June 2025

Improving AI-Assisted Decision-Making: Insights into Example-Based Explanations and Cognitive Load in Sales Forecasting

Niklas Weller
University of St. Gallen, niklas.weller@unisg.ch

Anuschka Schmitt
The London School of Economics and Political Science, a.schmitt2@lse.ac.uk

Tobias Benjamin Fahse
University of St.Gallen, tobias.fahse@mail.ch

Ivo Blohm
University of St. Gallen, ivo.blohm@unisg.ch

IMPROVING AI-ASSISTED DECISION-MAKING: INSIGHTS INTO EXAMPLE-BASED EXPLANATIONS AND COGNITIVE LOAD IN SALES FORECASTING

Completed Research Paper

Niklas Weller, University of St. Gallen, Switzerland, niklas.weller@unisg.ch

Anuschka Schmitt, The London School of Economics and Political Science, United Kingdom, a.schmitt2@lse.ac.uk

Tobias Fahse, University of St. Gallen, Switzerland, tobias.fahse@unisg.ch

Ivo Blohm, University of St. Gallen, Switzerland, ivo.blohm@unisg.ch

Abstract

AI-assisted decision-making often underperforms due to users' difficulties in effectively interacting with AI-based systems. This study investigates how example-based explanations—specifically factual and counterfactual explanations—impact users' decision-making performance and their tendency to overrule algorithmic advice in a sales forecasting task. We also examine the mediating role of cognitive load. By analyzing 1330 forecasts made in an online lab experiment, we find that factual explanations significantly enhance forecasting performance by enabling users to more effectively overrule algorithmic advice. While counterfactual explanations also result in performance gains, the increase is smaller and operates primarily through reduced deviation from AI advice due to cognitive overload. Our findings suggest that factual explanations align well with human cognitive processes, facilitating better decision outcomes, while counterfactuals may overwhelm users cognitively. This study contributes to a deeper understanding of explainable AI design in decision-making contexts, emphasizing the importance of aligning explanations with users' cognitive capacities.

Keywords: AI-Assisted Decision-Making, Explanations, Sales Forecasting, Cognitive Load

1 Introduction

Machine learning (ML) systems are leveraged to augment and complement humans in decision-making tasks. However, improved decision-making through this type of artificial intelligence (AI) often fails (Vaccaro et al., 2024). Research has turned towards potential design mechanisms that can help overcome interaction challenges in AI-assisted decision-making, including explanations.

Explanations are a well-established design concept to equip users' understanding of a task and domain, as well as the system that is assisting them (Gregor & Benbasat, 1999). Explanations have become particularly promising in the context of ML-based systems, commonly referred to as explainable AI or XAI. As ML-based systems and their underlying mechanism are marked by high complexity and inscrutability, XAI methods can make them transparent and understandable—with the ultimate goal of improving decision-making quality.

Findings regarding XAI's effectiveness in AI-assisted decision-making are inconclusive, though. Converging empirical evidence suggests that decision-making does not necessarily improve through XAI. This is because humans only selectively engage with explanations, e.g., when the explanation confirms humans' preconceptions (e.g., Abdel-Karim et al., 2023; Bauer et al., 2023) and because humans are easily induced to follow algorithmic advice as they view explanations as cue for an algorithm's trustworthiness without properly engaging with the information such explanations provide (Lakkaraju & Bastani, 2020; Schoeffer et al., 2024).

Explanations are not equal in their nature. Work on XAI has put forward that explanations can provide users with information on a global or local level, leading to different types of understanding of a system (Stepin et al., 2021). While global explanations provide more generic information on how a model works, local explanations incorporate details to specific output values of a ML model to make its working mechanisms more tangible, in particular to laypeople. For example, in a neural network performing an image classification task, global explanations illustrate the general patterns or features that the neural network has learned to recognize across many images, while local explanations highlight specific areas or pixels in an individual image that influenced the network's classification decision (Adadi & Berrada, 2018).

An important consideration regarding the previously mentioned interaction challenges is the alignment of explanations with human cognitive processes (Miller, 2019). Here, local explanations, specifically example-based explanations, promise to enrich users' understanding of a task by aligning with existing human cognitive evaluation processes. Rather than focusing on features or rules, example-based explanations provide examples of previous situations with similar (*factual*) or diverging (*counterfactual*) outcomes. While this is more similar to human sense-making, the implications in AI-assisted decision making remain unclear.

Ultimately, there is no concise understanding of how different XAI features are beneficial to improve AI-assisted decision-making. Such an understanding would be important: If we know what type of information is helpful for human decision-making, we can leverage ML-based systems and their theoretical benefits not only to convince humans to unquestiongly rely on the algorithmic advice, but rather to equip humans with helpful information that enables them to make decisions that exceed both the performance of the algorithm and the one that the human could achieve on their own (Murray et al., 2021; Vaccaro et al., 2024). As such, we aim to explore the following research questions (RQ) in our study:

> ***RQ1:*** *How do factual and counterfactual XAI features affect individuals' overruling of algorithmic advice and decision performance in a sales forecasting task?*
>
> ***RQ2:*** *What is the role of users' cognitive load in explaining the effectiveness of factual and counterfactual XAI features for overruling and decision performance?*

We conduct an exploratory analysis of the effect of example-based XAI features, namely *factual and counterfactual explanations*, on users' decision-making performance and overruling of algorithmic forecasts in an ML-assisted sales forecasting task. We further explore the role of users' cognitive load in explaining differences in decision-making performance and overruling behavior. As part of an online lab experiment (N = 95), users interact with a forecasting interface to make in total 1330 sales forecasting decisions.

We find that factual explanations help users perform significantly better in the forecasting task. Interestingly, users exposed to counterfactual explanations can also improve their forecasting performance but to a lesser extent. Two mechanisms related to users' overruling behavior help explain these findings: one, when users overrule algorithmic forecasts, we find that users exposed to factuals overrule significantly more in the right instances (i.e., improving the algorithmic forecast for better forecast accuracy). Two, when users are exposed to counterfactuals, their forecasting performance is mediated by their cognitive load and subsequent overruling. In other words, exposure to counterfactuals increases users' cognitive load, which, in turn, reduces overruling of algorithmic forecasts. This sequential mediation points towards the idea that counterfactuals induce users to unquestioningly adhere to the algorithmic forecasts as a result of counterfactuals imposing too much information for users to effectively process. We do not find a mediation of factuals' effect on users' forecasting performance via cognitive load, suggesting that factuals might be more aligned with individuals' sensemaking processes to make more informed decisions.

