Journal of Management Information Systems

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/mmis20

Human–Artificial Intelligence Collaboration in Prediction: A Field Experiment in the Retail Industry

Elena Revilla, Maria Jesus Saenz, Matthias Seifter & Ye Ma

To cite this article: Elena Revilla, Maria Jesus Saenz, Matthias Seifter & Ye Ma (2023) Human–Artificial Intelligence Collaboration in Prediction: A Field Experiment in the Retail Industry, Journal of Management Information Systems, 40:4, 1071-1098, DOI: 10.1080/07421222.2023.2267317

To link to this article: https://doi.org/10.1080/07421222.2023.2267317

Published online: 11 Dec 2023.

JOURNAL OF MANAGEMENT INFORMATION SYSTEMS
2023, VOL. 40, NO. 4, 1071–1098
https://doi.org/10.1080/07421222.2023.2267317

Routledge
Taylor & Francis Group

Human–Artificial Intelligence Collaboration in Prediction: A Field Experiment in the Retail Industry

Elena Revilla[a], Maria Jesus Saenz[b], Matthias Seifter[a], and Ye Ma[b]

[a]IE Business School, Madrid, Spain; [b]MIT Digital Supply Chain Transformation (MIT CTL), Cambridge, MA, USA

ABSTRACT

This study investigates the role of human intervention in artificial intelligence/machine learning (AIML)-driven predictions. By doing so, we distinguish between three different types of human-AIML collaboration: *automation, adjustable automation*, and *augmentation*. We theorize that prediction uncertainty and time horizon represent two critical determinants of forecast accuracy. Based on a field experiment involving AIML-driven demand forecasts approximately 1,888 stock-keeping units in the retail industry, we rely on a multivalued treatment effect methodology to measure the effects of human-AIML collaboration on forecast accuracy. Our findings show that human intervention complements AIML-driven forecasts most effectively (augmentation) in predictions with long time horizons and low uncertainty. However human intervention is least likely to contribute to the effectiveness of AIML predictions (automation) in environments with short time horizons and high uncertainty. We discuss implications for extant theory and propose a framework outlining the conditions in which human intervention is most likely to add predictive value to human-AIML collaborations.

KEYWORDS

Human-AI collaboration; AIML; algorithmic prediction; uncertainty; time horizon; demand forecasting; machine learning

Introduction

The rapid growth of big data and artificial intelligence (AI) technologies in recent years showcases the critical role that AI can play in supporting managerial tasks across a wide range of industry sectors [14]. Indeed, companies rely increasingly on AI for medical diagnostics, talent acquisition, demand forecasting, credit scoring, logistic task scheduling, allocation of resources, and the general diagnosis and resolution of managerial problems, among others. AI differs from traditional information technologies in that it learns autonomously [78] and is increasingly able to make decisions without human intervention.

One of the main drivers for promoting AI in the managerial arena is rooted in the realization that humans have limited cognitive capabilities [19, 28, 70] whereas AI has the potential to compensate for these limitations [18, 54]. Yet when AI is applied indiscriminately, particularly to support knowledge-based functions, it is subject to criticism as some tasks remain challenging for machines to complete [48, 7, 8, 15]. Some scholars warn that AI-based models may introduce systematic biases into the decision-making process (e.g., Kliegr et al. [46] and Li and Chai [53]) and highlight the proficiency of human experts to identify and correct them [4, 18, 71].

CONTACT Elena Revilla elena.revilla@ie.edu IE Business School, Maria de Molina 12, Madrid 28006, Spain.

From a scholarly perspective, much of the human-AI debate has taken place in the Information System (IS) literature (e.g., Tofangchi et al. [76], Wang and Benbasat [80], and You et al. [83]). For example, Berente et al. [9], state that "the interaction between humans and AI is perhaps the key managerial issue of our time" (p. 1440). When applying AI technologies to decisions that have an impact at the socioeconomic level, scholars call for humans to remain in the loop [84] and work together [26]. Humans are essential to AI because they provide contextualization [36]. While AI operates by applying rules without accessing the meaning of the data (i.e., contextual knowledge) [54], humans do this by means of reason and judgement. Thus, human skills augmented by AI technology promise to result in powerful intelligence that is likely to exceed the performance of either party separately [26]. Using augmented intelligence, humans continue to do what humans do best—create, imagine, and collaborate—while AI helps with the quantitative, rule-based capability to consistently generate predictions and provides decision-making speed and scalability.

At the management level, a study involving 1,500 companies found that firms achieved the most significant performance improvements when humans and machines worked together [81]. Yet, recent research [62] indicated that currently only 10 percent of companies obtain significant financial benefits from AI technologies and concluded that companies leveraging human-machine collaboration will likely be best positioned for success. The return on investment in AI therefore remains clearly behind expectations, which is largely attributable to the fact that companies often fail to effectively manage the relationship between AI and humans as a result of oversimplifying the complex interactions between technology and humans [57].

In the present article, we argue that the usefulness of AI depends not only on the technology itself but also on the ability to integrate it systematically with human expertise. Central to this line of inquiry is the question of how companies should determine the level of human intervention in human-AI collaboration [26]. Recent studies have proposed that the effectiveness of human intervention is task-dependent, in that the type of situation influences the degree to which humans and AI are likely to collaborate effectively [65]. However, despite various calls for further research addressing the specific role of task context in human-AI collaboration, it continues to remain an under-researched topic [9, 26, 61, 69, 74].

The goal of this study is therefore to seek answer to the following research question: *How do task characteristics influence the effectiveness of human intervention in human-AI collaboration?* To address this question, we focus on machine learning in prediction tasks (artificial intelligence machine learning [AIML]), which is at the core of building contemporary predictive AI [9, 77]. AIML detects systematic patterns in big data to generate predictions that go beyond the ability of human experts [58]. It represents a specific application of AI setting formal rules that autonomously adapt the model to changes in the environment [6]. Previous literature points to three reasons for studying human-AIML collaborations in predictions [40]. First, predictions are ubiquitous. Managers generate predictions as part of more generic organizational tasks that involve problem solving, matching, design, policy formulation, and guidance [45]. Second, a substantial proportion of managerial predictions nowadays rely on an assemblage of humans and algorithms. Third, recent advances in the field of AIML have significantly transformed managerial

practice, creating the need for a better understanding of the complementarities between humans and machines.

Thus, to address our research question, we begin by identifying three different degrees of human intervention. *Automation* (lowest level of intervention) refers to full delegation of decision responsibility to the predictive AIML system. In this case, human intervention is limited to the design stage, in which the predictive algorithm is configured and trained, but not yet utilized to generate predictions. A*djustable automation* (medium level of intervention) involves both human and AIML, but humans intervene only during the main steps of the prediction. In A*ugmentation* (maximum level of intervention), human intervention occurs during each step of the prediction process.

Second, pertaining to the need to understand the context of the predictive task of AIML, we rely on Metcalf and colleagues [57] who propose and discuss how human capability varies as a function of task uncertainty and time horizon, two important dimensions often used to differentiate the strengths and weaknesses of humans and AI. Uncertainty relates to the difficulty of predicting the future due to a lack of information. Time horizon refers to the temporal distance that a decision-maker considers when evaluating the consequences of a proposed action. Accordingly, we argue that the effectiveness of human intervention critically depends on the interaction along two dimensions of the prediction context: the task uncertainty and the time horizon until the prediction materializes.

In the present study, we draw on data obtained from demand forecasting in the retail sector as one example of a commonly performed prediction. Demand forecasting refers to the process of predicting future sales over a predetermined time to optimize subsequent decisions regarding production plans, inventory management, purchasing, logistics, and manufacturing [79]. More specifically, we rely on empirical data from an intelligent demand adjustment (IDA) system based on AIML, which is part of a collaborative program between a multinational supplier and its customer. The IDA system integrates customer inputs into the supplier demand forecast process. The supplier is one of the largest multinational fast-moving consumer goods (FMCG) companies that has an end-to-end supply chain synchronization strategy. Its most recent annual revenue was more than USD $70 billion, representing between 8 percent to 10 percent of revenue in the geographic area of our study. The customer is one of the largest online retailers in China, and one of the supplier's most important customers. Its most recent annual revenue was more than USD $100 billion, supplying a market of more than 500 million active end customers. The size and complexity of the customer base require forecasting adjustments from the supplier.

Using a field experiment applying AIML to demand forecasting, we estimate the causal effects of the various levels of human-AIML collaboration on predictive performance by relying on a multivalued treatment effect methodology. Our approach allows us to compare the predictive accuracy of different human-AIML collaborations by using—as matching criteria—product turnover, price, product segmentation, and product category. In addition, our cross-unit study controls for sources of extraneous variance that may affect the relationship between human-AIML collaboration, uncertainty, and performance, and thus improves our capacity to isolate the examined effects [32].

Our study contributes to the extant literature in the following way. First, responding to prior calls to develop a better understanding of human-AI collaboration in organizations [9, 26, 74], we use field data to examine how the level of human intervention affects the effectiveness of AIML-based predictions. Our analysis reveals significant performance

differences, extending our knowledge of how the phenomenon of AIML is related to organizational design. Second, our study answers present calls for better contextualization of the relationship between human-AI collaboration and performance [69]. Our research identifies uncertainty and time horizon as two determinants of the effectiveness of human-AIML collaboration. We show that human intervention complements AIML-driven forecasts most effectively when predicting with long time horizons and low uncertainty, whereas reliance on human intervention can be detrimental in predictions with high uncertainty and short time horizons. Third, we add to the scarce but growing literature regarding the role of augmented intelligence in demand forecasting. Our findings demonstrate how and when judgment contributes to improved forecasting, extending the literature that seeks to understand the role of judgmental forecasting in predicting demand [1, 67]. Finally, our study contributes to forecasting practice by increasing managerial cognizance about the different collaborative structures involving humans and AIML that may be used to achieve superior demand-forecasting performance.

