates the strength of the contribution based on Shapley values. The tabular representation of the SHAP explanations was inspired by SHAPTable [16] (see Fig. 1 D, abductive explanation).

Deductive explanations. We used an algorithm called Collection of High Importance Random Path Snippets (CHIRPS) [29], which generates a rule-based local explanation having high precision and coverage, enriched with a contrastive explanation [45]. CHIRPS extracts a rule explaining the prediction outcome in a tabular form as follows: each row includes an indicator tested against a threshold value (higher or lower), contributing the most to the RF classification. The column called "Contrast" shows how much the precision deteriorates if we exclude the indicator considered in the table row from the rule (counterfactual case). The last column "Decision" contains the RF classification result (see Fig. 1 D, deductive explanation).

3.2 Procedure

To verify our hypotheses, we carried out an online user study for the stock market trading task using the Prolific platform[2]. First, participants read a document containing a brief description of the study and filled the informed consent form. Then, the test introduced participants trading tasks, asking them to take their time when completing the tasks, and to imagine owning the stock shares and make profit as the goal of trading session. To encourage the commitment in this goal, we have included real profit for participants by setting a bonus payment for every correct answer. After the introduction, the test included a short tutorial video (2min 30s) describing each part of the XAI interface, including the goal of the buying and selling tasks and the meaning of the technical indicators. The tutorial was available also during the tasks.

Next, each participant completed four trading tasks. This number allowed us to balance the tradeoff between the number of participants and the time required for completing the test. Two tasks were of type *buy*, and two of type *sell* in a randomized order. In a buying task, participants are supposed to have a budget of $100. They have to decide whether or not to invest them into buying stocks of the considered company (see Section 3.1), considering past information on price, the indicators, the AI's advice and the explanation (if provided by the experimental condition). In the *sell* task, participants are supposed to own stocks worth $100 of the considered company. By using the same information provided in a buying task, they have to decide whether to sell the stocks or keep them and wait for a price increase. After each task, we placed an attention question (4 per user) where the answer was explicitly reported in the question text for ensuring the quality of the collected data. Further, we considered only the instances with a minimum of 2 positive attention checks for the user study. Each participant completed a task (either buy or sell) for each of the Explanation Style independent variable in a randomised order. Each task considered a different stock. Furthermore, each of the

[2]https://www.prolific.co

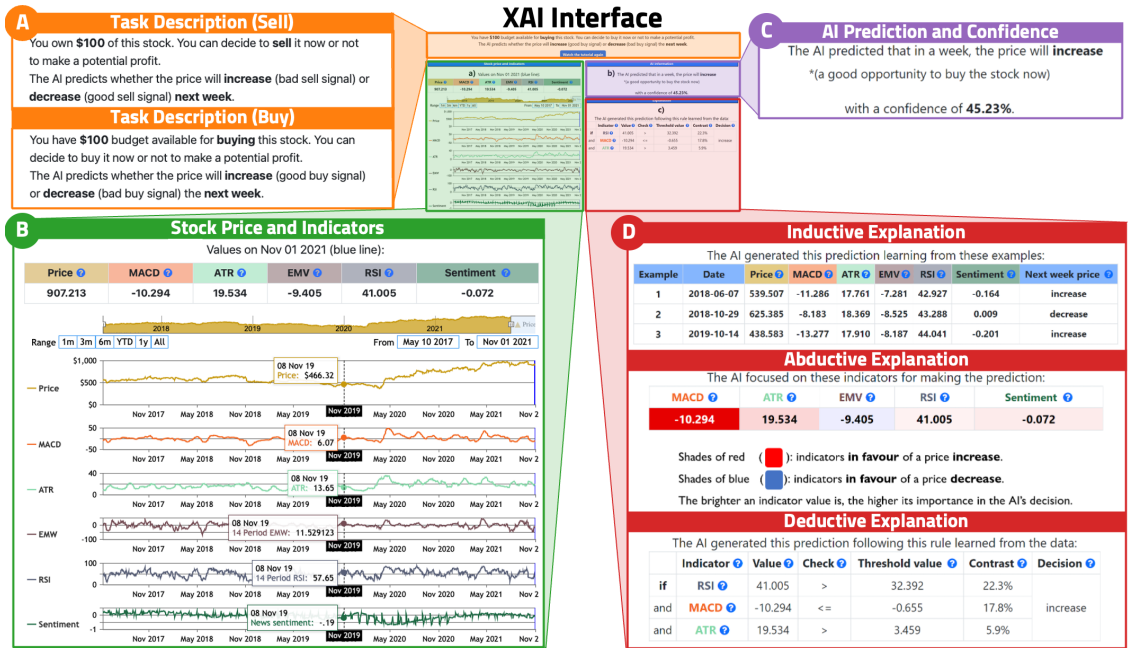

Fig. 1. Interface of the stock trading tasks. (A) Task description of the *buy* and *sell* tasks. (B) Stock price with indicators where users can explore the chart timeline via specific time intervals using the range panel, filter in between two dates, and the slider. Users can further display the technical indicators' meaning by hovering the info buttons. (C) AI prediction, suggestions on buying/selling, and confidence (in this case, a low confidence is shown). (D) Logic-style explanations: INDUCTIVE, ABDUCTIVE, and DEDUCTIVE (the "no explanation" condition is obtained by hiding the explanation box). Users can use the info buttons to obtain more information about the column names for each explanation style.

trading tasks was counter-balanced between participants on the AI CORRECTNESS and AI CONFIDENCE levels. Fig. 1 shows structure of the interface for making the decision, including an example stock price and indicators chart and one explanation per type. The protocol has been formally approved by the Ethics Committee of the University of Cagliari[3].

We recruited about 250 participants through Prolific, collecting 1000 decisions from users. We set this number considering the results of the power analysis, indicating the need for 735 instances (see Section 3.4) and considering that, in previous studies, we received about 30% of tasks having a failed attention check. We paid each participant £5 for completing all the tasks. On average, the four tasks lasted 25 minutes, with a reward per hour of £12, which the platform recognizes as a fair payment for participants. We rewarded participants with £0.5 for every correct classification. Once we discarded the instances having a faulting attention check, we considered 734 instances for the analysis.