Our findings contribute to research on AI-assisted decision-making and the role of XAI for improved decision quality, specifically. We provide a richer understanding of the different features of XAI. While the information systems (IS) literature on XAI treated the specific design of an explanation as given and predominantly focused on feature-based XAI, we find that some XAI features, namely counterfactuals, tend to exacerbate prevalent interaction challenges in AI-assisted decision-making, whereas other XAI features, namely factuals, help realize theorized performance gains. We also provide insights on the human cognitive evaluation of XAI: some XAI features, namely counterfactuals, tend to impose too much cognitive load on users, which does not seem to align well with human sensemaking processes when processing explanations (Miller, 2019).

2 Conceptual Background

In this section, we first review work on AI-assisted decision-making. Subsequently, we discuss the role of explanations and cognitive load for effective decision-making.

2.1 AI-Assisted Decision-Making

Leveraging the complementary skills of humans (i.e., general, contextual reasoning capabilities) and discriminative ML-based models (i.e., processing and computational capabilities) offers to improve performance in a variety of tasks, including decision-making (Fügener et al., 2021; Murray et al., 2021).[1]

Despite these potential complementarities, converging empirical evidence suggests that in presence of both human judgement and algorithmic output, task performance is significantly worse than human or algorithmic output would perform on their own (Vaccaro et al., 2024). This is due to several challenges, including human overreliance on algorithmic output, and humans' selected engagement with algorithmic output. One, in AI-assisted decision-making, individuals tend to rely too heavily on algorithmic output, including instances in which such output is harmful to task performance. The output is hereby unquestioningly adhered to—potentially due to a lack of personal involvement with the task, due to ascribing ultimate decision-making power to the system, or due to an insufficient engagement

[1] As we explore the complementary role of algorithmic output for human decision-making, we study an interaction where algorithmic outputs assist the individual in a closed task type with a defined solution space. We recognize scenarios and tasks where the automation of tasks through AI or where the complementary role of algorithmic output for open-ended, creative tasks are appropriate. While these are comparably interesting research avenues, we view the complementary role of algorithmic output in a decision-making context as given for the research aim of our study.

with additional information (Jussupow et al., 2021; Lebovitz et al., 2022). Two, individuals tend to engage with algorithmic output in a biased and selected manner. That is, individuals tend to ignore output if it diverges from their own judgement and tend to engage with output if it does confirm their own judgement (Abdel-Karim et al., 2023; Bauer et al., 2023).

These findings raise important questions about *what kind of information* humans should be confronted with, and how. As previously mentioned, seeking and processing complementary information on a task is an important mechanism to improved decision-making (Miller, 2019).

2.2 Explanations

Explanations are a well-established design mechanism in the IS literature to address issues of understandability and inscrutability associated with digital artifacts. Explanations offer to make transparent a system's problem-solving logic as well as to provide task- and domain-specific information, which can improve a user's understandability (Gregor & Benbasat, 1999). In the context of ML-based systems, specifically, the accuracy of models often comes at the cost of explainability, as algorithms can be difficult for humans to interpret (Liao et al., 2020; Townsend et al., 2024). This challenge has driven the development of XAI, which aims to make ML models and their output more understandable and transparent to users (Gunning et al., 2019).

The effectiveness of XAI for improved decision-making is unclear, however. While XAI methods might drive transparency and trustworthiness, their impact on performance-related outcomes is less obvious. Explanations become effective as they offer to mitigate cognitive biases and reduce the likelihood of errors (Wang et al., 2019). While addressing users' explanation needs can improve decision quality (Liao et al., 2020), empirical work on XAI is converging on whether explanations and what kind of explanations help users in arriving at this improved decision-making.

A prevalent form of XAI studied in the literature, namely feature-based explanations, provide insights into model behavior, e.g., by explaining how contributions of input variables are allocated to predictions (Apley & Zhu, 2020). Feature-based explanations, however, appear to serve as a cue for a system's trustworthiness without necessarily helping humans to improve their decision quality. Users struggle to leverage feature-based explanations to appropriate rely on correct versus erroneous ML-based output (Schoeffer et al., 2024) and consider these explanations to confirm their established perceptions (Bauer et al., 2023; Schoeffer et al., 2024).

Thinking more explicitly about how to design explanations, might allow to align XAI with human reasoning and thereby enable theorized performance benefits (Dandl et al., 2020). While useful for experts, feature-based explanations often fail to address the needs and properties of human sensemaking, especially for ML forecasting that involve abstract, multidimensional features such as lag variables (Abdel-Karim et al., 2023; Bauer et al., 2023; Schoeffer et al., 2024). Even white-box models like linear regression can yield outputs that are unintuitive for users with low technical literacy (Hagras, 2018).

Another form of XAI is that of example-based explanations. Example-based explanations can offer *factual* examples, or evidence, of a given instance or similar inputs leading to similar predictions. In the context of a sales forecast, a factual could illustrate how similar sales inputs lead to similar sales forecasts. *Counterfactuals* are also example-based explanations that illustrate how a hypothetical, alternative choice of inputs lead to a different prediction or outcome (Stepin et al., 2021). In the context of a sales forecast, a counterfactual could illustrate how contrastive sales inputs lead to a different sales forecast (for more details, see the Methods section). Both factuals and counterfactuals might align well with how individuals make sense of and evaluate explanations (Miller, 2019).

2.3 Cognitive Load

Taking into consideration that the evaluation of an explanation is a human cognitive process, we return to the previously raised question of what kind of information aligns well with human sensemaking. Several IS papers investigated how different forms of explanations impact individuals' cognitive processes and cognitive load, specifically (Blohm et al., 2016; Herm, 2023; Riefle, 2024).

Cognitive load theory explores the mental efforts required for a task, suggesting that there is a limit to how much information an individual can process at a time (Buller & Burgoon, 1996; Paas et al., 2010). As such, the limited amount of processable information in the human short-term memory is responsible for cognitive load (Sweller, 1988). If this amount of processable information is exceeded, user satisfaction (Nadkarni & Gupta, 2007) but also task performance such as decision quality (Blohm et al., 2016) can decrease. IS literature has thus been concerned with how to design digital artifacts, including websites, navigation systems and e-commerce applications, so that users interacting with these artifacts are likely to find information in a time-efficient and effortless manner (Hu et al., 2017; Paas et al., 2010; Schmutz et al., 2009).