Literature Review

Human-AI Collaboration

AI relates to the ability of a machine to simulate cognitive processes and perform tasks commonly conducted by human beings, such as learning, thinking, and interacting with the environment, as well as recognizing patterns, making decisions, and solving problems, and even demonstrating creativity [60]. At times, it is argued that AI can outperform human cognition. Simon [70] introduced the widely used term *bounded rationality* to argue that human beings are often incapable of optimizing decisions due to their limited memory and processing capacity and, therefore, settle for satisfactory solutions. In consequence, unaided decision strategies frequently rely on heuristic processes that are likely to suffer from judgmental biases and learning myopia [4]. In contrast, AI evaluates large sets of historical data more accurately than humans can, thanks to its greater and faster data processing capacity [46]. In addition, its superior analytic processing power leads to increased objectivity, which, in turn, confers higher status and legitimacy to AI-generated outcomes [54]. In sum, AI technology is likely to overcome some of the human biases associated with information processing and create insights that may not otherwise have been conceived [77].

In contrast, AI has limitations that can negatively affect performance. The performance of AI is influenced by both the quality of data used in the learning process and the biases of human creators during the process of determining which pieces of information can be considered relevant [17, 18]. Moreover, AI relies on rules that, to be effective, must be specific to the context of application. Scholars thereby refer to contextual knowledge as encompassing rules about nonhistorical information that is often of a subjective nature [10, 66]. Thus, AI works effectively in narrow contexts in which constraints and objectives are quantifiable and well-defined, whereas dynamic, incomplete, and unstructured information may at times be processed in an erroneous way [85]. Examples of this limitation are frequently found in the stock market, where a minor price decrease of a highly volatile share can trigger a reaction by an "alarmist" algorithm to sell the stock and, in so doing, trigger a disproportionate response from

other trading systems that are programmed with wider limits. As such systems also react to selling stocks, they may provoke unjustified falls in the market [50].

Another source of concern is the lack of transparency of complex algorithms. It may be unclear how AI abstracts and generalizes from data, or how exactly it fulfills its performance criteria. As a result, the AI-driven system may evolve in ways that are difficult to understand for human experts. Algorithms therefore are at risk of turning into "black boxes" [2] that work well under clear parameters but, from time to time, may result in nonsensical outputs that are difficult to predict, explain, or curtail before they lead to major ripple effects. Moreover, AI does not have the capacity to reason based on socially relevant factors such as fairness or morality. For example, a previous study has found a systematic bias in AI-based recruiting processes of companies when the algorithm generated recommendations by relying on success criteria that had been intentionally discarded during the model building phase (such as gender bias) [20].

Considering the limitations of both agents, scholars have increasingly set out to explore how AI can complement human cognition without replacing it [21, 42, 54, 61, 81, 85]. On the one hand, AI technologies are faster and more proficient in identifying complex interrelations among variables in large volumes of data, leading to new insights. On the other hand, human experts are required to provide the blueprint of AI systems, set the goals of the system, and provide contextual meaning. This includes the derivation of starting hypotheses, the selection of performance criteria based on contextual constraints, the evaluation of results, and the creation of perform scenario analyses to mitigate uncertain future outcomes. Beyond the design of the AI system, Balasubramanian and colleagues [4] highlight the importance of human intervention in detecting and correcting errors in the AI decision-making process that may arise from the lack of contextual knowledge. For example, financial analysts may decide to grant mortgages to customers by considering soft factors (e.g., related to social status) that indicate the customer's potential to access additional funds, but may go beyond the nominal wealth commonly estimated by an AI algorithm. In addition, AI may also miss intertemporal dependencies in dynamically changing contexts in which historical data quickly becomes obsolete. Humans, in contrast, are often efficient when it comes to diagnosing sudden changes in the characteristics of the task environment and can judge the appropriateness of the algorithm considering the new situation [67]. In short, despite its many benefits, it may be risky to blindly trust in AI-based systems if they perform unsupervised.

The aforementioned discussion calls for further research focusing on the collaboration between humans and AI [17, 26, 59, 78], in which both are seen as team members with different, complementary capabilities. At the heart of this problem lies the debate between automation and augmentation. Although automation (machines take over human tasks) installs the logic of formal rationality in decision-making, augmentation (humans collaborate with AI algorithms to enhance decision-making) maintains the logic of substantive rationality [54]. Through formal rationality, automation makes inferences from historical data and operates under rules that are explicitly formulated and aimed at optimizing and maximizing results. It rejects arbitrariness, thus relevant contextual data such as implicit knowledge may be excluded from the decision-making process. Substantive rationality, in turn, is about goal-driven rational actions and the qualitative analysis of the situation. It is concerned with acting in accordance with community and underlying values.

Therefore, augmentation relies on human intuition and commonsense reasoning [81], which requires humans to be part of the intelligent system. This approach recognizes the importance of considering various types of reasoning and ensures human knowledge is used effectively [85].

Some scholars describe the human–AI relationship as a continuous (rather than dichotomous) collaborative process that evolves into machine "augmentation" of human capabilities [61], implying that there is more than one way to define human-AI interactions [24]. Moreover, it means that greater levels of augmentation do not always result in superior performance. Our research emphasizes the need to adapt the relative importance of augmentation to the characteristics of the decision-making task.

Human-AIML collaboration in prediction

The objective of a prediction task is to estimate the unknown future state of a variable based in extracting patterns, trends, and causalities in available data. When it comes to applying AI techniques, Benbya and colleagues [6] indicate that the task as well as the outcome to be produced determine the particular AI system to be used. Predictions represent one of the most popular application types of machine learning in AI (AIML) [63]. In this context, one of the most advanced features of machine learning for generating predictions is to test and validate the performance of multiple models to match the dynamically changing environment [3]. Importantly, AIML models are not static; they not only produce outputs for a given set of parameters, but also possess the ability to learn and improve over time without the need to be re-programmed [58].

Past research has studied the performance of various levels of human interventions in human-AIML collaboration, ranging from automation to augmentation [12, 22, 69]. However, empirical findings regarding the effectiveness of these collaboration types are mixed [49]. To shed more light on the effectiveness of human-AIML collaborations in predictions, we propose that the degree of human intervention should vary as a function of the uncertainty underlying the prediction and the time horizon until its materialization.

Uncertainty describes the difficulty of predicting the future due to incomplete information or changing conditions [5]. In Galbraith´s words: "[. . .] the greater the task uncertainty, the greater the amount of information that must be processed among decision-makers during task execution in order to achieve a given level of performance" [28, p. 4]. In cases of high uncertainty, AIML can process large amounts of data to detect causal relationships between variables or recognize systematic patterns for generating predictions.

When the parameters of the prediction task are well-defined, AIML dynamically learns through direct interaction with the environment. Consequently, there is minimal need for human perception and anticipation. In these situations, where the level of uncertainty is low, AIML can fully play out its capability to predict future states [42]. In contrast, when the prediction task is associated with non-numerical information that is difficult to quantify or that requires knowledge of the world beyond historical data, expert judgment frequently helps to identify new variables and interdependences and recognize abnormal "broken-leg cues" [10]. AIML algorithms may not make these assessments efficiently due to their limited ability to make sense of qualitative information [4]. Thus, predictions that are based on large amounts of contextual, qualitative information call for more judgmental assessments by human experts.

Time horizon refers to the temporal distance that the decision-maker considers in the prediction. From the perspective of human cognition, short time horizons are extremely challenging to manage in predictions, due to the difficulty of disentangling information signals from random noise. Past research has shown that humans tend to misinterpret information signals consistently by reacting primarily to signals they observe and only secondarily to the environmental system that produced the signal. As a result, in environments with short time horizons, humans are likely to systematically under- and over-react in their judgments [47, 55]. In addition, humans have also been found to persistently perceive patterns in data where none exist [27] and to provide judgments that systematically dampen observed linear trends [37]. In environments with short time horizons this may pose a problem because humans mistake noise for an informative signal and thus inappropriately anchor their own judgment on the AIML-based model output and make insufficient adjustments from that anchor. These small adjustments frequently lower predictive accuracy and are motivated by an attempt to "tinker at the edges" [52].

However, as the time horizon lengthens, human forecasters have been shown to anchor their judgments to a lesser extent on recent (noisy) information signals and to a greater extent on the long term mean of the data series [51]. Hence, human intervention in AIML-driven models is likely to add predictive power, particularly if historical data is sufficiently available.

In sum, uncertainty and time horizon represent two important features of a prediction context whose interaction is likely to influence the effectiveness of human-AIML collaboration. We believe that the performance of human intervention in AIML-driven systems critically hinges on the ability to recognize the properties of the prediction at hand as well as to anticipate potential changes in the prediction context. Thus, our contextualization of prediction allows for a systematic analysis of how the interaction between humans and machines influences performance outcomes in one specific application that managers frequently face. It enables a better understanding of the strengths and weaknesses that humans and machines bring to the table in this collaborative relationship when the objective is to effectively cope with incomplete information about the task and/or rapidly changing conditions in the time available to complete it. Our study acknowledges the performance limitations of both humans and AI without favoring either input at the outset. Instead, we suggest that different degrees of human intervention should be used in accordance with the characteristics of the prediction context [61].

Hypothesis Development

In the following section, we describe the mechanism according to which uncertainty and time horizon moderate predictive performance and propose a set of testable hypotheses.

Role of Uncertainty in AIML-Driven Predictions

Uncertainty makes it difficult to predict future events [31]. For example, when a new product is introduced into the market, only limited historical sales data are available, which poses a challenge both for the development of AIML-based algorithms as well as for human cognition to generate accurate sales predictions. One often used assessment approach is to rely on simple heuristics (for instance, by drawing on past data of similar products) to make

inferences about the demand for the new product. However, previous research has demonstrated that human judges do not perform well as "intuitive statisticians" in noisy predictions [11, 38, 43, 44]. In such environments, humans find it difficult to differentiate between valuable information signals and noise introduced by random variation [67].

