Exploring the Synergies in Human-AI Hybrids: A Longitudinal Analysis in Sales Forecasting

Completed Research Full Paper

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

Anuschka Schmitt
University of St. Gallen
anuschka.schmitt@unisg.ch

Abstract

Despite the promised potential of artificial intelligence (AI), insights into real-life human-AI hybrids and their dynamics remain obscure. Based on digital trace data of over 1.4 million forecasting decisions over a 69-month period, we study the implications of an AI sales forecasting system's introduction in a bakery enterprise on decision-makers' overriding of the AI system and resulting hybrid performance. Decision-makers quickly started to rely on AI forecasts, leading to lower forecast errors. Overall, human intervention deteriorated forecasting performance as overriding resulted in greater forecast error. The results confirm the notion that AI systems outperform humans in forecasting tasks. However, the results also indicate previously neglected, domain-specific implications: As the AI system aimed to reduce forecast error and thus overproduction, forecasting numbers decreased over time, and thereby also sales. We conclude that minimal forecast errors do not inevitably yield optimal business outcomes when detrimental human factors in decision-making are ignored.

Keywords

Machine learning, sales forecasting, AI-assisted decision making, digital trace data, reliance, longitudinal.

Introduction

Advances in artificial intelligence (AI) research combined with increased commercial accessibility have led contemporary organizations to seek to integrate AI capabilities into existing human tasks and core business processes (Mendling et al., 2018; Von Krogh, 2018). Such integration can, on the hand, automate labor-intensive, repetitive tasks which would be failure-prone under human execution (Dixon et al., 2020). On the other hand, AI systems offer to augment organizational tasks that underlie human decision-making, such as creative design (e.g., Lysyakov and Viswanathan, 2022) or forecasting tasks (e.g., Jussupow et al., 2021). Such collaboration between AI and human jointly arriving at a superior outcome than one of the two on their own is commonly referred to as human-AI hybrids (Rai et al., 2019). In the context of a forecasting task, AI-assisted decision-making promises to improve human forecasting alone by considering relevant decision-making factors in a more holistic and exhaustive manner (Fildes et al., 2022a).

However, the implementation of AI and reaping potential complementarities of human-AI hybrids is not as straightforward: Information systems in organizations can be viewed as complex sociotechnical systems in which human and technological elements interact to process information (Mendling et al., 2018; Haki et al., 2020). In the context of AI-assisted decision-making, previous studies have illustrated how human assessment of and miscalibrated reliance on AI output can potentially derail AI augmentation (Logg et al., 2019; Fügener et al., 2022; Dietvorst et al., 2015). Successfully incorporating AI systems into the business processes of organizations is a major challenge for enterprises and can become costly and damaging when systems are not appropriately used (Dellermann et al., 2019). Extant studies have demonstrated how human overriding of AI output led to greater forecast error, i.e., the difference between forecasted and actual results. More so, users often underestimated the AI system's accuracy and overestimated their own potential to improve upon the AI forecasts (Fügener et al., 2022). We see that most research has studied human-AI hybrids in hypothetical, controlled laboratory settings. While these studies have put forward important causal relationships on the implications of AI augmentation, little is known about human reliance on AI output in a real-life context—where the decision to rely on AI and related costs of system error can have actual, detrimental implications to the human user. In addition, a plethora of cross-sectional studies leave unsettled how human-AI hybrids develop over a longer period.

This study seeks to fill these gaps by studying a human-AI hybrid where an AI system is introduced for a forecasting task in a bakery sales context. Access to digital trace data on employees' forecasting decisions before and after implementation of the AI sales forecasting system provides us an opportunity to study human reliance on the AI output, i.e., to what extent AI forecasting recommendations are overridden by the human decision-maker. In this study we seek to answer the following research questions (RQs):

RQ1: How does human overriding of AI forecasts impact hybrid sales forecasting performance?

RQ2: How do human overriding of AI forecasts and hybrid human-AI forecasting performance develop over time?

This study builds upon computational and process-driven theory development (Berente et al., 2019); Miranda et al., 2022; Grisold et al., 2020). We study a bakery business which introduced an AI sales forecasting system in 2020, providing sales data for 27 branches for the period of 42 months before the AI system launch and 27 months after the launch. The dataset includes 1.4 million forecasting decisions and the complete history of forecasts for 528 products for 69 months (5.7 years). Informed by AI reliance literature, this study's iterative empirical-inductive approach uses AI forecasts, hybrid forecasts, and actual sales data to calculate forecast errors in human-AI hybrids. Based on the analysis of forecast errors, we focus on the impact of human overriding on hybrid system performance and find that the overriding behavior changed over time. In a subsequent step, an additional variable (waste ratio) is analyzed to better understand observed behavior.

