5-11-2023

Modern Centaurs: How Humans and AI Systems Interact in Sales Forecasting

Jannis Beese
SAP Switzerland, jannis.beese@gmail.com

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

MODERN CENTAURS: HOW HUMANS AND AI SYSTEMS INTERACT IN SALES FORECASTING

Research Paper

Jannis Beese, SAP Switzerland, Tägerwilen, Switzerland, jannis.beese@gmail.com

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

Abstract

Recent achievements of artificial intelligence (AI) have caused organizations to increasingly bring AI capabilities into their core business processes. Such AI-supported business processes often result in human-AI hybrid systems, which consist of an AI system, which performs most of the execution, and humans, who monitor this execution and occasionally provide additional inputs and overrides. Using sales data from Walmart, we conduct an online study to investigate if human supervision can improve upon state-of-the-art AI forecasts. Furthermore, we analyze the perceptions and behavioral intentions of the human participants over time. We find that human interventions consistently lead to less accurate forecasts and that participants initially underestimate the AI system's accuracy and overestimate their own potential to improve upon AI forecasts. However, perceptions quickly shift over the course of the study, causing the participants to perceive the AI system increasingly favorably, which also leads to behavioral changes and better hybrid system performance.

Keywords: Hybrid intelligence, machine learning, AI-assisted decision making, sales forecasting.

1 Introduction

Recent advances in artificial intelligence (AI) research in combination with increasingly accessible technology and highly publicized landmark achievements of AI — such as the 2016 Go tournament between AlphaGo and Lee Sedol (Silver et al., 2016) or the success of large language models like ChatGPT (van Dis et al., 2023) — have led modern organizations to try integrating AI capabilities into increasingly complex tasks that occur in their core business processes (Mendling et al., 2018; Jordan and Mitchell, 2015; Von Krogh, 2018). However, information systems (IS) in large organizations are complex socio-technical systems, in which highly varied human and technological elements interact in non-linear ways to process information (Mendling et al., 2018; Haki et al., 2020). Consequently, although large language models can communicate with users in natural language (van Dis et al., 2023), the success of projects that aim to bring AI systems into the business processes of large organizations is far from guaranteed (Dellermann et al., 2019).

In general, three different organizational uses of AI can be differentiated: knowledge creation, task augmentation, and autonomous agents (Shollo et al., 2022). This paper focuses on task augmentation and the performance of hybrid systems ("human-AI centaurs"), in which both human and AI components work together to solve a task, compared to pure AI systems, which operate without human oversight and intervention (i.e., autonomous agents). In analogy to the Greek mythological creature, hybrid human-AI centaurs comprise the body of an AI system, which performs most of the execution, and the head of a human, which monitors the execution and occasionally provides additional inputs and overrides. However, implementing human-AI hybrids and leveraging the potential synergies between humans and AI systems to make better decisions is not trivial. In this context, previous studies have shown how human judgment and misplaced trust in AI system performance can jeopardize the performance of human-AI hybrids (Logg et al., 2019; Fügener et al., 2022; Dietvorst et al., 2015) and how algorithm aversion can lead to human resistance to use AI based forecasts, and thus to a reluctance to accept the benefits of AI (Burton et al., 2020)

Since the potential impact of AI in business processes is expected to depend heavily on the specific task and context (Von Krogh, 2018; Duan et al., 2019), we focus our investigation on sales forecasting in the retail industry as one archetypal example (Syntetos et al., 2016). Retailers need to ensure that each store holds exactly the right amount of product at any given point in time, so that the shelves are neither too empty nor too full and inventory and wastage costs are minimized. Orchestrating these deliveries along the supply chain of large retailers is a challenging endeavor, which fundamentally requires to know how much of a specific product is expected to be sold at a specific store and at a specific point in time (Fildes, Ma, et al., 2022; Sagaert et al., 2018). Therefore, retailers rely on algorithmically generated forecasts to automatically generate procurement orders that trigger the shipment of goods from large warehouses (distribution centers) to the individual stores (Syntetos et al., 2016). Almost all retail organizations employ human demand and replenishment planners who look over these forecasts for critical products, such as high-revenue products, promotional products, and products that frequently have issues (e.g., out-of-stock situations) (Fildes, Ma, et al., 2022). The degree to which such human oversight and interventions occur varies significantly between different retailers and across different countries. It is also not clear how much human oversight is really needed and, more fundamentally, if human interventions actually help improving upon state-of-the-art AI sales forecasts. In this paper we specifically consider the effects of direct human overrides of AI sales forecasts, asking:

RQ1: Can humans improve upon sales forecasts of state-of-the-art artificial intelligence systems?

Furthermore, we are interested to understand if the involved humans perceive the situation correctly. Previous research has indicated that humans tend to distrust AI systems disproportionately (Dietvorst et al., 2015) and having to compete with such AI systems for the same business task is expected to further skew judgment against the algorithms. If the people who design, implement, monitor, and work within current human-AI hybrid sales forecasting systems have false perceptions about the performance of the technological components and their own performance, the overall system is expected to operate suboptimally (Szajna and Scamell, 1993). Consequently, we aim to understand these perceptions as well and therefore ask:

RQ2: How do humans perceive their own performance and the AI system performance in hybrid human-AI sales forecasting systems?

To answer both research questions, we conduct an online study with 53 participants, recruited via MTurk (Paolacci and Chandler, 2014). Based on a large set of actual sales data from Walmart, which was publicly released for the M5 forecasting competition (Makridakis et al., 2022a), we use LightGBM (Ke et al., 2017) to create a high-quality forecast for four weeks of 2016. LightGBM is a gradient boosting decision tree framework for machine learning (ML) that has been used in four out of five top-placing forecasting models in the M5 forecasting competition (Makridakis et al., 2022b). Despite being a framework for ML, we use the general term AI to refer to the outputs of the LightGBM model to be consistent with the term human-AI hybrids.

We provide study participants with AI forecasts for 30 purposefully selected products, further relevant data, and suitable visualizations and ask them to provide their own sales forecasts. After participants provide a forecast, they are shown the actual sales data and receive feedback on their own performance as well as the AI system performance. We also survey the participants about their perception of the AI system performance and their own performance three times (after every ten forecasts).

Overrides of AI forecasts by our study participants consistently led to significantly less accurate forecasts compared to the original AI forecasts. Furthermore, participants initially significantly underestimated the AI system's accuracy and overestimated their own potential to improve upon the AI system's forecasts. However, after having received direct feedback several times, this perception shifted, so that in the later rounds the participants increasingly perceived the AI system more favorably. In line with this result, participants also perceived their own potential to improve upon the AI forecasts to be lower and started to make smaller adjustments to the AI forecasts.