The consideration of individuals' cognitive load also becomes relevant to the context of explainability for AI-assisted decision-making. The design of such explanations can be viewed as a balancing act between inducing individuals' cognitive deliberation so that relevant information is properly processed and understood rather than overwhelming individuals with complexity and density of information. Empirical findings suggest that different types of explainability features can help reduce users' cognitive load and increase task performance as compared to no explanation given (Herm, 2023). For forecasting decisions particularly, an excessive amount of cognitive load induced by an interface can lead to worsened decision forecasting quality (Blohm et al., 2016).

Despite the potential of these explainability features, it is unclear how different example-based explanations compare in their effect on users' cognitive load, overruling of algorithmic forecasts, and ultimately decision-making performance. Example-based explanations, including factuals and counterfactuals, might be well aligned with human reasoning, particularly lay users, without stressing cognitive efforts. Gaining a richer understanding in users' evaluation and interaction processes with factuals and counterfactuals offers to address prevalent challenges of overreliance and selected engagement in AI-assisted decision-making.

3 Methods

Our study empirically explores the role of example-based explanations, namely factuals and counterfactuals, in human tendencies to overrule algorithmic advice and in task performance in AI-assisted decision-making. We conduct an online lab experiment as part of which participants interact with an ML-based explanation interface (XI) to complete a sales forecasting task.

3.1 Experimental Design and Explanation Interface

This study employed a balanced three-group between-subjects experimental design to investigate the effects of example-based explanations—consisting of factual and counterfactual explanations—on participants' decision-making performance in an AI-assisted sales forecasting task. Forecasting plays a critical role in business strategy, and ML models have shown potential to significantly enhance forecast accuracy and often surpass human capabilities (Kelleher et al., 2015; Mentzer & Bienstock, 1998). The custom interactive XI as represented in Figure 1 was used in the experiment to explore the influence of different types of explanations.

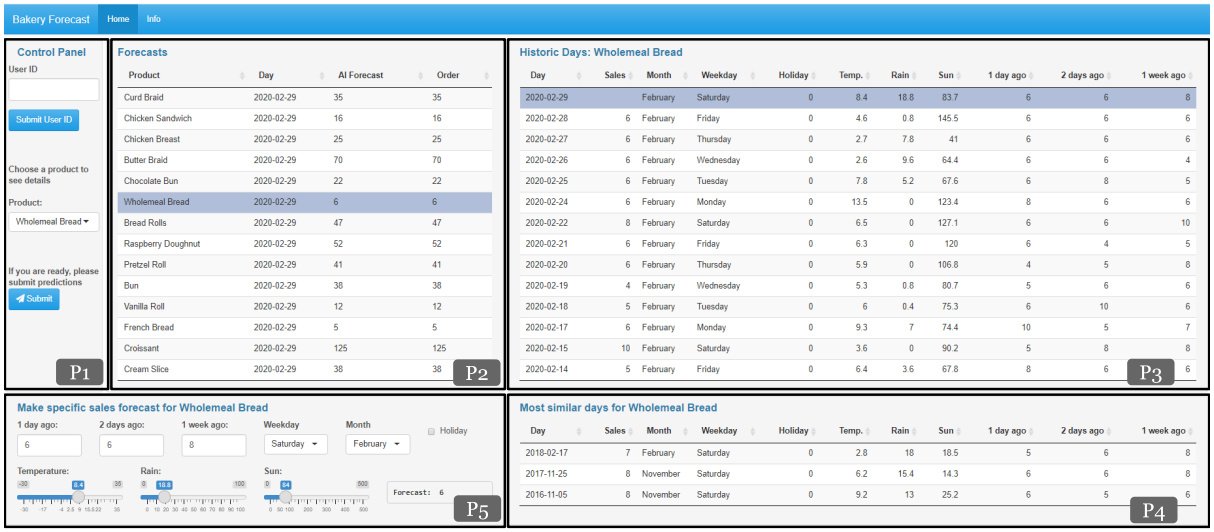

Figure 1. Exemplary screenshot of the interactive ML-based XI interface used in the experiment, with panel 4 (P4) and 5 (P5) containing the counterfactual and factual explanations, respectively.

Participants were randomly assigned to one of three experimental groups, each with access to different configurations of the XI:

1. **No Explanations (Control Group):** Participants interacted with a basic XI containing a control panel (P1), an ML forecast panel (P2), and a historical sales data panel (P3), but no example-based explanations.
2. **Counterfactuals Group:** Participants had access to the basic XI (see Control Group) plus the counterfactual explanations panel (P5). In P5, users could manipulate input features to observe how changes would alter the ML forecast. This allowed them to explore alternative scenarios and better understand the model's decision-making process.
3. **Factuals Group:** Participants were provided the basic XI plus the factual explanations panel (P4), which displayed three similar historical cases based on a K-Nearest Neighbors (KNN) algorithm (MacQueen, 1967). These examples allowed users to contextualize the ML forecast by drawing on analogous past data.

Experimental Condition	**Number of Participants**	**Number of Forecast Decisions**	**Average Age**	**Gender** Female (male)	**Time spent on task (minutes)**
1) No Explanations (Control)	32	448	28.78	44% (56%)	16.34
2) Counterfactuals	31	434	28.61	39% (58%)	19.50
3) Factuals	32	448	28.66	47% (44%)	13.72
Total	95	1330	28.78	44% (52%)	16.59

Table 1. Overview of experimental groups.

Participants were tasked with forecasting one-day-ahead sales for bakery products, using ML forecasts derived from a real-world dataset of over 400,000 transactions from a Swiss bakery. The dataset, spanning five years, was enriched with features such as weather conditions and holiday schedules. An XGBoost model (Chen & Guestrin, 2016) served as the basis for generating ML forecasts,

outperforming human branch managers' predictions by 1.7% in RMSE. This setup allowed us to systematically investigate how different types of example-based explanations influence users' ability to improve upon algorithmic forecasts. For further details on the XI and the ML model used in this experiment, we refer to a preliminary study by Fahse et al. (2022), which employed the same setup to examine *whether* example-based explanations are effective. In contrast, the present experiment focuses on exploring the underlying reasons *why* this might be the case.

3.2 Participants

Study participants were recruited via the online experiment platform Prolific in 2023. Requirements for study participation included a certain approval threshold for previous experiment participation, fluency in English, and technical requirements in order to interact with the XI. We removed participants from the final dataset if they failed attention checks, did not complete the forecasting task or completed the experiment in an unusually short time. Across the three experimental groups, a total of 1,330 forecast decisions (see Table 1 for an overview) were made. We deem this sample size as appropriate for the exploratory nature of our study as well as sufficient to run comparative analyses between the experiment conditions.