In previewing the results, we find that human intervention deteriorates sales forecasting performance. The more humans overrode AI forecasts, the worse forecast errors became. This finding is in line with a study by Fügener et al (2022), who found that humans have little idea of their own capabilities, i.e., when performing better than the AI. Interestingly, we find that decision-makers initially overrode the AI, yet quickly started to rely on the AI for a longer period of time. This phase of reliance was followed by increasing overriding again, which plateaued 18 months after the AI forecasting system was introduced. We explain these dynamics in overriding considering the AI's underlying reinforcement-based nature: Users' beginning reliance led to ultimately lower forecast error, yet also a drop in sales. While forecast error improved and overproduction could be reduced, total sales also decreased. At first glance, these results confirm the notion that AI outperform humans in forecasting tasks. As forecasting accuracy is improved by having a high match of forecasted and sold products, the reinforcement model of the AI learns to avoid high forecasts which exceed the actual number of products sold. In turn, however, bakeries lost important purchasing opportunities by forecasting a lower number of products in total. The increasing overruling by decision-makers can be viewed as a response to the decreasing production numbers forecasted by the AI system.

These findings have both theoretical and practical implications. First, we contribute to research exploring human-AI hybrids in work-related domains. We warrant a fresh perspective using longitudinal sales and forecasting data to quantitatively examine the hybrid forecasting error and human reliance on AI forecasts in response to the introduction of an AI sales forecasting system in a real-life organizational setting. In this context, we study the effects on the sales forecasting process. We view the launch of the AI system as a shock on this process and explore the effects by measuring different metrics, i.e., digital trace data that is produced by the process. Second, by showing the temporal progression of these metrics, we contribute to a deeper understanding of the resulting dynamics when AI systems are introduced in socio-technical systems. These findings can be used to derive general theories to explain AI system introduction and longitudinal human-AI hybrids. Our results highlight that purely considering accuracy and related performance metrics in human-AI hybrids is harmful as it fails to consider the complexity and human factors detrimental to such real-life contexts. Organizations can use our findings to better plan and vet against the unintended consequences of AI forecasts. Understanding how decision-makers respond to AI forecasts may help designing holistic forecasting strategies to optimize performance and revenue beyond forecasting accuracy, since minimal forecast error does not inevitably equal optimal real life decision-making.

Theoretical Background

Human-AI Hybrids in Organizations

AI is defined as techniques for machines to imitate human behavior deployed to match or surpass human decision making (Russel and Norvig, 2021). Seminal information systems literature discusses the effects of AI on human work and augmenting human decision-making (Raisch and Krakowski, 2020; Gregory et al., 2021; Lebovitz et al., 2022). While some of this research examines the impact of AI on the automation of human labor, i.e., operating without human oversight or intervention, more recent research focuses predominantly on how to arrive at the best collaboration of humans and AI to augment human labor. The goal of such collaboration is to leverage combined performance or to perform tasks that the AI system or the human decision-makers could not solve alone (Bansal et al., 2019, Chakraborti and Kambhampati, 2018). In line with this thinking, Rai et al. (2019) introduced the concept of human-AI hybrids, coined as a joint task allocation between AI and human agents according to corresponding competencies. Relatedly, literature around hybrid intelligence (e.g., Dellermann et al., 2019) discusses this concept of joint decision-making to be more satisfactory or even necessary than having humans or AI work in isolation.

Prior research on human-AI hybrids in organizations, however, points towards the difficulties of reaping the potential benefits of human-AI hybrids. First, the AI system might not be used for the appropriate task instances or at the right point in a decision-making scenario. Although a collaboration of humans and AI system might be more desirable in theory, e.g., by achieving greater accuracy in forecasting or classification tasks, humans have difficulties to assess for which task instances relying on the AI might be beneficial. Relatedly, previous studies have shown that humans struggle to recognize when an AI is erroneous, i.e., in which task instances relying on the AI would be harmful for task performance (Bansal et al., 2019). Fügener et al. (2022), for instance, demonstrate that humans have difficulties assessing for which task instances they are confident to rely on their own judgement and for which it would make sense to consider the AI output. In that sense, combined performance in AI-advised human decision making depends on the degree to which human decision-makers understand the AI system's frontier. When the frontier is misjudged, decision-makers (1) trust the AI system although its recommendation is erroneous or (2) do not trust the AI system although it makes a correct recommendation (Bansal et al., 2019; Schmitt et al., 2021). These cooperation errors can lead to poor decisions and economic harm. Second, the introduction of an AI system into a decision-making process can have unintended consequences on human behavior. Lebovitz et al. (2022) studied the introduction of an AI system in radiology departments for diagnostic judgement. They found that the introduction of the AI system not only introduced uncertainty as the AI output differed from decision-makers' initial judgement without offering an explanation, but that decision-makers became unengaged with the AI by ignoring its output or relying on it without consciously engaging with it.