The Prolific Platform provides information about Age, sex, level of education and task completion time for each participant. In addition, we collected the following information through specific questions:

- Stock trading experience: we asked participants their experience in trading stocks with the following statement: " Do you have any experience in trading stocks?". The available answers were "No experience", "Little experience", "Good experience", and "Vast experience".

[3]Received on 4 October 2022, Prot. 0213930

- **Reliance:** a ranking of the information included in the XAI interface, namely the instance (stock charts with indicators), AI information (prediction and confidence) and explanation. Participants responded to the statement: "Please rank the following information in terms of how much it helped you in making a decision: a) charts with indicators, b) AI information, c) explanation".
- **Task performance:** whether the participant's final decision is correct or not. The possible values are "correct" when the participant's answer is correct and "wrong" otherwise.
- **Agreement:** whether the participant's final decision agrees with the AI prediction or not. The possible values are "yes" when the decision matches the AI prediction and "no" otherwise.

3.3 Hypotheses

For studying the user's Reliance, we asked the participants to rank the types of information displayed in Figure 1: i) the stock charts and indicators, ii) the AI prediction and confidence, iii) the explanation. This ranking was limited to the experimental conditions including explanations.

The study considers participants having low domain expertise. Previous research shows that non-expert users may show overconfidence in their ability to analyse a problem and make decisions in an AI-assisted context [51, 65], or delegate the decision to the automatic support, without activating analytical cognitive processes [12]. Considering these facts, we expect that overconfident study participants under-rely on the system to trust their own judgement first, and use the stock chart and indicators as their primary information source. In contrast, when the participant over-relies on the AI, its prediction would be expected to be the primary information source. In addition, previous literature suggests that using explanations increases over-reliance [51]. We believe that the AI Confidence is likely to influence whether over- or under-reliance occurs, and thus which information the participants' use in the first instance. Such an influence should occur in particular when the user inspects the explanations. A high AI Confidence results in consistent explanations, which may persuade the user to follow the AI's advice. So the AI prediction and the instance presentation (i.e., charts and indicators) should have a comparable reliance. Instead, low AI confidence values may lead to weak explanations, raising some doubts on the suggested decision. Thus, the participant should rely on his/her ability to evaluate the information about the stock, indicating this part of the interface as the primary source for deciding.

In summary, we formulated Hypothesis 1 as follows:

H1: The user's Reliance on the information provided in XAI interfaces depends on AI Confidence level:

H1a: When the AI Confidence is high, the user will primarily rely on the charts with indicators or the AI prediction, then on the explanation.

H1b: When the AI Confidence is low, the user will primarily rely on the charts with indicators, then on the explanation or the AI prediction.

The reasoning on the reliance also guides our hypothesis on Task Performance. The different levels of AI Confidence should impact how participants use the information provided by the AI and the explanation. A high level of AI Confidence results in explanations that better "argue for" the AI's suggestion. The participant should, via the explanation, get useful insight that enables them to accept or reject the suggestion. Such insights are not available in case of low confidence predictions, which should use provide weaker or contradictory arguments. So, we do not have particular expectations on the explanation effect when the confidence is low, or when explanations are not available.

In addition, we expect that the Explanation Style impacts the user's interpretation. The *inductive* style provides a set of similar examples, but their interpretation requires an effort similar to the instance inspection for the user, who

should analyse the stock price and indicators for previous points in time. Instead, abductive and deductive explanations provide an interpretation of these values, which may be convincing or not for the user. So, we believe that Explanation Style moderates the Task Performance only in case of a high AI confidence. Other studies [61] demonstrated that reporting AI Confidence failed to improve the user's task performance. We deepen this analysis by considering explanations that, in our opinion, are a more informative way to present AI Confidence.

In summary, we formulate the Hypothesis 2 as follows:

H2: Users' Task Performance is moderated by the interaction between AI Confidence and the Explanation Style:

H2a: When the AI Confidence is high, abductive Explanation Style leads to a higher Task Performance if compared against the inductive.

H2b: When the AI Confidence is high, deductive Explanation Styles leads to a higher Task Performance if compared against the inductive.

We expect that users' agreement may depend on the interaction between AI confidence, the explanation style, and also AI correctness. Specifically, we believe that users' agreement may increase in the presence of abductive and deductive explanation styles with high AI confidence and AI correct predictions. The reasoning is similar to the one we described for H2. Abductive and Deductive styles are more suited to convey relevant arguments for understanding the AI prediction, particularly when the AI confidence is high, which results in consistent explanations. So, if we suppose that the user performs better (H2a and H2b) in such a case, this means that s/he is more likely to agree with the AI when it is correct. We think this is actually the only configuration where non-experts have relevant and sufficient information for recognising AI predictions as correct.

We formulated Hypothesis 3 as follows:

H3: Users' Agreement is moderated by the interaction between AI correctness, AI Confidence and Explanation Style:

H3a: When then AI Confidence is high, abductive Explanation Style leads to a higher Agreement if the AI Correctness is correct.

H3b: When then AI Confidence is high, deductive Explanation Style leads to a higher Agreement if the AI Correctness is correct.

3.4 Analytical Approaches

For H1 (reliance), we assess the results with the Friedman test [21, 22], analyzing AI confidence values (low and high) separately to find significant differences in the factors' distributions. We conduct the Nemenyi posthoc analysis when we discover significant factors in the Friedman test. To assess the number of participants required to validate this hypothesis, we carried out a power analysis using G*Power3 [19]. We set the analysis for medium effects (effect size with Cohen's d=0.16), an alpha of 0.05 and power of 0.80 for hypothesis 1. We used the Friedman test and a within-subjects design, using two levels of AI confidence (low and high) on the three ranked measurements (charts with indicators, AI prediction and confidence, and explanation). The results showed that we needed a sample size of 56 people to catch medium effects.