With this study, we contribute an in-depth analysis of an archetypal example of a central business task (retail sales forecasting) in organizations, where human-AI hybrids are consistently outperformed by pure AI systems without human intervention. Practitioners can use our results and analyses to support the design and development of hybrid human-AI decision systems for complex business processes. We show that the overall performance of hybrid human-AI systems does not just depend on the AI system performance, but also strongly depends on how the AI system is perceived by the involved human actors, whose perceptions may fundamentally change over time. Consequently, managers should not just measure and optimize the overall system performance but should also try to enable the involved human actors to gain accurate perceptions about the hybrid human-AI system.

2 Related work

This section covers related work in two areas. First, we briefly cover the status quo of retail sales forecasting and elaborate how related business processes work. Second, we discuss related work on how humans perceive and use technology.

2.1 The state of the art in retail sales forecasting

Retailers require highly accurate sales forecasts in most of their core business processes, impacting areas such as production and procurement, supply chain optimization, marketing, and personnel planning (Fildes, Ma, et al., 2022). Since retailers typically manage assortments comprising thousands of different products, manually handling procurement for all products in all stores is impossible. Thus, retailers rely on automated sales forecasting systems (Fildes, Ma, et al., 2022).

In the past, significant efforts have been made to build and further refine mathematical models and algorithms for retail sales forecasting and corresponding sophisticated IS. Exponential smoothing (e.g., Holt-Winters) and ARIMA models were historically frequently employed for sales forecasting, which are still relatively performant with little computational costs (Ramos et al., 2015). Such traditional models tend to be accurate when macroeconomic conditions are relatively stable (Zhang and Qi, 2005). In contrast, when customer behavior is unsteady, current studies indicate that nonlinear models can show improved forecast accuracy compared to traditional approaches (Fildes, Ma, et al., 2022). Examples include various neural networks (Veiga et al., 2016), support vector machines (Di Pillo et al., 2016), and decision trees (Gür Ali et al., 2009). The downside of such approaches are high resource requirements, since model training is computationally expensive, opaqueness, and data hunger (Kolassa, 2020; Gür Ali and Yaman, 2013). Consistent with the academic results, the top-placing algorithms in public sales forecasting competitions over the past few years have gradually shifted from traditional time series methods (e.g., ARIMA) towards artificial neural networks and gradient boosted decision trees (Makridakis et al., 2022b). A particularly notable trend in recent competitions is the prominent reliance of top-placing forecasting models on fast gradient boosting decision tree libraries (e.g., LightGBM), which enabled participants to easily experiment with their data without incurring infeasible computational costs.

Building on top of such algorithmically generated forecasts, several empirical studies have investigated the impact of human judgmental adjustments on sales and demand forecasts in supply chain planning (Petropoulos, 2022; Fildes et al., 2009; Fildes and Goodwin, 2007). Overall, this stream of literature highlights the intricacies involved in adequately evaluating algorithmic forecasting performance from a human perspective, including specific human biases such as algorithm aversion (Prahl and Van Swol, 2017; Dietvorst et al., 2018) and general concerns about the role of humans in hybrid human-AI systems (Binns et al., 2018). Several older studies (e.g., Fildes et al., 2009) already conclude that, in general, human overrides tend to not improve forecast performance, but there may be specific scenarios (e.g., large-scale interventions based on additional expert knowledge or data issues) in which they do. For example, the covid-19 pandemic had large effects on retail sales forecasting and often required demand planners to include their judgment (Fildes, Kolassa, et al., 2022). From an IS perspective, this provides an interesting starting point to investigate if (and why/why not) humans are able to accurately perceive situations in which their intervention is beneficial. Thus, we now discuss existing research on the perception and use of information technology, adapted to the specific case of hybrid human-AI systems.

2.2 Research on the perception and use of information technology

Research on how humans perceive and use information technology has a deeply rooted tradition in IS that has yielded manifold research papers, which profusely cover a multitude of aspects and perspectives (Venkatesh et al., 2016, Venkatesh et al., 2012). In an endeavor to clarify our terminology and our understanding of the constructs that we employ in this study, we rely on the work of Venkatesh et al. (2016), who provide an overview of different research streams and discuss related attempts to synthesize different models within this research area. Figure 1 provides a visual overview of the subsequently discussed constructs (bold text in boxes), their operationalization in our survey (text in boxes, also see section 3.2), and their relations (arrows). Compared to the classical Technology Acceptance Model (TAM) (Venkatesh et al., 2016, Venkatesh et al., 2012), we additionally add a new construct to measure human perception of AI system performance, which is introduced in section 2.3.

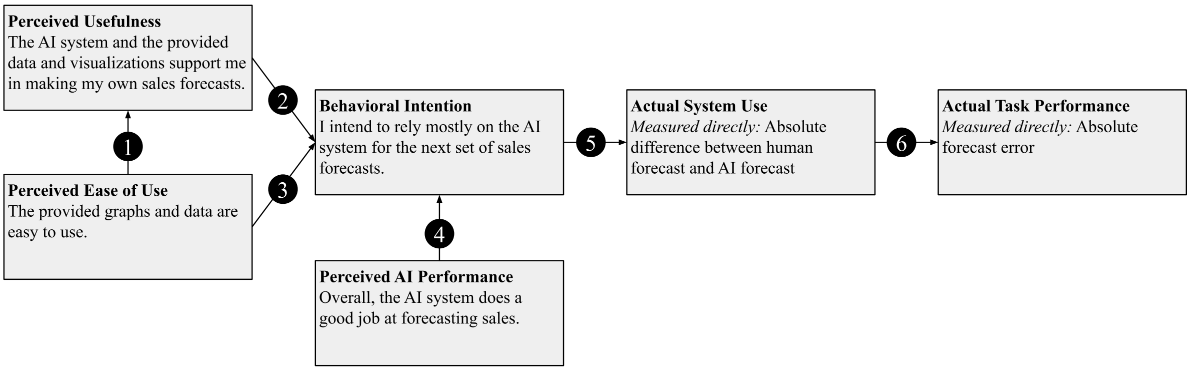

Perceived Usefulness
The AI system and the provided data and visualizations support me in making my own sales forecasts.
Perceived Ease of Use
The provided graphs and data are easy to use.
Behavioral Intention
I intend to rely mostly on the AI system for the next set of sales forecasts.
Perceived AI Performance
Overall, the AI system does a good job at forecasting sales.
Actual System Use
Measured directly: Absolute difference between human forecast and AI forecast
Actual Task Performance
Measured directly: Absolute forecast error
1
2
3
4
5
6

Figure 1. Overview of the conceptual model and operationalization.