As part of the final sample, participants were on average 29 years old, with 44% being female, 52% male, and 4% preferred not to say. Most of them held a higher education degree, with 57% having completed a Bachelor and 24% a Master's level studies. 18% reported high school as their highest level of education. 70% were employed at the time of taking the experiment, with 30% not being employed. Around half of the latter group was actively looking for a job. The average time spent on the forecasting task by the participants was a ranged between about 14 and 20 minutes, depending on the experimental condition.

3.3 Measures

Each participant forecasted sales for 14 products. Prior to the forecasting task, participants received instructions tailored to their assigned XI configuration to ensure task comprehension. To evaluate *task performance*, we constructed an error measure on the difference between the amount of bakery goods ordered by the user and actually sold goods for each of the 14 sales forecasts each user completed. We employed mean absolute percentage error (MAPE) as the main measure, as it deals well with large differences in order quantities and is intuitive to interpret as relative error. The absolute percentage error was calculated by taking the absolute difference between the user's order and the actual sales in each forecasting decision, and standardizing it by the amount of actual sales. If, for example, a participant would order 15 units of a specific product, but only 10 units actually would have been sold the next day, the resulting absolute percentage error for this forecast decision would be 50%. To aggregate the measure to user level, the arithmetic mean over the set of all products ($p \in P$) was calculated, resulting in the MAPE.

$$MAPE = \frac{1}{|P|} \sum_{p \in P} \left| \frac{Order_p - Sales_p}{Sales_p} \right|$$

For *overruling* behavior, the absolute deviation of participant's orders from the ML forecast was taken as key measure (hereafter referred to as absolute overruling). For aggregation to user level, we again calculated the arithmetic mean over all product forecasts. The mean absolute overruling by user was 5.35 units with a standard deviation of 4.99. Additionally, we measure the frequency of different types of overruling decisions. For this, we compare the user and ML forecast to see how often users followed the advice and how often they chose to overrule. In the cases where users chose to deviate from the ML forecast, we measure beneficial overruling as a human deviation that results in a more accurate forecast. Analogously, harmful overruling is referred to as a deviation that results in a less accurate forecast.

Post-task questionnaires measured demographics, users' cognitive load, and perceived task complexity. To capture users' *cognitive load* induced by the (lack of) XAI features, users were asked how effortful and complex they perceived the forecasting task to be (see Hong et al. (2004), measured on a Likert-scale from 1 to 7, sample item: "To make predictions, using the forecasting interface required too much effort", Cronbach's Alpha = 0.893).

4 Results

In this study, we aim to understand the role of the different explanation-based XAI features on task performance and the participants' inclination to follow or improve algorithmic advice (i.e., overruling). Therefore, we first derive an understanding of the XAI's effect on task performance (i.e., forecast error measured as MAPE). We then analyze the general effect of overruling algorithmic advice and explore differences in overruling behavior between the experimental groups. Finally, we explore if changes in cognitive load act as a cause of the effect.

4.1 Task Performance

Table 2 provides an overview of descriptive statistics for the forecast error (i.e., MAPE) measuring task performance and absolute overruling measuring users' deviation from the ML forecast. Both XAI approaches, factuals and counterfactuals, decreased the forecast error, i.e. increased users' task performance. This is in line with the results of the preliminary experiment using the same platform (Fahse et al., 2022). When exposed to factuals, users performed significantly better in the forecasting task compared to users in absence of an explantion ($M_{Factuals}$ = 0.177, M_{NoXAI} = 0.235, t = -2.844, p = 0.006). This average decrease in MAPE of nearly 6% was confirmed with a linear regression (β = -0.058, SE = 0.02, t = -2.844, p = 0.006, R^2 = 11.5%) and is robust to RMSE as error measure, although only marginally significant ($p < 0.07$). While users exposed to counterfactuals also show a lower forecast error compared to the users in the control group, we do not find a significant difference in forecast error between these two groups nor between the factuals and counterfactuals groups. The standard deviation measure, as reported in Table 2, is lowest in the counterfactuals condition. Thus, we can observe that, while the task performance was better with factuals, forecasts seemed more consistent when participants were exposed to counterfactuals.

Absolute overruling was lowest for the counterfactual condition. Users exposed to counterfactuals overrule ML forecasts significantly less as compared to users exposed to no explanation ($M_{Counterfactuals}$ = 3.583, M_{NoXAI} = 6.754, t = -2.529, p = 0.014). As overruling was lowest in the presence of counterfactuals and yet task performance was highest in the presence of factuals, it is unclear to what extent overruling behavior can explain task performance.

Experimental Condition	**MAPE**		**Absolute Overruling**	
	Mean	*SD*	*Mean*	*SD*
1) No Explanantion	23.46%	9.71%	6.75	5.84
2) Counterfactuals	20.26%	5.50%	3.58	3.88
3) Factuals	17.70%	6.08%	5.21	4.34

Table 2. Descriptive statistics for the task performance and overruling behavior in the different experimental groups.

To better understand the relationship between overruling behavior and task performance, we fit a linear regression predicting task performance measured by MAPE with the users' average absolute deviation from the ML forecast (absolute overruling) as the independent variable. The result confirms a highly significant positive correlation (β = 0.010, SE = 0.01, t = 6.693, $p < 0.001$, R^2 = 35.9%), indicating that higher overruling decreases task performance (see Figure 2). This regression result is robust to RMSE as an alternative error measure (β = 1.010, SE = 0.115, t = 8.806, $p < 0.001$, R^2 = 49.2%). We can

corroborate these findings on a decision-level beyond the previous analyses conducted on the user-level. Overall, these results suggest that we need to have a closer look at users' overruling behavior between different experiment conditions as the overall overruling effect does not seem to affect all experimental conditions in the same manner.

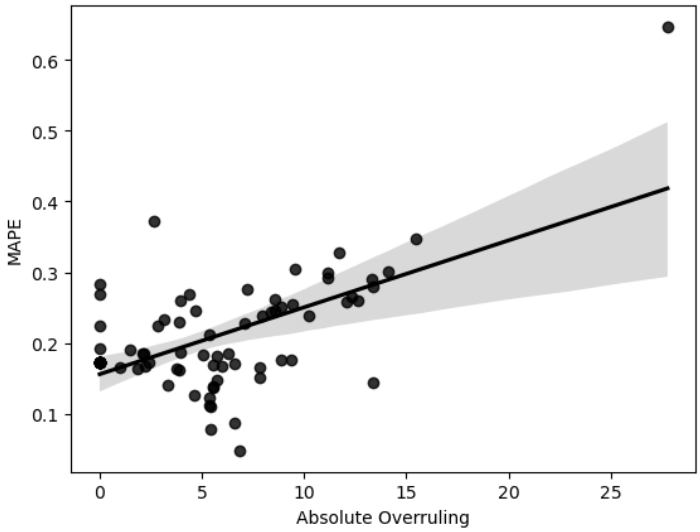

Figure 2. Regression plot showing the positive correlation between overruling and the forecasting error.