Judgmental accuracy of human experts has been found to improve as the length of the time series increases [75]. In the absence of historical data, human experts may misdiagnose the stationarity of the data series by perceiving it to be noisier than it is. Due to the behavioral tendency to overweigh recently observed time series values [47], a perception of nonexisting patterns or illusionary trends may result [27]. AIML-based models, conversely, critically require historical data for training and calibration purposes during the model building process. However, in comparison to the performance of human judges, AIML-driven models are likely to be superior in their predictive performance, as they are extremely efficient in extracting the signal from the noise by relying on multiple, real-time market signals and simultaneously processing masses of big data related to the firm's entire product portfolio [23].

Amar and colleagues [3] highlight four characteristics of AIML-based models that are likely to lead to superior predictive performance in environments with sparse historical data: First, AIML-driven models continuously test and compare multiple approaches to identify the best-performing model. Second, these models use data-smoothing methods to adjust for temporary, nonstationary changes in the data series, which may not be representative of the longer-term demand pattern. Third, AIML-based models can generate and test multiple scenario forecasts to allow managers to cope with future uncertainty. Fourth, they may utilize external data sources to detect seemingly unrelated causalities within the forecasting event. This means that even in environments with sparse data, AIML-driven models can produce outputs based on exploiting statistical interrelations with other, correlated predictions and are therefore likely to learn faster and more efficiently than their human counterparts. As a result, we expect model accuracy to be higher than the accuracy of human judges in the case of high task uncertainty.

Hypothesis 1: *In predictions with high uncertainty, the level of human intervention in human-AIML collaboration is negatively associated with accuracy.*

For predictions with low levels of uncertainty, such as in the case of established products that have been in the market for a long time, AIML-based algorithms are highly proficient in processing historical demand data and can extrapolate linear trends and seasonal patterns. At the same time, in these environments, human experts can use their contextual knowledge to diagnose nonlinear anomalies in the data series efficiently and make effective judgmental adjustments to model outputs [30, 52, 66]. While AIML-based models commonly use smoothing techniques to disregard temporary, nonstationary deviations from the regular demand pattern [3], human experts can use their domain knowledge to incorporate such deviations in their judgments. Specifically, expert judgment is often suggested to result from a pattern-matching process, during which humans cognitively compare the task at hand with other situations experienced in the past [67]. Human experts therefore hold contextual knowledge capturing nonhistorical information surrounding the prediction that is often of a subjective, qualitative nature

(e.g., new-product introduction plans, promotions, assortments, competitors, manufacturing, market, and macroeconomic conditions) [10, 44, 66].

Taking this into account, human-AIML collaborations based on available contextual knowledge, merged with a well-defined AIML-driven model, represent a highly effective combination. Specifically, when the parameters of the model are well-defined and calibrated using large historical data sets, the algorithm is likely to produce robust demand estimates. In this case, the role of the human expert reduces to monitoring the evolution of the demand pattern and intervening whenever an irregularity is diagnosed. Such irregularities may relate to changes in the prediction due to information that is difficult to quantify, or knowledge of the world outside the historical data that could relate to the emergence of new variables and interdependencies between them [10]. In sum, in predictions with low uncertainty, the need for human intervention arises when it is possible to draw on prior experience and contextual knowledge to identify disruptive changes in the data environment that the AIML-based algorithm could be missing. Accordingly, in the presence of large historical data, we expect human-AIML collaborations to be positively associated with predictive performance.

Hypothesis 2: *In predictions with low uncertainty, the level of human intervention in human-AIML collaboration is positively associated with accuracy.*

Role of Time Horizon in AIML-Driven Predictions

In addition to the role of uncertainty in predictions, a second major determinant of predictive performance concerns the length of the time horizon until the prediction materializes (i.e., short term vs. long term) [39]. Regardless of the level of uncertainty, predictions with short time horizons are difficult to make because contextual data related to special events such as sales promotions, which are frequently used in the retail sector, may trigger nonstationary changes in demand patterns [41]. For instance, promotional campaigns may be associated with temporary changes in marketing strategy, price reductions, inventory availability, display, and so forth, which makes short-term demand estimations one of the retailer's biggest challenges [56]. For example, a study among North American grocers [56] found that the grocers were not able to take all relevant aspects of a promotion into account when predicting sales. In addition, inaccurate promotional predictions can result in stock-outs and reduced customer satisfaction if the effect of the promotion is underestimated, or in costly spoilage and markdown losses if the effect of the promotion is overestimated. Beyond the effect of such contextual data, past studies have shown that human judgments with short time horizons are likely to suffer from systematic trend dampening when predictions are made based on historical data. This often results in severe underestimation of real and persistent changes in the demand pattern and hence in reduced predictive accuracy [34].

Considering this discussion, on the one hand, automation is likely to perform best for predictions with short time horizons. The ability to conduct extensive scenario analyses allows AIML-based models to consider a wide range of potentially related predictions and use real-time data on the fly as evidence to support or discard specific scenarios [3]. Similarly, as short time horizons require the model to be reactive and quickly adapt to

incoming data, machine learning has been proven to be highly effective in consistently comparing multiple candidate models to identify the best-fitting one [13]. Other studies also point to the superior predictive performance of AIML-based algorithms, which have been shown to improve both promotional forecasting capabilities and accuracy dramatically [56], even though models generating predictions with short time horizons may quickly become obsolete.

Hypothesis 3: *In predictions with short time horizons, automation will be associated with the highest predictive accuracy, relative to other forms of human-AIML collaboration.*

For predictions with long time horizons, on the other hand, a more nuanced perspective is needed to comprehend the relationship between human-AIML collaboration and predictive accuracy. Particularly in the presence of historical data (i.e., when uncertainty is low), AIML-based models can extrapolate values reliably within the range of observations based on which the models have been estimated. Predictions with long time horizons, however, also require more historical data to model the entire demand trajectory. In this regard, human experts have been shown to be proficient in using their prior knowledge and/or industry-specific experience related to the product's general life cycle to account for long-term patterns in demand behavior [66].

In addition, it has been shown that human judges anchor their predictions to a lesser extent on recent periods of the time series and to a greater extent on the long-term mean of the data series [51], which implies that humans are likely to be proficient in filtering out random noise in stable data environments and focus on underlying trends when generating predictions with longer time horizons. Hence, we expect human-AIML collaboration to be positively associated with predictive accuracy in predictions with low uncertainty. However, this may not be true in the case of high uncertainty, for which historical data is typically more variable and of shorter duration. In fact, past research has found that more variable data series also tend to result in more variable random judgment error [34]. As a result, the performance of human experts is likely to be reduced, as they may be distracted by the level of noise in the data environment and fail to apply their contextual knowledge effectively to generate judgments. When data from external, seemingly unrelated, sources are available, AIML-based models are likely to perform better than human experts due to their ability to scan and evaluate millions of data points to identify interrelationships with the prediction at hand, which would be impossible with human involvement [69].

Hypothesis 4: *In predictions with long time horizons, augmentation will be associated with the highest predictive accuracy under low uncertainty, while it will be associated with the lowest predictive accuracy when uncertainty is high.*

Methodology

We designed a field experiment in the context of AIML predictions. Our empirical setting studies predictions in terms of AIML-driven demand forecasting in the retail industry. The increasing availability of big data sets and recent advancements in AIML technologies have revolutionized the practice of demand forecasting and machine learning is increasingly used

in supply chain management, particularly for performing predictions [3]. Demand forecasts serve as the basis for making decisions regarding supply chain planning, inventory management, purchasing, production and distribution [79]. Gartner [29] argued that AIML technologies are likely to affect demand forecasting more than any other components of the supply chain processes—particularly in the retail industry. This is because the rapidly changing retail landscape, characterized by e-commerce sales and shoppers who choose from new fulfillment options, complicates forecasting and replenishment processes. This complexity is often exacerbated by sales promotions, weather, disruptions, market trends, and seasonal demands.

We study demand forecasting in a specific retail company and its key customer. We estimate the causal effects of human intervention in human–AIML collaboration on demand forecast accuracy under two features that define the prediction context: demand uncertainty and time horizon. We then employ a multivalued treatment effects methodology to test our hypotheses.

Sample Collection and Description

The data underlying this study were provided by an international fast-moving consumer goods (FMCG) company. We collected data from its demand forecasting process from both sides of the dyad: from the supplier's (the FMCG) intelligent demand adjustment system and from the FMCG's key e-commerce customer. This collaborative forecasting system integrated other sources of demand signals from the e-commerce customer. The empirical setting selected for this research was suitable, considering the high level of uncertainty inherent in e-commerce retail markets, especially in a fast-growing environment such as China.

With the main aim of developing a better understanding of the engagement between the AIML forecasting system and the human forecasters involved in this process, we conducted several interviews with demand forecasting experts *a posteriori* of the experimental study. As part of the interview, we elicited the type of contextual information the forecaster could contribute to each step of the forecasting process, the dynamics of interactions between the forecasting algorithm and the forecaster, their collaboration with other functions to enrich the forecasting process, and other types of expert knowledge that could be contributed to the forecasting process. The insights obtained thereby enhanced our understanding of their particular role in improving demand forecasting performance, as well as their particular role in the dynamics of human-AIML collaboration. Our sample was composed of 1,888 stock-keeping units (SKUs), the total number of SKUs that integrate the customer's portfolio of products. During a 50-week period (ending in April 2020), an AIML algorithm (see the following description) was trained dynamically in the actual environment of this field experiment to forecast the weekly demand for each SKU according to the data flow depicted in Figure 1. For each SKU, one of the three levels of our treatment condition—automation, adjustable automation, or augmentation—was assigned randomly, and the paired SKU-treatment was held constant for the entire 50 weeks to guarantee the robustness of the experimental design. Given the empirical purpose of our research, we focused on the most recent result (Week 50) for each of the 1,888 SKUs.