Our primary concern is that the actual task performance of hybrid human-AI systems does not only depend on the technological components, but also on how humans use and interact with these technological components (actual system use). Even if a perfect AI system generates accurate sales forecasts, the hybrid human-AI system may produce suboptimal forecasts if the involved human actors incorrectly use these AI forecasts. How humans use and interact with technological components depends on their behavioral intentions, which in turn depend on how the respective technological components are perceived. This split between individual human perceptions, behavioral intentions, and actual behavior was proven to be useful through multiple previous studies, which confirmed that these are separate but related constructs (Venkatesh et al., 2012; Davis, 1989).

We focus on two specific types of human perceptions to capture the individual beliefs of the participants about the AI system. We examine if the participants perceive the AI system and the resulting data and visualizations to be (1) easy to use (perceived ease of use), and (2) useful for fulfilling their task (perceived usefulness). Previous studies that build on the TAM have confirmed that these are two separate, but related constructs (Venkatesh et al., 2016; Davis, 1989).

Perceived ease of use measures the degree to which participants perceive the provided data and visualizations of the AI forecasts easy to use. Perceived usefulness captures the extent to which participants perceive the provided technology (i.e., the AI forecast in the form of data and visualizations) to be useful for the task at hand (i.e., generating accurate sales forecasts) (Davis, 1989). Behavioral intention targets the participants' intention to use the provided technology (Venkatesh et al., 2008). In our study, this refers to the extent to which the participants intend to rely on the forecasts provided by the AI system. Related studies based on the classical TAM have shown that perceived usefulness is influenced by perceived ease of use (Relation 1), and that behavioral intention is influenced by both perceived usefulness (Relation 2) and perceived ease of use (Relation 3) (Venkatesh et al., 2016; Venkatesh et al., 2008).

Finally, we measure actual system use directly by calculating the human overriding of the AI forecast and we measure actual task performance directly by calculating the actual forecast error of the submitted forecast. Thus, as in typical retail organizations, we assume that the final decision on the sales forecasts for critical products remains a human responsibility. Consequently, the overall system performance (i.e., actual task performance) is directly influenced by the actual system use (Relation 6), which in turn is influenced by the user's behavioral intention (Relation 5).

2.3 Perceived AI performance

Human-AI systems differ from other IS in that the AI system could, in principle, also perform the task without humans. Consequently, we separately measure the human perception of AI system performance without human intervention. To this end, we add an additional construct to measure this perception in our study, which we call perceived AI performance. We expect that if the participants perceive the AI system to perform well even without their interventions, they have a stronger behavioral intention to follow the AI system's forecasts more closely (Relation 4).

3 Methodology and data

We conducted an observational online study in a controlled environment, in which we observed how 53 study participants behaved while trying to improve upon AI sales forecasts. The design of this study aims to emulate how employees within retail organizations interact with and perceive AI forecasting systems. In the following, we give an overview of the M5 data set, which is the basis for our sales forecasts and explain how the online study was conducted and how we analyzed the data.

3.1 Dataset: M5 retail sales data from Walmart

The M5 competition is the fifth iteration of a series of influential public forecasting competitions, named after Spyros Makridakis, which started in 1979 and whose results have fundamentally influenced the academic forecasting community (Nikolopoulos et al., 2020). The M5 competition employs a carefully curated and rich dataset from Walmart, one of the world's largest corporations. Thus, we expect the M5 to not only have a large academic impact, but also to be closely watched by the retail, wholesale, and distribution industry. The M5 dataset comprises daily sales of 3'049 household and food products from seven departments of ten different Walmart stores across Texas, Wisconsin, and California, covering a timeframe from January 29th, 2011 to June 19th, 2016. The products in the dataset are representative of a typical retailer, so that the time series display high variance and intermittency. Furthermore, the sales data is augmented with additional master data that is considered critical for retail sales forecasting. This includes hierarchical information on departments, product categories, and geographical areas (California, Texas, and Wisconsin), information about special events (e.g., Super Bowl, Valentine's Day), as well as price and promotion data.

We used LightGBM (Ke et al., 2017) to build a competitive forecasting system. The M5 uses the Weighted Root Mean Squared Scaled Error (WRMMSE) forecasting error measure to rate submitted forecasts (Hyndman and Koehler, 2006). The WRMMSE indicates how much the forecast improves compared to a naïve baseline forecasting. Using a separate hold-out test, we estimate the WRMMSE score of our system to be slightly below 0.56, indicating that our forecasting model reduces the weighted error of a naïve one-day-ahead forecast by 44%. This would place the forecasting system we employed in the top 50 scoring entries of the M5 competitions, which brings it close to the current state-of-the-art.

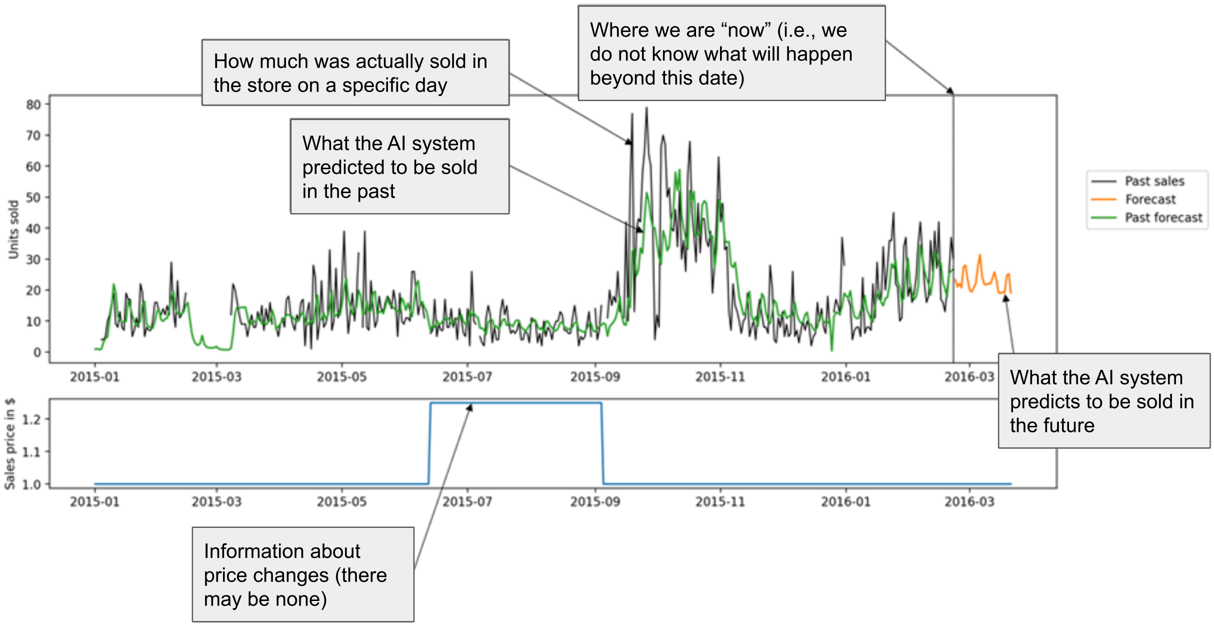
3.2 Online study design and operationalization

After ensuring that our forecasting model performed well when tested on the official hold-out test set, we proceeded to retrain this model using only data leading up to and including February 22nd, 2016.