For H2 (task performance) and H3 (agreement), we used logistic regression. For H2, the model includes these factors: AI correctness (wrong, correct) and the interaction between the explanation (noexp, inductive, abductive, deductive) and the AI confidence (low, high). For H3, we consider the interaction between the explanation (noexp, inductive, abductive, deductive), the AI confidence (low, high) and the AI correctness (wrong, correct) as factors. The baselines for

the logistic regression factors are: "noexp" for the explanation, "low" for the AI confidence, and "wrong" for the AI correctness. For both H2 and H3, the results showed that we needed a sample size of 735 instances for medium effects (A priori χ^2 test with effect size d=0.16, alpha= 0.05, power=0.80, Df=15). Since each user sees four different instances (one for each Explanation Style), we divide the sample size of 735 by four, thus obtaining 184 participants needed for H2 and H3, considering that for H1 56 people are sufficient.

4 RESULTS

4.1 Participants

The 184 participants that successfully passed the attention checks consists of 94 females and 90 males, aged between 19 and 62 years old ($\bar{x}$ = 28.1, $\tilde{x}$ = 25, s = 8.4). We have ensured that participants had a good level of English to understand the meaning of technical indicators through the pre-screening supported by the Prolific platform. The results concerning the stock trading experience show that 52.2% of users (96) had no experience in trading stocks, and the remaining 47.8% (88) had little experience. Consequently, no expert users participated in the stock trading tasks.

4.2 H1: Reliance

For making a decision, the user relies on the information provided by the XAI interface, including the stock chart with indications, the AI prediction, and the explanations. Studying the relative importance of the different information types in making the decision (i.e., the Reliance) is relevant for establishing the causes of opposite phenomena like overconfidence [51, 65] and overreliance [12]. In H1, we suppose that the AI Confidence impacts the process of establishing such importance.

To ensure a fair comparison, we excluded participants assigned to the no explanation condition, resulting in 139 users. The Friedman test for the Reliance shows a significant difference between three information types when the AI confidence is high (H1a, $\chi^2(2) = 65.13$, df=2, $p < .05$). The same happens when the AI confidence is low (H1b, $\chi^2(2) = 82.41$, df = 2, $p < .05$).

The pairwise comparisons using Nemenyi post-hoc test for mean rank considering a high AI Confidence (H1a) highlights no significant differences between the stock chart with indicators and the AI information on rank 1. The bottom-left side of Figure 2 shows a significant difference between the explanation compared to AI information and the stock chart with indicators, placing the explanation at rank 2. Hence, we *reject the null hypothesis* for H1a and for high AI confidence and we conclude that users interchangeably rely on the stock chart with indicators or AI information as a primary source of information (rank 1), only then followed by the explanation. In contrast, the pairwise comparisons considering a low AI confidence (H1b) show that users rely the most on the stock chart with indicators (rank 1), followed by the AI prediction (rank 2) and the explanation (rank 3). So, we *reject the null hypothesis* for H1b, concluding that users primarily rely on the stock chart with indicators, and only then on the AI information followed by the explanation (see top-left side of Figure 2). *In summary, for high AI confidence users rely more on the AI prediction and charts equally, and less on the explanation. For low AI confidence users rely more on the charts followed by the AI prediction and lastly, the explanation.*

Different studies highlighted a significant impact of a correct AI prediction on the user's decision [13, 35]. So, as additional analysis, we further investigated how AI Correctness impacts users' Reliance in ranking of the interface information. The Friedman test highlights a significant difference between factors considering correct ($\chi^2(2) = 83.34$, df = 2, $p < .05$) and wrong predictions ($\chi^2(2) = 45.13$, df=2, $p < .05$). We proceeded with a Nemenyi post-hoc test, which