A promising application domain for human-AI hybrids is that of forecasting. Several empirical studies have investigated the impact of human judgmental adjustments on AI sales forecasts, for instance in supply chain planning (Petropoulos, 2022; Fildes et al., 2009; Fildes and Goodwin, 2007). Studies highlight the difficulties of properly assessing algorithmic prediction performance from a human perspective, including specific human behavior such as algorithm aversion (Prahl and Van Swol, 2017) and general investigations regarding the role of humans in human-AI hybrids (Binns et al., 2018). In a seminal work on AI-augmented forecasting, human interventions on AI sales forecasts were harmful as forecast error increased (e.g., Fildes et al., 2009). However, human overriding can benefit forecasting performance, e.g., in scenarios where additional expert knowledge (e.g., a new construction site next to the store) complements the AI models or data issues require the decision-maker to intervene in the AI forecasts. For example, the Covid-19 pandemic had a significant impact on sales forecasts and regularly required input from domain experts (Fildes et al., 2022a), as AI systems did not have access to information about the progression of the pandemic that would have helped the AI system to adapt to the special situation (Fahse, 2022).

In theory, human-AI hybrids hold much potential by overcoming the limitations both decision-making parties hold when executing the task individually. However, extant research findings offer limited insights into how to best design such hybrids and what their long-term implications are. Studies commonly look at a hypothetical decision-making scenario where costs of wrong decision-making making have little to no implications to users' real life. More so, previous research has largely studied cross-sectional and initial interaction with an AI system. This raises questions about the dynamics of continuous AI output, and phenomena of habit, trust formation, and algorithmic familiarity (McKnight et al., 2020).

Artificial Intelligence Based Sales Forecasting

Decisions based on sales forecasts are especially important for organizations as central processes like production, procurement, supply chain management, marketing, and personnel planning require highly accurate sales forecasts (Fildes et al., 2022b). Consequently, accurate sales forecasting is of crucial importance to a company's profitability (Sun et al., 2008). It is very time-consuming for decision-makers to perform sales forecasting manually for all products. In addition, decision-makers need to trade off two conflicting goals: On the one hand, food waste shall be minimized for cost reduction and sustainability reasons. On the other hand, sales volume and therefore revenue are to be maximized with empty shelves often posing a risk to the retailer's reputation. Simultaneously, human biases like loss aversion can result in less successful decisions as a study on real-life angel investment decision illustrates (Blohm et al., 2020). In sales forecasting, loss aversion may lead to inadequately low forecasts and therefore revenue loss.

Consequently, decision-makers increasingly rely on automated sales forecasts (Fildes et al., 2022b). In recent years, much research on advanced forecasting algorithms has been conducted to deal with unsteady market conditions, for which nonlinear models are expected to outperform traditional approaches like exponential smoothing or ARIMA models (Fildes et al., 2022b, Ramos et al., 2015). Given the importance of sales forecasting and the potential to outperform human decision-makers, sales forecasting is therefore among the most promising application areas for AI (Cam et al., 2019).

Research Context, Data and Measures

Following computational and process-driven theory development (Berente et al., 2019; Miranda et al., 2022; Grisold et al., 2020) we rely on trace data, i.e., digital records from actual forecasting activities to understand human interaction with AI forecasts in a real-life decision-making context. According to computational theory development, studying trace data can provide a more accurate understanding of human-AI interaction by offering a "form of unobtrusive measure" (Berente et al., 2019).

The studied sales forecasting context encompasses domain experts to predict next-day sales of a bakery enterprise. Next-day sales forecasting constitutes a numerical estimate for the quantities to be produced of each product category. Sales forecasting is crucial to the bakery enterprise as it affects production costs, food waste of unsold goods, and (missed) sales opportunities. A mid-sized bakery enterprise with over 200 employees provides point-of-sales data that consists of 1.4 million daily sales transactions over 2078 days (69 months, 5.7 years) and covers a timeframe from January 2017 until September 2022. The bakery enterprise consists of 27 branches which sell 528 different products. A state-of-the-art AI model was trained to make one-day-ahead predictions. The AI model uses, among others, information about the day ("weekday", "month") and information about previous sales (e.g., "sales of previous day", "sales of two days previous", "mean sales of prior week", "mean sales of prior month") as input variables. Moreover, the dataset was augmented with additional information about weather indicators ("temperature", "rainfall", "radiation/sunshine"). Mirroring the human forecasting process, the AI provides next-day sales forecast for each product category and was integrated into the extant enterprise system of the bakery. The AI forecasting system was introduced after 1247 days (42 months), which equals 61% of the available time frame.

