Bauer, Kevin; Nofer, Michael; Abdel-Karim, Benjamin M.; Hinz, Oliver

Working Paper

The effects of discontinuing machine learning decision support

SAFE Working Paper, No. 370

Provided in Cooperation with:
Leibniz Institute for Financial Research SAFE

Suggested Citation: Bauer, Kevin; Nofer, Michael; Abdel-Karim, Benjamin M.; Hinz, Oliver (2022) : The effects of discontinuing machine learning decision support, SAFE Working Paper, No. 370, Leibniz Institute for Financial Research SAFE, Frankfurt a. M., https://doi.org/10.2139/ssrn.4299664

This Version is available at:
https://hdl.handle.net/10419/266691

Kevin Bauer | Michael Nofer | Benjamin Abdel-Karim | Oliver Hinz

The Effects of Discontinuing Machine Learning Decision Support

SAFE Working Paper No. 370 | December 2022

Leibniz Institute for Financial Research SAFE
Sustainable Architecture for Finance in Europe
info@safe-frankfurt.de | www.safe-frankfurt.de

The Effects of Discontinuing Machine Learning Decision Support

Kevin Bauer[1], Michael Nofer[2], Benjamin Abdel-Karim[2], Oliver Hinz[2]

Abstract: Advances in Machine Learning (ML) led organizations to increasingly implement predictive decision aids intended to improve employees' decision-making performance. While such systems improve organizational efficiency in many contexts, they might be a double-edged sword when there is the danger of a system discontinuance. Following cognitive theories, the provision of ML-based predictions can adversely affect the development of decision-making skills that come to light when people lose access to the system. The purpose of this study is to put this assertion to the test. Using a novel experiment specifically tailored to deal with organizational obstacles and endogeneity concerns, we show that the initial provision of ML decision aids can latently prevent the development of decision-making skills which later becomes apparent when the system gets discontinued. We also find that the degree to which individuals "blindly" trust observed predictions determines the ultimate performance drop in the post-discontinuance phase. Our results suggest that making it clear to people that ML decision aids are imperfect can have its benefits especially if there is a reasonable danger of (temporary) system discontinuances.

1. Introduction

Organizations increasingly rely on Machine Learning (ML) systems such as Random Forests, Support Vector Machines, or Neural Networks to augment their employees' performance on specific tasks (Jordan and Mitchell 2015; Berente et al. 2019). These systems typically process available information and produce predictions about unknown (future) states of the world. Predictions then feed as an input factor into employees' decision-making processes helping them make better and more nuanced decisions under uncertainty by improving their judgment (Agrawal et al. 2019). Contemporary examples include a wide range of decision aids: Procurement Systems leveraging historical data to forecast required future resources, Sales Systems using information about customers to predict cross-selling opportunities (Loureiro et al. 2018), and Marketer Systems harnessing economic indicators to predict the efficacy of specific marketing strategies (Singh et al. 2017).

Yet, following previous insights from the automation literature on expert system decision aids, the provision of such predictive decision aids might be a double-edged sword. While it can inform business decisions and, thereby, improve organizational efficiency (Shang and Seddon 2002; Stallkamp et al. 2012; Janssen et al. 2013), the provision of decision aids generally comes at the expense of negatively affecting employees' skill development (see, e.g., Orlikowski 1991; Dzindolet et al. 2002; Skinner 2004; Goddard et al. 2012, Alavi and Leidner 2001; McCall et al. 2008). When employees overly rely on computerized decision aids instead of engaging in vigilant information seeking and processing

[1] Leibniz Institute for Financial Research SAFE.
[2] Goethe University Frankfurt, Frankfurt am Main, Germany
KB and OH gratefully acknowledge research support from the Leibniz Institute for Financial Research SAFE.

themselves, they may not learn how to make a decision in the absence of computerized support (see, e.g., Mosier et al. 1998; Skitka et al. 1999; Carr 2014).

However, insights from the literature stem from settings where employees interact with expert systems (comprehensive, declarative instructions explicitly devised by human experts), not contemporary predictive decision aids (unintelligible predictions based on self-learning ML models) that differ in dimensions important to the development of knowledge. Given the growing prevalence of ML-based systems, it is imperative to explore whether insights into the relationship between the provision of computerized decision support and the development of skills are still valid (Teodorescu et al. 2021). So far, we lack a thorough understanding of how predictive systems shape employees' knowledge development, and, relatedly, their consequences for employee performance under system discontinuance. The potential problem of system discontinuance is particularly relevant to ML applications because these systems are more likely to exhibit discontinuances than expert systems. That is because they stop working appropriately and thus become unusable when disruptions in the environment evoke fundamental changes to the data generating processes, i.e., concept drifts (Widmer and Kubat 1996; Gama et al. 2014). In these situations, knowledgeable human experts need to navigate decision-making on their own until they have produced sufficient new data to retrain the disrupted system. Therefore, somewhat ironically, the development (and maintenance) of human decision-making skills is a complement to the successful implementation of ML-based decision support.

The purpose of the study at hand is to experimentally test whether the provision of decision-supporting ML predictions causally affects the development of decision-making skills. More specifically, we intend to answer two research questions:

(i) What is the impact of providing unintelligible ML decision support on the development of decision-making skills and performances when systems get discontinued?

(ii) Does the degree of "blind" trust shape the occurrence of such effects?

Exploring the interplay between contemporary ML decision support and individuals' development of decision-making skills is a crucial endeavor since organizations increasingly employ ML-based systems to aid their employees. Since individuals' skills constitute one of the most valuable assets of organizations, it is pivotal to understand how the ML decision support system may have on the development of skills, as well as its ramifications under a system discontinuance.

There are several requirements for studying the causal effects of ML decision support systems on skill development. First, we need to ensure that individuals do not have prior experience with the decision task so that they develop any decision-making skills only during the study. Second, individuals may not have distinct access to additional information relevant to developing skills. Third, we need to eliminate the possibility that individuals possess unobserved (strategic) motives affecting their willingness to develop skills. Finally, we need to exogenously (and unanticipatedly) discontinue the ML decision support system for a random subset of employees. Meeting these requirements in a field setting is particularly difficult, if not outright impracticable. Against this background, we address our research

questions by designing and implementing an incentivized online experiment specifically tailored to cope with the outlined obstacles.

In our treatment condition of the experiment, participants solve a series of logical puzzles. For the first half, participants solve the puzzle with the help of an ML decision support system. For the second half, we discontinue the system without a warning. We track treatment participants' decision-making performance over time and compare it to one of the baseline participants who work on identical puzzles, however, always without the aid of an ML decision support system. Our analyses reveal that the provision of the ML decision aid impedes treatment participants' latent development of skills which becomes visible once the system discontinuance occurs. We find evidence that the degree to which treatment participants "blindly" trust the system shapes their decision-making performance in the post-discontinuance phase.

Our paper relates to two streams of literature. First, our paper complements previous work that examines the impact of expert system decision support on deskilling (e.g., Fitts 1951; Johnson et al. 2010; Ranz et al. 2017; Mateus et al. 2019). Evidence on the impact of expert system decision support on employee skill is mixed. Several studies indicate that the long-term use of such decision support systems (e.g., Knowledge Management Systems) decreases users' business process knowledge. (Dowling et al. 2008; McHall 2008; Axelsen 2012; Triki and Weisner 2014; Rinta-Kahila et al. 2018). Other studies suggest that there can be positive skill effects as well (see, e.g., Millman and Hartwick 1987; Orlikowski and Barley 2001; Schuppan 2014). Our paper complements this work by producing novel evidence on how ML decision support affects the development of decision-making skills. Specifically, despite considerable differences between expert and ML decision support systems (see, e.g., Berente et al. 2021; Teodorescu et al. 2021), the insights previous IS research generated appear to remain valid.

Second, we contribute to relatively nascent literature that studies how the discontinuance of decision support systems affects users. When computerized decision support systems create benefits, organizations typically keep them in place more or less indefinitely. Therefore, in the absence of system failures, adverse effects on people's skill development may not come to light and only occur latently. However, for different reasons, systems may, at least temporarily, get discontinued, e.g., due to updates or in the case of ML-based systems retraining, so that decision support is absent (see, e.g., Power & Gruner 2015; Rinta-Kahila et al. 2018). In general, only a few studies have explored the discontinuance of decision support systems in organizations (see, e.g., Tully 2015; Rinta-Kahila et al. 2018; Soliman & Rinta-Kahila 2020; Rinta-Kahila et al. 2021). Existing studies typically focus on the antecedents of discontinuance, while ignoring potential downstream ramifications. One notable example is the paper by Rinta-Kahila et al. (2018) who explore the impact of discontinuing an expert system on employee performance. We add to this literature by exploring the ramifications of discontinuing an ML decision support system for the post-continuance decision-making performance of users.

The paper proceeds as follows. In section 2, we outline the theoretical background motivating our research hypotheses. Section 3 presents the experimental design, while we report our empirical results

in section 4. We conclude with a discussion of results and provide an outlook on future research avenues in section 5.

2. Theoretical background

Skill development in the ACT-R framework

From an organizational perspective, skills are a pivotal asset to firms that can help them gain a competitive advantage if utilized and passed on effectively among employees (Barney 1991, Grant 1996, Alavi and Leidner 1999). Therefore, it is imperative to shed light on how the provision of contemporary ML decision support affects employees' development of decision-making skills. When there exists the chance of a (temporary) discontinuance of the predictive system, e.g., because managers decide to quit licensing a system, or because the system is inoperative due to malfunctioning (Soliman and Rinta-Kahila 2020), latent adverse effects on employees' skills will come to light and may lead to considerable disruptions in business activities. Importantly, with the increasing implementation of ML-based decision support, the chance of system discontinuances grows. That is due to the nature of modern ML applications. Typically, models learn from historical data that comes in the form of input-output pairs. However, in dynamically changing, nonstationary environments, the data generating process, and thereby the data distribution, can change over time so that the pattern learned from historical data is no longer accurate – a phenomenon referred to as concept drift (Widmer and Kubat 1996). When a concept drift occurs, predictive models require an updating or retraining on novel data sets that encode the changed data distribution (Gama et al. 2014). To adapt to disrupted ML systems, however, organizations require knowledgeable employees to navigate decision processes themselves and thus produce adequate new data. Somewhat ironically, the development and maintenance of human knowledge is, therefore, a vital complement to the implementation of ML-based decision support.

To examine the interplay between predictive decision support and skill development, one must first understand conceptually how people develop it. The starting point of skill is knowledge. Going back to Plato, knowledge is often broadly defined as a "justified true belief" (Nonanka 1994, Boghossian 2007) that can take on one of two forms: procedural knowledge or "know-how" that is directly applicable to specific tasks, i.e., skills, and declarative knowledge or "know-what" about the nature of things, i.e., facts and rules (Norström 2015). Notably, procedural knowledge is tacit, meaning that it is intuitive in nature and difficult to articulate (e.g., knowing how to ride a bike). By contrast, declarative knowledge is explicit. People possess a conscious awareness of this type of knowledge, meaning it is easy to articulate declarative knowledge (e.g., knowing what the technology adoption theory is). Both types of knowledge are naturally interdependent (Fantl 2008) with procedural knowledge being seen as the ability to apply declarative knowledge as condition-action pairs. Put differently, declarative knowledge is the antecedent for developing procedural knowledge, i.e., decision-making skills (Anderson and Fincham 1994).