The forecasting process consisted of a series of six steps supported by AIML-driven algorithms (Referred to as Intelligent Demand Forecasting System in Figure 1). Before

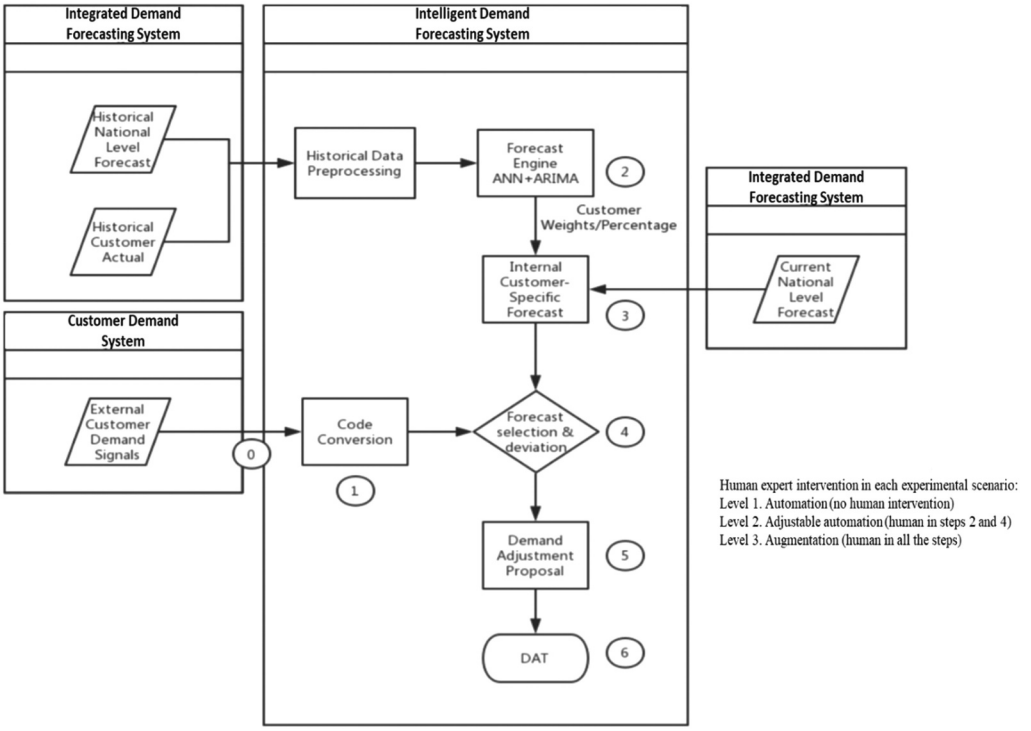

Figure 1. The six steps of the demand forecasting process are circled.

starting the process, the customer sent demand indicators for the upcoming 13 weeks. Additionally, the historical data was preprocessed; FMCG collected both the historical national-level forecast and the historical customer demand actuals internally in the company's integrated demand forecasting system. During Step 1, customer forecast codes were converted to assess missing master data or mismatches, and experts reviewed the data conversion to ensure consistency. During Step 2, the percentage of internal customer demand and demand per region associated with a particular distribution center were calculated per SKU by the forecast engine (customer ratio), before the AIML Artificial Neuro Network (ANN) and Autoregressive Integrated Moving Average (ARIMA) methods began to be applied (see as follows). In this step, human expert forecasters could review the percentages proposed by the AIML system and consider additional effects related to demand, including promotions, special events, or the potential behavior of a particular product in certain regions or distribution centers. The forecaster could receive informational signals to make corrections, or he/she could proactively check pieces of information to review this forecasting step more meticulously. During Step 3, the data were transformed into a particular forecast for the customer and the SKU, using the customer ratio developed in the previous step. If necessary, experts could make corrections in Step 3 based on their product category expertise. During Step 4, the final forecast was determined based on the balance (weights) between the internal forecast and the customer forecast. In this step, human forecasters could adjust the deviation and balance the ratio of customer needs against the ideal (future) scenario in terms of product manufacturing versus distribution.

These adjustments may be based on external information, such as available inventory, customer priorities, promotions, nonstationary changes in demand, or unexpected disruptions. During Step 5, a forecast was produced (with the right format) based on the 13-week demand adjustment. In Step 6, the adjusted demand forecast results were reviewed and released to guide further production and/or distribution scheduling and decision-making. The results from Steps 5 and 6 could be assessed by expert forecasters considering certain signals regarding expected gaps, data formats or final recommendations from marketing and, if needed, further adjusted.

The supervised AIML forecasting system was based on a machine learning algorithm grounded, as noted above, in ANN and ARIMA technologies [83, 86]. It was implemented using a multilayer perceptron model, a feed-forward (fully connected) neural network trained with a backpropagation algorithm [41]. Particularly, ANN techniques represent a variety of AIML deep learning technologies that assess the performance of multiple predictive models to adjust to the changing environment [3]. They mimic how human neurons signal among themselves to perform a learning dynamic [41].

The same algorithm forecasting method was applied to all SKUs in considering the transactional data shown in Figure 1. The input variables to the AIML-driven model include factors such as the transactional demand forecast, customer (retailer) demand signals, scheduled promotions, and actual historical customer demand or shipment records.

The human forecasters were a team of 15 experts working for the FMCG company; each had more than 4 years of forecasting experience in the e-commerce retail industry. Each forecaster had expertise in particular product categories. This expertise served as the basis for their contextual knowledge when intervening in the forecasting process. Expert forecasters processed the same SKU throughout the overall forecasting process. The initial and interim engagement between the algorithm and the forecaster was triggered either by the treatment allocated to the SKU in this study or by the initiative of the forecaster to review the corresponding forecasting steps (This did not occur in the case of Treatment 1-automation).

Forecasters were not informed about the various experimental conditions or the general goal of the study when providing forecasting judgments. However, they were aware that their intervention in the demand forecasting process was recorded for the purpose of improving the AIML demand forecasting system. All the experts had access to the same sources of information from different functions in the company via their Enterprise Resource Planning (ERP) system, Warehouse Management Systems, and a repository of historical forecasting data. During the forecasting process, they could identify contextual data from these sources and use web-based interfaces to interact with the forecasting process and add their key insights.

In the overall forecasting process, the human forecaster could discuss specific situations of certain SKUs with other functions in the FMCG company, including, for example, recent changes in the marketing strategy for that SKU, changes in displays or marketing channels, priorities for other key customers that eventually could affect the availability of that SKU, and changes in the production strategy, such as new equipment or re-routing the product to another distribution facility. Sometimes, the regular data in their ERP system is not available with the expected quality; for example, in terms of the granularity or the expected

frequency in refreshing the data. In this case, the forecaster must interpret the context to mitigate this lack of data quality. At times, the forecaster receives signals from the e-commerce customer via their commercial department that require parameters to be adjusted.

Experimental Design

We studied human–AIML collaboration in a field experiment, in which we analyzed the effect of human intervention along the six steps of the demand forecasting process adopted by the company (Figure 1). Each of the steps could be adjusted by human judgment-based intervention.

Consistent with our conceptualization, we distinguished between three levels of human intervention in human–AIML collaboration that represented our experimental conditions, and each SKU in the sample was randomly assigned to one of the conditions. The first level is *automation*, in which the AIML-based model is fully responsible for generating demand forecasts and operates without human intervention in any of the 6 steps of the process. The second level is *adjustable automation*, a medium level of human intervention in which both human forecasters and AIML-based models interact only for the two main steps (2 and 4) of the six-steps in the demand forecasting process. The third level, *augmentation*, is characterized by the highest level of intervention, such that the work is distributed more evenly between the forecaster and AIML model. In this case, human forecasters interact with the AIML during all six steps of the demand forecasting process.

Considering the additional workload for the expert forecasters in the augmentation condition, who had to work in the experiment setting and participate in all forecasting steps, we placed an additional restriction of a maximum threshold of SKUs under this condition. After having randomly assigned SKUs to one of the three experimental conditions in our study, our sample included 873 SKUs for the automation condition, 746 SKUs for the adjustable automation condition, and 269 SKUs for the augmentation condition.

Empirical Methodology

As mentioned before, the goals of our research were to estimate the causal effect of human intervention in human–AIML collaboration on demand forecast accuracy and to determine how this effect differs in terms of demand uncertainty and time horizon. Consequently, we applied the treatment effects models based on the three levels of human intervention defined above. We assigned value 1 to the no intervention treatment level (automation), value 2 was assigned to the medium level of intervention (adjustable automation), and value 3 was assigned to the higher level of intervention (augmentation). The analyzed outcome was demand forecast accuracy, in the form of forecast errors.

To examine the data obtained from the field experiment, we followed the analytical strategy outlined in Revilla and Rodríguez-Prado [64] and Cattaneo and colleagues [16]. We used the following two indices: treatment level j ($j = 1, 2, 3$) and SKU i ($i = 1, 2, 3, \ldots, n$). Our model is fully specified based on random vector $z_i = \left(y_i, w_i, x_i^{'}\right)$, where y_i is the outcome variable (forecasting accuracy error for short- and long-time horizons), w_i denotes the experimental condition, and x_i represents the $k_x \times 1$ vector of covariates (characteristics of each SKU).

Based on the causal effects of each treatment level on predictive accuracy, we built a counterfactual model. Only the actual outcomes, y_i, were observed (forecast accuracy error for the short- and long-time horizon); other potential outcomes after treatment, $y_i(j)$, could not be measured. Therefore, we created an independent variable to represent the outcomes for every SKU i:

$$[y_i(1), y_i(2), \ \ldots, y_i(J)], y_i = d_i(1)y_i(1) + \ldots \ + d_i(J)y_i(J),$$

where $d_i(j) = 1(w_i = j)$, which implies that SKU i received treatment j; otherwise, $d_i(j) = 0$.

Following Winship [82], we wanted to observe the aggregated effect of each treatment level j by calculating the mean of potential outcomes with their distribution $E[Y(j)]$. Note that the outcome (forecast accuracy) for a particular SKU could also be observed if other experimental conditions had been applied. Thus, we calculated the average treatment effect on forecast errors as follows:

$$E[Y(j) \ - Y(j')],$$

where j is the experimental condition applied, instead of another condition j'.