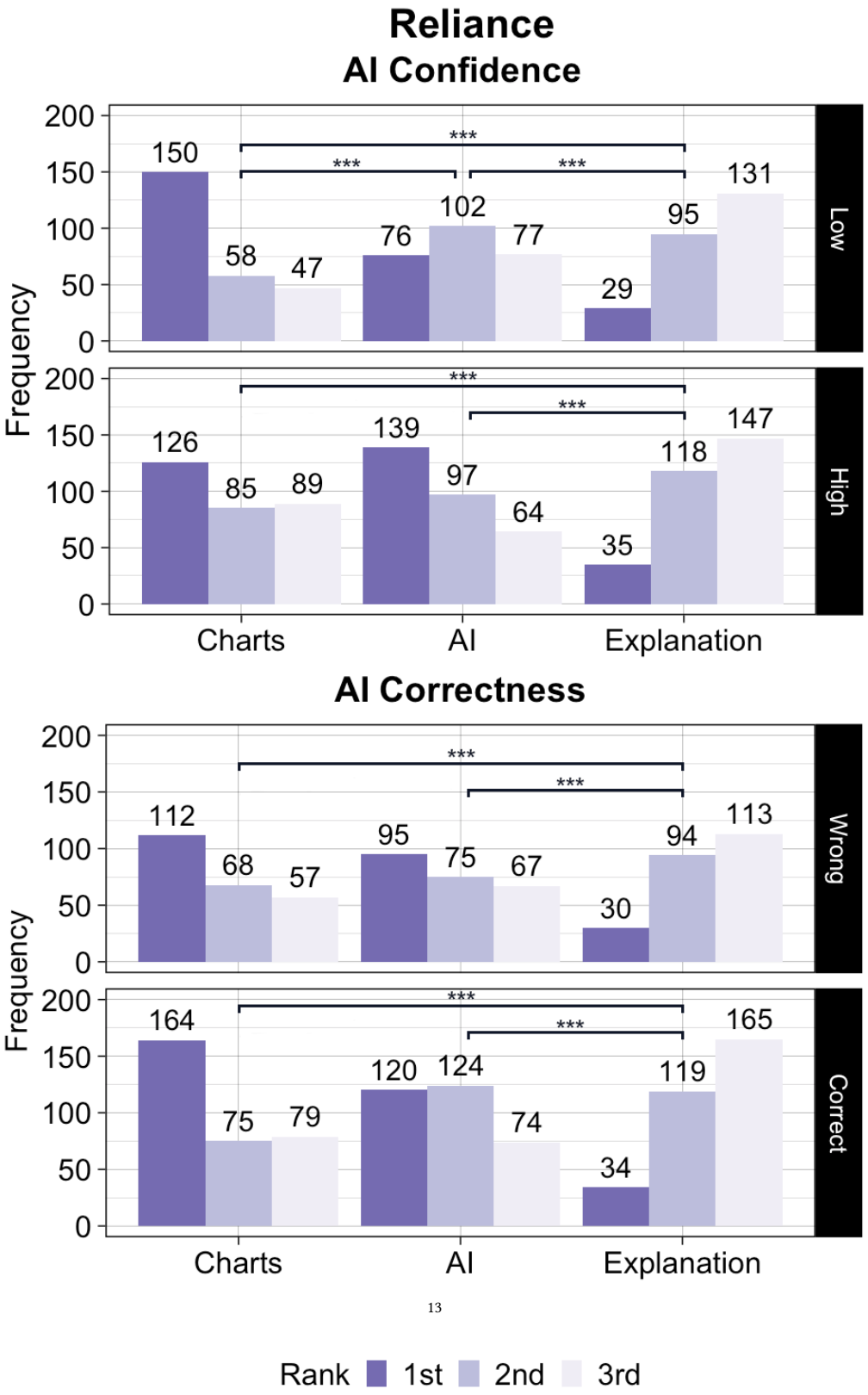

Fig. 2. Rank frequencies for users' reliance split by AI confidence (top) and AI correctness (bottom). Each line indicates whether exists a significant difference between a pair of levels, and the asterisks highlight the degree of the significance based on p-value (*$p < .05$; **$p < .01$; ***$p < .001$).

highlighted in both conditions a significantly higher ranking for charts compared to explanation, and a higher ranking for the AI prediction and compared to the explanation. We do not register any significant difference between the stock charts and the AI prediction. Considering such results, the levels of AI Correctness set the same Reliance ranking: the primary information types are the stock charts and the AI prediction, while the explanation is secondary.

Table 1. Logistic regression results on Task Performance (H2).

Predictor	Log-Odds	Std. error	z-value	p
AI correctness [correct]	-0.263	0.152	-1.728	.083
Explanation style [inductive]	-0.163	0.314	-0.521	.602
Explanation style [abductive]	-0.572	0.313	-1.824	.068
Explanation style [deductive]	-0.463	0.310	-1.494	.135
AI confidence [high]	*-0.625	0.306	-2.044	.041
Explanation style [inductive] * AI confidence [high]	0.370	0.428	0.865	.387
Explanation style [abductive] * AI confidence [high]	**1.168	0.427	2.735	.006
Explanation style [deductive] * AI confidence [high]	*0.980	0.425	2.305	.021

*$p < .05$; **$p < .01$; ***$p < .001$

4.3 H2: Task Performance

Which information users rely on should ultimately increase their ability to make correct decisions. Unfortunately, the literature to date suggests that users perform worse when supported by AI compared to the users or AI working alone [7, 12, 26]. The information the system supplies can potentially be misleading. Therefore, it is relevant to assess how the user performs for different levels of AI confidence and when they are exposed to different explanation styles.

Recall that in H2, we suppose an effect of the interaction between the AI confidence and Explanation Style on task performance. We report the results of the logistic regression for users' Task Performance in Table 1, considering the interaction between AI Confidence and the Explanation Style.

We found a significant interaction between the AI confidence and the Explanation style: when the AI confidence is high, abductive and deductive Explanation styles positively affect Task Performance while we do not register such an effect on the inductive style (or for low AI confidence). We would expect a "good" explanation style to increase task performance when the confidence is high. In the case of high AI confidence, we see a task performance of 43.0% for the inductive, 52.5% for the abductive, and 50.5% for the deductive explanation styles, respectively. Hence, we *reject the null hypothesis* for H2a and H2b since abductive and deductive Explanation styles resulted in a higher task performance compared with the inductive style when AI confidence is high (see Figure 3). In case of low confidence, the best option is avoiding to show any explanation (52.4% for the noexp style in low AI confidence). The task performance is also generally low, as expected considering the low expertise of the participants and the balancing of the experimental conditions.

As we did for H1, we investigated whether AI correctness may have an impact on users' task performance, but did not register any significant difference (see Table 1).

4.4 H3: Agreement

In establishing the conditions potentially leading to overreliance, it is relevant to study which factors lead to an agreement between the final user's decision and the AI prediction. Specifically, for assessing H3, we inspect whether