4.2 XAI and Overruling of Algorithmic Advice

Out of all decisions made across the three experimental conditions, 45.73% of orders adhered to the suggestion of the ML forecast, while 54.27% of orders exhibited overruling behavior. Out of these, 60.83% (39.17%) of ML forecasts were overruled with a higher (lower) order. While users more often overruled in the correct direction than not (55.70%), the placed order was closer than the ML forecast to the actually sold amount in only 41.73% of the overruled cases. This implies that, even when users correctly decided that more products will sell the next day than the ML forecast suggested, their adjustment was too extreme in about 25% of decisions. However, the considerable amount of overruled cases that improved upon the ML forecast (>40%) could explain why task performance increased together with absolute overruling for users exposed to factual explanations, although this stands against the overall trend as shown with the linear regression.

We next investigate the effect of the two different approaches to example-based explainable AI on overruling behaviour and related task performance. Figure 3 presents a summary of the performance of human participants compared to the theoretical ML forecast across the three different experimental conditions.

As seen in the left plot, the presence of factuals results in participants correctly overruling the ML forecast more often. The median participant in this group achieved a performance better than the algorithmic forecast in around a third of all forecasts. Comparing the differences between all pairs of conditions (ANOVA), we find that beneficial overruling occurred significantly more frequently when factuals were present compared to when counterfactuals were present ($M_{Counterfactuals} = 0.171$, $M_{Factuals} = 0.297$, $t = -2.424$, $p = 0.018$).

The distributions in Figure 3 moreover show that participants exposed to counterfactuals were more inclined to follow the ML forecast, as people achieved results equally as good as the ML forecast most frequently (see center plot). A linear regression analysis confirms that, compared to the control group,

exposure to counterfactuals led to an increase of 16.55% of participants performing equally as good as the ML forecast (β = 0.166, SE = 0.087, t = 1.910, p = 0.061, R^2 = 5.6%).

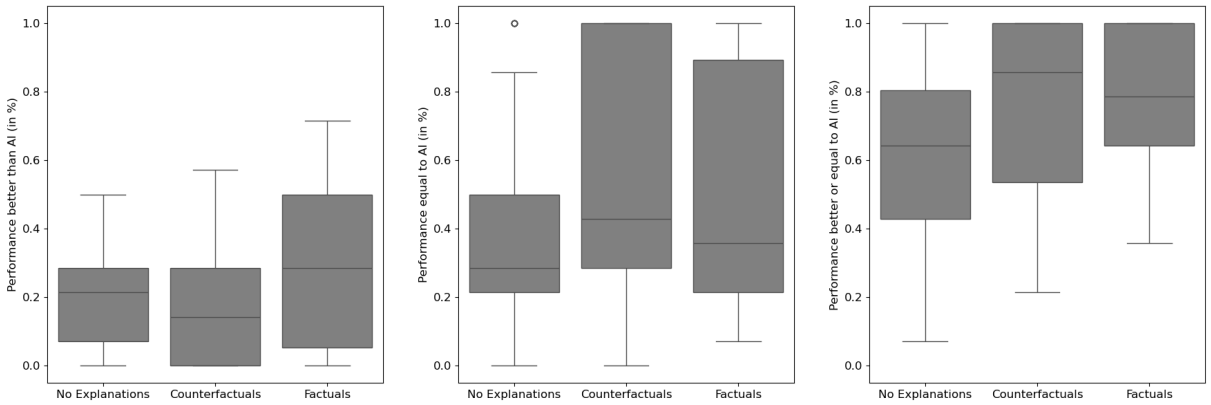

Figure 3. *Boxplots illustrating the distributions of user performance compared to the ML forecasts by experimental group. Left: Comparison of experiment conditions along their performance exceeding ML forecasts. Center: Comparison of experiment conditions along their performance being equal to ML forecasts. Right: Comparison of experiment conditions along their performance exceeding or being equal to ML forecasts.*

The median person in the counterfactuals condition also achieved a performance equally as good or better than the AI (right plot) most frequently. This, however, was not sufficient to achieve the best task performance as successful overruling was rare. Fitting linear regressions showed that the frequency of forecasts where participants were at least as good as the ML forecast, increased significantly in both XAI conditions by 11.95% (β = 0.120, SE = 0.062, t = 1.918, p = 0.06, R^2 = 5.7%) and 14.06% (β = 0.141, SE = 0.057, t = 2.457, p = 0.017, R^2 = 8.9%) for counterfactuals and factuals, respectively. We can thus conclude that the cases of better overruling that were most common within the factuals group achieving the best results, are overcompensating for less adherence to the forecast.

In line with these findings, presence of counterfactuals also reduced the frequency of overruling by 19.2% (β = -0.192, SE = 0.090, t = -2.139, p = 0.036, R^2 = 7%). When exposed to factuals however, the frequency of overruling did not decrease in a statistically significant way.

Next, we analyze the effect of the conditions on the absolute overruling measure. Here, a direct negative effect in the presence of counterfactuals can be observed, meaning that overruling was not only less frequent but also less severe when participants were exposed to counterfactuals (β = -3.172, SE = 1.254, t = -2.529, p = 0.014, R^2 = 9.5%). For the presence of factuals, no significant direct effect on overruling could be identified, which corroborates the previous findings suggesting no significant decrease in frequency of overruling.

4.3 The Mediating Role of Cognitive Load

To gain a better understanding of the effect and effectivess of different XAI types, we analyze cognitive load as a possible mediator helping explain the effect of the explanation types on users' overruling behavior and ultimately task performance.

We can observe that, within both XAI conditions, cognitive load increased significantly compared to the control group, with the average increase in the presence of counterfactuals (β = 7.235, SE = 1.889, t = 3.830, p < 0.001, R^2 = 19.4%) being about 55% stronger than with factuals (β = 4.656, SE = 2.054, t = 2.267, p = 0.027, R^2 = 7.7%).