We computed an augmented inverse propensity weighted (AIPW) estimator to estimate the experimental condition effect. This method combines an adjustment of the regression model with a propensity score. Thus, according to Tan [73], we get the benefit of doubling the robustness property because either the model for estimating the effect, or the regression model for our outcome, is specified correctly; therefore, the AIPW estimator will be consistent.

According to Cattaneo and colleagues [16] and Glynn and Quinn [33], to determine the AIPW estimator we first computed a multinomial logit model for the generalized propensity score from the estimation of the treatment model. To estimate each treatment's outcome for each SKU, we used regression analysis. The potential means were subsequently calculated, weighting predicted outcomes means, using the inverse propensity weight calculated at the beginning of the process.

Implementing the AIPW method implies consideration of both the unconfoundedness (conditional independence) and common support assumptions [16]. Unconfoundedness can be satisfied if we control covariates which demonstrate that potential outcome distributions are not related to specific treatment levels. We relied on the multiple covariates analyzed in our research, which cover all the potential factors relevant to treatment level on the outcomes. Several empirical research studies have relied on this assumption [64, 72]. Common support assumptions required us to observe effects for each type of SKU for all treatment levels, which are determined by predicting the previously mentioned generalized propensity scores.

Outcome and Contextual Variables

A common method for quantifying forecast accuracy is to study the quality of forecasts in terms of the error component, which is determined by the difference between actual demand and forecast demand and expressed as a percentage of the actual demand.

To analyze the effects of the factors that define the context of the prediction (forecasting demand) in the retail environment, two key variables were considered: demand uncertainty, which is operationalized in terms of product innovativeness, and the time horizon.

Scholars have proposed that innovativeness of the product for which sales are predicted represents a major cause of demand uncertainty [31]. In fact, previous research has distinguished between two types of consumer products that differ primarily according to their associated demand patterns: basic (or functional) products and innovative (or fashion) products [25]. Basic products satisfy basic needs. These products are bought regularly and are considered staples, with a predictable, stable demand and long-life cycles. Because the characteristics of basic products do not change frequently, large amounts of historical data are usually available when forecasting future demand. In contrast, new, innovative products have short life cycles and a more fluctuating, uncertain demand. Because of their short time in the market, these products have little historical data prior to a selling period, which makes demand forecasting extremely difficult. In addition, these products may also be associated with new attributes on which AIML-based models need to be trained to predict future sales.

Consequently, product innovativeness is measured based on the product's launch date. For FMCG products, if forecasting occurs within 6 months of product launch, items are defined as innovative products with new life cycles and scant, if any, historical data. If forecasting occurs more than 6 months from product launch, products are considered well-established in the market. Therefore, this variable (product innovativeness) is divided into two groups of SKUs: 1) when the launch time until the time of forecasting is less than 6 months (innovative product, so high uncertainty), or 2) when the launch time until the time of forecasting is more than 6 months (established product, so low uncertainty).

Our second key variable, time horizon, is determined by means of discriminating the demand forecast accuracy error for long-term versus short-term time horizons. The long-term demand forecast, also known as the *layout accuracy error*, is a 13-week forecast time horizon. Therefore, when determining the demand forecast over actual demand, the value would be more than 100 percent, but it would be less than 100 percent if determining actual demand over the forecast demand. The short-term demand forecast, also known as the *mean absolute percentage error*, uses a forecast horizon of 2 weeks.

Pretreatment Variables

To fulfill the conditional independence assumption, we used a wide range of pretreatment variables. They were selected to control the potential influence of SKU-specific product features on predictive accuracy. We selected the following pretreatment variables based on the features of each product: price, turnover rate, product segmentation, and product category. First, we controlled for the price of the particular SKU using a binary variable, differentiating between high and low prices according to the customer's assessment. We considered the customer dynamics for each SKU to assess the potential impact of the turnover rate to determine whether sales were moving quickly or slowly. We also created a binary variable for turnover rate: 1 for a high turnover rate (<40 days), 0 for a low turnover rate (≥40 days). We differentiated product segmentation, again using binary variables, from six different segments, from A to F, according to the customer's classifications, which represent particular operational practices and reflect the relevance of a particular SKU each week. Finally, we considered the effect of product category based on which of the seven categories each SKU belongs to, as mandated by the FMCG industry, and included binary variables according to the type of end user of the product. Note that different resources and

Table 1. Sample sizes and descriptive statistics.

Total Observation Treatment Groups (N = 1888)			n
Automation (1)			873
Adjustable automation (2)			746
Augmentation (3)			269
Variables	Mean	SD	Obs.
Outcome variables			
Accuracy error for long-term forecast	1.24	0.43	1888
Accuracy error for short-term forecast	0.58	0.03	1888
Pretreatment variables			
Turnover	71.72	161.99	1888
Turnover_dummy	0.48	0.50	1888
Basic price	58.03	61.55	1888
Price_dummy	0.47	0.50	1888
New product_dummy	0.27	0.45	1888
Established product_dummy	0.73	0.45	1888
Product segmentation (a–f)			1888
Product category			
Category_1	0.06	0.24	116
Category_2	0.48	0.50	905
Category_3	0.01	0.11	24
Category_4	0.06	0.24	117
Category_5	0.19	0.40	365
Category_6	0.04	0.20	81
Category_7	0.15	0.36	280

Abbreviations: Obs., observed; SD, standard deviation.

management teams could have been assigned to a particular product category, which would have resulted in differences in performance.

Table 1 summarizes our pretreatment variables, in terms of sample sizes and descriptive statistics. We conducted our analysis using Stata/IC and the command group "teffects aipw" to calculate the average treatment effects results.

Results

Table 2 shows the estimated potential mean forecast error of long-term demand forecasts for each of the three treatment levels: *automation, adjustable automation*, and *augmentation*. It also reports the average treatment effects when comparing one experimental condition versus another. Standard errors and levels of significance are also noted. We were interested in analyzing how the different experimental conditions varied in terms of forecast error based on the contextualization variables (product innovativeness and forecast time horizon).

First, when analyzing the effects of product innovativeness, we observe significant levels of accuracy error for long-term forecasts; this error increases when comparing treatment level 3 to level 2, and treatment level 3 to level 1, as shown in Table 2 (left column). A similar representation is seen in Table 3 (left column) but, in this instance, the outcome variable is forecast accuracy error for a short-term demand forecast. For innovative products in the market, better accuracy for a short-term horizon is obtained when the AIML system generates forecasts without human intervention. The worst scenario (when the error increases) occurs when we compare treatment level 3, augmentation, with level 1, automation. Therefore, based on the significant results presented in Tables 2 and 3, we conclude

Table 2. Average treatment effect by product innovativeness for long forecast time horizon.

	Accuracy error for long-term time horizon				
	Innovative Product			Established Product	
Human Intervention in Human-AIML Collaboration	Potential Mean (Percent)	SE	Significance of Two-Sided Test for Equality	Potential Mean (Percent)	SE
Automation (1)	1.05	0.07	***	1.59	0.09
Adjustable Automation (2)	0.92	0.05	***	1.35	0.08
Augmentation (3)	2.23	0.39	***	0.78	0.06
	Average treatment effect	SE		Average treatment effect	SE
2 vs. 1	–0.12	0.07	***	–0.15**	0.07
3 vs. 1	1.13***	0.39	***	–0.51***	0.05
3 vs. 2	1.41***	0.44	***	–0.42	0.05

Notes: Augmented inverse propensity weighted estimators controlling for product difference in product segmentation category, turnover, product tier, and product price.
Abbreviations: AIML, AIML, artificial intelligence/machine learning; SE, standard error.
*, **, ***Significant at 10 percent, 5 percent, and 1 percent, respectively.

that H1 is fully supported, verifying the negative effect of human intervention in human-AIML collaboration on forecast accuracy for innovative retail products (high uncertainty).

When we focus on established retail products, we observe in Table 2 that lower levels of forecast errors are observed when augmentation is applied to products that have been in the market for a long-time horizon, and demand forecasts for a long-term horizon were elicited (upper part of Table 2, right column). In this context, there is a significant reduction in forecast error for long-term predictions as we move from treatment level 2 (adjustable automation) to level 1 (automation), and the same happens when moving from treatment level 3 (augmentation) to level 1 (automation) (Table 2, lower part, right column). For established retail products the forecast error for short-term predictions (Table 3, right column) is reduced when we compare treatment level 2, adjustable automation, to level 1, automation. Conversely, this error increases when we compare more human-centered decision-making in forecasting (treatment level 3, augmentation) versus medium intervention (level 2, adjustable automation). The difference between augmentation (treatment level 3) versus automation (level 1) is not significant. For established retail products,

Table 3. Average treatment effect by product innovativeness for short forecast time horizon.

	Accuracy Error for Short-Term Time Horizon				
	Innovative Product			Established Product	
Human Intervention in Human-AIML Collaboration	Potential mean (Percent)	SE	Significance of Two-Sided Test for Equality	Potential Mean (Percent)	SE
Automation (1)	0.81	0.09	***	0.50	0.03
Adjustable automation (2)	1.00	0.17	***	0.31	0.03
Augmentation (3)	1.99	0.50	***	0.55	0.08
	Average treatment effect	SE		Average treatment effect	SE
2 vs. 1	0.25	0.24	***	–0.39***	0.06
3 vs. 1	1.47**	0.66	***	0.10	0.17
3 vs. 2	0.98*	0.59	***	0.81***	0.31

Notes: Augmented inverse propensity weighted estimators controlling for product difference in product segmentation category, turnover, product tier, and product price.
Abbreviations: AIML, artificial intelligence/machine learning; SE = standard error.
*, **, ***Significant at 10 percent, 5 percent, and 1 percent, respectively.

lower accuracy errors for the short-term horizon are obtained under treatment level 2, adjustable automation (Table 3, right column).

Consequently, we conclude that, in predictions with low uncertainty (i.e., established retail products), human intervention in human-AIML collaboration is positively associated with forecasting accuracy. Therefore, H2 is supported.

When analyzing time horizon, augmentation clearly shows the highest level of forecast error for short-term forecasts regarding innovative products (high uncertainty). However, for established products (low uncertainty), forecast errors improve when we compare augmentation and adjusted automation, but errors worsen for adjusted automation versus automation. Consequently, H3 is partially confirmed.