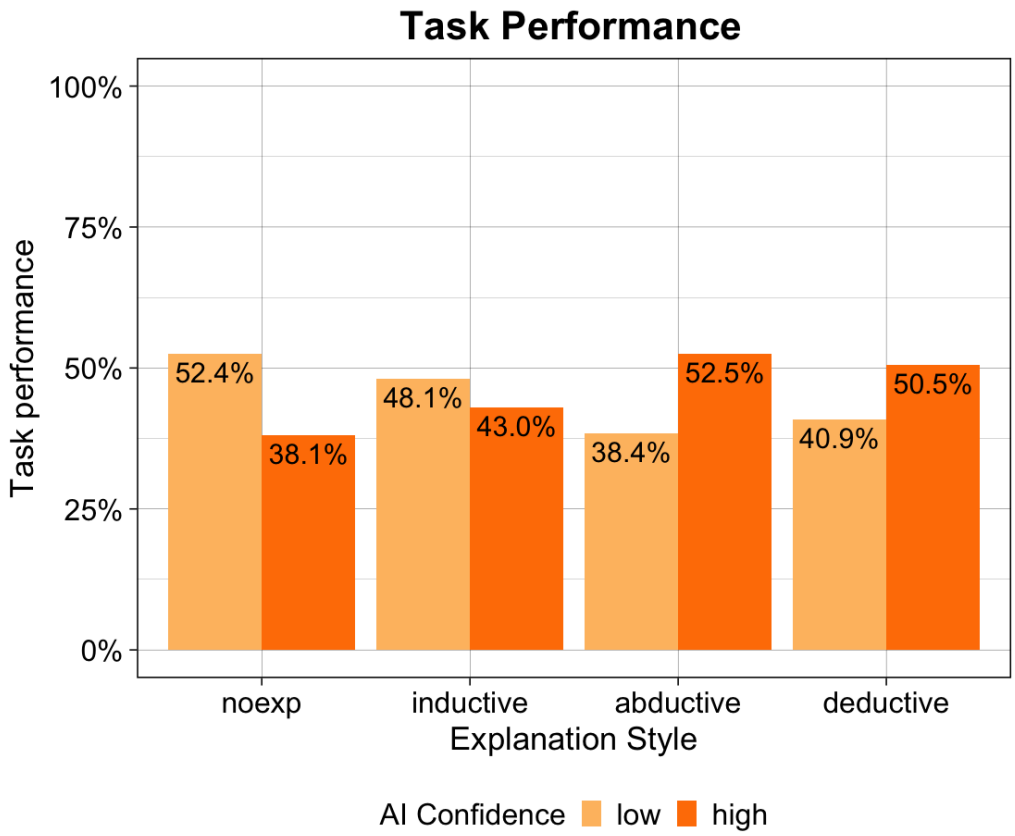

Fig. 3. Task Performance results considering explanation styles and different levels of ai confidence.

Table 2. Logistic regression results on Agreement (H3).

Predictor	Log-Odds	Std. error	z-value	p
AI correctness [correct]	0.392	0.470	0.835	.403
Explanation style [inductive]	0.472	0.540	0.874	.381
Explanation style [abductive]	0.611	0.522	1.170	.241
Explanation style [deductive]	0.741	0.523	1.416	0.157
AI confidence [high]	*1.067	0.476	2.239	.024
AI correctness [correct] * Explanation style [inductive]	-0.466	0.666	-0.701	.483
AI correctness [correct] * Explanation style [abductive]	-1.177	0.657	-1.790	.073
AI correctness [correct] * Explanation style [deductive]	-1.051	0.652	-1.612	.107
AI correctness [correct] * AI confidence [high]	*-1.355	0.629	-2.154	.031
Explanation style [inductive] * AI confidence [high]	-0.876	0.674	-1.299	.193
Explanation style [abductive] * AI confidence [high]	*-1.503	0.664	-2.265	.023
Explanation style [deductive] * AI confidence [high]	-1.094	0.664	-1.648	.099
AI correctness [correct] * Explanation style [inductive] * AI confidence [high]	0.855	0.885	0.966	.333
AI correctness [correct] * Explanation style [abductive] * AI confidence [high]	**2.358	0.877	2.688	.007
AI correctness [correct] * Explanation style [deductive] * AI confidence [high]	*2.0436	0.873	2.339	.019

*$p < .05$; **$p < .01$; ***$p < .001$

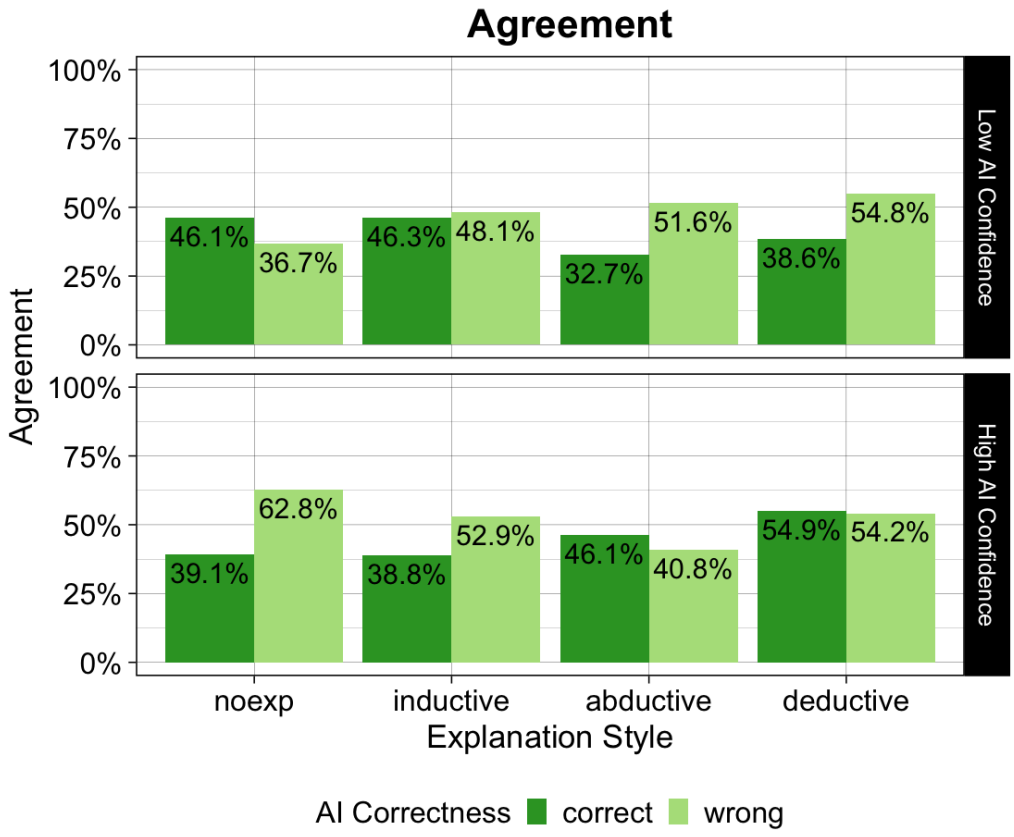

Fig. 4. Agreement results considering explanation styles and different levels of ai confidence and AI correctness.

users' Agreement is affected by high AI confidence coupled with abductive and deductive Explanation Styles considering AI correct predictions. We report the results of the logistic regression for users' Agreement in Table 2 considering the interaction between AI Correctness, AI Confidence and the Explanation Style. We found significant interactions among abductive and deductive explanations styles, high AI confidence and AI correct predictions, so we *reject the null hypothesis* for H3a and H3b (see Figure 4). In particular, we registered a positive effect (more agreement) for abductive and deductive explanations when the AI confidence is high and its prediction is correct, as expected in H3a and H3b. The agreement increases for abductive explanations from 32.7% registered for a correct and low-confident AI to 46.1% when it is correct and high-confident. The deductive style explanations in contrast had a higher level of appropriate agreement (when AI is correct); with an agreement of 38.6% for low confidence, and 54.9% for high confidence.