Since the AI system was introduced at different time points in the different branches, we focus on the largest branch and use the others to confirm the analyses' robustness. The dataset of the largest branch consists of 110.000 daily sales and covers the same timeframe mentioned above. This branch sells 351 different products, of which many are seasonal or only sporadically sold. To obtain an unbiased assessment of the effects of the AI system introduction, we focus on the 20 most sold products of the branch, as these are sold continuously and with high frequency. For these products, the potential for AI-assisted decision-making is much higher compared to sporadically sold products like wedding cakes, which are mostly pre-ordered and do not need to be forecasted. We report averages for different measures over these products to base our analyses not only on individual products, but on a large part of the branch's sales. Since these products are relatively homogeneous in sales, generating averages does not distort the weighting of the individual products when absolute values are used. In addition, in many cases, we compute relative (percentage) values, which can be aggregated without distortion and hence be jointly evaluated (Koutsandreas et al., 2022). In some cases, values are standardized before calculating averages to ensure comparability.[1]

To initiate a forecast, a domain provides a numerical forecast of next-day sales for each product. This forecast is then used to determine the quantities to be produced. Before AI introduction, decision-makers had access to historical forecasts and could thus use previous forecasts and own experience for orientation when making a forecast. Next day's forecasts are made every day for each product in each branch. We track forecasted and actual sold quantities to measure forecast errors. With the introduction of the AI system, decision-makers now had the opportunity to consider the AI predictions in their sales forecasts, i.e., by exactly following or adjusting the AI forecasts based on their own perceptions. Decision-makers were not forced to consider or follow the AI forecast and could place any integer forecast number.

Name	**Description**	**Overall Mean (SD)**	**Pre-AI Mean (SD)**	**Post-AI Mean (SD)**
Human forecast	Decision-makers' sales forecast for the next day without aid of the AI forecast.	-	41.95 (20.92)	-
AI forecast	Sales forecast made by the AI system.	33.85 (18.95)	36.04 (20.23)	30.72 (16.47)
Hybrid forecast	Sales forecast made by human decision-makers with aid of the AI forecast.	39.25 (20.07)	-	35.38 (18.1)
Human forecast error	Difference between the human forecast and actual sales [MAPE].	-	0.31 (0.21)	-
AI forecast error	Difference between the AI forecast and actual sales [MAPE].	0.22 (0.12)	0.22 (0.12)	0.22 (0.12)
Hybrid forecast error	Difference between the hybrid forecast and actual sales [MAPE].	0.28 (0.2)	-	0.24 (0.18)
Overriding	Deviation between hybrid forecast and AI forecast [MAPE].	0.36 (0.13)	0.38 (0.13)[2]	0.31 (0.1)
Actual sales	Actual mean number of products sold per day in the bakery.	33.43 (19.23)	35.41 (20.34)	30.59 (17.14)
Waste	Mean number of products produced but not sold per day.	5.41 (3.25)	6.03 (3.2)	4.53 (3.11)

Table 1. Descriptive statistics for the main variables in the dataset

Following an iterative empirical-inductive approach, we measure the isolated AI forecast (AI forecast), the final forecast submitted by the user (hybrid forecast), and actual sales for the forecasted products to calculate forecast errors in human-AI hybrids (see Table 1). Based on the analysis of forecast errors, we focus on the impact of human overriding on hybrid system performance. In a subsequent step, an additional variable (waste ratio) is analyzed to better understand observed behavior.

To address the first research question, we measure decision-makers' "reliance" (the degree to which human decision-makers override the AI forecasts, called "overriding") and "task performance" (the degree to which

[1] The Covid-19 pandemic containment measurements started after 1163 days and ended after 1224 days (duration of 61 days, which equals a share of 2.9%). During this time, the bakery closed its branches and only carried out contract sales. Spontaneous sales of customers were not possible. Consequently, the Covid-19 pandemic period is excluded from our analyses, as it is not representative of the pre- and post-pandemic periods. Reported metrics in this study do not refer to the pandemic period unless otherwise stated.

[2] This is a theoretical value as decision-makers were not provided with AI forecasts at this time.

the hybrid human-AI system makes good decisions in terms of forecast error, called "hybrid forecast error") as well as the effect of "reliance" on "task performance". We operationalize "reliance" directly in terms of relative difference between human forecast and AI forecast and "task performance" directly in terms of mean absolute percentage forecast error (MAPE). To measure the influence of "overriding" on "hybrid forecast error", an Ordinary Least Squares (OLS) regression is conducted. This way, we analyze if overriding of the AI forecast is beneficial for the hybrid system's forecast error. To address the second research question, we compare the degree of overriding and hybrid system forecasting performance over time after the introduction of the AI sales forecasting system.