But how exactly do people develop skills to solve problems? A widely used conceptualization of the skill development process originates from the ACT-R theory – a framework outlining the workings of human cognition (Anderson and Lebiere 1998, Anderson 2013). The ACT-R framework proposes two subsequent stages of skill development: a declarative and a procedural stage. The development of skill begins in the declarative stage by observing declarative information such as rules, definitions, instructions, or examples from the environment. Individuals encode this information as experiences into memory – either working or long-term memory (Anderson et al. 1997). Initially, individuals address the problem at hand by analogizing from the stored experience, i.e., they engage in interpretive problem-solving. Over time, individuals observe additional information either through obtaining endogenous feedback on the successes and failures of their problem-solving attempts or through exogenous events in the environment that reveal further problem-relevant facts. The newly encountered information is considered in the working memory where individuals can increasingly refine their high-level understanding of the problem resulting in the development of declarative knowledge that is eventually committed as "chunks" in the (declarative) long-term memory (Anderson 1997, Anderson et al. 2004). When stored in long-term memory, the retrieval of chunks to engage in analogizing becomes simpler as the number of retrievals and related chunks increases (Grimaldi and Karpicke 2012).

Once individuals have developed and stored declarative knowledge in their long-term memory, the procedural knowledge development process begins. In this second stage, individuals gradually convert declarative knowledge into procedural knowledge through a process referred to as production compilation. The compilation results in explicit production rules that one can think of as condition-action pairs (Anderson 1997). These condition-action pairs are at the heart of decision-making skills. Production rules are stored in the (procedural) long-term memory and allow solving problems without the cognitively strenuous retrieval of declarative knowledge for guidance. With additional practice, individuals can further refine or expand compiled production rules, leading to improved skills. Utilizing (optimized) condition-action pairs instead of more complex analogizing both, speeds up performance and frees cognitive resources by reducing the required level of attentional capacity (Anderson 2014).

In our experimental study, we incentivize participants to develop a problem-solving strategy for logical puzzles. We provide feedback on a constant basis so that they can engage in a trial-and-error process to come up with successful condition-action pairs, i.e., develop decision-making skills. Correctly recognizing the logical pattern involved allows participants to maximize their earnings in our study (see section 3 for details on the experimental design).

Computerized decision support and human knowledge

Studies from automation research across disciplines frequently indicate that the broad implementation of expert system decision support, i.e., systems that perform recommendations for tasks based on an understanding of how human experts behave, can contribute to deskilling (see, e.g., Mosier et al. 1998; Skitka et al. 1999; Mascha and Smedley 2007; Dowling et al. 2008; McHall et al. 2008; Parasuraman

and Manzey 2010; Hoff 2011: Axelsen 2012; Carr 2014; Rinta-Kahila et al. 2018). Mosier et al. (1998) study omission and commission errors by pilots resulting from using expert systems (automated communication and electronic checklists) in flight simulators. They find that automation biases constitute a significant factor when pilots engage with automated aids and that pilots do not consider all available information when making decisions in conjunction with computerized decision aids. Dowling et al. (2008) study the long-term effects associated with the use of audit support systems finding that it decreases employees' existing declarative knowledge over time. Using a qualitative research method, Axelsen (2012) finds that the long-term use of audit support systems decreases auditors' procedural knowledge. In a case study with an IT service company, Rinta-Kahila et al. (2018) show that using software that automates fixed assets accounting and reporting leads to a latent deskilling of accountants that comes to the surface when users lost access to the system. While these studies depict how computerized decision support can entail the loss of human knowledge, none of them considers the development of new skills, which is what we look at.

Smedley and Sutton (2007) study the impact of expert systems on procedural knowledge development. They find that the development of knowledge and response to the presented information differs considerably based on users' initial expertise, with 'intentional learners'' developing relatively high levels of knowledge. McCall et al. (2008) examine the impact of knowledge-management systems on performance and skill development. They find that the availability of knowledge-management systems as decision aids enhances user performance and does not impede the development of declarative knowledge. Instead, using this expert decision aid incites the development of different knowledge. Hoff (2011) finds decreased clinical knowledge and a lower willingness to learn about medical trends such as innovative methods for diagnosis following the introduction of electronic medical records. Arnold et al. (2018) find that the automated provision of explanations in expert decision support systems is an important design feature influencing the skill development processes. Insights on the development of knowledge from these studies stem from cases where the decision aid comes from expert systems.

Overall, many previous studies depict a potential dark side of the employment of computerized decision aid, where machines exercise real authority (Agion and Tirole 1997) over human users who merely do what the machine tells them to do without thinking for themselves (Hirschheim et al. 1991). However, there are also studies, especially in the context of automated business decisions, that provide contradicting evidence. For instance, Sayed (2006) finds that the implementation of an ERP system in Egypt did not cause a deskilling of employees. Instead, employees recognize that their knowledge is central to the proper functioning of the technology. Other studies show that deskilling caused by expert system decision support is not as pervasive as broadly suggested. Instead, the occurrence of deskilling effects from implementing such systems depends on the system's design (e.g., Orlikowski and Barley 2001). Relatedly, Orellana (2015) suggests that the provision of computerized decision support does not result in de- but reskilling. In sum, previous studies examining the relationship between computerized decision support and user knowledge focus on expert systems that provide users with comprehensive

information on how to do their tasks (e.g., Cummings 2004; McCall 2008, Rinta-Kahila et al. 2018), reports (e.g., Dowling et al. 2008, Verghese 2008, Hoff 2011), and even take over decision-making completely (see, e.g., Skitka et al. 1999).

Complementing previous studies, we consider contemporary ML decision aids that aggregate and transform available information to produce probabilistic predictions. These systems are increasingly in use in organizations (see, e.g., McAfee et al., 2012; Hoffman et al. 2018; De Spiegeleer et al., 2018, Leo et al., 2019) and, for instance, include decision aids that forecast procurement prices, predict chances of sales conversions, predict the probability of customer churn, and assess the fit of an applicant for a vacancy. Importantly, this form of decision support is fundamentally different from previous generations of expert system decision support (see, e.g., Teodorescu et al., 2021; Berente et al., 2021). In general, "[…] ML may require a massive rethinking of significant portions of the corpus of IS research in light of these differences [to previous generations of AI-based systems]" (Teodorescu et al., 2021, p.1483). Rethinking and retesting accepted IS theories in the light of ML requires that we once again start at the ground level and address things at the most basic level of analysis. Doing so will be an essential step toward the efficient and beneficial usage of this contemporary generation of AI technologies (Berente et al., 2021). ML-based decision support systems are typically machines that have learned, more or less all by themselves, high-dimensional, non-linear relations between different variables from large, labeled data sets to produce a label estimate for unseen data without a label (LeCun et al. 2015; Jordan and Mitchell 2015). Put differently, ML decision support systems represent true machine knowledge. This characteristic is in stark contrast to expert systems that comprise expert knowledge developed by humans. Given the differences between traditional expert systems and modern ML support systems, it remains an open question to what extent previous insights in deskilling are still valid. The field of Information Systems must have the ambition to uncover the changing influence of constantly evolving technologies.

3. Study design

We designed a novel laboratory experiment to identify the relation between ML decision support and skill development. Our study design leverages the strengths of the controlled laboratory environment in ways that would be difficult to replicate in a field setting, if not outright impracticable. The study design ensures that participants (i) have no previous experience of the logic governing the task, (ii) do not have different access to information that is relevant to solving the task, and (iii) have no unobserved (strategic) motives affecting task performance, and (iv) participants do not anticipate the discontinuation of the ML decision aid. We control the structure of the task, the material consequences of decisions, and the information flow, allowing us to detect the development of decision-making skills and how predictive decision support affects it.

Design Overview

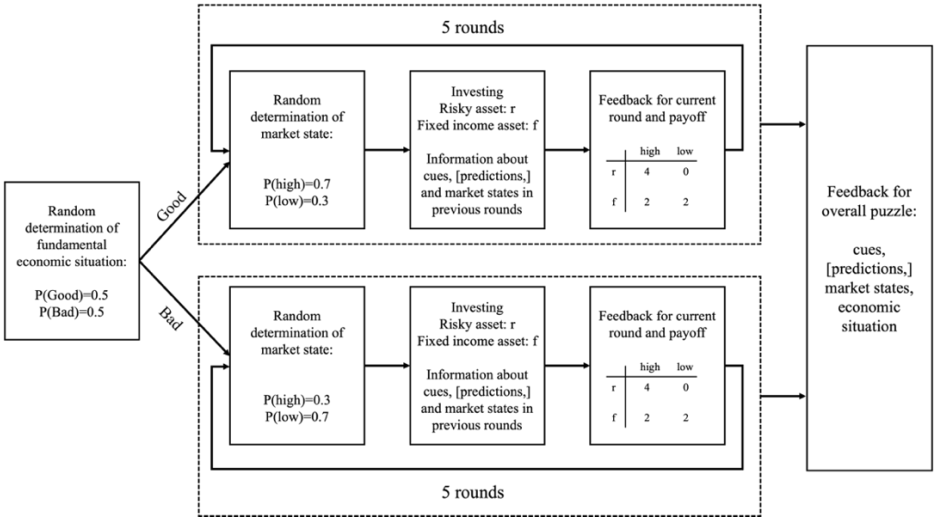

Figure 1: Illustration of one market simulation. Participants in our study engaged in 12 subsequent market simulations all following the depicted structure. We denote the treatment variation, i.e., the presence of predictions about the current market state in the first six market simulations, via squared brackets.

In our experimental study, participants repeatedly work on a logic puzzle (see figure 1 for an illustration of an individual puzzle). To make the setting more accessible to participants, we frame the puzzles as market simulations, each comprising five consecutive rounds of making investment decisions with intermediary feedback. However, the structure of the task more fundamentally mirrors aspects that many decision-making tasks under uncertainty share: (i) there exists an unknown state of the world that affects the personal consequences of decision making, and (ii) people observe (conditionally independent) environmental cues that relate to the hidden state and are, therefore, relevant to decision making. In each of the five rounds of a market simulation, participants receive four conditionally independent cues that help them identify which of two possible decisions maximizes their payoff, i.e., the best investment option. The design is related to Kuhnen (2015) and Kuhnen and Miu (2017).

Most importantly for our research objective, there exists a fixed relation between the cues and the market state that we do not reveal to participants. However, we always give participants feedback about their investment performance after deciding, so that they can learn to interpret the cues correctly over time. We even provide an overview of the cues and actual market state of all previous rounds for a given market simulation. Participants learn about the existence of the pattern and its stability throughout the experiment. We explicitly tell participants that they can benefit from figuring out the pattern (see Appendix for the exact instructions). In addition to the four cues, participants observe the output of a machine learning model that uses the four cues to make a highly accurate prediction about the market state in the current round. Again, we are completely transparent and provide participants with detailed information about the machine learning model, how we trained it, and how accurate the model is on a representative test set.

The experiment comprises two stages. In stage 1, participants work on six market simulations one after another, always observing both cues and a machine learning prediction. In stage 2, participants again work on six consecutive market simulations that follow the exact same structure. The only difference is that participants no longer observe machine learning predictions but only the four cues in stage 2. We did not inform participants about them losing access to the predictive aid before so it comes exogenously. This abrupt discontinuance enables us to observe any latent effects on the skill development that are unobservable while treatment participants can rely on the prediction. Everything else, including the relationship between cues and optimal decisions, is identical.