5 DISCUSSION

For discussing the implications of the findings we presented in this paper, it is worth summarising the differences between the expected effects and the actual results in our study. Table 3 shows the list of hypotheses, the results of their verification in the study data and additional insights highlighted by the data analysis. Overall, the results met

Table 3. Hypothesis summary

	Hypotheses	Notes
	H1: *Reliance*	
✓	**H1a:** When the AI Confidence is high, the user will primarily rely on the charts with indicators or the AI prediction, then on the explanation.	Same as expected.
✗	**H1b:** When the AI Confidence is low, the user will primarily rely on the charts with indicators, then on the explanation or the AI prediction.	User rely on 1) charts, 2) AI prediction, 3) explanations
	H2: *Task Performance*	
✓	**H2a:** When the AI Confidence is high, abductive Explanation Style leads to a higher Task Performance if compared against the inductive.	No positive effect for noexp and inductive explanations
✓	**H2b:** When the AI Confidence is high, deductive Explanation Styles leads to a higher Task Performance if compared against the inductive.	
	H3: *Agreement*	
✓	**H3a:** When then AI Confidence is high, abductive Explanation Style leads to a higher Agreement if the AI Correctness is correct.	Same as expected.
✓	**H3b**: When then AI Confidence is high, deductive Explanation Style leads to a higher Agreement if the AI Correctness is correct.	Same as expected.

our expectations, implying some advances in our knowledge about the decision process we discuss in Section 5.1. Our results also have limitations that we acknowledge in Section 5.2.

5.1 Implications

The study results identify implications useful for creating AI-powered decision supports and XAI interfaces. All of them should be related to the task, which is difficult for humans and AI, and to the user type since we considered people with low domain expertise.

We should use explanations when the AI confidence is high. There is converging evidence in our results on the combined impact of the AI Confidence and the Explanation Style when the AI confidence is high. The explanations communicate such confidence by providing consistent arguments supporting the AI prediction, independently of its correctness. Even though we do not consider expert users, our participants made good use of such information, increasing the number of correct decisions. Besides the results on the Task Performance, our initial idea was confirmed by the Reliance results, where a difference in the confidence level resulted in different rankings between the types of information in the XAI interface. When the confidence is low, the ranking is 1) stock charts and indicators, 2) AI prediction, and 3) explanation. The high confidence "overshadows" significant differences between the charts and the AI prediction, making them equally important for the final decision. In addition, the Agreement increases in case of high confidence and correct AI prediction, which aligns with the increased Task Performance: if the AI is correct and the user agrees, the final decision would be correct.

When the AI confidence is low, our results suggest that it would be better not to explain the AI's prediction: the condition without explanation registered the highest performance for a low AI confidence. Such conclusion is supported by data depicted in Figure 3, showing that users perform better overall without an explanation and inductive explanations in the AI low confidence condition and, most importantly, in the results reported in the logistic regression in Table 1 considering the significant positive impact of high AI confidence coupled with abductive and deductive

explanations. Such configuration (deductive or abductive explanations with high AI confidence) also led users to over-rely on AI predictions when the AI confidence was *low*, thus *lowering the performance* (see Fig. 3). This evidence of overreliance is also confirmed in the agreement hypothesis (H3) since using abductive and deductive explanations resulted in a higher agreement with wrong AI predictions when the AI confidence is low.

So, XAI interfaces may use AI CONFIDENCE as a criterion for selecting whether or not to show the explanations. While other relevant factors are unknown in the general case (such as the AI CORRECTNESS), a model needs only the current instance to classify for evaluating its confidence.

We should carefully select the explanations reasoning style. We registered all the interesting effects we discussed in the previous implication considering the abductive and the deductive EXPLANATION STYLES. This highlights the relevance of an overlooked aspect of XAI. To be understood by the user, explanations must trigger effective reasoning processes, which we should select considering the current task and, most importantly, the data types describing the instance. In our study, we used tabular time-series data. Unlike image or text classification tasks, which usually require low effort for users, the stock prediction requests cognitive effort for comparing indicators and finding trends. Triggering an inductive inference process for explaining the AI prediction is not optimal for this task because using such information would multiply the user's effort, who will ignore the explanation. Instead, deductive and abductive explanations provide a key for reading the relevant part of the instance description that leads to the AI prediction. This resulted in a higher understanding of the AI's "arguments" and a more effective acceptance or rejection of the AI's suggestion.

We believe that this effect depends on both task and data type. We would expect that for decision tasks that are easier for humans, such as image classification, the inductive style would be more effective than in stock trading. In this case, the inspection of an example set requires a low effort for the user, and establishing a visual similarity between the image to classify and the examples identified by the AI could be a more effective way of establishing trust in the AI prediction (or not).

AI correctness does not change the user's performance. Providing correct suggestions does not make a significant difference in the correctness of the final user's decision in our experiment. The high uncertainty of the stock market prediction task and the lack of domain expertise of the study participants make them equally likely to accept or reject both correct and wrong AI predictions. Therefore, for such a high uncertainty task, the AI correctness does not explain the over or the under-reliance registered in the literature motivating our work [12, 51, 65]. Instead, relevant factors for correctly considering the AI suggestion in our setting are the explanation style and the AI confidence. Overall, our results suggest that for guiding non-experts through AI support, it may be more relevant to be able to estimate and communicate *confidence* in predictions through specific *explanation styles* (abductive and deductive).

5.2 Limitations

This section discusses some limitations in our work, which may lead to further research.

One limitation to the generalization of the results concerns the selection of representative elements in our study among the many available options. This includes the selection of the stock trading domain, the selected stocks and time frame, the definition of the buying and selling scenarios, the classification model, and the technical indicators selected for the evaluation. For each option, we selected options which balanced the study's relevance and feasibility. On the one hand, we tried to replicate a realistic stock trading scenario, but we also tried to minimize the interface burden for users with no experience in stock trading. Additionally, we attempted to mitigate the lack of responsibility for trading using "fake" money by introducing a bonus reward for correct decisions, motivating participants to put real effort

into the task. Also, we believe that the choice of technical indicators was a critical component of the task. Although some fundamental indicators like RSI, MACD, and news sentiment were present in the XAI interface, other essential indicators like stock volume or moving average rates would probably have guided users into different interpretations of the price movement.