Analyzing the results from Table 2, we observe the significant results of Average Treatment Effect(ATE) for long forecast time horizons (lower part). For established products, the highest reduction in accuracy error occurs in high levels of human intervention (both augmentation and adjusted automation) as compared to automation alone. For innovative retail products (high uncertainty), the highest increase in forecast errors occurs when we compare augmentation (treatment 3) with both treatments 1 and 2, which suggests that the best performance is obtained by reducing human intervention in the collaboration when predicting the demand for innovative products. Therefore, H4 is fully confirmed.

Tables 2 and 3 summarize the main results extracted from our data. The results allow us to generally confirm that differences in human intervention in human-AIML collaboration impact predictive accuracy depending on the factors that define the context of the prediction, such as uncertainty (in terms of product innovativeness) and time horizon.

Discussion

This study extends emerging research on human-AI collaboration in the information systems (IS) literature by investigating human and AI complementarities and identifying opportunities to overcome biases in the predictions of both humans and machines. We aim to understand the heterogeneity of relationships that the context can create around human-AI collaboration. Toward this broader objective, we focus our research on AIML predictions and investigate effective ways of combining humans and AIML, considering the limitations and strengths of both agents. We argue that the two key dimensions of the prediction context—uncertainty and time horizon—have a critical impact on the effectiveness of human intervention in AIML-based predictions. Our findings suggest that human intervention complements AIML-driven models most effectively in predictions with low uncertainty and long-time horizons, where humans are proficient in relying on past information to diagnose anomalous changes in the data series. In contrast, AIML-driven models appear to perform best autonomously in environments with high uncertainty and short time horizons, thanks to their ability to detect and exploit interrelations with variables beyond the prediction at hand. Between these two extremes, we find that a moderate level of human intervention is likely to result in the best predictive performance in human-AIML collaboration.

Theoretical Implications

Our research answers the recent call for obtaining a better understanding of human-AI collaboration [9, 26, 61, 69, 74], by showing that the effectiveness of human-AIML collaboration depends on the characteristics of the prediction context. Specifically, we identify task uncertainty and the time horizon available until the prediction materializes as two key factors that explain the relationship between human–AI collaboration and performance.

Our study has also allowed us to shed light on how the predictive context might require changes in the configuration of human-AIML collaborations. We introduce uncertainty and time horizon in retail demand forecasting as two key factors and identify forecasting challenges for various levels of human-AIML collaboration. Product innovativeness, as the main source of uncertainty, is critical to our understanding of how the human intervention in the human-AIML collaboration affects demand forecast accuracy because, to varying degrees, human judges evolve their predictions based on available data. Two key considerations here are historical data and contextual data, because they are often integral components of demand forecasting. AIML-based models perform well with historical data and eliminate variants that do not conform to these data [4]. A lack of contextual data may result in AIML models making poor decisions, particularly when unexpected changes that might alter underlying cause-effect relationships are not fully appreciated.

Our research indicates that human expertise can complement AIML-driven prediction models when the input of the algorithm lacks contextual data. In addition, AIML predictive models handling historical data affect forecasters' effectiveness significantly when they analyze contextual data, which also suggests that humans and AI complement each other [35, 42, 81, 85]. Building on this finding, we emphasize the emergence of a hybrid human-AIML collective intelligence and offer a vision of augmentation as an evolutionary process during which humans learn from machines and machines learn from humans. This requires repeated interactions between the independent judgment of AIML predictive models and the contextual, practically relevant criteria of domain experts [78]. Going further, companies should regularly question the acquisition of knowledge that is being produced in this co-evolutionary process over the course of its development. Knowledge acquired through this process should serve to maximize AI capabilities and minimize automated rational judgment biases. Simultaneously, mutual learning through human and AI interaction should help revise humans' preconceived biases and augment decision-making.

The time horizon is also a key consideration for how companies should distribute the work between humans and AIML-driven prediction models. In our theorization, we mention that humans are limited in terms of both information processing speed and capacity, which creates biases in decision-making for short-term demand forecasts. Moreover, humans need time to perceive and understand changes in the environment that drive sound decisions. In contrast, AIML-driven prediction models' speed and ability to discover systematic patterns by analyzing large volumes of historical data compensate temporally for the lack of managerial judgment. In part, the retail data from our study support the insights of prior research since it demonstrates the positive impact of AIML-based decision-making (automation) on short-term demand forecasting performance for innovative products [69].

As a result of this work, we propose a framework for understanding the human intervention in human-AIML collaboration. Consistent with our empirical findings, Figure 2 illustrates how the effectiveness of human intervention varies along two critical dimensions

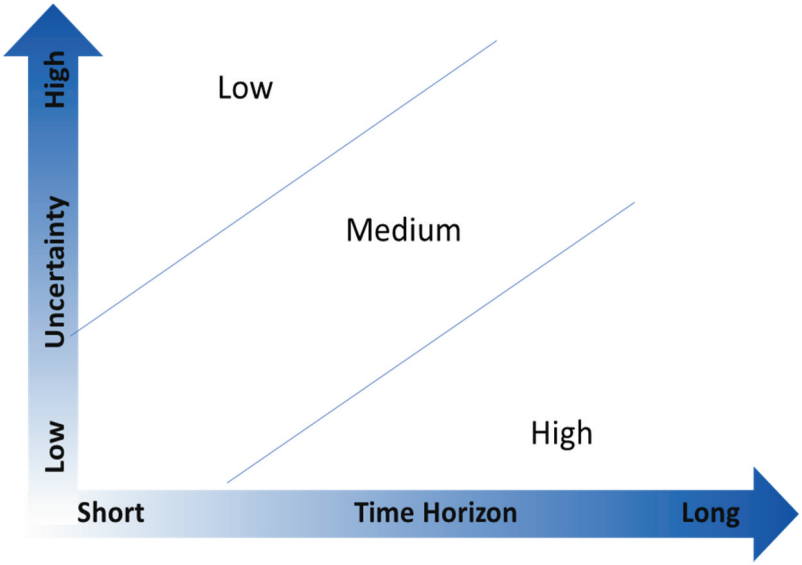

Figure 2. The level of human intervention in Human-AIML collaboration in predictions. Abbreviations: AIML, artificial intelligence/machine learning.

of the predictive context: uncertainty and the time horizon. It shows that human judgment is likely to complement AIML-driven models when task environments are associated with long time horizons and low uncertainty. In contrast, our research suggests that this is not the case when the task requires coping with short time horizons and high uncertainty, in which case human judgment is unlikely to improve AIML-driven models. In between these two extremes, a moderate degree of reliance on the human judgmental component in human-AIML collaboration will likely result in the best predictive performance.

Our findings can be extrapolated to managing the demand instability resulting from the COVID-19 pandemic. In some markets, such as tourism, sales dropped to levels that had previously been unknown. In other markets, such as grocery stores, demand outweighed products on hand. Because we had never seen the impact of a global pandemic on demand, human forecasters who relied on prior knowledge and experience often made poor forecasts. Human thinking may not be fast enough to incorporate new and anomalous information, and models rely heavily on previous behavior [68]. Therefore, our research proposes the use of AIML automation to create forecasts based on information about anomalous situations. Human forecasters needed knowledge of the expected impact of the virus to make quick adjustments to the AIML system to accurately predict demand in the second wave of COVID-19. In this scenario, the positive effect of human-AIML collaboration on forecast accuracy is enhanced until augmentation is again required. In conclusion, we suggest that the richness of contextual knowledge associated with human learning ensures more accurate predictions. By establishing the interdependences between humans and AIML, we show that contextual data complement historical data and enable a more reliable collaboration for long-term forecast horizons.

Managerial Implications

Managers should be aware that the implementation of AIML-driven systems requires a profound understanding of the context in which AIML is applied. Some AIML applications assume that adapting to the context can be done by the AIML system itself. But this can result in unexpected biases and undesirable consequences. To obtain quality performance, contextual data that describe how humans and AIML make decisions together are essential.

The difficulties inherent in generating predictions in cases of uncertainty, such as in the case of demand forecasting, have always been a focus of attention for supply chain managers and solutions vendors, who have progressed toward using advanced and highly sophisticated algorithms that promise to dispense with the singularities of uncertainty. Our research may help to improve their AIML predictive analytics by demonstrating that the key issues relate not only to the configuration of the algorithm but also to how it is adaptively applied and integrated with human expertise in different contexts to reach the best performance.

Another important managerial implication relates to the different levels of collaboration in supply chain decision-making. Collaboration between supply chain actors to establish greater levels of performance in demand predictive analytics has been a long-standing practice. However, the new trends of incorporating AIML-driven algorithms, applied to both forecasting demand and lead times, as well as their contextualization based on supply chain operations, can benefit from this research.

This research also has implications for the levels of efficiency that managers can expect when AI technology is introduced to digitize business processes. In this instance, forecasting represents one of the most popular tasks in which AIML is being used. This is due primarily to the broad availability of historical demand data and the need to make effective decisions associated with production and distribution. In addition, given how expensive an AIML algorithm can be to implement and scale up, managers often expect to extract high returns on investment. However, our findings suggest that managers who implement AIML algorithms should temper these expectations. Since predictive performance is highly context-dependent, expert forecasters should consistently account for factors that define a particular forecasting environment in terms of product features and the characteristics that define the context of the decisions to be made after the forecasting is generated. The same recipe cannot be used for all forecasting circumstances; large investments in AIML implementation may not deliver the desired results.

Limitations and Future Research Directions

Our study is subject to several limitations that offer opportunities for future research. First, although the field experiment underlying our study has the advantage of offering a better understanding of the relationship between human-AIML collective intelligence and performance by controlling for sources of extraneous variance and ensuring consistent definitions and measurements, additional studies in alternative industries could further generalize our findings.