To identify how the availability of the ML decision support in stage 1 affects participants' capability to learn the hidden relation between cues and the optimal decision, i.e., develop decision-making skills, we implement a baseline condition of the experiment. The only difference between the outlined treatment condition and the baseline is that participants in the latter do not receive predictive decision support during the first six market simulations. By comparing the performance of participants who initially observe machine learning predictions (Treatment) to those who do not (Baseline), we are able to isolate the causal treatment effects. In the following, we fill in the important details of the market simulations that participants encountered.

Details on market simulations

In both stages of our experiment, participants encounter 6 market simulations each. After every market simulation, participants receive a detailed overview of all relevant information (cues, market states, predictions, and the fundamental economic situation) and their decisions, so that they can detect potential faults in their interpretation of cues (and predictions) and revise their strategy accordingly in the next simulation. Doing so is possible because the underlying randomization process and mechanisms are identical across all twelve simulations. The process works as follows. A market simulation comprises five consecutive rounds of decision making. At the beginning of a simulation, we flip a fair coin to determine the fundamental economic situation $E \in \{G, B\}$ that can either be good (G) or bad (B). The randomization occurs on the individual level so that we can mitigate concerns about ordering effects. The economic situation E is fixed for the five consecutive rounds of decision making comprised in a market simulation. For every round, we randomly draw one of two mutually exclusive market states $m \in \{h, l\}$ from the distribution P_E where $p_G(h) = 0.7$, $p_B(h) = 0.3$, and h,l indicate respectively that the market is in a high or low state in a given round.[3] Participants only become aware of the fundamental economic state at the end of a simulation, when they observe all information from the previous five rounds of decision making. Participants had an unlimited amount of time to examine the overview. Obtaining this information allows them to develop a more thorough understanding about how to interpret cues, especially after they have seen some market states in a given simulation.

Details on rounds

[3] As the two possible market states are mutually exclusive, it holds that $p_G(l) = 0.3$ and $p_B(l) = 0.7$.

In every round of a given market simulation, participants' objective is to figure out whether the unobserved market state m is high or low, allowing them to identify the payoff maximizing investment decision $d \in \{r, f\}$. Participants have unlimited time to make a decision. If the market in a given round is in the high state h, the optimal decision $d^*(.)$ is to invest in the risky asset r, i.e., $d^*(h) = r$ yielding the payoff $\Pi(\mathrm{r}|\mathrm{h}) = 4$ monetary units (MU). By contrast, when the market is in the low state l, the optimal decision is to invest in the fixed income asset f, i.e., $d^*(l) = f$, yielding the payoff $\Pi(\mathrm{f}|\mathrm{l}) = 2$ MU. The payoffs for the opposite decision equal $\Pi(\mathrm{f}|\mathrm{h}) = 2$ MU and $\Pi(\mathrm{r}|\mathrm{l}) = 0$ MU, respectively. With this specification, both decisions (ex-ante) yield the same expected payoff, i.e., are mutually attractive to risk neutral individuals.[4] After making their decision, we inform participants about the actual market state in the given round. Apart from the extensive feedback participants receive after each decision and at the end of each simulation, they also observe all results from previous rounds of the current simulation in a table format while deciding in a given round (see screenshot in figure 2). Specifically, they always observe the expert cues, (the machine learning prediction,) and the actual market state of all prior rounds of the current simulation. By providing this information, we intend to facilitate learning without adding additional sources of noise such as participants capabilities to recall information from previous rounds. Notably, to ensure that participants are familiar with the task at hand before making their first payoff relevant decision, we let them engage in two trial rounds of decision making that we told them would not influence the remainder of the study. Participants engaged in the trial rounds right after reading through the instructions and before starting the first complete market simulation

Details on cues

While participants neither observe the current economic situation nor the current market state directly when making a decision, they observe four conditionally independent cues that can be accurate or inaccurate with different probabilities.[5] Each cue can be high or low $e_1, e_2, e_3, e_4 \in \{h, l\}$, indicating the state of the market. The cues are correlated with the market state, conditional on the fundamental economic situation so that they help participants anticipate the payoff maximizing decision. Note that the aggregate informational content of cues, and by extension the difficulty of correctly interpreting them, is by design the same in both economic situations. We only interchange the reliability of individual cues so that the logic is not too simple to detect as it is arguably the case if it is always the same cue that is accurate with the highest probability. The following table 1 outlines the probabilities with which certain cues are accurate. As illustrated, the cue 1 (Expert 1) is most reliable when the fundamental economic situation is good, whereas cue 2 (Expert 2) is most reliable if fundamental situation is bad.

Cue by	Fundamental Economic Situation	
	Good	Bad
Expert 1	80% accurate	50% accurate
Expert 2	50% accurate	80% accurate

[4] E(f) = 0.5*2 + 0.5*2 = 2 = 0.5*0.7*4 + 0.5*0.3*4 = E(r)

[5] The cues are independent given the fundamental economic situation and the market state.

Expert 3	38% accurate	65% accurate
Expert 4	65% accurate	38% accurate

Table 1: Summary of cue accuracies conditional on the fundamental economic situation. Note that we presented the cues to participants as opinions of human experts.

We present these cues to participants as opinions from four fixed human experts, who give different good advice. To differentiate the cues sufficiently from other information on the screen, especially the ML prediction in the treatment, we assigned each expert an image of a fictitious person that we generated using a generative adversarial network by Karras et al. (2019). While we fixed the image-expert assignment across simulations for an individual participant, the order could randomly differ across participants.

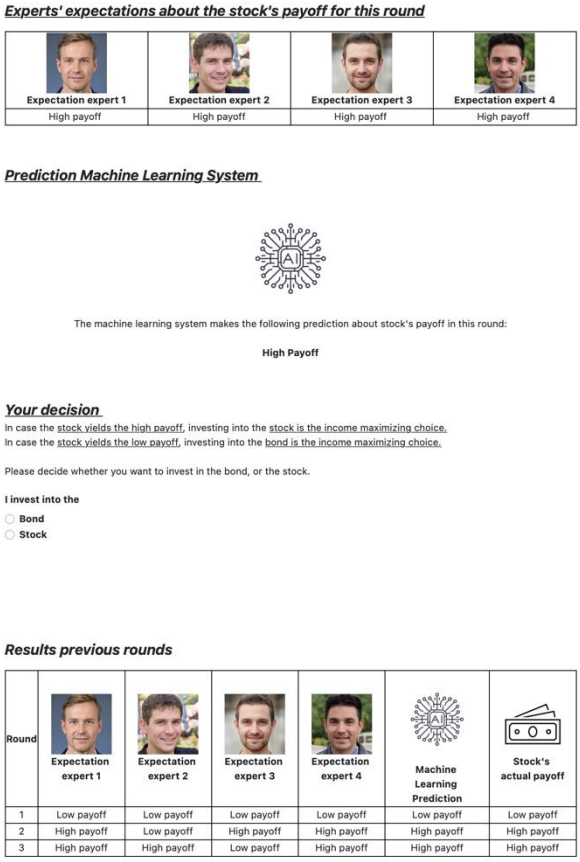

Figure 2a: Illustration of the decision-making interface in the treatment condition.

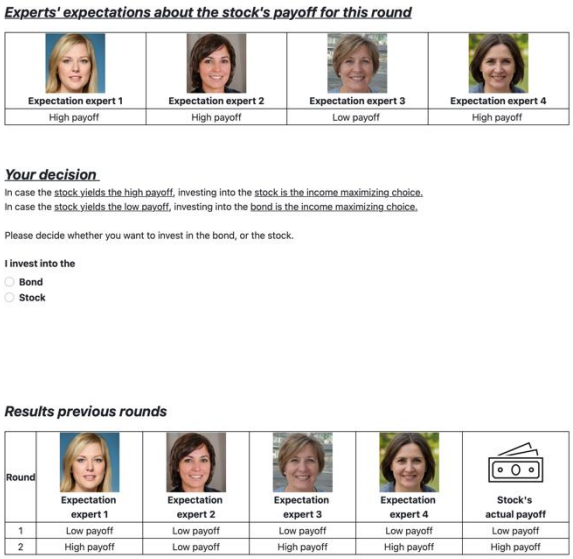

Figure 2b: Illustration of the decision-making interface in the baseline condition.

We also randomized whether participants observed four male or female images. With equal probability, participants either always observed the same images of four women or four men (we show screenshots in figures 2a and 2b). Notably, while the order might differ across participants as outlined, it is always the same four women and men that each participant could randomly see. For additional information on the generation of expert cues in our software refer to the appendix.

With the overall structure of the logic involved and design, we are confident that the relationship between cues and the market states that participants are incentivized to learn is sufficiently cognitively

challenging but, at the same time, follows simple enough rules so that participants can develop the knowledge over time if they exert effort.

Details on predictive decision support system

The predictive decision aid treatment participants have access to is a Random Forest classifier using the popular Python library Scikit-learn (Pedregosa et al. 2011). The final model comprises 200 individual learners, each with a depth of 3. We trained this model using a simple 2 step procedure. First, we randomly generated 1.000.000 data points each comprising the fundamental economic situation, the market state, and the four cues. The correlations in the generated data set are identical to the ones implemented in the experiment. Second, we optimized (5-fold cross-validation) and trained the Random Forest on a random subset of 750.000 data points and tested its performance on the remaining 250.000 observations. The final model's accuracy equals 85.4%. We further tested the efficacy of the model by letting it compete with pilot participants (other researchers) unfamiliar with the underlying logic in controlled mock sessions of the experiment. The model consistently outperformed the pilot participants (9 out of 9 cases; 6 simulations each).

To ensure that participants had the same perception and beliefs about the nature of the ML model, they received a detailed description of the model including the data we trained it on, its accuracy rate, the fact that it outperformed other humans, the type of model implemented, and examples of domains where Random Forests are frequently used.

Experimental procedure

We conducted an online experiment and recruited participants from Great Britain using the commercial platform Prolific. We ran the study at the end of October 2021. The study procedure consisted of three parts.

First, we randomly assigned participants to the treatment or baseline condition with equal probability. The study began with a pre-experimental survey containing questions about participants' socio-demographics and several items on their familiarity with and attitude towards predictive software. The answers to these questions serve as a randomization check and as additional co-variates in some analyses (see table 4 in the appendix for an overview).

Second, participants received instructions on the upcoming task. Participants had unlimited time to read the instructions carefully. After reading the instructions and before the first market simulation started, participants engaged in two trial rounds of investment decision-making so that they would already be familiar with the interface and structure of decision-making. Subsequently, participants started with stage 1 as outlined above. Notably, we explicitly informed subjects that there would be another part of the experiment – even though we did not tell them the exact nature of it – and that learning to understand how cues relate to market states will be materially beneficial. To make sure that participants'

concentration did not drop, we introduced an attention check after three market simulations, which we control for in our analyses.

Third, after they finished with the first six rounds, we informed participants that they had to work on another six puzzles. The instructions explicitly state that the next puzzles have the identical structure as before and that participants can use the learned pattern to identify the payoff maximizing decision in a given round. Again, we implemented a simple attention check after three puzzles. After stage 2, the experiment ended with informing participants about their overall earnings.