To test a mediation effect through cognitive load, we employed PROCESS Model 4 (Hayes, 2017). Initially, a significant direct effect of counterfactuals on average absolute overruling could be observed

(β = -3.172, SE = 1.254, t = -2.529, p = 0.014, R^2 = 9.5%). With the exposure to counterfactuals inducing a significantly higher level of cognitive load, it was tested whether this in turn is mediating the users' reduced overruling of ML advice. The results indeed reveal an indirect effect of exposure to counterfactuals through cognitive load on absolute overruling ($\beta_{Indirect}$ = -0.962, SE = 0.585, 95% CI [-2.503, -0.127]). The initially significant direct effect of the presence of counterfactuals on absolute overruling becomes non-significant when including cognitive load as a mediator (β_{Direct} = -2.210, SE = 1.380, 95% CI [-4.914, 0.494]). This pattern suggests a full mediation explaining how counterfactuals affect overruling of algorithmic advice primarily through its impact on cognitive load.

The same result can be observed on decision level using clustered standard errors to account for cognitive load being measured only once per user ($\beta_{Indirect}$ = -0.949, p = 0.040, 95% CI [-2.180, -0.050]), although the direct effect after mediation is still marginally significant (β_{Direct} = -2.215, p = 0.058, 95% CI [-4.449, 0.090]). This suggests a partial mediation with a still considerable 29.06% of the total effect being mediated through cognitive load ($p < 0.05$).

In combination with the previously established effect that higher overuling leads to worse task performance, we can infer that the exposure to counterfactual XAI features should also lead to a better performance. To prove this, we conducted a serial mediation analysis with cognitive load and absolute overruling as sequential mediators and error measures as the dependent outcome variables. The model shows that the effect of counterfactual explanations on task performance is fully mediated by cognitive load and overruling. The indirect effect is significant ($\beta_{Indirect}$ = -0.017, SE = 0.008, p = 0.036), while the direct effect is non-significant (β_{Direct} = 0.015, p = 0.205). This confirms that counterfactuals improve task performance indirectly through the reduction in overruling, which in itself is mediated by higher cognitive load. A summary of these effects in the serial mediation model is illustrated in Figure 4.

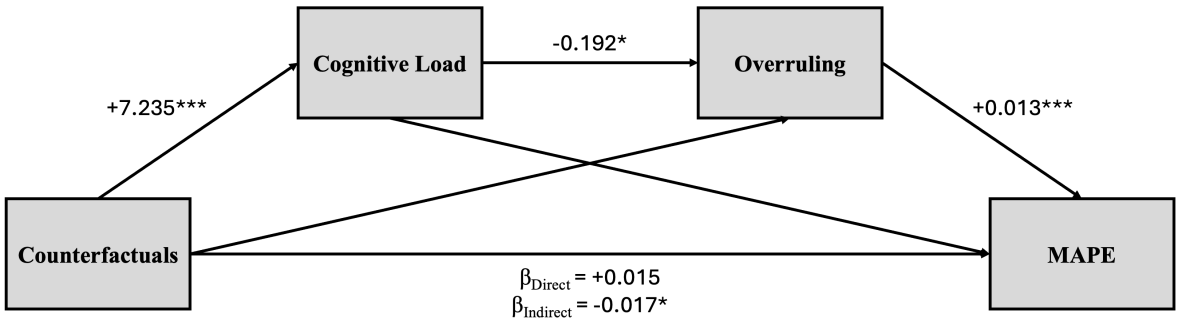

*Figure 4. Illustration of the serial mediation model for the counterfactuals condition (user level), including coefficients and significance levels (*p<0.05, **p<0.01, ***p<0.001).*

In contrast, in the presence of factuals, no significant mediation can be demonstrated ($\beta_{Indirect}$ = -0.006, SE = 0.004, p = 0.213), but a significant direct effect on task performance (β_{Direct} = -0.035, SE = 0.014, p = 0.013) can be observed. This suggests that the observed direct improvement is independent of cognitive load or overruling. This effect–together with all other effects reported in the mediation analyses–is robust to a decision level analysis ($\beta_{Indirect}$ = -0.003, SE = 0.003, p = 0.243; β_{Direct} = -0.042, SE = 0.015, p = 0.004).

5 Discussion

Our results demonstrate that factual explanations significantly enhance forecasting performance by enabling users to overrule algorithmic advice more effectively. This suggests that they align well with users' natural sensemaking processes. Factual explanations provide concrete examples of similar past instances, which help users contextualize the algorithmic output within their domain knowledge. This finding resonates with the assertion by Miller (2019) that explanations are more effective when they

align with how humans naturally seek and process information. By facilitating a better understanding of the task without causing cognitive overload, factual explanations enable users to identify instances where overruling the algorithm leads to better outcomes. This aligns with prior studies emphasizing the importance of explanations that are intuitive and easily interpretable by users (Dandl et al., 2020; Miller, 2019). Unlike feature-based explanations, which can be complex and unintuitive for lay users (Hagras, 2018), factual explanations offer accessible, case-based information that users can readily understand and apply. Our findings extend this understanding by demonstrating that factual explanations not only enhance user understanding but also lead to improved decision-making performance by facilitating effective human-AI collaboration, allowing users to leverage both the algorithmic insights and their domain expertise to improve decision outcomes

Counterfactual explanations, while informative, increased users' cognitive load, leading to reduced overruling behavior and increased adherence to algorithmic forecasts. Our mediation analysis highlights the critical role of cognitive load in the relationship between counterfactual explanations and decision-making behavior, while no significant mediation effect of cognitive load was observed with factual explanations. The increased cognitive load associated with counterfactual explanations reduces users' propensity to overrule the algorithm, leading to performance improvements due to the generally superior accuracy of the ML forecasts. The increased cognitive demands imposed by counterfactual explanations may hinder users' ability to engage in the deep processing required to evaluate and potentially overrule the algorithmic advice. This underscores the importance of designing explanations that are not only informative but also cognitively manageable for users. This suggests that counterfactuals may overwhelm users with information, resulting in cognitive overload (Sweller, 1988). Consequently, users may default to accepting the algorithmic advice without critically evaluating it—a phenomenon observed in prior studies where users unquestioningly adhere to algorithmic output due to insufficient engagement with the information provided (Jussupow et al., 2021; Lebovitz et al., 2022). While recognizing the cost of increased cognitive burden, Lee and Chew (2023) observed that counterfactual explanations in a different context helped reduce adherence to incorrect AI outputs, which stands in contrast to our findings. Although their study did not compare counterfactuals to other forms of example-based explanations, it highlights potential context-specific variations in the effectiveness of counterfactual explanations.