Second, the AIML algorithm used in the field experiment is specific to the task of demand prediction. There is an ample variety of AIML-based algorithms and AI-driven technologies applied to multiple functions in a company for which the insights from our study may not hold. Thus, our findings should be interpreted with caution since they have been tested in the specific domain of demand forecasting.

Third, to ensure the reliability and independent condition of the experimental analysis, our research focuses only on how human intervention affects the predictive performance of human-AIML collaboration for a limited demand forecasting outcome. This focus did not enable us to empirically analyze potential longitudinal effects that may emerge when humans and AIML learn from each other over time. For example, as previous studies suggest [83], it will be important to understand how human trust variations related to AIML prediction influence the predictive performance of human-AIML collaboration. For example, understanding how performance varies ranging from an initially fragile interaction, based on a lack of transparency in AIML prediction or algorithm aversion, to a deep understanding of the value provided by the AIML model, could be insightful.

Fourth, due to our study's scope and our resource limitations, only three human-AIML decision-making structures are addressed in our research. The implications of different scenarios and other ways to operationalize human-AIML collaboration should be explored.

Finally, although we propose two important characteristics of predictions (uncertainty and time horizon) to analyze the influence of the context on human-AIML collaboration, other studies could examine alternative influence factors related to human cognition (e.g., creativity, common sense, introspection). Another avenue for future research may be to examine in more depth the implications of human-AIML collaboration for other types of organizational outcomes, such as penalty-based versus reward-based consequences or trust in AI. While the current literature recognizes the potential benefits of human-AIML collaboration in these contexts, more research is needed to comprehensively study these variables [61].

Conclusions

AI and humans are growingly seen as highly complementary in terms of their capabilities. However, there is still a lack of know-how in making the most of this collaboration. To address this issue, we investigate the role of human intervention in AIML-driven prediction and show that the uncertainty and the time horizon of the prediction determine the forecast accuracy for three different types of collaboration: automation, adjustable automation, and augmentation. As a result of this work, we propose a framework that explains the level of human intervention required in human-AIML collaboration, extending the knowledge of how the phenomenon of AIML is related to organizational design. Specifically, it shows that human intervention complements AIML-driven forecasts most effectively when predictions are associated with long time horizons and low uncertainty, whereas too much reliance on human intervention is likely to be detrimental in prediction with high uncertainty and short time horizons. In between these two extremes, a moderate degree of reliance on the human intervention in human-AIML collaboration will likely result in the best predictive performance.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Elena Revilla is a professor of Operations Management in Operations & Business Analytics at IE Business School, Spain. She is also a post-doctoral fellow at the University of North Carolina, Chapel Hill and visiting researcher at MIT. She received her PhD in Business Administration (Operations) from the Universidad de Valladolid, Spain. Dr. Revilla specializes in organizational behavior in the operational and supply-chain management, specifically in the emerging digitalization context. Her work has appeared in the academic journals as Journal of *Management, Journal of Operations Management, Research Policy, Journal of Supply chain, Decision Science, Human Resources Management*, and *Sloan Management Review*, as well as in managerial outlets such as *Harvard Business Review* and *MIT Sloan Management Review*. Her research has received funding from the European Union.

Maria Jesus Saenz is the Director of the Digital Supply Chain Transformation Lab at the Massachusetts Institute of Technology, Center for Transportation and Logistics. Dr. Saenz also serves as the Executive Director of the MIT Supply Chain Management Master Programs. She received her PhD in Manufacturing and Design Engineering from the University of Zaragoza, where she obtained her tenure as Associate Professor. The primary research examines new collaborative paradigms that arise while implementing new digital technologies in data-driven supply chains. Dr. Saenz is co-author of more than 100 publications, including in *Journal of Supply Chain Management, European Journal of Operational Research, MIT Sloan Management Review, Forbes*, and *Financial Times Press*, among others.

Matthias Seifert is an Associate Professor of Decision Sciences in the Operations & Business Analytics Department at IE Business School. Previously, he was affiliated with the London Business School, Cambridge University and the London School of Economics and Political Science. Dr. Seifert's research focuses on decision making under risk and uncertainty, and managerial forecasting. His work has been published in academic journals including *Management Science, Organizational Behavior and Human Decision Processes, Journal of Operations Management, Nature Human Behavior, Personality and Social Psychology Bulletin*, and others, as well as in practitioner outlets such as *Harvard Business Review* and *MIT Sloan Management Review*.

Ye Ma is Assistant Manager at McKinsey & Company. He obtained his Master of Engineering in Supply Chain Management at the Massachusetts Institute of Technology. His expertise has developed in the areas of digitalization, strategy, operations and technology in consulting and fast-moving consumer good companies.

References

1. Agrawal, A.; Gans, J.S.; and Goldfarb, A. Exploring the impact of artificial Intelligence: Prediction versus judgment. *Information Economics and Policy*, 47 (2019), 1–6.
2. Al-Amoudi, I.; and Latsis, J. Anormative black boxes: Artificial intelligence and health policy, In E. Lazega and I. Al-Amoudi (eds.), *Post-Human Institutions and Organisations: Confronting the Matrix*. Oxfordshire, UK: Taylor & Francis, 2019, pp. 119–142.
3. Amar, J.; Rahimi, S.; Surak, Z.; and von Bismarck, N. AI-driven operations forecasting in data-light environments. *McKinsey & Company*. 2022. https://www.mckinsey.com/capabilities/operations/our-insights/ai-driven-operations-forecasting-in-data-light-environments (accessed October 2022).
4. Balasubramanian, N.; Ye, Y.; and Xu, M. Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, 47, 3 (2022), 448–465.
5. Beckman, C. M.; Haunschild, P. R.; and Phillips, D. J. Friends or strangers? Firm-specific uncertainty, market uncertainty, and network partner selection. *Organization Science*, 15, 3 (2004), 259–275.

6. Benbya H.; Davenport T.H., and Pachidi S. Artificial intelligence in organizations: Current state and future opportunities. *MIS Quarterly Executive*, 19, 4 (2020), 9–21.
7. Benbya, H.; Nan, N.; Tanriverdi, H.; and Yoo, Y. Complexity and information systems research in the emerging digital world. *Mis Quarterly*, 44, 1 (2020), 1–17.
8. Benbya, H.; Pachidi, S.; and Jarvenpaa, S. Special issue editorial: Artificial intelligence in organizations: implications for information systems research. *Journal of the Association for Information Systems*, 22, 2 (2021), 10.
9. Berente, N.; Gu, B.; Recker, J.; and Santhanam, R. Managing artificial intelligence. *MIS Quarterly*, 45, 3 (2021), 1433–1450.
10. Blattberg, R. C.; and Hoch, S. J. Database models and managerial intuition: 50% model + 50% manager. *Management Science*, 36, 8 (1990), 887–899.
11. Brehmer, B. Response consistency in probabilistic inference tasks. *Organizational Behavior and Human Performance*, 22, 1 (1978), 103–115.
12. Brown, A.; Chouldechova, A.; Putnam-Hornstein, E.; Tobin, A.; and Vaithianathan, R. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. (2019), 1–12.
13. Brown, S. Machine learning, explained. 2021. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained (accessed October 2022).
14. Brynjolfsson, E.; and Mitchell, T. What can machine learning do? Workforce implications. *Science*, 358, 6370 (2017), 1530–1534.
15. Brynjolfsson, E.; Rock, D.; and Syverson, C. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. In *The Economics of Artificial Intelligence: An Agenda*. Chicago: University of Chicago Press, 2018, pp. 23–57.
16. Cattaneo, M.D.; Drukker, D.M.; and Holland, A.D. Estimation of multivalued treatment effects under conditional independence. *The Stata Journal*, 13, 3 (2013), 407–450.
17. Chandra, S.; Shirish, A.; and Srivastava, S.C. To be or not to be . . . Human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. *Journal of Management Information Systems*, 39, 4 (2022), 969–1005
18. Choudhury, P.; Starr, E.; and Agarwal, R. Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal*, 41, 8 (2020), 1381–1411.
19. Cyert, R.M.; and March, J.G. *A Behavioral Theory of the Firm*. Englewood Cliffs, NJ: Prentice-Hall, 1963, pp. 169–187.
20. Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (accessed October 20, 2020).
21. Davenport, T.H.; and Kirby, J. Beyond automation. *Harvard Business Review*, 93, 6 (2015), 58–65.
22. De-Arteaga, M.; Fogliato, R.; and Chouldechova, A. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. (2020), 1–12.
23. E2open. 2021 Forecasting and Inventory Benchmark Study. 2021. https://cdn.brandfolder.io/IDKCKNW5/at/68qzkp9n3ttw39ncvxhvcx7n/E2open_2021_Forecasting_and_Inventory_Benchmark_Study_Full_Report.pdf (accessed October 2022).
24. Farwell, D.C. Dss mechanisms for judgmental flexibility: An exploratory survey. *Journal of Management Information Systems*, 1, 2 (1984), 72–79.
25. Fisher, M.; and Rajaram, K. Accurate retail testing of fashion merchandise: Methodology and application. *Marketing Science*, 19, 3 (2000), 266–278.
26. Füegener, A.; Grahl, J.; Gupta, A.; and Ketter, W. Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. *Information Systems Research*, 33, 2 (2021), 678–696.
27. Gaissmaier, W.; and Schooler, L. The smart potential behind probability matching. *Cognition*, 109, 3 (2008), 416–422.
28. Galbraith, J. *Designing Complex Organizations*. Reading, MA: Addison-Wesley, 1973.