Overall, 291 different individuals participated in our experiment. The randomized treatment assignment resulted in 149 treatment and 142 baseline participants. We observe the behavior of each individual across 12 simulations each comprising 5 rounds, i.e., 17.460 individual decisions. On average, it took participants 31 minutes to complete the experiment. In addition to a fixed participation fee of 2 Euro, we paid participants a bonus according to their investment performance across stages 1 and 2. Specifically, we added up the entire amount of MU participants earned through their decisions. For every MU, we paid participants 2 Euro cents. The average bonus amounted to 2.97 Euro (approx. 148 MU) so the overall average income equaled 4.97 Euro (9.62 Euro/hr).

4. Results

In this section, we present our results in two parts. First, we begin our analyses on an aggregate level examining how the availability of predictive decision support affects participants' decision-making performance in the two parts of the experiment. Second, we study how participants' decision-making performance evolves across market simulations. Third, we test for the presence of treatment heterogeneities to gain insights into the origins and circumstances under which ML decision support affects participants' development of decision-making skills.

Decision making performance across parts

Figure 1 depicts the average frequency with which baseline and treatment participants made a payoff-maximizing decision in the two stages of our experiment. A decision maximized a participant's payoff if she invested her endowment in the stock whenever the market was in a high state and the bond otherwise. Participants' performance arguably depended on recognizing the hidden relationship between the unobserved market state and observed cues from experts. Therefore, we argue that the trajectory of this measure across market simulations reflects participants' decision-making skills. This argumentation is conceptually in line with the learning goal orientation leading an individual to increase the level of performance in a given activity (Button et al. 1996).

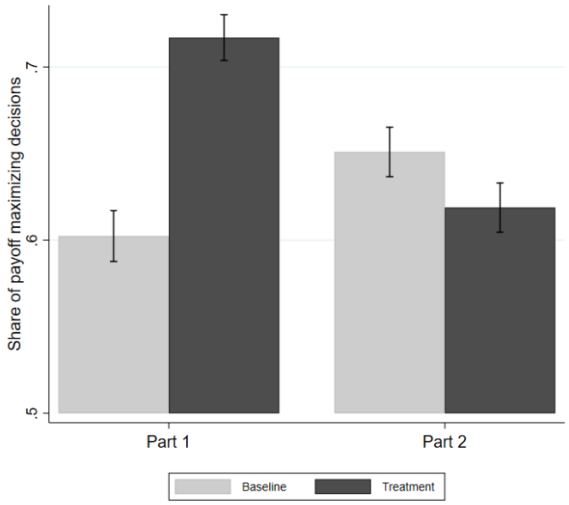

Figure 3: *Participants' investment decision performance in the two parts of the experiment rounds. Bars represent average shares of payoff maximizing decisions. Error bars depict 95% confidence intervals of averages. We show results separately for baseline and treatment participants.*

On average, in stage 1 of our baseline condition, participants chose the payoff-maximizing investment decision 60.2% of the time. Thus, on average, participants chose the option with the highest payoff in 3 out of 5 decisions they had to make per market simulation. This ratio is significantly better than random guessing, suggesting that baseline participants took advantage of the cues provided by the experts ($p<0.01$, Wilcoxon signed-rank test). In addition to the expert cues, treatment participants observed ML predictions about the market state in stage 1. On average, they made a payoff-maximizing decision 71.7% of the time. Therefore, the availability of the ML decision led to an average increase in investment performance by 11.5 percentage points in stage 1. This difference is economically (+19.1%) and statistically highly significant ($p<0.01$, Chi2-test), emphasizing that having access to a predictive decision aid did make treatment participants considerably better off. Additional regression analyses reveal that treatment participants placed significantly more weight on the observed ML prediction than on the expert cues, providing evidence that they did not see the ML output as another expert.

Note that treatment participants' performance is significantly lower than the prediction accuracy, which equals 84.8% ($p<0.01$, Wilcoxon signed-rank test), implying two things. On the one hand, at least some treatment participants did not always follow observed predictions but overruled them from time to time. On the other hand, they could have performed even better had they relied on the ML decision aid more frequently. Hence, treatment participants did not use the available predictions to their full potential.

Turning to participants' performance in the second stage, we find that baseline participants' average likelihood of making a payoff maximizing decision increased significantly by 4.9 percentage points to 65.1% from stage 1 to stage 2 (+8.1%; $p<0.01$, Wilcoxon signed-rank test). Because the randomization of market states and the relation between expert cues and market states are identical in the two stages, the enhanced performance arguably reflects an improvement in their decision-making skill. When looking at the performance of treatment participants, we find that they, on average, made a payoff-maximizing decision in 61.9% of the cases. Relative to their baseline counterparts, treatment participants

were, on average, 3.2 percentage points less likely to choose the most profitable investment option. This difference amounts to 5.2% and is statistically significant ($p<0.01$, Chi2-test). Put differently, while observing ML predictions in stage 1 made treatment participants significantly better off, they performed significantly worse than baseline participants once we discontinued the additional decision support in stage 2. This observation suggests that the availability of the ML decision aid in stage 1 latently impeded treatment participants' skill development. This adverse effect came to light once we discontinued the ML decision support system. Result 1 summarizes the findings outlined thus far:

Result 1: *The availability of ML decision support in stage 1 causes treatment participants to exhibit a significantly higher investment performance than baseline participants in stage 1 (while using the decision aid), however, a significantly lower one in stage 2 (post-discontinuance).*

Before moving on to a more nuanced analysis at the individual level, one final comment seems appropriate. Figure 3 shows that treatment participants' investment performance averaged across all simulations is significantly higher than that of baseline participants (respectively 66.8% vs. 62.7%; $p<0.01$, Chi2-test). Even though this observation depicts the efficacy of the ML decision support, outlined results show that the initial decision support harmed participants' performance when the discontinuance occurred. This documented mechanism can be particularly harmful in situations in which a suboptimal decision is hard to reverse or has considerably negative consequences, e.g., undetected money laundering by supervisory agents, misdiagnosed serious diseases by physicians, or inadequate financial risk assessments by banks.

Skill development across simulations

We next examine the development of participants' performance across consecutive market simulations. These more nuanced analyses provide insights into the development of participants' decision-making skills.

Figure 2 shows the average frequency with which participants made payoff-maximizing decisions in the twelve consecutive market simulations. We present results separately for the baseline and treatment conditions. The dashed vertical line separates stages 1 and 2, which comprise simulations one through six and seven through twelve, respectively. In other words, the dashed vertical line illustrates the point in time when we discontinue the decision aid for treatment participants.

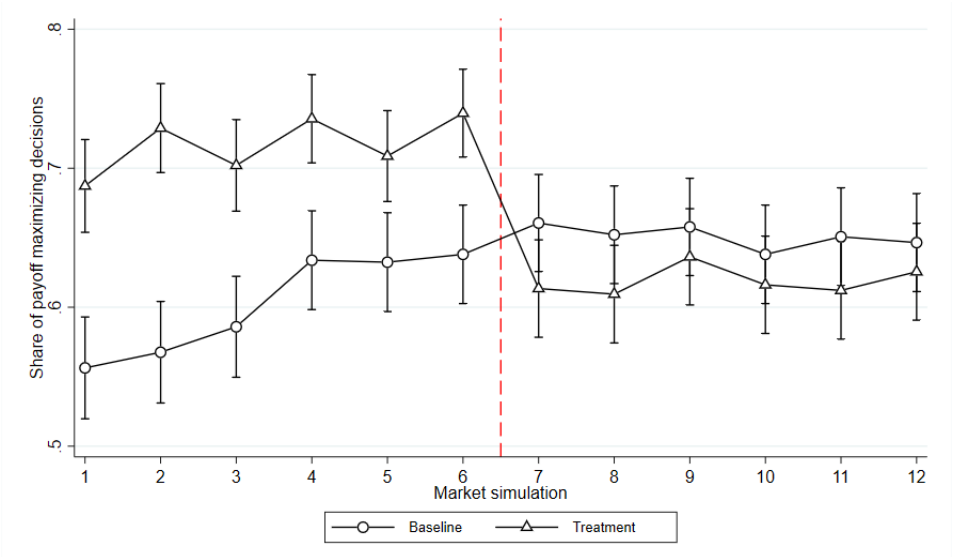

Figure 4: *Participants' investment decision performance across consecutive market simulations. Marked lines show the average share of payoff maximizing decisions per simulation. The line marked with a circle shows results for baseline participants, while the line marked with a triangle depicts results for treatment participants. The dashed red line separates stage 1 and stage 2 (respectively left and right). Error bars depict 95% confidence intervals of averages.*

Figure 4 indicates that the frequency with which baseline participants' chose the payoff maximizing investment increased gradually from 55.6% to 63.8%. (+14.7%) across the six consecutive market simulations in stage 1. This observation strongly supports the notion that baseline participants increasingly understood and exploited the underlying relation between expert cues and market states over time, i.e., their decision-making skills improved over time. Notably, the steepest increase in performance occurred across the first four market simulations. From simulation four onwards, the performance was rather steady at an average of about 64%.

Dep. Variable: Payoff maximizing decision	(1) Baseline	(2) Treatment
No. simulation	0.019***	0.007*
	(0.007)	(0.005)
Stage 2	0.122***	-0.080***
	(0.035)	(0.026)
No. simulation*Stage 2	-0.021***	-0.005
	(0.008)	(0.007)
p-value of F-test: No. simulation + No. simulation *Stage 2	0.6	0.71
Expert and state controls	YES	YES
Observations	8520	8940
p	0.000	0.000
R-squared	0.046	0.060

Table 2: *OLS regression with individual and round fixed effects. Column (1) shows estimates for the subsample of baseline participants, whereas column (2) shows results for treatment participants. We report robust standard errors in parentheses. The dependent variable is a dummy indicating whether a participant chose the payoff maximizing option in a given investment scenario. We also include, but do not explicitly report, control variables for the observed expert cues and the unobserved market state. We denote significance levels as *p<0.1, **p<0.05, ***p<0.01.*

Fixed effects OLS regressions corroborate the patterns observed in figure 4 (see table (2)). In both columns, we regress a dummy indicating a payoff maximizing choice on the simulation number, a stage dummy, and their interaction term. We further include individual and round fixed effects, controls for

expert cues, and unobserved market state controls. Estimates in column (1) show results for baseline participants, confirming statistically that their performance increased significantly by about two percentage points per market simulation in stage 1. However, the interaction term reveals that the time trend in stage 2 was significantly lower. An F-test confirms that the overall time trend in stage 2 is not significantly different from 0 (p=0.6), i.e., the development of skills stagnated.

When examining the performance development of treatment participants, we find that their performance increased from 68.7% in simulation one to 73.9% in simulation six (+7.6%). Regression analyses reveal that this time trend is weakly statistically significant (see column (2) of the table (2)) but significantly smaller than the trend we observe for baseline participants ($p<0.05$, F-test). In stage 2, we do not find treatment participants' decision performance to change significantly over market simulations. The frequency with which they made payout-maximizing decisions appears to have leveled off at around 61%. Notably, the average performance of treatment participants per market simulation was consistently lower than that of baseline participants. The difference in performance is especially evident in simulations seven and eight (66.1% versus 61.3% and 65.2% versus 60.9%, respectively). This gap, however, gradually declined across subsequent simulations (see table 3 for a statistical overview). Hence, post-discontinuance, treatment participants caught up to the performance of their baseline counterparts. This trajectory may indicate that treatment participants developed decision-making skills gradually once they could no longer rely on ML predictions.