Results

Introducing the AI forecasting system has decreased overall forecast error [MAPE] by 6.7 percentage points (21.7%). If decision-makers had strictly followed AI forecasts after AI system launch, the performance gain would have been 9.1 percentage points (29.3%). Overall, humans were able to improve AI forecasts in 6 out of 20 products. For these products, humans improved AI forecasts by 18.2% on average. When humans deteriorated AI forecasts, they did so by 15.5% on average. Figure 1 shows the hybrid forecast error (MAPE) in black and mean absolute percentage overriding in red. The horizontal lines depict the human forecast error prior to AI system introduction on the left and the hybrid forecast error after AI system introduction on the right. Both graphs have been centralized by subtracting their means to provide more visibility of the correlation. The pandemic period has been hidden.

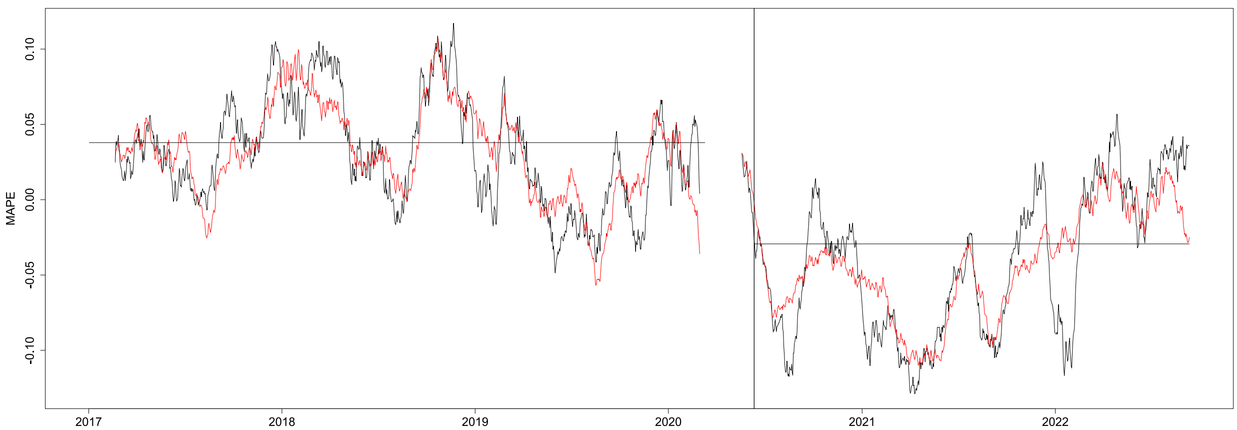

Figure 1. Hybrid forecast error (black) and mean absolute percentage overriding (red)

As can be observed in Figure 1, human overriding of the AI forecasts led to higher forecast error in general. To answer RQ1, Table 2 depicts the results of two regressions (R1 and R2) where the hybrid system's forecast error is the dependent variable. In R1, the independent variable is the degree to which human decision-makers (i.e., the domain experts in the branch) overrode the AI forecasts in terms of MAPE (always >=0). In R2, instead of mean absolute percentage values, mean percentage values are used to consider whether decision-makers increased or decreased AI forecasts (may be <0).

	Independent variable	Estimate	t-value	R^2	F-value	p-value
R1	Intercept	-0.0091	-0.477	0.1931	192.6	0.634
	Mean absolute percentage overriding	0.8045	13.88			<2e-16 ***
R2	Intercept	0.1212	12.37	0.2195	226.4	<2e-16 ***
	Mean percentage overriding	0.6465	15.05			<2e-16 ***

Table 2. Results of the two regression analyses R1 and R2

The estimates for "Mean absolute percentage overriding" and "Mean percentage overriding" are positive and highly significant. For R1, this means that 1% of decision-makers' overriding leads to an increase of forecast error in terms of MAPE of 0.8%. This is consistent with R2, where negative overriding was considered.

To answer RQ2, Figure 2 shows the progression of human overriding over the entire available period. The vertical line represents the start of the AI system. Left of the vertical line, the overriding is only theoretical, as decision-makers did were not provided with AI forecasts at this time. Right of the vertical line, human overriding roughly follows a U-shape. In the beginning overriding occurs, however, orientation and influence by AI system was already present. In the middle phase, decision-makers' overriding of AI forecasts was at its minimum. In this phase, overall system forecast error is the lowest. At the end of the covered period, decision-makers increasingly overrode again. More specifically, AI forecasts were mainly increased, leading to higher waste ratios. However, this might be beneficial from a real-life business perspective, as empty shelves are disadvantageous for the reputation of the branch. As the forecast error in the end phase of the covered period is still lower than before AI system launch, the branch might have found the sweet spot between forecast error and avoiding empty shelves.