Our goal was to generate a forecast for the next 28 days, covering February 23rd to March 21st, 2016, which we can then use in the subsequent online study. We used MTurk qualifications to carefully select participants, limiting the access to our study to people who (1) currently reside in the US, (2) have a US high school degree, (3) have a history of employment within the retail, wholesale, or distribution industry, and (4) have completed at least 500 MTurk tasks with an acceptance rate > 95%.

Since we could not ask participants to provide forecasts for all 30'490 sales timeseries — similar to how retailers cannot manually check millions of forecast each day — we narrowed down our focus to a set of 30 product-location combinations. A common approach in practice is to select only particularly critical products, which are characterized by having both a high sales volume and inaccurate forecasts in the past (Fildes, Ma, et al., 2022). Adopting this idea, we first selected the top 100 selling products for each of the ten stores. Then we simulated how our model would have forecasted these products during the year before the hold-out timeframe (February 22nd, 2015 to February 22nd, 2016). Next, we compared these forecasts with the actual sales during this period to calculate the past forecasting error. Finally, we selected the top three worst products per store, i.e., the three products that had the highest forecast errors. This resulted in a set of 30 products, which are both within the top-100 selling products per store and have high AI forecast errors. We then calculated forecasts for these 30 products for the targeted 28-day hold-out timeframe and generated corresponding visualizations to use in our online study.

Participants were first introduced to the experiment and asked demographic questions on their age, gender, education level, mathematical and statistical proficiency, and their experience with Walmart and the general retail industry. Participants then received an in-depth explanation of the task they are asked to perform (forecasting the sales of various products over 28 days), including several examples and test questions to ensure that the task is understood clearly. We also used the introductory part to explain participation remuneration, including a specific bonus incentive scheme: For each forecast, participants could gain 1 additional cent for every 10 units of improved accuracy to keep participants engaged throughout the study. Using pilot tests, we estimated the study to take 30-45 minutes and the total bonus payments to vary between 0.2$ and 0.6$, depending on the degree to which the participants managed to improve upon the AI forecasts. Participants were required to forecast 30 products, being provided (1) the average sales per 28 days in the past two years, (2) the forecast of the AI system for the next 28 days, and (3) a graphic visualization of the past sales since 2015, the past AI forecasts since 2015, the AI forecast for the next 28 days, and of price information (see Figure 2).



How much was actually sold in the store on a specific day

Where we are "now" (i.e., we do not know what will happen beyond this date)

What the AI system predicted to be sold in the past

What the AI system predicts to be sold in the future

Information about price changes (there may be none)

Figure 2. Example and explanation of forecast visualization provided to study participants.

All 30 products were shown to each participant in a random order. After a forecast was submitted, participants directly received feedback, including the actual sales, the AI forecast error, the error that participants themselves made, by how much they improved the forecast, and what their corresponding bonus is. After every 10 forecasts (three times in total), participants were asked a set of questions regarding their own perceived performance and that of the AI system (see Figure 1). To measure these perceptions and intentions, we questioned the participants about their agreement to a single statement per construct (see the text in the boxes in Figure 1) by using a 5-point Likert scale, covering 5 - Strongly agree, 4 - Agree, 3 - Neither agree nor disagree, 2 - Disagree, and 1 - Strongly disagree. Actual system use and actual task performance are measured directly. Actual system use is measured directly by calculating the difference between the participants' forecasts and AI forecasts, with lower differences indicating higher actual system use (i.e., participants follow the AI system's suggestions). Actual task performance is measured directly by calculating the absolute difference between the actual sales and the participants' forecasts. Lower values indicate higher actual task performance (since smaller differences indicate that the participant more accurately predicted the actual sales). Hence, actual system use and actual task performance are not based on participants' perception, but are measured directly by the difference between participants' forecasts and AI forecasts (respectively actual sales). Since AI systems perform, in general, better than humans in sales forecasting, following the AI forecast is expected to a better strategy for humans compared to making own forecasts. Nevertheless, it is generally possible for humans to improve upon AI forecasts. The possibility that human overriding could be beneficial is not excluded in this study: "actual system use" measures the degree of human overriding without prior assumptions.

Our research approach, a controlled online study, allows disregarding or controlling for several other factors, which have previously been hypothesized to affect the relationships in Figure 1. First, all participants completed the study in the same controlled online environment. We therefore do not consider any higher-level contextual factors (e.g., environmental or organizational attributes) or unexpected external events (Venkatesh et al., 2016). Second, all study participants were required to complete the same tasks with the same technology, therefore we do not consider task or technology attributes in our study. Third, we know what an ideally performed task would be since actual sales data is available. Thus, we focus our analysis on measuring actual task performance as the target construct, instead of using any indirect proxy. Furthermore, we required participants to provide information on personal attributes, allowing to control for age, gender, education level, mathematical and statistical expertise, and on the participants' familiarity with the context of the data (i.e., Walmart). Hence, these constructs are excluded in the conceptual model of Figure 1 (Venkatesh et al., 2016).

We aimed to receive 50 fully usable data sets for analysis (i.e., 50 x 30 = 1'500 total forecasts) and estimated that we might need to discard up to 20% of responses, based on recommended best practices to obtain valid and reliable data from Amazon MTurk (Rouse, 2015; Paolacci and Chandler, 2014). Therefore, we initially recruited 60 participants and later discarded 7 data sets, leading to a final number of 53 study participants and 1'590 forecasts. All analyses were performed in Python by using the scipy (version 1.4.1) and statsmodel libraries (version 0.11.1).

4 Results

The responses of all 60 participants were first manually inspected for outliers and particular response patterns to ensure that we only consider valid responses in our subsequent analyses (Rouse, 2015; Paolacci and Chandler, 2014). We checked that the study participants spent a reasonable amount of time (about 30-60 seconds) on each forecast, leading to the exclusion of four participants from our analysis, who proceeded much too quickly through the study. Furthermore, we excluded three participants, who responded in clear patterns (e.g., always alternating between 150 and 200 units for every forecast). After dropping a total of seven participants from the original set of 60, we describe our results for the remaining 53 participants in the following (see Table 1 for a demographic overview).