	Number of market simulation											
Avg.	Stage 1						Stage 2					
performance	1	2	3	4	5	6	7	8	9	10	11	12
Treatment	0.56	0.57	0.59	0.63	0.63	0.64	0.66	0.65	0.66	0.64	0.65	0.65
Baseline	0.69	0.73	0.7	0.74	0.71	0.74	0.61	0.61	0.64	0.62	0.61	0.63
Est. difference	0.14***	0.18***	0.14***	0.11***	0.07***	0.09***	-0.05*	-0.06**	-0.03	-0.02	-0.04	-0.01
OLS regression	(0.026)	(0.026)	(0.026)	(0.025)	(0.026)	(0.025)	(0.027)	(0.027)	(0.026)	(0.026)	(0.026)	(0.026)

Table 3: *Summary statistics of participants' investment performance across simulations and OLS estimates of treatment differences. The dependent variable of OLS regressions is a dummy indicating whether a participant chose the payoff maximizing option in a given investment scenario for the given market simulation. We also include, but do not explicitly report, control variables for the observed expert cues, the unobserved market state, and personal characteristics. We report robust standard errors in parentheses and denote significance levels as *p<0.1, **p<0.05, ***p<0.01.*

At this point, it is worthwhile to elaborate on the weakly significant, yet positive time trend for treatment participants in stage 1. While one may be inclined to interpret this finding as evidence that their decision-making skills improve, further analyses of treatment participants' propensity to overrule observed ML predictions point in another direction (see table 4 in the appendix). First, we find that participants' inclination to overrule the ML prediction decreased significantly with the number of market simulations (see column (1) of table 5 in the appendix). Given that the average prediction accuracy equals 84.8%, following predictions more frequently naturally increased the performance. Second, treatment participants did not become more capable of detecting when a prediction was inaccurate. If anything, the opposite appears to be true. Regression analyses show that the probability of overriding incorrect ML predictions decreased over the course of the simulations (see column (2) of table 5 in the appendix). Hence, it does not seem to be the case that treatment participants' higher inclination to rely on ML predictions resulted from a more general understanding of how to use expert cues to determine when to

follow or overrule an ML prediction. Together, these insights suggest that treatment participants' performance increase in stage 1 did neither stem from an increased understanding of the relation between expert cues and the market state, nor from an increased understanding of when the decision aid was inaccurate. Instead, it appears more plausible that they increased their reliance on ML prediction across market simulations.

In sum, these results suggest that participants exhibit lower decision-making performance in Stage 2 because the initial availability of ML predictions impeded the development of decision-making skills. Notably, treatment participants caught up to the performance of baseline participants over the course of stage 2. In practice, however, the time required to catch up may not be available to decision-makers – or only at considerable costs – who unexpectedly lose access to predictive decision support in a consequential decision scenario outside the safe space of a controlled experiment.

Result 2: *The availability of ML predictions impeded the development of decision-making skills. Treatment participants increasingly adhered to observed ML predictions. The lowered decision-making skills came to light when the ML system got discontinued.*

Skill development and trust in the ML decision support system

In the final part of our analyses, we study whether there exist heterogeneities in the extent to which the initial availability of ML predictions impedes the development of decision-making skills. Conceptually, our starting point is the notion that the degree to which an algorithmic decision aid impedes learning depends on users' (over)reliance on the system.

Previous research documents that people try to exert the least amount of cognitive effort that is still acceptable in a given situation (see, e.g., Fiske and Taylor 1991; Garbarino and Edell, 1997). As outlined before, the predictions provide treatment participants with a shortcut for identifying a production rule indicating the payoff maximizing decision. Put differently, predictions constitute a decision heuristic that economizes cognitive effort and developing skills. Naturally, the more participants blindly rely on the heuristic instead of processing expert cues, the less likely they gain a thorough understanding of their relation to the optimal decision. In the context of our study, we would, therefore, expect to find stronger treatment effects for participants who rely more on ML predictions (see figure 5 for an overview).

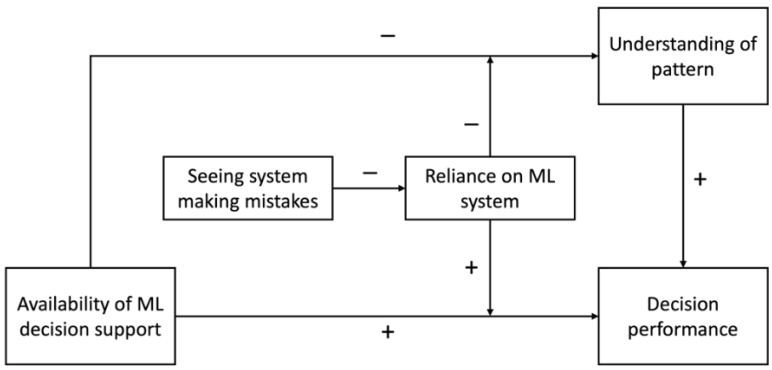

Figure 5: *Illustration of research model that motivates analyses of treatment heterogeneities.*

The question, however, is how to operationalize such a test given that reliance on the system arguably develops endogenously over time. In our analyses, we exploit a combination of two facts: (i) people's documented propensity to be especially averse to ML predictions after seeing them being incorrect (Dietvorst et al. 2015; 2018), and (ii) the fact that participants observe errors randomly due to our individual-level randomization of market states and expert cues. Specifically, we very conservatively distinguish treatment participants based on whether the very first prediction they observe turns out to be incorrect and examine the impact of this random event on their performance in stages 1 and 2.

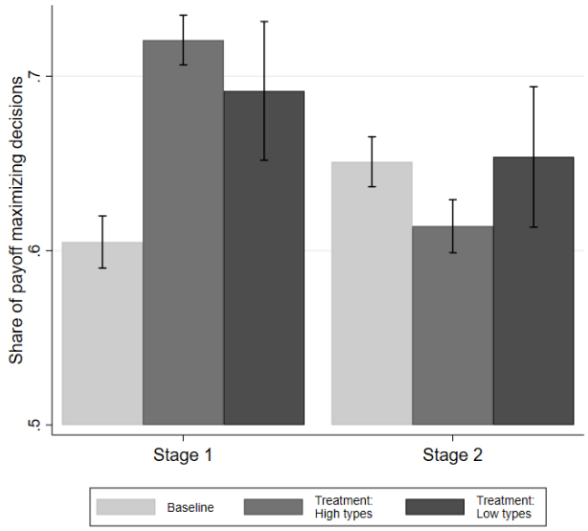

Figure 6: *Participants' investment decision performance in the two parts of the experiment rounds. Bars represent average share of payoff maximizing decisions. Error bars depict 95% confidence intervals of averages. We show results separately for baseline participants, high type treatment participants, and low type treatment participants.*

In our treatment condition, 12.1% of participants encountered an incorrect prediction during their very first investment decision. We refer to these participants as low types and the remaining ones as high types. Importantly, a prediction error in the first decision is not correlated with future errors. After the first decision, low and high types observe prediction errors in 15.7% and 15.2% of the cases, respectively ($p=0.77$, Chi2-test). Despite prediction error rates being virtually identical for subsequent rounds in part 1, we find that low types (25.9%) were significantly more likely than high types (20.4%) to overrule a prediction ($p<0.01$, Chi2-test). Therefore, we argue, that low types indeed relied significantly less on the prediction. Did this reliance indeed affect participants' development of skills? Figure 6 provides insights into this question, depicting the average decision performance for baseline participants, and high and low type treatment participants.

In stage 1, high and low treatment participant types made a payoff maximizing investment decision in 72.1% and 69.2% of the cases, respectively. The difference is statistically insignificant ($p=0.17$, Chi2-test). Hence, even though low types relied significantly less on the highly accurate prediction, their performance, if anything, only decreased marginally in stage 1. We further find that high and low type treatment participants performed significantly better than baseline participants ($p<0.01$ for both, Chi2-test).

When looking at the second stage, we find evidence for the existence of treatment heterogeneities. While high type treatment participants, on average, made a payoff maximizing decision in 61.4% of the cases, their low type counterparts did so 65.4% of the time (+6.5%). A Chi2-test shows that this difference is statistically significant ($p<0.08$). Therefore, our results provide some evidence that the lower reliance on observed predictions in stage 1 appears to be associated with a higher decision-making performance in stage 2. We further find that the performance of low type treatment participants in stage 2 is not significantly different from the one of baseline participants ($p=0.9$, Chi2-test). By contrast, high type treatment participants performed significantly worse than baseline participants in stage 2 ($p<0.01$, Chi2-test).

In sum, the outlined evidence suggests that treatment participants who saw the decision aid make a mistake right at the beginning developed more decision-making skills than their high type counterparts. Our analyses on treatment heterogeneities support the notion that the impediment of skill development increased with participants' blind adherence to predictions. One implication of this finding is that one may mitigate the negative impact of (predictive) decision aids on skill development processes by enhancing (and maintaining) users' awareness of the system's potential to be wrong.

Result 3: *By reducing treatment participants' reliance on the predictive decision aid, initially observing an incorrect ML prediction mitigated the impediment of the skill development process.*

5. Discussion and Conclusion

While there is evidence that ML-based predictive decision support can effectively improve employee performance in organizations (see, e.g., Berente et al. 2021), there is reason to suspect that the implementation of such systems adversely affects the development of decision-making skills. With the growing use of such systems in organizations, it becomes ever more imperative to understand whether, and if so when, such negative ramifications come to light. The paper at hand contributes to this endeavor by examining the influence of ML-based decision support on skill development in a controlled experimental setting.

We produce evidence that the provision of ML predictions impedes the development of decision-making skills so that ML system users exhibit relatively lower decision-making performance once a system discontinuance occurs. Notably, we find that users increasingly follow ML predictions over time, without learning when to overrule them. We further provide evidence that the degree to which users "blindly" follow observed predictions shapes the ultimate lack of skill in the post-discontinuance phase. Our findings complement previous studies from the automation and information system discontinuance literature that document deskilling effects and drops in post-discontinuance performance caused by expert system decision support (Stone 2007; McCall et al. 2008; Soliman and Rinta-Kahila 2020;). We provide evidence that the effects previous research has documented in the domain of expert systems largely transfer to ML-based decision support characterized. Specifically, our results relate to McCall et al. (2008) who find that the use of knowledge-management systems affects the development of

declarative knowledge of users. Our results also relate to Rinta-Kahila et al. (2018) who observe deskilling effects for accountants who had access to a knowledge management system that functioned as a decision aid. Complementing their finding on expert systems, we also observe that ML decision support impedes the development of skill right from the beginning which comes to light once human users lose access to the decision aid. This evidence further emphasizes the importance of studying IS discontinuance effects.

Our results further suggest that participants gradually increase their reliance on the ML system over time, without getting better at detecting when predictions are incorrect. Hence, individuals do not seem to invest cognitive efforts in developing knowledge about the system but become increasingly complacent in blindly adhering to what the system says. This observation accords with previous studies reporting declining cognitive efforts of people who can constantly rely on technological support (Mendoza 2018; Lee et al. 2021). Sparrow (2011), for instance, coined the term "Google effect" after investigating the recall of words people typed into a computer. People knowing that the word is retrievable later performed worse in remembering the facts than those aware that the computer will not store the information. In a similar vein, the Kaspersky Lab (2017) described the tendency to forget information stored on a trusted device as "digital amnesia". Relatedly, our analyses on treatment heterogeneities suggest that such adverse effects depend on individuals' perceptions about the performance of the computerized decision aid. In our experiment, participants who saw the system make a mistake right at the beginning were less likely to rely on the support and seemed to develop better decision-making skills. This evidence complements previous research on the factors influencing the occurrence of deskilling effects (see, e.g., Orlikowski and Barley 2001, Orellana 2015) and contributes to reconciling contradicting evidence on the absence of deskilling in certain domains (see, e.g., Sayed 2006).

