Understanding Choice Independence and Error Types in Human-AI Collaboration

Erlei, Alexander; Sharma, Abhinav; Gadiraju, Ujwal

DOI
10.1145/3613904.3641946

Publication date
2024

Document Version
Final published version

Published in
CHI 2024 - Proceedings of the 2024 CHI Conference on Human Factors in Computing Sytems

Understanding Choice Independence and Error Types in Human-AI Collaboration

Alexander Erlei
alexander.erlei@wiwi.uni-goettingen.de
Georg-August-Universität Göttingen
Germany

Abhinav Sharma
abhinav.sharma@iiitg.ac.in
IIIT Guwahati
India

Ujwal Gadiraju
u.k.gadiraju@tudelft.nl
Delft University of Technology
Netherlands

ABSTRACT

The ability to make appropriate delegation decisions is an important prerequisite of effective human-AI collaboration. Recent work, however, has shown that people struggle to evaluate AI systems in the presence of forecasting errors, falling well short of relying on AI systems appropriately. We use a pre-registered crowdsourcing study ($N = 611$) to extend this literature by two underexplored crucial features of human AI decision-making: *choice independence* and *error type*. Subjects in our study repeatedly complete two prediction tasks and choose which predictions they want to delegate to an AI system. For one task, subjects receive a decision heuristic that allows them to make informed and relatively accurate predictions. The second task is substantially harder to solve, and subjects must come up with their own decision rule. We systematically vary the AI system's performance such that it either provides the best possible prediction for both tasks or only for one of the two. Our results demonstrate that people systematically violate choice independence by taking the AI's performance in an unrelated second task into account. Humans who delegate predictions to a superior AI in their own expertise domain significantly reduce appropriate reliance when the model makes systematic errors in a complementary expertise domain. In contrast, humans who delegate predictions to a superior AI in a complementary expertise domain significantly increase appropriate reliance when the model systematically errs in the human expertise domain. Furthermore, we show that humans differentiate between error types and that this effect is conditional on the considered expertise domain. This is the first empirical exploration of choice independence and error types in the context of human-AI collaboration. Our results have broad and important implications for the future design, deployment, and appropriate application of AI systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *User studies*; *Empirical studies in collaborative and social computing*; • **Applied computing** → *Economics*; Psychology.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3641946

KEYWORDS

Human-AI Collaboration; Interaction; Algorithm Aversion; Errors; Complementary AI Systems; Decision Support System; Crowdsourcing Study

ACM Reference Format:
Alexander Erlei, Abhinav Sharma, and Ujwal Gadiraju. 2024. Understanding Choice Independence and Error Types in Human-AI Collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3641946

1 INTRODUCTION

Humans collaborate with AI in many important decision domains, ranging from everyday product recommendations to critical workplace predictions in fields like medicine, law or financial services [1–4, 15, 33, 34]. Researchers and policy makers regularly stress the importance of human agency in these situations, e.g., for ethical, legal and safety reasons [9, 17, 20, 64, 65, 71, 96, 107, 108]. Following that principle, this article focuses on appropriate delegation as a crucial instantiation of human-AI collaboration. A decision maker faces multiple tasks, and decides for which ones to rely on an AI system. Ideally, this process involves carefully considering the predictive or diagnostic accuracy of each choice alternative. For example, a consumer could rely on recommender systems in so far as they have produced better outcomes for specific product types in the past or demonstrate capabilities that suggest desirable outcomes. Similarly, many judges would benefit from delegating bail decisions to predictive algorithms [7], and a physician may want to outsource certain parts of the diagnostic process when AI models can leverage vast and representative amounts of historical data [8, 80, 106]. If implemented appropriately, delegation to superior AI systems can create more effective workflows and produce better consumer outcomes (i.e, optimal *human-AI team performance* [10]).

However, there are at least three factors that impede such a scenario. One, humans struggle to consistently enforce good delegation rules in the presence of AI [67]. For example, recent work on algorithm aversion shows that humans over-weigh errors by automated decision systems, leading to substantial under-utilization [16, 30, 84]. Two, humans may not identify when a problem should be delegated to an AI system because of inadequate self or task assessments [39, 47, 94]. Three, effective delegation requires task-based *choice independence* from the human decision maker. Crudely, the independence axiom states that if a decision maker prefers to delegate task A to an AI system when the AI system makes good predictions for tasks A and B, they should also prefer to delegate task A to an AI system when the system makes good predictions

for task A but bad predictions for the unrelated task B. This axiom underlies the assumption that human-AI collaboration benefits particularly from AI systems optimized to assist humans in their weaker domains, i.e., *complementary AI*. For example, a physician using an AI system to augment their own diagnosis may recognize that the model provides useful information for common illnesses such as allergies or the flu but is less reliable for rare conditions like epilepsy. In that case, the physician should be able to judge the model's usefulness for common diseases independently of its other shortcomings. Despite the importance and relevance of this assumption, choice independence has not been empirically investigated within the broad context of human-AI collaboration.