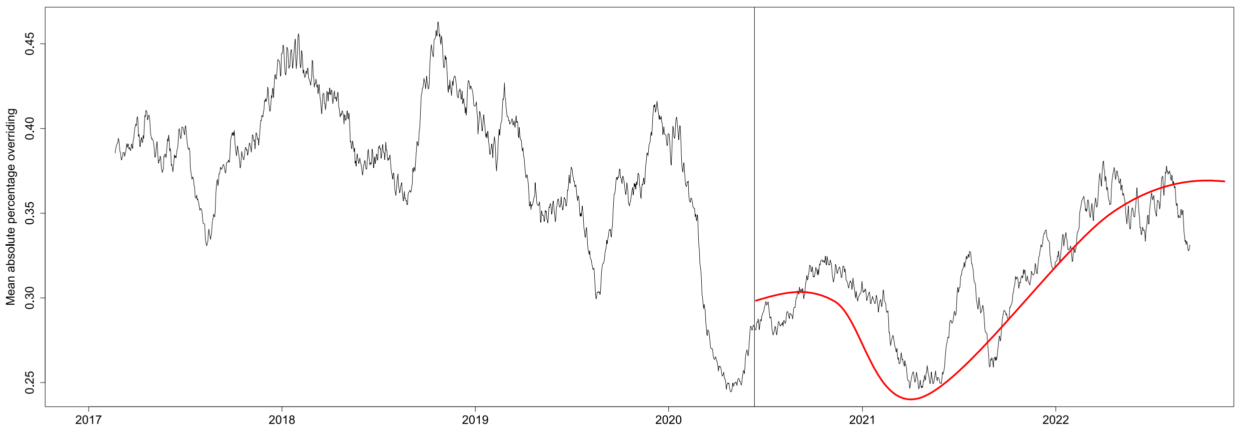

Figure 2. Progression of human overriding over time (red line depicts approximation)

The spread between human decisions (i.e., human forecasts) and AI decisions (i.e., the forecasts generated by the AI forecasting system) is significantly higher prior to AI system introduction than after. This also applies to the initial phase - directly after the AI system's introduction - and indicates that human decision-makers have followed AI forecasts to some degree, but still overrode when they thought it is necessary.

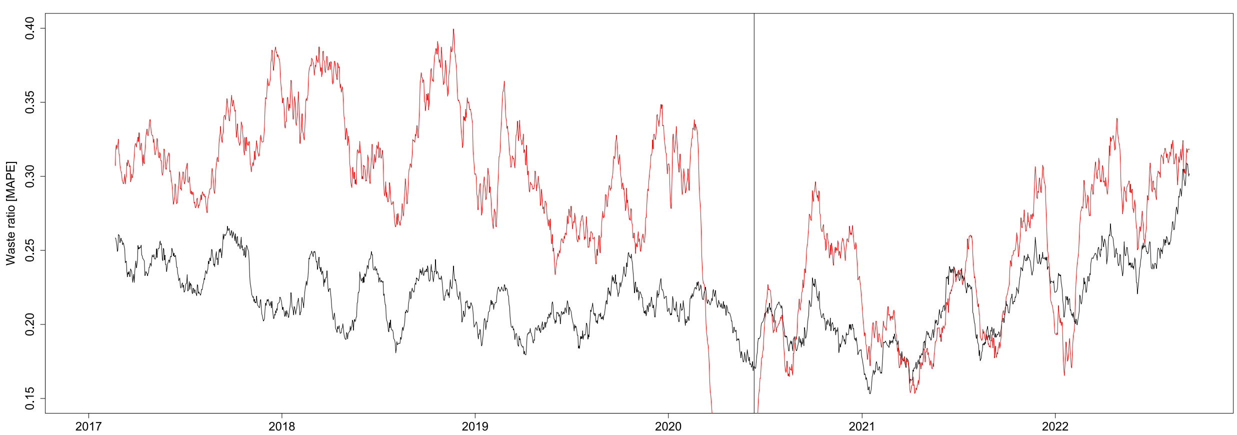

Figure 3. Theoretical waste if AI forecasts had been used (black) and actual waste ratio (red)

In this study, when overriding, decision-makers increase AI forecasts on average. This leads to higher waste ratios and consequently higher forecast errors and can thus explain the negative impact of human overriding on forecast error (see Table 2). Figure 3 shows the pure AI system's waste ratio (black), the actual, hybrid system's waste ratio (red line, right of vertical line) and the pure human system's waste ratio (red line, left of line). It shows that the waste ratio would have been lower had the decision-makers strictly followed the AI forecast (waste ratio of 21.9% versus 24.2%).