Gender:		**Age:**	
Male	28 (52.8%)	51 or higher	8 (15.1%)
Female	25 (47.2%)	30 to 50	31 (58.5%)
		29 or lower	14 (26.4%)
How often do you go to Walmart?			
Never	0 (0.0%)	**Highest completed education level:**	
Rarely (less than 3 visits per year)	2 (3.8%)	Graduated from high school	16 (30.2%)
Sometimes (about 3-6 times per year)	12 (22.6%)	Graduated from college	27 (50.9%)
Often (about once a month)	24 (45.3%)	Completed graduate school	10 (18.9%)
Very often (multiple times per month)	15 (28.3%)		

Table 1. Participant overview (53 participants total).

4.1 Effects of human interventions on forecast accuracy

In general, human adjustments to the AI forecasts led to lower forecast accuracy in our study (see Figure 3). The left graph in Figure 3 shows the distribution of how the forecasts of the 30 products were affected through human intervention. The horizontal axis displays the average improvement over the AI system's sales forecast that was achieved across all study participants. Negative numbers indicate that, on average, participants made the forecast less accurate. For only 7 out of 30 products the forecasts were improved, whereas for 23 products the forecasts were made worse. On average, human intervention made the sales forecast less accurate by 16.64 units (vertical dashed line in left depiction of Figure 3) when compared to the original AI forecast. The resulting differences (i.e., human improvement over AI system) approximately follow a normal distribution with (slightly) negative mean. The fact that humans were able to improve AI predictions for seven products is thus due to ordinary random influences.

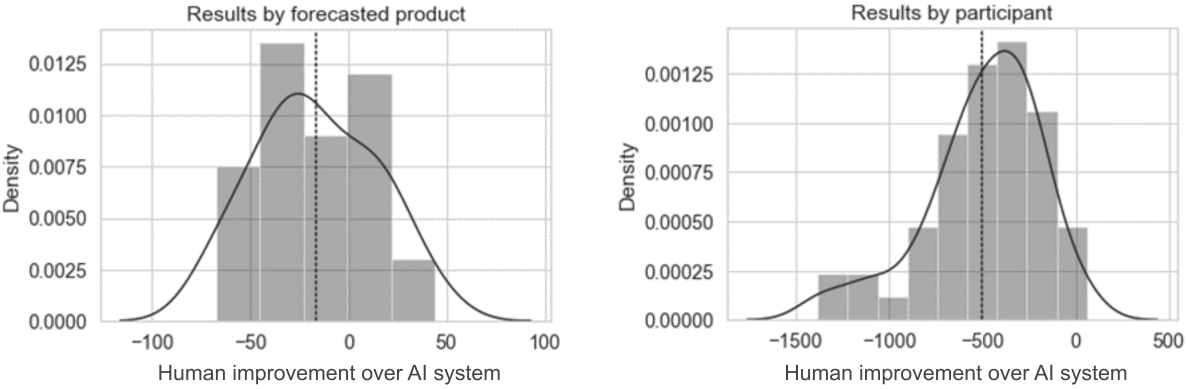


Figure 3. Human improvement over AI system by product (left) and participant (right).

The right graph in Figure 3 shows the distribution of the individual performance of the 53 participants. The horizontal axis displays the total improvement that each participant achieved, summing up the changes in accuracy between their forecast and the AI forecast for all 30 products. Only one out of 53 participants managed to have an overall positive impact, and on average the combined forecasts of each study participant were less accurate by 499 units (vertical dashed line in right depiction of Figure 3) when compared to the AI forecasts. These differences approximately follow a normal distribution with a longer left-hand tail (left skewed). We checked the negative outliers manually, confirming that these are indeed valid sales forecasts, where participants simply predicted a trend to go in the wrong direction.

We conducted regression analyses to conclude that no participant characteristics had a significant impact on participants' performance in this study. Specifically, we performed simple linear regressions of several participant characteristics as independent control variables (see Table 2) on their overall improvement in forecast accuracy over the AI system (i.e., the dependent variable).

Dependent variable	Control / Independent variable	R^2	F-value	p-value
Participant improvement over AI	Age	0.040	2.124	0.151 (p>0.1)
	Gender	0.002	0.092	0.763 (p>0.1)
	Frequency of Walmart visits	0.001	0.047	0.829 (p>0.1)
	Highest education level	0.006	0.283	0.597 (p>0.1)
	Self-perception of mathematical skill	0.032	1.669	0.202 (p>0.1)
	Self-perception of statistical skill	0.013	0.672	0.416 (p>0.1)

Table 2. Regression analysis between participant characteristics and forecast accuracy.

No significant regression equation was found (all p > 0.1). We therefore conclude that neither gender, nor age, nor education level, familiarity with Walmart, or statistical and mathematical proficiencies of a participant allow to make any prediction on how a participant will perform. Consequently, we attribute the few positive results (on average only 7 out of 30 products were improved and only 1 out of 53 participants achieved an overall positive impact) to random chance and conclude that, in general, participants have failed to improve upon the AI forecasts. We also calculated the average forecasts per product over all 53 participants. These forecasts are much closer to the AI forecasts, as some participants correct upwards and some participants correct downwards. However, no averaged participant forecast managed to significantly improve the AI forecasts.

4.2 Changes over time and human perceptions

The order of products was randomized for each participant, so any observed differences over time cannot be attributed to specific products or their order in the study. To understand changes in perception, behavior, and performance over time, the mean values and standard deviations of the observed variables in three different rounds (i.e., after participants provided 10, 20, and 30 total forecasts respectively) were calculated (see Table 3). For example, the average agreement to the perceived usefulness statement (see Figure 1) was 3.53 (5 = strongly agree, 1 = strongly disagree) in round 1 (survey after 10 forecasts), which then increased to 4.08 in round 3 (survey after 30 forecasts).

	Round 1		Round 2		Round 3		t-test (round 1 vs. round 3)	
Variable	Mean	SD	Mean	SD	Mean	SD	t	p-value
Participant improvement over AI	-193.3	164.0	-189.3	161.1	-116.8	181.1	-2.21	0.031*
Perceived ease of use	3.74	0.88	3.72	0.77	3.72	0.91	0.14	0.886
Perceived usefulness	3.53	0.89	3.92	0.65	4.08	0.64	-4.37	0.000***
Perceived AI performance	2.83	1.12	3.76	0.87	3.68	0.89	-4.73	0.000***
Behavioral intention	3.38	0.97	3.53	0.89	3.74	0.96	-2.31	0.024*
Actual system use	327.7	238.4	291.2	190.0	219.4	193.0	3.04	0.004**
Actual task performance	1154.3	164.0	840.5	257.8	701.8	181.0	13.09	0.000***
Perception of own performance	3.30	0.93	2.89	1.12	2.64	1.15	3.74	0.000***
Perceived potential to improve upon AI forecast	3.43	0.84	2.60	1.13	2.57	1.05	6.08	0.000***

Table 3. Over time analysis and t-tests (: p < 0.05; **: p < 0.01; ***: p < 0.001).*

Additionally, since the same population is queried multiple times at different points in time, we conducted dependent paired sample t-tests to confirm that the observed differences are statistically significant (last two columns in Table 3). Note that participant improvement over AI is always negative, since participants generally made the forecasts less accurate. Furthermore, actual system use is measured as the sum of the absolute differences between participant forecasts and the AI forecasts, so that lower values correspond to higher system use (i.e., participants more closely follow the AI system's suggestions). Similarly, actual task performance is measured as the sum of absolute forecast errors, so that lower numbers correspond to lower errors and thus to better task performance.