5.1 Theoretical Contributions and Practical Implications

Our study makes several theoretical contributions to the fields of AI-assisted decision-making and explainable AI. First, we provide empirical evidence that the specific design of XAI features significantly impacts decision-making performance, supporting calls for a more nuanced understanding of different forms of XAI, especially in the context of human cognitive processes (Dandl et al., 2020; Miller, 2019). While prior IS research has often treated the design of XAI as given and focused predominantly on feature-based explanations (Apley & Zhu, 2020), our findings highlight the distinct effects of example-based explanations—specifically factual and counterfactual explanations—on user behavior and performance. We hereby extend the understanding of how users make sense of explanations in AI-assisted decision-making contexts. Our results suggest that factual explanations enhance users' understanding and decision-making performance by aligning with natural human reasoning processes (Miller, 2019). Conversely, counterfactual explanations may impose higher cognitive demands, leading to less effective use of the information provided. This contributes to the broader discourse on the importance of aligning XAI designs with users' cognitive capacities and information processing styles (Abdul et al., 2018).

Second, our study contributes to the ongoing discourse on the unintended challenges in AI-assisted decision-making preventing theorized performance gains to materialize. Converging empirical evidence illustrates that in AI-assisted decision-making scenarios, human and AI working together rarely surpass human or AI only performance (Vaccaro et al., 2024). Findings in the IS literature attribute these challenges to an overreliance on algorithmic advice and only selectively engaging with the information provided (Abdel-Karim et al., 2023; Bauer et al., 2023; Jussupow et al., 2021; Lebovitz et al., 2022). Our findings suggest that an explicit consideration of *how* to design algorithmic output can help

overcome these challenges. Similar to XAI studies we find that the use of counterfactuals serves as a cue of trustworthiness where users do not fully engage with the counterfactuals due to the cognitive load they induce but adhere to the ML forecast. The use of factuals may suggest a path forward, though: users exposed to factuals in our study overruled the ML forecast successfully, which led these users, on average, to perform best in the forecasting task.

Third, our study contributes to cognitive load theory in the context of AI-assisted decision-making. By demonstrating that cognitive load mediates the relationship between counterfactual explanations and decision-making behavior, we highlight the need to consider cognitive load when designing XAI features. Our findings suggest that explanations that are too cognitively demanding may inadvertently lead users to rely more heavily on algorithmic advice without critical evaluation. This finding contributes to the literature on cognitive load in forecasting (Blohm et al., 2016; Sweller, 1988) by illustrating how cognitive demands can mediate the effectiveness of XAI features in decision-making contexts. Our findings hereby align with the notion that cognitive load influences users' ability to process and utilize explanations effectively (Hu et al., 2017; Paas et al., 2010; Schmutz et al., 2009).

Our findings have managerial and strategic implications for the design and implementation of AI-assisted decision-making systems in, e.g., the workplace. For practitioners aiming to enhance decision quality through effective human-AI collaboration, it is crucial to consider the type of explanations provided to users. Simply providing explanations is not sufficient; the design and presentation of these explanations must be tailored to support users' ability to understand and utilize the information. As suggested by XAI literature, this highlights the need for user-centered design approaches in developing XAI features that effectively support decision-making (Herm, 2023; Riefle, 2024).

Factual explanations appear to be particularly effective for users with domain expertise, as they facilitate the integration of algorithmic insights with users' own knowledge to improve decision outcomes. Designers should consider incorporating factual, example-based explanations to augment users' sensemaking processes and enable appropriate overruling of algorithmic advice. This design appears to be promising for workers' long-term learning and skill expertise, too.

In contexts where the automation of a task through AI is unfeasible, e.g., due to legal or safety reasons, yet adherence to algorithmic advice is preferred, we present the counterintuitive finding that counterfactual explanations may still be beneficial despite causing cognitive overload. By leading users to rely more heavily on the ML forecasts, counterfactual explanations can improve performance in tasks where the algorithm outperforms human judgment. We urge practitioners to consider the long-term implications of this design, though. As users seem to not properly engage with the explanation, workers might be limited in their learning and agency.

5.2 Limitations and Future Research

Our study's findings on the impact of factual and counterfactual explanations on decision-making performance should be interpreted in the light of certain limitations. First, it is important to note that we disregarded the sociotechnical contextualization of AI-assisted decision-making. Decision-making in the workplace, for instance, is largely influenced by organizational factors, including managerial expectations, norms and values in the workplace, or workers' level of expertise, just to name a few. The main aim of our exploratory study was to ensure the psychological realism of human sensemaking processes when evaluating explanations, hence the choice for an online lab experiment. Going further, we find it important to study the effect of AI-assisted decision-making and the impact of XAI methods in the workplace, where workers' sensemaking might be confounded by external factors, such as time pressure. Similarly, the measure of decision-quality could be extended in future research by encompassing not only the forecasting error, but also accounting for other measures like the time taken or the subjective decision confidence.

Second, our participant sample, drawn from an online platform, may limit the generalizability of our findings to professional decision-makers or experts with domain-specific knowledge. The impact of different explanation types may vary based on users' expertise, familiarity with AI systems, or cognitive styles (Riefle, 2024). Future studies should explore the differential effects of explanations among users

with varying levels of expertise, as well as the potential for tailored explanation designs to cater to diverse user needs. Additionally, the robustness of our evaluation and thereby its generalizability could be improved by testing multiple scenarios with more diverse datasets in subsequent experiments.

Third, while we focused on example-based explanations, our study did not compare them with feature-based explanations, which remain a prevalent form of XAI in the literature (Apley & Zhu, 2020). Investigating the interaction between different explanation types and their combined effects on decision-making could provide a more comprehensive understanding of how to optimize XAI designs for various contexts. Additionally, even when only focusing on the example-based explanations present in our setting, evaluating the effects of combined access to both factual and counterfactual explanations could yield interesting insights. In our study, a few participants had access to the full XI, but, due to the exclusion of some users due to validity checks and incomplete data, we used these 13 observations to augment the other experimental conditions, focusing only on the presence of counterfactual or factual explanations. While we conducted several robustness checks to ensure that this consolidation of groups is not affecting our results, it might be interesting in the future to gather a larger sample of participants for a fourth condition with access to multiple explanation approaches (example-based or feature-based). This way, it could for example be explored whether cognitive overload exacerbates or whether users are able to focus on the most useful information. By addressing these limitations in future research, we can develop a more holistic understanding of how XAI features influence human-AI collaboration in diverse and complex real-world settings.