Another limitation regards the generation of logical reasoning explanations. Although we used well-known state-of-the-art methods frequently used in other evaluations, we acknowledge that different XAI techniques using an equivalent reasoning style and could lead to different results. For example, we rendered the deductive explanation style using a rule-extraction method, which generates only one set of rules. We acknowledge that many rule-extraction techniques exist from RF models, which may extract more than one set of rules and prioritizes other metrics compared to CHIRPS, possibly leading to different outcomes.

The last limitation concerns the methods used for estimating and splitting AI confidence into low and high. We used the MACEst algorithm since estimates calibrated confidence values, and it was an appropriate method for our Random Forest models. In addition, we split AI confidence into low and high levels using the second quartile (Q_2) as a threshold. We employed this approach since each stock had similar confidence distributions and the accuracy of the RF models was very close to each other. Further studies are needed to find more generalizable approaches.

6 CONCLUSION

This paper contributes to advancing the knowledge in the AI support to user decisions by investigating the effects of AI confidence and the explanation reasoning styles on 1) the reliance on the information types included in an XAI interface, 2) the task performance (i.e., making the correct decision) and 3) the agreement between the AI suggestion and the final user's decision. We focus on the stock market domain for studying high-uncertainty tasks for both human beings and AI. We conducted a user study including 184 participants making selling and buying decisions on four stocks. The results show that users primarily rely on charts and AI predictions equally when AI confidence is high, while low confidence values lead users to rely the most on charts.Abductive and deductive explanation styles positively impact users' task performance when the AI is confident and contribute to a higher agreement when the AI is correct and with high confidence.

In future work, we aim to investigate open questions not covered by the results of this study. First, we will investigate the effectiveness of the different explanation styles in different domains and their relationship with the data type presenting the classification instance. In addition, we will try to understand if other XAI techniques (including counterfactual reasoning) leveraging the same reasoning style have similar effects. Finally, we want to investigate the relationship between domain expertise, AI confidence, and correctness in task performance.

ACKNOWLEDGMENTS

Federico Maria Cau gratefully acknowledges the "CRS4 - Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna" for the PhD funding and the collaboration on the RIALE (Remote Intelligent Access to Lab Experiment) Platform. The work has been partially supported by the Sardinia Regional Government and by Fondazione di Sardegna, ASTRID project (FdS 2020) - CUP F75F21001220007.

REFERENCES

[1] 2012. *Price Rate of Change.* John Wiley & Sons, Ltd, Chapter 5, 51–60. https://doi.org/10.1002/9781119204428.ch5 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119204428.ch5

[2] Rebecca Abraham, Mahmoud El Samad, Amer M. Bakhach, Hani El-Chaarani, Ahmad Sardouk, Sam El Nemar, and Dalia Jaber. 2022. Forecasting a Stock Trend Using Genetic Algorithm and Random Forest. *Journal of Risk and Financial Management* 15, 5 (2022). https://doi.org/10.3390/jrfm15050188

[3] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[4] Md Manjurul Ahsan and Zahed Siddique. 2021. Machine learning based disease diagnosis: A comprehensive review.

[5] Gerald Appel. 2005. *Technical Analysis: Power Tools for Active Investors.*

[6] Harit Bandi, Suyash Joshi, Siddhant Bhagat, and Dayanand Ambawade. 2021. Integrated Technical and Sentiment Analysis Tool for Market Index Movement Prediction, comprehensible using XAI. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*. 1–8. https://doi.org/10.1109/ICCICT50803.2021.9510124

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

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Dan Weld, Walter Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *AAAI Conference on Artificial Intelligence*. AAAI. https://www.microsoft.com/en-us/research/publication/updates-in-human-ai-teams-understanding-and-addressing-the-performance-compatibility-tradeoff/

[9] Suryoday Basak, Saibal Kar, Snehanshu Saha, Luckyson Khaidem, and Sudeepa Roy Dey. 2019. Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance* 47 (2019), 552–567. https://doi.org/10.1016/j.najef.2018.06.013

[10] Eric Benhamou, Jean-Jacques Ohana, David Saltiel, and Beatrice Guez. 2021. Explainable AI (XAI) Models Applied to Planning in Financial Markets. *SSRN Electronic Journal* (2021). https://doi.org/10.2139/ssrn.3862437

[11] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 807–819. https://doi.org/10.1145/3490099.3511139

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[13] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Mar 2020). https://doi.org/10.1145/3377325.3377498

[14] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-Based Explanations in a Machine Learning Interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 258–262. https://doi.org/10.1145/3301275.3302289

[15] Federico Maria Cau, L. D. Spano, and N. Tintarev. 2020. Considerations for Applying Logical Reasoning to Explain Neural Network Outputs. In *XAI.it@AI*IA*.

[16] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[17] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. https://doi.org/10.48550/ARXIV.1702.08608

[18] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7 Trust and Technology.

[19] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2013. G*Power 3.1.7: A flexible statistical power analysis program for the social, Behavioral and Biomedical sciences, Beh. *Res. Meth.s* 39 (01 2013), 175–191.

[20] Peter Flach and Antonis Kakas. 2000. Abductive and Inductive Reasoning: Background and Issues. (01 2000). https://doi.org/10.1007/978-94-017-0606-3-1

[21] Milton Friedman. 1937. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Amer. Statist. Assoc.* 32, 200 (1937), 675–701. https://doi.org/10.1080/01621459.1937.10503522 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/01621459.1937.10503522

[22] Milton Friedman. 1940. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *Annals of Mathematical Statistics* 11 (1940), 86–92.