The results from Table 3 allow to confirm that even in the later rounds, participants did not manage to significantly improve the AI forecasts, since they still performed significantly worse in round 3 (116.8 units higher forecast error on average). While participants significantly increase their improvement over the AI system (t = -2.21, p = 0.031) from round 1 (M = -193.3, SD = 164.0) to round 3 (M = -116.8, SD = 181.1), this improvement also coincides with a significantly increased actual system use (t = 3.04, p = 0.004) from round 1 (M = 327.7, SD = 238.4) to round 3 (M = 219.4, SD = 193.0). Consequently, the increased accuracy of the human forecasts over time is not due to the participants being able to outperform the AI system consistently but a result of the participants choosing to follow the AI system more closely in later rounds. To analyze this change in behavior and associated perceptions over time, we first confirm that the hypothesized relations from our basic conceptual model (see Figure 1) generally hold in our study. Table 4 shows the results of regression analyses, using the same numbering as Figure 1. Note that these regressions cover three responses per participant (After 10, 20, and 30 forecasts each), so that the sample size is n=159 (= 3 x 53). For each relation we also used simple moderation analyses to confirm that none of the control variables age, gender, education, mathematical and statistical expertise, and familiarity with the context (i.e., Walmart) had a significant influence.

ID	Predictor variable	Dependent variable	Coefficient	R^2	F-value	p-value
1	Perceived ease of use	Perceived usefulness	0.308	0.116	20.64	0.000***
2	Perceived usefulness	Behavioral intention	0.330	0.072	12.09	0.000***
3	Perceived ease of use	Behavioral intention	0.335	0.090	15.54	0.000***
4	Perceived AI performance	Behavioral intention	0.387	0.179	34.21	0.000***
5	Behavioral intention	Actual use	290.543	0.195	12.32	0.000***
6	Actual use	Actual performance	-0.323	0.246	16.65	0.000***

*Table 4. Regression analysis for the relations in Figure 1 (***: p < 0.001).*

The typical TAM-relations (1, 2, 3, and 5) are confirmed: If the tool (i.e., the AI forecasts and related visualizations) is perceived to be easy to use, it is also perceived to be useful (1). If the forecasting tool is perceived to be useful, participants intend to increasingly rely on it for their own forecasts (2). If the forecasting tool is perceived to be easy to use, participants intend to increasingly rely on it for their own forecasts (3). If participants intend to rely on the forecasting tool for their own forecasts, they tend to differ less from the AI forecast, which was employed to measure actual use (5). Furthermore, the perceived accuracy of the AI system strongly influenced the participants intention to increasingly rely on the AI system (4). Additionally, a consequence of the preceding analysis is confirmed: The AI system tends to produce more accurate sales forecasts than the human study participants, so that participants who deviate less from the AI forecasts (corresponding to actual use) tend to perform better (actual performance) in terms of overall forecast accuracy (6). The negative coefficient (-0.323) means that higher actual use corresponds to lower errors, i.e., better actual performance.

These perceptions and behavior patterns change over time. Table 3 shows that while there is no significant difference for the perceived ease of use (t = 0.14, p = 0.886) between round 1 (M = 3.74, SD = 0.88) and round 3 (M = 3.72, SD = 0.91), there is indeed a significant improvement in perceived usefulness (t = -4.37, p < 0.001) from round 1 (M = 3.53, SD = 0.89) to round 3 (M = 4.08, SD = 0.64). Similarly, we find a significant increase in perceived AI performance (t = -4.73, p < .001) from round 1 (M = 2.83, SD = 1.12) to round 3 (M = 3.68, SD = 0.89). Over the course of the study, participants generally found the provided AI forecasts to be significantly more useful in later rounds, when they also tended to rate the AI system performance significantly better.

As depicted in the regression analyses in Table 4, changes in perceived usefulness and perceived AI performance also lead to changes in behavioral intention. The over-time analysis from Table 3 confirms this hypothesis: Significant increases in the participants behavioral intention (t = -2.31, p = 0.024) to follow the AI forecasts more closely from round 1 (M = 3.34, SD = 0.97) to round 3 (M = 3.74, SD = 0.96) can be observed. This intention is realized through significant increases in actual use (t = 3.04, p = 0.004) from round 1 (M = 327.7, SD = 238.4) to round 3 (M = 219.4, SD = 193.0), leading to significant increases in actual performance (t = 13.09, p < 0.001) from round 1 (M = 1154.3, SD = 164.0) to round 3 (M = 701.8, SD = 181.0).

Additionally, participants were asked to directly provide feedback on how they judge their own performance (perception of own performance) and if they believe they can improve the AI forecast (perceived potential to improve upon AI forecast). Consistent with the preceding analysis, we find that initially humans overestimate their own performance and their potential to improve the AI forecasts (see Table 3). Statistically, this corresponds to a significant decrease in the participants perception of own performance (t = 3.74, p < 0.001) from round 1 (M = 3.30, SD = 0.93) to round 3 (M = 2.64, SD = 1.15) along with a significant decrease in the perceived potential to improve upon the AI forecast (t = 6.08, p < 0.001) from round 1 (M = 3.43, SD = 0.84) to round 3 (M = 2.57, SD = 1.05).