6 Conclusion

Our study advances the understanding of how different types of explanation-based XAI features impact users' decision-making performance in AI-assisted decision-making tasks. By demonstrating that factual explanations enhance performance through effective overruling, we highlight the importance of aligning XAI designs with human cognitive processes. Conversely, counterfactual explanations also improved performance but to a lesser extent. This improvement is primarily mediated by increased cognitive load and subsequent decrease in overruling, which points to cognitive overload causing users to adhere to the provided ML forecast without improving upon them. Our findings contribute to both theory and practice by elucidating the mechanisms through which explanations affect AI-assisted decision-making and offering guidance for designing AI systems that–by considering cognitive demands–effectively support human users.

References

Abdel-Karim, B. M., Pfeuffer, N., Carl, K. V., & Hinz, O. (2023). How AI-Based Systems Can Induce Reflections: The Case of AI-Augmented Diagnostic Work. *MIS Quarterly*, *47*(4), 1395-1424. https://doi.org/10.25300/MISQ/2022/16773

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). *Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal QC, Canada. https://doi.org/10.1145/3173574.3174156

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Apley, D. W., & Zhu, J. (2020). Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1059-1086. https://doi.org/10.1111/rssb.12377

Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research*, *34*(4), 1582-1602. https://doi.org/10.1287/isre.2023.1199

Blohm, I., Riedl, C., Füller, J., & Leimeister, J. M. (2016). Rate or Trade? Identifying Winning Ideas in Open Idea Sourcing. *Information Systems Research*, *27*(1), 27-48. https://doi.org/10.1287/isre.2015.0605

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal Deception Theory. *Communication Theory*, *6*(3), 203-242. https://doi.org/10.1111/j.1468-2885.1996.tb00127.x

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-Objective Counterfactual Explanations. In (pp. 448-469). https://doi.org/10.1007/978-3-030-58112-1_31

Fahse, T., Blohm, I., & van Giffen, B. (2022). Effectiveness of Example-Based Explanations to Improve Human Decision Quality in Machine Learning Forecasting Systems. Proceedings of 43rd International Conference on Information Systems (ICIS 2022), Copenhagen, Denmark.

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly*, *45*, 1527-1556. https://doi.org/10.25300/MISQ/2021/16553

Gregor, S., & Benbasat, I. (1999). Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly*, *23*(4), 497-530. https://doi.org/10.2307/249487

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI-Explainable artificial intelligence. *Sci Robot*, *4*(37). https://doi.org/10.1126/scirobotics.aay7120

Hagras, H. (2018). Toward Human-Understandable, Explainable AI. *Computer*, *51*(9), 28-36. https://doi.org/10.1109/MC.2018.3620965

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Herm, L. V. (2023). Impact of Explainable AI on Cognitive Load: Insights from an Empirical Study. Proceedings of 31st European Conference on Information Systems (ECIS 2023), Kristiansand, Norway.

Hong, W., Thong, J. Y. L., & Tam, K. Y. (2004). The Effects of Information Format and Shopping Task on Consumers' Online Shopping Behavior: A Cognitive Fit Perspective. *Journal of Management Information Systems*, *21*(3), 149-184. https://doi.org/10.1080/07421222.2004.11045812

Hu, P. J.-H., Hu, H.-F., & Fang, X. (2017). Examining the Mediating Roles of Cognitive Load and Performance Outcomes in User Satisfaction with a Website. *MIS Quarterly*, *41*(3), 975-988. https://doi.org/10.25300/MISQ/2017/41.3.14

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research*, *32*(3), 713-735. https://doi.org/10.1287/isre.2020.0980

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press.

Lakkaraju, H., & Bastani, O. (2020). *"How do I fool you?"* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society,

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. *Organization Science*, *33*(1), 126-148. https://doi.org/10.1287/orsc.2021.1549

Lee, M. H., & Chew, C. J. (2023). Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proc. ACM Hum.-Comput. Interact.*, *7*(CSCW2), Article 369. https://doi.org/10.1145/3610218

Liao, Q. V., Gruen, D., & Miller, S. (2020). *Questioning the AI: Informing Design Practices for Explainable AI User Experiences* Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA. https://doi.org/10.1145/3313831.3376590

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press,

Mentzer, J. T., & Bienstock, C. C. (1998). *Sales Forecasting Management: Understanding the Techniques, Systems and Management of the Sales Forecasting Process.* SAGE Publications.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, *267*, 1-38. https://doi.org/https://doi.org/10.1016/j.artint.2018.07.007

Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Humans and Technology: Forms of Conjoined Agency in Organizations. *Academy of Management Review*, *46*(3), 552-571. https://doi.org/10.5465/amr.2019.0186

Nadkarni, S., & Gupta, R. (2007). A Task-Based Model of Perceived Website Complexity. *31*(3), 501-524. https://doi.org/10.2307/25148805

Paas, F., Renkl, A., & Sweller, J. (2010). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, *38*(1), 1-4. https://doi.org/10.1207/s15326985ep3801_1

Riefle, L. H., P.; Benz, C.; Vössing, M. . (2024). The Role of Cognitive Styles for Explainable AI. Proceedings of 32nd European Conference on Information Systems (ECIS 2024), Paphos, Cyprus.

Schmutz, P., Heinz, S., Métrailler, Y., Opwis, K., & Kline, R. (2009). Cognitive Load in eCommerce Applications—Measurement and Effects on User Satisfaction. *Advances in Human-Computer Interaction*, *2009*(1). https://doi.org/10.1155/2009/121494

Schoeffer, J., De-Arteaga, M., & Kuehl, N. (2024). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. Proceedings of the CHI Conference on Human Factors in Computing Systems,

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Farina, M. (2021). A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, *9*, 11974-12001. https://doi.org/10.1109/access.2021.3051315

Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, *12*(2), 257-285. https://doi.org/https://doi.org/10.1207/s15516709cog1202_4

Townsend, D. M., Hunt, R. A., Rady, J., Manocha, P., & Jin, J. h. (2024). Are the Futures Computable? Knightian Uncertainty and Artificial Intelligence. *Academy of Management Review*. https://doi.org/10.5465/amr.2022.0237

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat Hum Behav*. https://doi.org/10.1038/s41562-024-02024-1

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). *Designing Theory-Driven User-Centric Explainable AI* Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland Uk. https://doi.org/10.1145/3290605.3300831