The twofold effect that ML decision support initially enhances individuals' performance but comes at the expense of developing decision-making skills constitutes a dilemma for organizations that consider implementing ML decision support. Contemporary ML models trained on historical data become inaccurate if disruptions and changes in the data distributions, i.e., concept drifts, occur (Widmer and Kubat 1996; Gama et al. 2014). In such a situation, employees can no longer rely on the predictive decision aid which means that there is an unintended, exogenous system discontinuance disrupting the decision process (Soliman and Rinta-Kahila 2020). Employees must cope with the absence of the ML decision aid and make informed decisions on their own until enough new data encoding the novel data distribution is collected so that the ML model can adapt itself, e.g., via retraining or updating. Only after such a transmission phase of human expert decision-making and an update of the ML model, the ML decision aid can again reliably support the corresponding business process. Put differently, the ML models' reliability, in the long run, requires human expertise. According to our results, however, the growing implementation of ML decision aids in organizations might reduce the expertise within organizations. As a consequence, the improvement of existing ML support systems may be inherently

limited due to the adverse effects on the development of decision-making skills. Contrary to suggestions by prior research that contemporary (AI-) technologies help us improve our cognitive potencies (e.g., Wilson and Daugherty 2018), our results depict a potential pitfall and advocate for more careful use of these technologies. Thereby we contribute to a nascent stream of papers that report challenges related to using ML applications. Fügener et al. (2021), for instance, show that the use of ML applications may reduce decision-making individuality, creating a convergence toward similar behaviors in the aggregate. Kane et al. (2021) even foreshadow the loss of human control over ML systems, painting a dark picture of the dawn of a machine-dominated age in which humans have no decision autonomy (see also Hirschheim et al. 1991). However, two observations may provide a possible silver lining to this rather dim perspective. First, it appears as if treatment participants' post-discontinuance performance slightly recovers over time. In cases where suboptimal decisions do not weigh strongly, employees may constructively adapt to the disrupted process and (re-)develop their skills. Second, we find that participants' blind reliance on predictions permanently decreases if they initially see the system err. This pattern implies that there is a value to increasing and maintaining individuals' awareness of the ML systems' error susceptibility to mitigate skill development concerns. In that regard, organizations might benefit from implementing training programs, guidelines, or instructions emphasizing the support system's imperfection. Moreover, participants who initially see the system make a mistake still reap considerable benefits from having access to the predictions. Hence, there might exist such a thing as "the right amount" of predictive decision support, where short-term productivity is elevated, and long-term adverse effects on knowledge development are contained.

Our results also have practical implications for managers who decide about the implementation of ML decision support. On the one hand, the insight that ML decision aids can impede the development of skills suggests that managers are well-advised not to implement such systems in domains where employees have to execute an assortment of distinct yet inherently related and complementary tasks. Providing employees with an ML decision aid for one of the tasks can impede the development of skills and thus negatively affect the performance across tasks. On the other hand, the impediment to skill development under ML decision support is naturally relevant to the training of novices, young professionals, and career starters. In domains where these novel employees need to develop a comprehensive understanding of the workings of interlinked organizational processes, it might be advisable to train them with limited access to ML decision aids so that they can make better decisions from the start, however, also develop fundamental decision-making skills.

As with any study, there are limitations to this paper. While decreasing the generalizability of our results, these limitations can serve as an inspiration for future research. One limitation naturally stems from the controlled experimental design. Even though our design allows us to isolate effects, it may not be representative of the development of decision-making across the entire range of the professional environment. That is because we use a relatively simple logic puzzle and recruit a pool of participants from the Prolific platform to study causal deskilling effects. One might argue that this setting is

reasonably representative of gig-workers or low-skill workers in routine environments (e.g., Uber drivers, or clerical office workers) but not for high-skilled professionals such as managers. Against this background, one fruitful avenue for future research is to study the existence of deskilling effects on the employment of predictive systems in the managerial domain (e.g., considering predictive due diligence assessments before mergers). Another direction to go is to test how ML-based decision support interacts with previously developed skills. That is, how do trained individuals' skills and decision performance change in response to providing ML-based decision support, and what happens if they encounter a system discontinuance?

Another limitation of our study is that we cannot directly measure reskilling effects, i.e., to what extent participants can become more proficient in understanding the workings of the decision support system. For the most part, that is because the ML decision support offered is a black box system. The black-box nature of many modern ML systems has adverse effects in many domains, for instance, on user trust or users' ability to detect errors (see, e.g., Gregor and Benbasat 1999; Pearl 2019; Rosenfeld and Richardson 2019). Current developments in the field of eXplainable Artificial Intelligence provide methods designed to counter these problems (see, e.g., Bauer et al. 2021). The implementation of explainability measures may be paramount to counter adverse skill development effects. Future research should put this notion to the test, especially due to the growing number of regulations prohibiting the use of black-box ML systems in certain areas.

Finally, our study cannot provide any guidance regarding an optimal level of ML decision support. Our results depict the existence of meaningful treatment heterogeneities when it comes to system reliance. These insights may suggest that there is such a thing as "the right amount" of predictive decision support, where short-term productivity is elevated, and long-term deskilling effects are contained. Future research may move beyond examining the repercussions of dichotomously deciding about implementing computerized decision support and consider the provision of decision support on a more continuous scale. One possible avenue would be to test the effectiveness of computerized decision support provided only at the explicit request of the user, i.e., when the effort required to obtain the support varies.

References

Aghion, P., & Tirole, J. (1997). Formal and real authority in organizations. Journal of political economy, 105(1), 1-29.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. Information Economics and Policy, 47, 1-6.

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. MIS quarterly, 107-136.

Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological review, 89(4), 369.

Anderson, J. R., & Fincham, J. M. (1994). Acquisition of procedural skills from examples. Journal of experimental psychology: learning, memory, and cognition, 20(6), 1322.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. Journal of experimental psychology: learning, memory, and cognition, 23(4), 932.

Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Hillsdale, NJ: Lawrence Erlbaum Associates. ISBN 0-8058-2817-6.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. Psychological review, 111(4), 1036.

Anderson, J. R. (2013). The architecture of cognition. Psychology Press.

Anderson, J. R. (2014). Rules of the mind. Psychology Press.

Arnold, V., Leech, S. A., Rose, J., & Sutton, S. G. (2018). Can Knowledge Based Systems be Designed to Counteract Deskilling Effects. Working paper, The University of Melbourne.

Asatiani, A., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2019). Implementation of automation as distributed cognition in knowledge work organizations: Six recommendations for managers. In 40th international conference on information systems (pp. 1-16).

Axelsen, M. Continued use of intelligent decision aids and auditor knowledge: Qualitative evidence. 18th Americas Conference on Information Systems 2012, AMCIS 2012 5, (2012), 3860–3869.

Barney, J.B. (1991). Firm Resources and Sustained Competitive Advantage. Journal of Management. 17 (1): 99–120.

Barney, J.B. (2001). Is the Resource-Based "View" a Useful Perspective for Strategic Management Research? Academy of Management Review. 26 (1): 101.

Bauer, K., Hinz, O., van der Aalst, W., & Weinhardt, C. (2021). Expl (ai) n it to me–explainable ai and information systems research. Business & Information Systems Engineering, 1-4.

Beaudry, P., Green, D. A., & Sand, B. M. (2016). The great reversal in the demand for skill and cognitive tasks. Journal of Labor Economics, 34(S1), S199-S247.

Berente, N., Gu, B., Recker, J., and Santhanam, R. (2019). Call for Papers MIS Quarterly Special Issue on Managing AI. MIS Quarterly, 43(1), 1-5.

Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing Artificial Intelligence, MIS Quarterly (45:3), pp. 1433–1450.

Boghossian, P. (2007). Fear of knowledge: Against relativism and constructivism. Clarendon Press.

Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation. Organizational behavior and human decision processes, 67(1), 26-48.

Carr, N. (2014). The Glass Cage: How Our Computers Are Changing Us. W. W. Norton & Company; 1 edition.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. Applied cognitive psychology, 10(2), 151-170.

Cummings, M. (2004). Automation Bias in Intelligent Time Critical Decision Support Systems. In AIAA 1st Intelligent Systems Technical Conference, 6313.

De Spiegeleer J, Madan DB, Reyners S, Schoutens W (2018) Machine learning for quantitative finance: Fast derivative pricing, hedging and fitting. Quantitative Finance 18(10):1635–1643.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms if They Can (Even Slightly) Modify Them. Management Science, 64(3), 1155-1170.

Dowling, C., Leech, S. A., & Moroney, R. (2008). Audit support system design and the declarative knowledge of long-term users. Journal of Emerging Technologies in Accounting, 5(1), 99-108.

Downey, M. (2021). Partial automation and the technology-enabled deskilling of routine jobs. Labour Economics, 69, 101973.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The Perceived Utility of Human and Automated Aids in a Visual Detection Task. Human Factors, 44(1), 79-94.

Fantl, J. (2008). Knowing-How and knowing-that. Philosophy Compass, 3(3), 451-470.

Fiske, S. T., & Taylor, S. E. (1991). Social cognition. Mcgraw-Hill Book Company.

Fitts PM (ed) (1951) Human Engineering for an Effective Air Navigation and Traffic Control System. National Research Council, Washington, DC.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI, MIS Quarterly (45:3), pp. 1527–1556.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 1-37.

Garbarino, E. C., & Edell, J. A. (1997). Cognitive effort, affect, and choice. Journal of consumer research, 24(2), 147-158.

Goddard, K., Roudsari, A., and Wyatt, J.C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 19, 1, 121–127.

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. Strategic management journal, 17(S2), 109-122.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. MIS quarterly, 497-530.

Grimaldi, P. J., & Karpicke, J. D. (2012). When and Why do Retrieval Attempts Enhance Subsequent Encoding? Memory & Cognition, 40(4), 505-513.

Hirschheim, R. and Newman, M. Symbolism and information systems development: Myth, metaphor and magic. Information Systems Research, 2, 1 (1991), 29–62.

Hoff, T. (2011). Deskilling and Adaptation Among Primary Care Physicians Using Two Work Innovations. Health Care Management Review, 36(4), 338-348.

Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in hiring. The Quarterly Journal of Economics, 133(2), 765–800.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. ACM Transactions on Computer-Human Interaction (TOCHI), 7(2), 174-196.

Hutchins, E. (2000). Distributed cognition. International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science, 138.

Janssen, J. H., Tacken, P., de Vries, J. J. G., van den Broek, E. L., Westerink, J. H., Haselager, P., & IJsselsteijn, W. A. (2013). Machines Outperform Laypersons in Recognizing Emotions Elicited by Autobiographical Recollection. Human-Computer Interaction, 28(6), 479-517.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Van Riemsdijk, B., & Sierhuis, M. (2010). The Fundamental Principle of Coactive Design: Interdependence Must Shape Autonomy. International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems, 172-191, Springer, Berlin, Heidelberg.

Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. Science, 349(6245), 255-260.

Kane, G. C., Young, A. G., Majchrzak, A., and Ransbotham, S. (2021). Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants, MIS Quarterly (45:1), pp. 371– 396.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4401-4410).