5 Discussion

The results indicate that human interventions have a significant potential to actively impair the overall system performance of hybrid human-AI systems when compared to fully autonomous AI forecasting systems without any human supervision. This result contributes to the debate on hybrid human-AI systems and the effects of AI systems on human work and AI-assisted decision making (Gregory et al., 2021; Rai et al., 2019). Multiple recently published research projects describe fully or partially automated AI systems that perform very well in complex business tasks (e.g., Mehdiyev and Fettke, 2020; Lebovitz et al., 2022). One central question for any such system is whether human supervision and intervention still positively contribute to the overall performance. On the one hand, modern AI systems often surprisingly manage to consistently outperform humans in complex tasks (Silver et al., 2016). These results question the general need to involve human actors in complex decision systems. On the other hand, Dellermann et al. (2019) argue that human intelligence and machine intelligence have complementary strengths, which in principle allow human-AI hybrids to outperform pure AI systems. Research on demand forecasting in supply chain planning supports this perspective, finding that humans do not tend to outperform AI systems in general, while simultaneously arguing that there exist specific circumstances in which human overrides and other interventions (e.g., manual inputs to the model training data) are beneficial (Fildes, Kolassa, et al., 2022; Fildes et al., 2009; Fildes and Goodwin, 2007; Prahl and Van Swol, 2017). Against this backdrop, our findings are in line with studies from other areas like medicine, where performance issues occurred when well-performing AI systems were implemented into the organization and combined with human judgment (Lebovitz et al., 2022; Lebovitz et al., 2021).

Our contribution to this debate is an in-depth analysis of an archetypal example of a central business task in organizations. In our study, hybrid human-AI systems are consistently outperformed by pure AI systems without human intervention. Instinctively, this may be surprising, since human decision makers could always choose not to override the AI system's decision at all. Consequently, the hybrid system performance is partially attributable to human misperceptions. The new theoretical contribution of our research is the simultaneous investigation of overall system performance and human perceptions over time. Our analysis confirms that humans underestimate the performance of AI systems and overestimate their own potential to improve upon AI forecasts, building on previous studies that have shown human judgment to be skewed against the AI system (Dietvorst et al., 2015; Prahl and Van Swol, 2017).

Furthermore, we show that the initial false perceptions of the AI system performance are not set in stone but can quickly change even throughout the course of a short 45-minute study. Our results also demonstrate that the TAM can be extended, by adding the perceived AI performance construct, to trace human perception and behavior in hybrid human-AI systems over time.

The observed change in perceptions over time also entails an important methodological implication of our results: When studying hybrid human-AI systems, researchers need to account for potentially rapidly changing perceptions over time. Our results would have been completely different had we just measured once after 10 rounds, compared to if we had measured just once after 30 rounds (see Table 3). Since human perceptions directly affect human actions, which in turn impact system performance, such fluctuations can also occur in studies that only measure overall system performance. Consequently, research on behavior in hybrid human-AI systems should sample and compare data from multiple points in time, since investigations will likely encounter biased and potentially changing human perceptions.

In terms of practical implications, organizations can use our results and analyses to support the design and development of hybrid human-AI decision systems for complex business processes. We show that the overall performance of such systems does not just depend on the AI system performance, but also on how this AI system is perceived by the involved human actors. Consequently, developers and managers should not just measure and optimize the hybrid system performance but should also try to measure perceptions of the involved human actors. Ideally, humans are aware of both their own and the AI system's strengths and limitations, allowing them to carefully identify situations in which they have a good chance to improve upon the AI system's decisions so that the respective strengths are combined effectively. However, it has been shown that humans have trouble evaluating for which tasks they can rely on their own judgment and for which it would be better to adopt the AI output (Fügener et al., 2022).

We expect activities such as frequent and transparent feedback on the results of human interventions to help decreasing extant human biases in human-AI hybrid systems and consider investigations in this field to be a promising avenue for future research. Human interventions can also prevent AI biases like feedback bias in human-AI systems by breaking the vicious circle in cases where the AI system's output affects its training data and amplifies an initial small bias (Fahse et al., 2021).

The preceding discussion should be interpreted within the idealized context of this study, which represents both the limitations of this study and an avenue for future research. The AI system and the humans had access to identical data in this study. In contrast, humans in real-world organizations can cooperate and access additional information (e.g., by calling a store manager) which might help improving their performance (Petropoulos et al., 2022; Fildes, Ma, et al., 2022). Thus, our results do not imply that humans should in general not be involved in sales forecasting processes in real-world organizations, where there may be much more information available to decision makers (e.g., local sports events, upcoming bulk orders, store renovations) compared to what is available to the AI system. Furthermore, the participants in our experiment differ from actual decision makers with domain-specific expert knowledge in terms of expertise, knowledge, motivation, and incentives. Finally, our results are limited by only considering one specific type of human intervention, that is, direct overriding of AI forecasts through human judgmental adjustments. Other inputs, such as manually curating and correcting training data for the AI system or manually tracking data quality issues are not considered but might lead to different results. In this context, it subsequent studies could investigate special scenarios in which human overriding is beneficial (e.g., large-scale interventions based on additional expert knowledge). For example, the covid-19 pandemic had large effects on retail sales forecasting and often required demand planners to include their judgment (Fildes, Kolassa, et al., 2022). Furthermore, the short temporal horizon of this study, which makes statements about long-term effects of human overriding in hybrid systems impossible, calls for further studies over a longer period of time.

Despite these limitations, our study highlights the significant threat of suboptimal human-AI system performance due to false human perceptions, which need to be specifically addressed. Having confirmed this issue in an abstract setting, future research could aim to replicate this study on a larger scale with actual domain experts in a real-world setting to overcome the discussed limitations. Conducting more in-depth investigations of real-world organizations to understand which types of human interventions are particularly helpful or particularly harmful (Fildes and Goodwin, 2007) are promising to overcome both human misperceptions (Dietvorst et al., 2018) as well as AI biases (van Giffen et al., 2022), and thus support the design of hybrid human-AI decision systems.

References

Binns, R., Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018). "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC, Canada.

Burton, J., Stein, M., and Jensen, T. (2020). "A systematic review of algorithm aversion in augmented decision making," *Journal of Behavioral Decision Making 33* (2), 220–239.

Davis, F. (1989). "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *Management Information Systems Quarterly* 13 (3), 318-340.

Dellermann, D., Ebel, P., Söllner, M., and Leimeister, J. M. (2019). "Hybrid Intelligence," *Business & Information Systems Engineering* 61 (5), 637-643.

Dietvorst, B., Simmons, J., and Massey, C. (2015). "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology-General* 144 (1), 114-126.

Dietvorst, B., Simmons, J., and Massey, C. (2018). "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Management Science 64* (3), 1155–1170.

Di Pillo, G., Latorre, V., Lucidi, S., and Procacci, E. (2016). "An application of support vector machines to sales forecasting under promotions," *4OR -A Quarterly Journal of Operations Research* 14 (3), 309-325.

Duan, Y., Edwards, J., and Dwivedi, Y. (2019). "Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda," *International Journal of Information Management* 48, 63-71.

Fahse, T., Huber, V., and van Giffen, B. (2021). "Managing Bias in Machine Learning Projects," in: *Wirtschaftsinformatik 2021 Proceedings*.

Fildes, R. and Goodwin, P. (2007). "Against Your Better Judgment? How Organizations Can Improve Their Use of Management Judgment in Forecasting," *Interfaces* 37 (6), 570-576.