29. Gartner. Current use cases for machine learning in supply chain solutions. 2018. https://www.gartner.com/en/documents/3875876/current-use-cases-for-machine-learning-in-supply-chain-p (accessed September 7, 2020).
30. Gaur, V.; Kesavan, S.; Raman, A.; and Fisher, M.L. Estimating demand uncertainty using judgmental forecasts. *Manufacturing & Service Operations Management*, 9, 4 (2007), 480–491.
31. Germain, R.; Claycomb, C.; and Dröge, C. Supply chain variability, organizational structure, and performance: The moderating effect of demand unpredictability. *Journal of Operations Management*, 26, 5 (2008), 557–570.
32. Glebbeek, A.C.; and Bax, E.H. Is high employee turnover really harmful? An empirical test using company records. *Academy of Management Journal*, 47, 2 (2004), 277–286.
33. Glynn, A.N.; and Quinn, K.M. An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18, 1 (2010), 36–56.
34. Goodwin, Paul, Brent Moritz, and Enno Siemsen. 2018. "Forecast Decisions." In The Handbook of Behavioral Operations, edited by Karen Donohue, Elena Katok, and Stephen Leider, 433–458. Hoboken, NJ: John Wiley & Sons, Inc.
35. Guszcza, J.; Lewis, H.; and Evans-Greenwood, P. Cognitive collaboration: Why humans and computers think better together. *Deloitte Review*, 20 (2017), 8–29.
36. Guthrie, P.M. *Supply chains need AI, but AI needs humans*. 2020. https://www.kinaxis.com/ko/node/2873 (accessed October 20, 2020).
37. Harvey, N.; and Reimers, S. Trend damping: Under-adjustment, experimental artifact, or adaptation to features of the natural environment? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 2 (2013) 589–607.
38. Harvey, N.; Ewart, T.; and West, R. Effects of data noise on statistical judgement. *Thinking & Reasoning*, 3, 2 (1997), 111–132.
39. Hogarth, R. M.; and Makridakis, S. Forecasting and planning: An evaluation. *Management science*, 27, 2 (1981), 115–138.
40. Hong, L.; Lamberson, P. J.; and Page, S. E. Hybrid predictive ensembles: Synergies between human and computational forecasts. *Journal of Social Computing*, 2, 2 (2021), 89–102.
41. Huber, J.; and Stuckenschmidt, H. Daily retail demand forecasting using machine learning with emphasis on calendric special days. *International Journal of Forecasting*, 36, 4 (2020), 1420–1438.
42. Johnson, M.; and Vera, A. No AI is an island: the case for teaming intelligence. *AI Magazine*, 40, 1 (2019), 16–28.
43. Kahneman, D.; and Tversky, A. On the psychology of prediction. *Psychological review*, 80, 4 (1973), 237.
44. Kim, C.N.; and McLeod Jr, R. Expert, linear models, and nonlinear models of expert decision making in bankruptcy prediction: A lens model analysis. *Journal of Management Information Systems*, 16, 1(1999), 189–206.
45. Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Obermeyer, Z. Prediction policy problems, A*merican Economic Review*, 105, 2 (2015), 491–495.
46. Kliegr T.; Bahník S.; and Fürnkranz J. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295 (2021), 103458.
47. Kremer, M.; Moritz, B.; and Siemsen, E. Demand forecasting behavior: System neglect and change detection. *Management Science*, 57, 10 (2011), 1827-1843.
48. Kumar, V.; Rajan, B.; Venkatesan, R.; and Lecinski, J. Understanding the role of artificial intelligence in personalized engagement marketing. *California Management Review*, 61, 4 (2019), 135–155.
49. Lai, V.; Chen, C.; Liao, Q. V.; Smith-Renner, A., and Tan, C. Towards a science of human-ai decision making: A survey of empirical studies. *arXiv preprint arXiv:2112*.11471 (2021). https://arxiv.org/pdf/2112.11471.pdf
50. Launay V. EN. The impact of high frequency trading. 2019. https://www.centralcharts.com/en/gm/1-learn/5-trading/16-automatic/312-the-impact-of-high-frequency-trading (accessed September 7, 2020).

51. Lawrence, M.; and O'Connor, M. Exploring judgmental forecasting. *International Journal of Forecasting*, 8, 1 (1992), 15–26.
52. Lawrence, M.; Goodwin, P.; O'Connor, M.; and Önkal, D. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting*, 22, 3 (2006), 493–518.
53. Li, W.; and Chai, Y. Assessing and enhancing adversarial robustness of predictive analytics: An empirically tested design framework. *Journal of Management Information Systems*, 39, 2 (2022), 542–572.
54. Lindebaum, D.; Vesa, M.; and Den Hond, F. Insights from "the machine stops" to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review*, 45, 1 (2020), 247–263.
55. Massey, C.; and Wu, G. Detecting regime shifts: The causes of under-and overreaction. *Management Science*, 51, 6 (2005), 932–947.
56. McLaren, J. 4 Keys to Better Retail Promotion Forecasting and Replenishment. 2020. https://www.relexsolutions.com/resources/4-keys-to-better-retail-promotion-forecasting-and-replenishment/ (accessed December 15, 2020).
57. Metcalfe, J. S.; Perelman, B. S.; Boothe, D. L.; and Mcdowell, K. Systemic oversimplification limits the potential for human-AI partnership. *IEEE Access*, 9 (2021), 70242–70260.
58. OECD Framework for the classification of AI systems. OECD Digital Economy papers. Feb 2022. No. 323
59. Peeters, M.M.M.; van Diggelen, J.; van den Bosch, K.; Bronkhorst, A.; Neerincx, M.A.; Schraagen, J.M.; and Raaijmakers, S. Hybrid collective intelligence in a human–AI society. *AI and Society*, 36 (2021), 217–238.
60. Rai A.; Constantinides, P.; and Sarker, S. Next-generation digital platforms: Toward human-AI hybrids. *Mis Quarterly*, 43, 1 (2019), iii–ix.
61. Raisch, S.; and Krakowski, S. 2021. Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review*, 46, 1 (2021), 192–210.
62. Ransbotham, S.; Khodabandeh, S.; Kiron, D.; Candelon, F.; Chu, M., and LaFountain, B. Expanding AI's impact with organizational learning. *MIT Sloan Management Review*. Research Report, (2020). https://web-assets.bcg.com/f1/79/cf4f7dce459686cfee20edf3117c/mit-bcg-expanding-ai-impact-withorganizational-learning-oct-2020.pdf
63. Rastogi, C.; Leqi, L.; Holstein, K., and Heidari, H. A unifying framework for combining complementary strengths of humans and ML toward better predictive decision-making. *arXiv preprint arXiv:2204.10806*, (2022). https://arxiv.org/pdf/2204.10806.pdf
64. Revilla, E.; and Rodríguez-Prado, B. Building ambidexterity through creativity mechanisms: Contextual drivers of innovation success. *Research Policy*, 47, 9 (2018), 1611–1625.
65. Saenz, M.J.; Revilla, E., and C. Simón. Designing AI systems with human-machine teams. *MIT Sloan Management Review*, 61, 3 (2020), 1–5.
66. Seifert, M.; and Hadida, A.L. On the relative importance of linear model and human judge(s) in combined forecasting. *Organizational Behavior and Human Decision Processes*, 120, 1 (2013), 24–36.
67. Seifert, M.; Siemsen, E.; Hadida, A.L.; and Eisingerich, A.B. Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management*, 36 (2015), 33–45.
68. Sheffi Y. *The new (Ab)normal: Reshaping Business and Supply Chain Strategy Beyond COVID-19*. MIT CTL Media, 2020. https://books.google.com/books?hl=en&lr=&id=lR8AEAAAQBAJ&oi=fnd&pg=PT13&dq=new+normal+early+childhood+management+education&ots=vAmPvnAcTw&sig=g-7OCoIMrlr1jAPoAYU7Kq2OMaE
69. Shrestha, Y.R.; Ben-Menahem, S.M.; and Von Krogh, G. Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61, 4 (2019), 66–83.
70. Simon, H.A. *Models of Man; Social and Rational*. New York: Wiley, 1957.
71. Smith, B.C. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: MIT Press, 2019.
72. Swink, M.; and Jacobs, B.W. Six sigma adoption: Operating performance impacts and contextual drivers of success. *Journal of Operations Management*, 30, 6 (2012), 437–453.
73. Tan, Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97, 3 (2010), 661–682.

74. Teodorescu, M.H.; Morse, L.; Awwad, Y., and Kane, G.C. Failures of fairness in automation require a deeper understanding of human-ML augmentation. *MIS Quarterly*, 45, 3 (2021), 1483–1500.
75. Theocharis, Z.; and Harvey, N. When does more mean worse? Accuracy of judgmental forecasting is nonlinearly related to length of data series. *Omega*, 87 (2019), 10–19.
76. Tofangchi, S.; Hanelt, A.; Marz, D.; and Kolbe, L.M. Handling the efficiency–personalization trade-off in service robotics: A machine-learning approach. *Journal of Management Information Systems*, 38, 1 (2021), 246–276.
77. Van Den Broek, E.; Levina, N.; and Sergeeva, A. In pursuit of data: Negotiating data tensions between data scientists and users of AI tools. *Academy of Management Proceedings*, 2022, 1 (2022), 16942.
78. Van Den Broek, E.; Sergeeva, A.; and Huysman, M. When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, 45, 3 (2021), 1557–1580
79. Van Donselaar, K.H.; Gaur, V.; Van Woensel, T.; Broekmeulen, R.A.; and Fransoo, J.C. Ordering behavior in retail stores and implications for automated replenishment. *Management Science*, 56, 5 (2010), 766–784.
80. Wang, W.; and Benbasat, I. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems*, 23, 4 (2007), 217–246.
81. Wilson, H.J.; and Daugherty, P.R. Collaborative intelligence: humans and AI are joining forces. *Harvard Business Review*, 96, 4 (2018), 114–123.
82. Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge, UK: Cambridge University Press, 2007.
83. You, S.; Yang, C.L.; and Li, X. Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems*, 39, 2 (2022), 336–365.
84. Zanzotto, F.M. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 1 64 (2019), 243–252.
85. Zheng, N.N.; Liu, Z.Y.; Ren, P.J.; Ma, Y.Q.; Chen, S.T.; Yu, S.Y.; Xue, J.R.; Chen, B.D.; and F. Y. Wang. Hybrid-augmented intelligence: collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering*, 18, 2 (2017), 153–179.
86. Ma, Y. Human-Machine Teaming for Intelligent Demand Planning. Thesis Master of Engineering in SCM. DSpace@MIT, Massachusetts Institute of Technology (2020).