Kaspersky Lab (2017). The rise and impact of Digital Amnesia. Retrieved from

https://media.kasperskycontenthub.com/wp-content/uploads/sites/100/2017/03/10084613/Digital-Amnesia-Report.pdf

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing Unsuccessfully: A Specification of the Underlying Mechanisms Supporting Its Influence on Retention. Journal of Memory and Language, 66(4), 731-746.

Kuhnen, C. M. (2015). Asymmetric Learning from Financial Information. Journal of Finance, 70(5), 2029-2062.

Kuhnen, C. M., & Miu, A. C. (2017). Socioeconomic Status and Learning from Financial Information. Journal of Financial Economics, 124(2), 349-372.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 275-284).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444.

Lee, S., McDonough, I. M., Mendoza, J. S., Brasfield, M. B., Enam, T., Reynolds, C., & Pody, B. C. (2021). Cellphone addiction explains how cellphones impair learning for lecture materials. Applied Cognitive Psychology, 35(1), 123-135.

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine learning in banking risk management: A literature review. Risks, 7(1), 29.

Loureiro, A. L., Miguéis, V. L., & da Silva, L. F. (2018). Exploring the Use of Deep Neural Networks for Sales Forecasting in Fashion Retail. Decision Support Systems, 114, 81-93.

Mascha, M. F., & Smedley, G. (2007). Can computerized decision aids do "damage"? A case for tailoring feedback and task complexity based on task experience. International Journal of Accounting Information Systems, 8(2), 73-91.

Mateus, J. C., Claeys, D., Limère, V., Cottyn, J., & Aghezzaf, E. H. (2019). A Structured Methodology for the Design of a Human-Robot Collaborative Assembly Workplace. The International Journal of Advanced Manufacturing Technology, 102(5), 2663-2681.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. Harvard Business Review, 90(10), 60–68.

McCall, H., Arnold, V., & Sutton, S. G. (2008). Use of knowledge management systems and the impact on the acquisition of explicit knowledge. Journal of Information Systems, 22(2), 77-101.

Mendoza, J. S., Pody, B. C., Lee, S., Kim, M., & McDonough, I. M. (2018). The effect of cellphones on attention and learning: The influences of time, distraction, and nomophobia. Computers in Human Behavior, 86, 52-60.

Millman, Z., & Hartwick, J. (1987). The impact of automated office systems on middle managers and their work. MIS quarterly, 479-491.

Mosier, K. L., Skitka, L. J., Burdick, M. D., & Heers, S. T. (1996). Automation bias, accountability, and verification behaviors. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 40, No. 4, pp. 204-208). Sage CA: Los Angeles, CA: SAGE Publications.

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation Bias: Decision Making and Performance in High-Tech Cockpits. The International Journal of Aviation Psychology, 8(1), 47-63.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. Organization science, 5(1), 14-37.

Norström, P. (2015). Knowing how, knowing that, knowing technology. Philosophy & Technology, 28(4), 553-565.

Orellana, E. Deskilling, Up-skilling or Reskilling? Effects of Automation in Information Systems Context. Twenty-first Americas Conference on Information Systems, Puerto Rico, 2015, (2015), 1–17.

Orlikowski, W. J. (1991). Integrated information environment or matrix of control? The contradictory implications of information technology. Accounting, management and information technologies, 1(1), 9-42.

Orlikowski, W.J. and Barley, S.R. Technology and Institutions: What Can Research on Information Technology and Research on Organizations Learn from Each Other? MIS Quarterly 25, 2 (2001), 145–165.

Paas, F., A. Renkl, and J. Sweller. (2003). Cognitive load theory and instructional design: Recent developments. Educational Psychologist 38 (1): 1–4.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. Human Factors, 52(3), 381-410.

Parliament & Council of European Union. (2016). Regulation (EU) 2016/679 of the european 22 parliament and of the council. (https://eur-lex.europa.eu/eli/reg/2016/679/oj)

Parliament & Council of European Union. (2021). Proposal for a regulation of the European parliament and of the council laying down harmonized rules on artificiaul intelligence (artificial intelligence act) and amending certain union legislative acts. COM/2021/206 final. (https://eurlex.europa.eu/legalcontent/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206)

Pearl, J. (2019). The Limitations of Opaque Learning Machines. In J. Brockman (Ed.), Possible Minds: 25 Ways of Looking at AI, 13-19, New York, NY: Penguin Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. The Journal of machine Learning research, 12, 2825-2830.

Power, D., & Gruner, R. L. (2015). Exploring reduced global standards-based inter-organisational information technology adoption. International Journal of Operations & Production Management.

Ranz, F., Hummel, V., & Sihn, W. (2017). Capability-Based Task Allocation in Human-Robot Collaboration. Procedia Manufacturing, 9, 182-189.

Rinta-Kahila, T., Penttinen, E., Salovaara, A., & Soliman, W. (2018). Consequences of Discontinuing Knowledge Work Automation : Surfacing of Deskilling Effects and Methods of Recovery. In Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS 2018) (pp. 5244-5253). University of Hawai'i at Manoa. Proceedings of the Annual Hawaii International Conference on System Sciences. https://doi.org/10.24251/hicss.2018.654

Rinta-Kahila, T., Penttinen, E., Kumar, A., & Janakiraman, R. (2021). Customer reactions to self-checkout discontinuance. Journal of Retailing and Consumer Services, 61, 102498.

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. Autonomous Agents and Multi-Agent Systems, 33(6), 673-705.

Sayed, H. ERPs and accountants' expertise: the construction of relevance. Journal of Enterprise Information Management 19, (2006), 83–96.

Schuppan, T. E-government at work level: Skilling or de-skilling? Proceedings of the Annual Hawaii International Conference on System Sciences, (2014), 1927–1934.

Singh, S., Hussain, S., & Bazaz, M. A. (2017). Short Term Load Forecasting Using Artificial Neural Network. In 2017 Fourth International Conference on Image Information Processing (ICIIP), 1-5, IEEE.

Shaffer, V. A., Probst, C. A., Merkle, E. C., Arkes, H. R., & Medow, M. A. (2013). Why Do Patients Derogate Physicians Who Use a Computer-Based Diagnostic Support System? Medical Decision Making, 33(1), 108-118.

Shang, S., & Seddon, P. B. (2002). Assessing and managing the benefits of enterprise systems: the business manager's perspective. Information systems journal, 12(4), 271-299.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does Automation Bias Decision-Making? International Journal of Human-Computer Studies, 51(5), 991-1006.

Soliman, W., & Rinta-Kahila, T. (2020). Toward a refined conceptualization of IS discontinuance: Reflection on the past and a way forward. Information & Management, 57(2), 103167.

Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2012). Man vs. Computer: Benchmarking Machine Learning Algorithms for Traffic Sign Recognition. Neural Networks, 32, 323-332.

Stone, G.D. Agricultural Deskilling and the Spread of Genetically Modified Cotton in Warangal. Current Anthropology 48, 1 (2007), 67–87.

Sparrow, B., Liu, J., and Wegner, D. M. (2011). Google effects on memory: cognitive consequences of having information at our fingertips. Science 333, 776–778. doi: 10.1126/science.1207745.

Spitzer, M. (2014). Information technology in education: Risks and side effects. Trends in Neuroscience and Education, 3(3-4), 81-85.

Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation require a deeper understanding of human-ML augmentation. MIS Quarterly, 45(3).

Triki, A., & Weisner, M. M. (2014). Lessons from the literature on the theory of technology dominance: Possibilities for an extended research framework. Journal of Emerging Technologies in Accounting, 11(1), 41-69.

van Merrienboer, and F. Paas. 1998. Cognitive architecture and instructional design. Educational Psychology Review 10 (3): 251–296.

van Merrienboer, J., and J. Sweller. 2005. Cognitive load theory and complex learning: Recent developments and future directions. Educational Psychology Review 17 (2): 147–177.

Verghese, A. (2008). Culture Shock-Patient as Icon, Icon as Patient. The New England Journal of Medicine, 359(26), 2748-2751.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. Machine learning, 23(1), 69-101.

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative Intelligence: Humans and AI Are Joining Forces. Harvard Business Review, 96(4), 114-123.

Appendix

Additional material

Category	Item	Overall	Treatment	Baseline	p-value
Socio-demographic information	How old are you?	26.2	25.6	26.8	p=0.23 (Wilcoxon ranksum test)
	Please indicate your gender.	Female: 0.64 Male: 0.36	Female: 0.67 Male: 0.33	Female: 0.6 Male: 0.4	p=0.24 (Chi^2-test)
	How many years of working experience do you have?	4.8	4.33	5.28	p=0.395 (Wilcoxon ranksum test)
	Please indicate your highest degree.	No degree: 0.01 Secondary school: 0.06 High school: 0.37 Bachelor: 0.39 Master: 0.15 Ph.D.: 0.02	Secondary school: 0.06 High school: 0.38 Bachelor: 0.39 Master: 0.16 Ph.D.: 0.01	No degree: 0.02 Secondary school: 0.06 High school: 0.37 Bachelor: 0.39 Master: 0.15 Ph.D.: 0.01	p=0.94 (Kolmogorov-Smirnov test)
Risk aversion	I enjoy being daring.	4.6	4.7	4.54	p=0.417 (Wilcoxon ranksum test)
	I take risks.	4.78	4.9	4.62	p=0.03 (Wilcoxon ranksum test)
	I am looking for danger.	2.32	2.2	2.42	p=0.31 (Wilcoxon ranksum test)
Technology anxiety	I feel apprehensive about using technology.	3.08	3.04	3.11	p=0.71 (Wilcoxon ranksum test)
	Technical terms sound like confusing jargon to me.	3.06	3.12	2.99	p=0.48 (Wilcoxon ranksum test)
	I have avoided technology because it is unfamiliar to me.	1.69	1.63	1.75	p=0.34 (Wilcoxon ranksum test)
	I hesitate to use most forms of technology for fear of making mistakes I cannot correct.	1.95	1.85	2..05	p=0.38 (Wilcoxon ranksum test)
Other	Do you have a background in computer science?	0.3	0.31	0.3	p=0.81 (Chi^2-test)
	Please tell us, in general, about your programming skills, using a scale from 0 to 10, where 1 means no skills at all and 10 means expert.	2.9	3.07	2.75	p=0.29 (Wilcoxon ranksum test)
	Please tell us, in general, about your machine learning expertise, using a scale from 0 to 10, where 1 means no expertise at all and 10 means expert.	3.18	3.46	2.89	p=0.04 (Wilcoxon ranksum test)
	Human trust	4.73	4.62	4.85	p=0.2 (Wilcoxon ranksum test)

Machine trust	4.58	4.49	4.68	p=0.36 (Wilcoxon ranksum test)
Experience with investing	2.77	2.93	2.61	p=0.01 (Wilcoxon ranksum test)

Table 4: Overview of survey items. We report overall measures, and measures separately for treatment and baseline participants. We provide test statistics on treatment differences as randomization checks..