Discussion

This study reports results of a quantitative analysis of digital trace data generated in sales forecasting. We thereby contribute to temporal digital trace data research (Grisold et al., 2020) and aim to explain organizational change in the context of an AI forecasting system's introduction in a bakery enterprise. RQ1 generally scrutinizes the performance of hybrid human-AI systems; RQ2 focuses on the progression of decision-makers' overriding of AI forecasts over time. To this end, we analyzed longitudinal sales data over a period of 69 months between 2017 and 2022. The results indicate that human decision-makers cannot improve upon AI sales forecasts in terms of forecast error. For each 1% of absolute percentage overriding, the forecast error increases by 0.8%. Extant studies (e.g., Fildes et al., 2009) already concluded that forecast performance tends not to be improved by human overriding, but an analysis with real sales data and a state-of-the-art ML model has – to the best of the authors' knowledge – still been missing. In addition, the longitudinal data covers forecasting decisions before and after the AI system's introduction and outlines the course of human overriding of AI forecasts. This course of overriding changed over time in a U-shape. Decision-makers' initial high overriding decreased after ~6 months, built a minimum after ~12 months, and formed a plateau after ~20 months.

With our findings, we contribute to a better understanding of human intervention in hybrid systems. We confirm previous research in the evaluation that humans tend to deteriorate AI forecasts when overriding (Beese and Fahse, 2023). We further shed light into some aspects of the Technology Acceptance Model (TAM) (Venkatesh et al., 2016). Specifically, our study focusses on the variables "use behavior" ("actual system use", operationalized through mean absolute percentage overriding), "outcome mechanism" ("actual task performance", operationalized through hybrid system forecast error), and their interaction, finding that 1% of absolute percentage overriding leads to an extra 0.8% of forecast error. The data analysis also indicates that human overriding tended to increase the quantities forecasted by the AI system, which led to higher waste ratios. This finding sheds a different light on the influence of humans on hybrid forecast performance: In five interviews with CEOs of different bakery enterprises, it was stated that empty shelves damage the store's reputation. Hence, optimal forecasts in terms of certain performance metrics do not inevitably yield optimal real-life business outcomes. Complex human behavior, such as feeling uncomfortable when shelves are sporadically filled, could theoretically be incorporated into AI systems, but are not usually considered from the start. Consequently, human overriding stagnates when the AI system has adapted to the special requirements of its application context. These considerations can be the first step towards theory building on generalizable patterns of human overriding after AI system introduction.

For practice, our study provides concrete implications for business owners and managers considering implementing AI systems. Our analysis of real-life data helps practitioners better set expectations by demonstrating the impact that can be expected from the AI forecasting system, thus providing a blueprint for similar endeavors. In addition, by analyzing the waste ratio and resulting empty shelves when waste ratios are too low, the results show that critical human thinking is still important in hybrid systems and should find its way into the system. AI systems are great for specific tasks, but do not always encompass all relevant features. Finally, by exemplarily identifying products for which human overriding was especially beneficial (6 out of 20 products), we provide initial guidance for the design of hybrid human-AI systems by recommending for which products humans should be kept in the loop. In summary, our study shows that neither the AI system nor human decision-makers alone can solve the task best, but joint effort is required (Bansal et al., 2019; Lundberg et al., 2018).

Nevertheless, our results and conclusions are subject to some limitations, which represent an avenue for future research. The empirical analyses of this study exclude competing alternative explanations such as general performance of the bakery enterprise or market dynamics (e.g., the Covid-19 pandemic occurred shortly before AI system introduction and might have permanently changed the business environment). We also did not systematically collect a priori user expectations and perceptions of the AI system. A qualitative analysis of human perception of the AI system over time would be useful to examine how human perception of the AI system affects its use and hence hybrid system performance. Further, this study is conducted in a specific context: We focused on one domain (baked goods) and one bakery enterprise. To derive more general conclusions and to examine whether similar effects can be observed, data from other domains (e.g., e-commerce) would be needed. Analyses on specific constellations in which human overriding is especially beneficial may help designing and improving hybrid systems. These constellations could be a full list of products that should (not) be automatically forecasted, specific times in which overriding is most beneficial (e.g., nearby construction sites), or certain orders of magnitude that overrides must (not) exceed. In addition, future research could consider human factors in AI model development, i.e., by providing the model with the time of the last sale.

Acknowledgements

We would like to thank the research partners (bakery enterprise and AI company) who have granted access to the data and the Swiss National Science Foundation (grant number 192718) for their support of this research.

REFERENCES

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. 2019. "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (7), pp. 2–11.

Beese, J. and Fahse, T. 2023. "Modern Centaurs: How Humans and AI Systems Interact in Sales Forecasting," in *Proceedings of the European Conference on Information Systems*.

Berente, N., Seidel, S., and Safadi, H. 2019. "Research commentary—data-driven computationally intensive theory development," *Information Systems Research (*30:1), pp. 50-64.