Fildes, R., Goodwin, P., Lawrence, M., and Nikolopoulos, K. (2009). "Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning," *International Journal of Forecasting* 25 (1), 3-23.

Fildes, R., Kolassa, S., and Ma, S. (2022). "Post-script-Retail forecasting: Research and practice," *International Journal of Forecasting* 38 (4), 1319–1324.

Fildes, R., Ma, S., and Kolassa, S. (2022). "Retail forecasting: Research and practice," *International Journal of Forecasting* 38 (4), 1283-1318.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2022). "Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation," *Information Systems Research 33* (2), 678–696.

Gregory, R., Henfridsson, O., Kaganer, E., and Kyriakou, H. (2021). "The Role of Artificial Intelligence and Data Network Effects for Creating User Value," *Academy of Management Review 46* (3), 534–551.

Gür Ali, Ö., Sayın, S., Van Woensel, T., and Fransoo, J. (2009). "SKU demand forecasting in the presence of promotions," *Expert Systems with Applications* 36 (10), 12340-12348.

Gür Ali, Ö. and Yaman, K. (2013). "Selecting rows and columns for training support vector regression models with large retail datasets," *European Journal of Operational Research* 226 (3), 471-480.

Haki, K., Beese, J., Aier, S., and Winter, R. (2020). "The Evolution of Information Systems Architecture: An Agent-Based Simulation Model," *Management Information Systems Quarterly* 44 (1), 155-184.

Hyndman, R. and Koehler, A. (2006). "Another look at measures of forecast accuracy," *International Journal of Forecasting* 22 (4), 679-688.

Jordan, M. I. and Mitchell, T. M. (2015). "Machine learning: Trends, perspectives, and prospects," *Science*, 349 (6245), 255-260.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in: Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (Eds.) *Advances in Neural Information Processing Systems* 30.

Kolassa, S. (2020). "Will Deep and Machine Learning Solve Our Forecasting Problems?," *Foresight: The International Journal of Applied Forecasting* 57, 13-18.

Lebovitz, S., Levina, N., and Lifshitz-Assa, H. (2021). "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," *Management Information Systems Quarterly 45* (3), 1501–1526.

Lebovitz, S., Lifshitz-Assaf, H., and 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.s

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes 151*, 90–103.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022a). "The M5 competition: Background, organization, and implementation," *International Journal of Forecasting* 38 (4), 1325–1336.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2022b). "M5 accuracy competition: Results, findings, and conclusions.," *International Journal of Forecasting* 38 (4), 1346–1364.

Mehdiyev, N. and Fettke, P. (2020). "Prescriptive Process Analytics With Deep Learning and Explainable Artificial Intelligence," in: *European Conference on Information Systems*, Marrakesh, Morocco.

Mendling, J., Decker, G., Hull, R., Reijers, H., and Weber, I. (2018). "How do Machine Learning, Robotic Process Automation, and Blockchains Affect the Human Factor in Business Process Management?," *Communications of the Association for Information Systems*, 297-320.

Nikolopoulos, K., Thomakos, D., Katsagounos, I., and Alghassab, W. (2020). "On the M4.0 forecasting competition: Can you tell a 4.0 earthquake from a 3.0?," *International Journal of Forecasting* 36 (1), 203-205.

Paolacci, G. and Chandler, J. (2014). "Inside the Turk," *Current Directions in Psychological Science* 23 (3).

Petropoulos, F., Apiletti, D., Assimakopoulos, V., et al. (2022). "Forecasting: theory and practice," *International Journal of Forecasting* 38 (3), 705–871.

Prahl, A. and Van Swol, L. (2017). "Understanding algorithm aversion: When is advice from automation discounted?," *Journal of Forecasting* 36 (6).

Rai, A., Constantinides, P., and Sarker, S. (2019). "Editor's comments: next-generation digital platforms: toward human–AI hybrids," *Management Information Systems Quarterly 43* (1), iii – ix.

Ramos, P., Santos, N., and Rebelo, R. (2015). "Performance of state space and ARIMA models for consumer retail sales forecasting," *Robotics and Computer-Integrated Manufacturing* 34, 151-163.

Rouse, S. V. (2015). "A reliability analysis of Mechanical Turk data," *Computers in Human Behavior* (43), 304-307.

Sagaert, Y., Aghezzaf, E., Kourentzes, N., and Desmet, B. (2018). "Tactical sales forecasting using a very large set of macroeconomic indicators," *European Journal of Operational Research* 264 (2), 558-569.

Shollo, A., Hopf, K., Thiess, T., and Müller, O. (2022). "Shifting ML value creation mechanisms: A process model of ML value creation," *The Journal of Strategic Information Systems 31* (3).

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). "Mastering the game of Go with deep neural networks and tree search," *Nature* 529 (7587), 484-489.

Syntetos, A., Babai, Z., Boylan, J., Kolassa, S., and Nikolopoulos, K. (2016). "Supply chain forecasting: Theory, practice, their gap and the future," *European Journal of Operational Research* 252 (1), 1-26.

Szajna, B. and Scamell, R. (1993). "The Effects of Information-System User Expectations on Their Performance and Perceptions," *Management Information Systems Quarterly* 17 (4), 493-516.

van Dis, E., Bollen, J., Zuidema, W., van Rooij, R., and Bockting, C. (2023). "ChatGPT: five priorities for research," *Nature 614* (7947), 224–226.

van Giffen, B., Herhausen, D., and Fahse, T. (2022). "Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods," *Journal of Business Research* 144, 93–106.

Veiga, C. P. D., Veiga, C. R. P. D., Puchalski, W., Coelho, L. D. S., and Tortato, U. (2016). "Demand forecasting based on natural computing approaches applied to the foodstuff retail segment," *Journal of Retailing and Consumer Services* 31, 174-181.

Venkatesh, V., Brown, S., Maruping, L., and Bala, H. (2008). "Predicting Different Conceptualizations of Systems Use: The Competing Roles of Behavioral Intention, Facilitating Conditions, and Behavioral Expectation," *Management Information Systems Quarterly* 32 (3), 483-502.

Venkatesh, V., Thong, J., and Xu, X. (2016). "Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead," *Journal of the Association for Information Systems* 17 (5), 328-376.

Venkatesh, V., Thong, J., and Xu, X. (2012). "Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology," *Management Information Systems Quarterly* 36 (1), 157-178.

Von Krogh, G. (2018). "Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing," *Academy of Management Discoveries* 4 (4), 404-409.

Zhang, G. P. and Qi, M. (2005). "Neural network forecasting for seasonal and trend time series," *European Journal of Operational Research* 160 (2), 501-514.