Dep. variable:	(1) Overrule prediction	(2) Overrule incorrect prediction
No. simulation	-0.016***	-0.003*
	(0.003)	(0.002)
Prediction	0.018	-0.012
	(0.018)	(0.011)
Expert and state controls	YES	YES
Observations	4470	4470
p	0.000	0.030
R-squared	0.194	0.063
Adj. R-squared	0.164	0.028

Table 5: *OLS regression with individual and round fixed effects. Both regression models are estimated on the subsample of treatment participants in stage 1. The dependent variable in column (1) is a dummy indicating whether a participant overruled the observed prediction in a given round. In column (2) the dependent variable is a dummy indicating whether a participant correctly overruled the observed prediction in a given round. The independent variables indicate the simulation round and the observed prediction. We also include, but do not explicitly report, control variables for the observed expert cues and the unobserved market state. We denote significance levels as *p<0.1, **p<0.05, ***p<0.01.*

Additional information on the experiment

To generate combinations of the economic situation, market states, and signals we used the python library *random*. The procedure for both training data generation and the data generation in the study worked as depicted in the following code example:

```python
import random

data = []
data_large = []
for i in range(1000000):
    # ECONOMIC STATE
    rs = (0 if int(random.uniform(0, 1) // 0.5)==0 else 1)
    if rs == 1:
        # MARKET STATE
        outcome = (int(0) if int(random.uniform(0, 1) // 0.3) == 0 else 1)
        # EXPERTS
        if outcome == 0:
            exp1 = (int(0) if int(random.uniform(0, 1) // 0.8) == 0 else 1)
        elif outcome == 1:
            exp1 = (int(0) if int(random.uniform(0, 1) // 0.2) == 0 else 1)
        exp2 = (int(0) if int(random.uniform(0, 1) // 0.5) == 0 else 1)
        exp3 = (_int(0) if int(random.uniform(0, 1) // 0.8) == 0 else 1)
        if outcome == 0:
            exp4 = 0
        if outcome == 1:
            exp4 = (int(0) if int(random.uniform(0, 1) // 0.5) == 0 else 1)

        l_= [exp1, exp2, exp3, exp4, outcome]
        l_large = [rs, exp1, exp2, exp3, exp4, outcome]

    elif rs == 0:
        # MARKET STATE
        outcome = (int(0) if int(random.uniform(0, 1) // 0.7) == 0 else 1)
        # EXPERTS
        exp1 = (int(0) if int(random.uniform(0, 1) // 0.5) == 0 else 1)
        if outcome == 0:
            exp2 = (int(0) if int(random.uniform(0, 1) // 0.8) == 0 else 1)
        if outcome == 1:
            exp2 = (int(0) if int(random.uniform(0, 1) // 0.2) == 0 else 1)
        if outcome == 0:
            exp3 = (int(0) if int(random.uniform(0, 1) // 0.5) == 0 else 1)
        if outcome == 1:
            exp3 = 1
        exp4 = (int(0) if int(random.uniform(0, 1) // 0.2) == 0 else 1)

        l_= [exp1, exp2, exp3, exp4, outcome]
        l_large = [rs, exp1, exp2, exp3, exp4, outcome]

    data.append(l_)
    data_large.append(l_large)
```

Figure 7: Python code for market state and expert signal generation.

Experimental Instructions

General instruction:

This experiment comprises two subsequent stages. In both parts, you will have to make a series of choices between investing into either a stock or bond. Contingent on your choice, you earn monetary units. The number of units you earn in every decision is summed up at the end of the experiment. For every monetary unit you have earned, you are paid 2 Eurocents. In addition, you will receive a fixed income of 2 Euros for your participation.

Example 1: Assume you earn 280 monetary units in the experiment. Your ultimate income in the experiment, which you are paid out, equals: 280*0.02 + 2 = 7.60 Euros.

Example 2: Assume you earn 220 monetary units in the experiment. Your ultimate income in the experiment, which you are paid out, equals: 220*0.02 + 2 = 6.40 Euros.

Please make your decisions carefully as they determine your ultimate income that you receive at the end of the experiment. The use of any decision support which is not provided on screen is strictly forbidden.

Instruction part 1:

Your task:

You participate in 6 asset market games. Each game consists of 5 consecutive rounds. In every round, you can invest in one of two assets:

- A risk-free bond that is always paying 2 monetary units for certain. A bond is a fixed income instrument that represents a loan made by an investor to a borrower (typically corporate or governmental) that traditionally pays a fixed interest rate (coupon) to investors.
- A risky stock which is either paying 0 (Low Payoff) or 4 (High Payoff) monetary units.

The probability that an investment into the stock yields the high payoff (4 units), depends on the stock's fundamental state, which you only observe at the end of the five rounds. With an equal probability (50% / 50%), the fundamental state can either be good, or bad. Depending on the market fundamental state, the probability of high and low payoffs looks as follows:

- Good state: If the stock is in a good state, the probability for the high payoff equals 70%. This is, with a chance of 70% you receive the payoff of 4, and with 30% you receive the payoff of 0.
- Bad state: If the stock is in a bad state, the probability for the high payoff equals 30%. This is, with a chance of 30% you receive the payoff of 4, and with 70% you receive the payoff of 0.

The fundamental state is the same across the 5 consecutive rounds and may only vary across the seven asset market games.

Available information

In every round, before you make your decision, you observe expectations about the stock's payoff in the given round from 4 independent experts.
Each expectation can either say High Payoff, or Low Payoff, indicating the expert's expectation about the stock's payoff in this round, which can be correct or incorrect.

Important: There exists a pattern behind the expert expectations, which does not change throughout the entire experiment. If you have figured out the underlying patterns, the expectations can help you identify when the stock will yield a high payoff. Understanding the pattern will be useful throughout the experiment by helping you to maximize your personal income.

[BEGIN: TREATMENT ONLY]
In addition to the expectations of the four experts, you observe the prediction of a Machine Learning System that was trained to predict the stock's payoff in a given round based on the

four experts' stated expectations. In other words: The Machine Learning System was designed to recognize the underlying pattern of expert expectations.

The Machine Learning System was trained and tested on 1.000.000 distinct observations. Each observation comprised the four expectations of the experts as independent variables and the stock's actual payoff as dependent variable. In a test, the trained Machine Learning System correctly predicts the stock's payoff in a given round in more than 85% of the cases. In several tests, the Machine Learning System systematically outperformed other researchers when it comes to correctly predicting the stock's payoff. Below you can find additional information about the structure of the system. We use a Random Forest Learning model.

Additional Information about the Machine Learning System

The Random Forest is one of the simplest, yet one of the most powerful machine learning models for classification. In the context of the experiment, classification refers to correctly predicting whether the stock will yield a high payoff or low payoff.

Random Forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time (Wikipedia). Each individual decision tree makes a single prediction. The majority prediction of the set of trees determines the ultimate prediction of the Random Forest. In other words: the forest uses the "wisdom of the crowd".

In this experiment, the Random Forest computes the probability that the stock will yield a high payoff. In other words, the model captures the underlying relationship between experts' expectations and the stock's actual payoff and can therefore help you make the payoff maximizing choice. You are only shown a High Payoff prediction if the probability for this event is the maximum. Otherwise, you are shown a Low Payoff prediction.

Some real-world applications:

E-Commerce
- Product Recommendation
- Price Optimization

Stock Market
- Stock Market Prediction
- Stock Market Sentiment Analysis

Healthcare and Medicine
- Cardiovascular Disease Prediction
- Diabetes Prediction

If you have further questions regarding the Machine Learning System or would like to obtain the entire code, please contact our research team on Prolific or via mail: [blinded for peer review]

[END: TREATMENT ONLY]

To ensure that you understand the task at hand, there will be two mock rounds of investment decisions, before the main part of the experiment starts. Note that the results of these mock rounds will not be paid out.

Instruction part 2:

You have finished part 1. Part 2 is about to start. Below you find instructions for the second part of the experiment.

Your task
You again participate in 6 asset market games. The games have the identical structure as before

[START: TREATMENT ONLY]
except for the fact that you do not observe predictions of the Machine Learning System anymore.
[END: TREATMENT ONLY]

Each game consists of 5 consecutive rounds. As before, you can always invest in one of two assets:

- A risk-free bond that is always paying 2 monetary units for certain. A bond is a fixed income instrument that represents a loan made by an investor to a borrower (typically corporate or governmental) that traditionally pays a fixed interest rate (coupon) to investors.
- A risky stock which is either paying 0 (Low Payoff) or 4 (High Payoff) monetary units.

The probability that an investment into the stock yields the high payoff (4 units), depends on the stock's fundamental state, which you only observe at the end of the five rounds. With an equal probability (50% / 50%), the fundamental state can either be good, or bad. Depending on the market fundamental state, the probability of high and low payoffs looks as follows:

- Good state: If the stock is in a good state, the probability for the high payoff equals 70%. This is, with a chance of 70% you receive the payoff of 4, and with 30% you receive the payoff of 0.
- Bad state: If the stock is in a bad state, the probability for the high payoff equals 30%. This is, with a chance of 30% you receive the payoff of 4, and with 70% you receive the payoff of 0.

The fundamental state is the same across the 5 consecutive rounds and may only vary across the seven asset market games.

Available information

As in the previous games, in every round, before you make your decision, you observe expectations about the stock's payoff in the given round from 4 independent experts.
The pattern behind the expert expectations, is exactly the same as in the previous games and does not change. If you have already figured out the underlying patterns in stage 1, you can apply your knowledge to better identify when the stock will yield a high payoff.

LEIBNIZ INSTITUTE SAFE Sustainable Architecture for Finance in Europe

Recent Issues

No. 369	Katja Langenbucher	Consumer Credit in The Age of AI – Beyond Anti-Discrimination Law
No. 368	Vanya Horneff, Raimond Maurer, Olivia S. Mitchell	How Would 401(k) 'Rothification' Alter Saving, Retirement Security, and Inequality?
No. 367	Jens Lausen, Benjamin Clapham, Peter Gomber, Micha Bender	Drivers and Effects of Stock Market Fragmentation - Insights on SME Stocks
No. 366	Satchit Sagade, Stefan Scharnowski, Christian Westheide	Broker Colocation and the Execution Costs of Customer and Proprietary Orders
No. 365	Caroline Fohlin	Short Sale Bans May Improve Market Quality During Crises: New Evidence from the 2020 Covid Crash
No. 364	Rachel Nam	Open Banking and Customer Data Sharing: Implications for FinTech Borrowers
No. 363	Kevin Bauer, Oliver Hinz, Johanna Jagow, Cristina Mihale-Wilson, Max Speicher, Moritz von Zahn	The Smart Green Nudge: Reducing Product Returns through Enriched Digital Footprints & Causal Machine Learning
No. 362	Tabea Bucher-Koenen, Andreas Hackethal, Johannes Kasinger, Christine Laudenbach	Disparities in Financial Literacy, Pension Planning, and Saving Behavior
No. 361	Ata Can Bertay, José Gabo Carreño Bustos, Harry Huizinga, Burak Uras, Nathanael Vellekoop	Technological Change and the Finance Wage Premium
No. 360	Alfons J. Weichenrieder	A Note on the Role of Monetary Policy When Natural Gas Supply Is Inelastic
No. 359	Spencer Yongwook Kwon, Yueran Ma, Niklas Kaspar Zimmermann	100 Years of Rising Corporate Concentration
No. 358	Matteo Bagnara, Ruggero Jappelli	Liquidity Derivatives
No. 357	Huynh Sang Truong, Uwe Walz	Spillovers of PE Investments
No. 356	Markus Eyting	Why do we Discriminate? The Role of Motivated Reasoning
No. 355	Stephan Jank, Emanuel Moench, Michael Schneider	Safe Asset Shortage and Collateral Reuse