[23] Shilpa Gite, Hrituja Khatavkar, Ketan Kotecha, Shilpi Srivastava, Priyam Maheshwari, and Neerav Pandey. 2021. Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Computer Science* 7 (01 2021), e340. https://doi.org/10.7717/peerj-cs.340

[24] Nikola Gradojevic and Dragan Kukolj. 2022. Unlocking the black box: Non-parametric option pricing before and during COVID-19. *Annals of Operations Research* (25 Feb 2022). https://doi.org/10.1007/s10479-022-04578-7

[25] J.E. Granville. 1976. *Granville's New Strategy of Daily Stock Market Timing for Maximum Profit.* Simon & Schuster.

[26] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 90–99. https://doi.org/10.1145/3287560.3287563

[27] Rhys Green, Matthew Rowe, and Alberto Polleri. 2021. MACEst: The reliable and trustworthy Model Agnostic Confidence Estimator. https://doi.org/10.48550/ARXIV.2109.01531

[28] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. *CoRR* abs/1706.04599 (2017). arXiv:1706.04599 http://arxiv.org/abs/1706.04599

[29] Julian Hatwell, Mohamed Medhat Gaber, and R. Muhammad Atif Azad. 2020. CHIRPS: Explaining random forest classification. *Artificial Intelligence Review* (2020), 1 – 42.

[30] Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. Trust in Automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88. https://doi.org/10.1109/MIS.2013.24

[31] Stephen C. Hora. 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. *Reliability Engineering & System Safety* 54, 2 (1996), 217–223. https://doi.org/10.1016/S0951-8320(96)00077-4 Treatment of Aleatory and Epistemic Uncertainty.

[32] Richard W. Arms Jr. 1990. *Ease of movement.* V.8:5 (187–190) pages.

[33] Taylan Kabbani and Fatih Usta. 2022. Predicting The Stock Trend Using News Sentiment Analysis and Technical Indicators in Spark. *ArXiv* abs/2201.12283 (2022).

[34] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf

[35] Eoin M. Kenny, Courtney Ford, Molly Quinn, and Mark T. Keane. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence* 294 (2021), 103459. https://doi.org/10.1016/j.artint.2021.103459

[36] Luckyson Khaidem, Snehanshu Saha, and Sudeepa Roy Dey. 2016. Predicting the direction of stock market prices using random forest. *CoRR* abs/1605.00003 (2016). arXiv:1605.00003 http://arxiv.org/abs/1605.00003

[37] Wasiat Khan, Mustansar Ali Ghazanfar, Muhammad Awais Azam, Amin Karami, Khaled Hamed Alyoubi, and Ahmed S. Alfakeeh. 2020. Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing* (2020), 1–24.

[38] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1885–1894. https://proceedings.mlr.press/v70/koh17a.html

[39] 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

[40] G. Lane. 1984. *Lane's stochastics.* Second issue of Technical Analysis of Stocks and Commodities magazine. pp 87–90 pages.

[41] Pedro Lopes, Eduardo Silva, Cristiana Braga, Tiago Oliveira, and Luís Rosado. 2022. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Applied Sciences* (2022).

[42] Rand Kwong Yew Low and Enoch Tan. 2016. The role of analyst forecasts in the momentum effect. *International Review of Financial Analysis* 48 (2016), 67–84. https://doi.org/10.1016/j.irfa.2016.09.007

[43] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[44] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777. https://doi.org/10.5555/3295222.3295230

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

[46] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2018. A Survey of Evaluation Methods and Measures for Interpretable Machine Learning. *CoRR* abs/1811.11839 (2018). arXiv:1811.11839 http://arxiv.org/abs/1811.11839

[47] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. 2019. Meaningful Explanations of Black Box AI Decision Systems. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 9780–9784. https://doi.org/10.1609/aaai.v33i01.33019780

[48] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 535, 14 pages. https://doi.org/10.1145/3491102.3501967

[49] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 97–101. https://doi.org/10.18653/v1/N16-3020

[50] Perry Sadorsky. 2021. A Random Forests Approach to Predicting Clean Energy Stock Prices. *Journal of Risk and Financial Management* 14, 2 (2021). https://doi.org/10.3390/jrfm14020048

[51] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[52] Mohammad Hossein Shaker and Eyke Hüllermeier. 2020. Aleatoric and Epistemic Uncertainty with Random Forests. *CoRR* abs/2001.00893 (2020). arXiv:2001.00893 http://arxiv.org/abs/2001.00893

[53] Adam Shapiro, Moritz Sudhof, and Daniel Wilson. 2017. Measuring News Sentiment. *Federal Reserve Bank of San Francisco, Working Paper Series* (01 2017), 01–A2. https://doi.org/10.24148/wp2017-01

[54] Si Shi, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau. 2022. Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications* 34 (09 2022). https://doi.org/10.1007/s00521-022-07472-2

[55] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. 2022. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human–Computer Interaction* 0, 0 (2022), 1–15. https://doi.org/10.1080/10447318.2022.2101698 arXiv:https://doi.org/10.1080/10447318.2022.2101698

[56] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404. https://doi.org/10.1016/j.artint.2020.103404

[57] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[58] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[59] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do You Trust Me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (Paris, France) *(IVA '19)*. Association for Computing Machinery, New York, NY, USA, 7–9. https://doi.org/10.1145/3308532.3329441

[60] J. Welles. Wilder. 1978. *New concepts in technical trading systems.* Trend Research.

[61] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[62] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* 10, 5 (2021). https://doi.org/10.3390/electronics10050593

[63] Jianlong Zhou, Huaiwen Hu, Zhidong Li, Kun Yu, and Fang Chen. 2019. Physiological Indicators for User Trust in Machine Learning with Influence Enhanced Fact-Checking. In *Machine Learning and Knowledge Extraction: Third IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2019, Canterbury, UK, August 26–29, 2019, Proceedings* (Canterbury, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 94–113. https://doi.org/10.1007/978-3-030-29726-8_7

[64] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. 2019. Effects of Influence on User Trust in Predictive Decision Making. 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/10.1145/3290607.3312962

[65] H.-J. Zimmermann. 2000. An application-oriented view of modeling uncertainty. *European Journal of Operational Research* 122, 2 (2000), 190–198. https://doi.org/10.1016/S0377-2217(99)00228-3