Binns, R., Kleek, M. V., 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.

Blohm, I., Antretter, T., Sirén, C., Grichnik, D., and Wincent, J. 2020. "It's a Peoples Game, Isn't It?! A Comparison Between the Investment Returns of Business Angels and Machine Learning Algorithms," *Entrepreneurship Theory and Practice*.

Cam, A., Chui, M., and Hall, B. 2019. "Global AI Survey: AI Proves Its Worth, but Few Scale Impact," *McKinsey Global Institute*, McKinsey. (http://dln.jaipuria.ac.in:8080/jspui/bitstream/123456789/1323/1/Global-AI-Survey-AI-proves-its-worth-but-few-scale-impact.pdf).

Chakraborti, T., and Kambhampati, S. 2018. "Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-ai collaboration," *arXiv preprint*.

Dellermann, D., Ebel, P., Söllner, M., and Leimeister, J. M. 2019. "Hybrid Intelligence," *Business & Information Systems Engineering* (61:5), pp. 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), pp. 114–126.

Dixon, J., Hong, B., and Wu, L. 2020. "The employment consequences of robots: Firm-level evidence," Ontario: Statistics Canada.

Fahse, T. 2022. "Do Forecasting Algorithms Need a Crisis-Mode? Machine Learning Based Sales Forecasting in Times of COVID-19," in *Proceedings of the Conference of the Italian Chapter of AIS*.

Fildes, R. and Goodwin, P. 2007. "Against Your Better Judgment? How Organizations Can Improve Their Use of Management Judgment in Forecasting," *Interfaces* (37:6), pp. 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), pp. 3-23.

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

Fildes, R., Ma, S., and Kolassa, S. 2022b. "Retail Forecasting: Research and Practice," *International Journal of Forecasting*.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. 2021. "Will Humans-in-The-Loop Become Borgs? Merits and Pitfalls of Working with AI," *Management Information Systems Quarterly* (45:3), pp. 1527-1556.

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), pp. 678-696.

Gregory, R. W., 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), pp. 534-551.

Grisold, T., Wurm, B., Mendling, J., and Vom Brocke, J. 2020. "Using Process Mining to Support Theorizing About Change in Organizations," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*.

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), pp. 155–184.

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. 2021. "Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence," *Information Systems Research* (32:3), 713-735.

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

Koutsandreas, D., Spiliotis, E., Petropoulos, F., Assimakopoulos, V. 2022. "On the selection of forecasting accuracy measures," *The Journal of the Operational Research Society* (73:5), pp. 937–954.

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), pp. 126-148.

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), pp. 90-103.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., and Lee, S.-I. (2018). "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering* (2:10), pp. 749–760.

Lysyakov, M., and Viswanathan, S. 2022. "Threatened by AI: Analyzing Users' Responses to the Introduction of AI in a Crowd-sourcing Platform," *Information Systems Research*.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. 2022. "M5 Accuracy Competition: Results, Findings, and Conclusions," *International Journal of Forecasting* (38:4), pp. 1346–1364.

McKnight, D. H., Liu, P., and Pentland, B. T. 2020. "Trust change in information technology products," *Journal of Management Information Systems* (37:4), pp. 1015-1046.

Mendling, J., Decker, G., Hull, R., Reijers, H. A., 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* (43:1), p. 19.

Miranda, S., Berente, N., Seidel, S., Safadi, H., and Burton- Jones, A. 2022. Editor's Comments: Computationally Intensive Theory Construction: A Primer for Authors and Reviewers. *Management Information Systems Quarterly* (46:2), pp. iii–xviii.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., et al. 2022. "Forecasting: theory and practice," *International Journal of Forecasting* (38:3), pp. 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. "Next generation digital platforms: toward human-AI hybrids," *Management Information Systems Quarterly* (43:1), pp. iii-ix.

Raisch, S., and Krakowski, S. 2021. "Artificial intelligence and management: The automation–augmentation paradox," *Academy of Management Review* (46:1), pp. 192-210.

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), pp. 151–163.

Russell, S., and Norvig, P. 2021. *Artificial Intelligence: A Modern Approach*, Pearson Education Limited.

Schmitt, A., Wambsganss, T., Janson, A., and Leimeister, J.M. 2021. "Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice," in *Proceedings of the International Conference on Information Systems*.

Veiga, C. P., Veiga, C. R. P., Puchalski, W., Coelho, L., and Tortato, U. 2016. "Demand Forecasting Based on Natural Computing Approaches Applied to the Foodstuff Retail Segment," *Journal of Retailing and Consumer Services* (31), pp. 174–181.

Venkatesh, V., Thong, J., 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), pp. 328–376.

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