This paper examines the efficacy of human-AI delegation when humans face multiple tasks. We use an online experiment in which subjects make a series of predictions based on three input numbers for two different outcomes of interest. For one task, subjects receive a simple decision heuristic and are thereby enabled to make very accurate predictions. We call this the *human expertise domain*. The second task is more complex, and subjects only learn through limited observation and experience, resulting in lower accuracy. This is the *complementary expertise domain*. Our setup reflects that most human decision makers have heterogeneous capabilities that map differently onto their various problem sets. Instead of relying on their own predictions, subjects can also choose to delegate each task to an AI system. We systematically vary the performance of the AI system for each outcome of interest. Depending on the treatment,[1] the AI system either (1) makes the best possible prediction for both outcomes, (2) makes systematic errors for the complex task, or (3) makes systematic errors for the easy task. This allows us to analyze two crucial elements of human-AI collaboration:

RQ1: Does the independence axiom of choice hold for delegation decisions in human-AI collaboration?
RQ2: How do humans condition their delegation choices on objective performance differences of an AI system between different prediction tasks?

Second, we vary both the error type caused by randomness in an uncertain forecasting environment and the error type caused by a systematic bias in the AI system's predictions. Our setup differentiates between continuous but relatively small inaccuracies, and rare but large prediction errors that may fall beyond the bounds of being reasonable. For example, in many financial decision domains or pricing predictions, AI models will almost never offer the "perfect" solution, instead exhibiting good and stable performances without any catastrophic deviations. On the other hand, even objectively "small" deviations in models used for self-driving cars or everyday medical diagnoses may result in large costs for the human delegator [5]. More generally, differentiating between different error types allows us to gauge which errors designers and developers should prioritize when training their models in order to maximize uptake.

[1]Note that we use the word 'treatment' interchangeably with 'experimental condition' in this paper.

RQ3: How do different prediction error types influence human reliance on a relatively more accurate AI system?

Our results show that humans consistently violate the choice independence assumption when delegating predictions to a superior AI system. Furthermore, the effect appears strongly conditional on the expertise domain. When an AI system makes the best-possible prediction for the easy task where humans receive a decision heuristic and are therefore relatively accurate, systematic AI errors in the complementary expertise domain *reduce* delegation shares for the easy task. In contrast, when the AI system functions as a complement and makes the best-possible prediction only for the complex task, systematic AI errors in the human expertise domain can *increase* delegation shares for the complex task.

Regarding error type, there is moderate evidence that participants are more likely to delegate their complex predictions to the best-possible AI system under continuous, rather than rare high-variance randomness. This pattern seems to be driven by lower subject self-confidence in prediction environments where perfect predictions are extremely rare.

Beyond that, we show that humans strongly condition their delegation behavior on objective AI system performance differences. In the human expertise domain, this leads to less delegation by humans who outperform a systematically erring AI system. In the complementary expertise domain, all participants significantly adjust their delegation shares downwards, irrespective of the performance level. This highlights the importance of expertise in building up the necessary meta-knowledge to utilize effective delegation rules. Lay populations may be less likely to tolerate more accurate but erring AI systems.

These results have strong implications for the design and application of AI systems. It is important to note that almost all documented effects depend on the considered expertise domain, despite the AI system outperforming almost every single human forecaster irrespective of treatment or problem. Humans appear to make very different choices depending on their self-confidence and the existence of helpful decision rules. This may be particularly important when thinking about designing systems for either experts or laypeople. Regarding our specific research questions, we provide strong evidence that humans do not evaluate AI systems task-independently. Whenever a system performs more than one function and exhibits performance differences between them, there could be implications for human utilization. For instance, a radiologist who observes the AI system's inaccuracies for complicated long-tail low-probability illnesses may reduce beneficial AI reliance in mainstream diagnoses [99]. On the other hand, truly complementary systems that strongly outperform humans in specific tasks may even benefit disproportionately from more fine-tuning that trades off their performance in the human expertise domain (see e.g. [53]). Further, our results suggest that error type can mediate the relationship between human delegation and AI performance. Areas that select for low-frequency but high-impact randomness, like the medical domain, may be particularly vulnerable to harmful algorithm aversion.

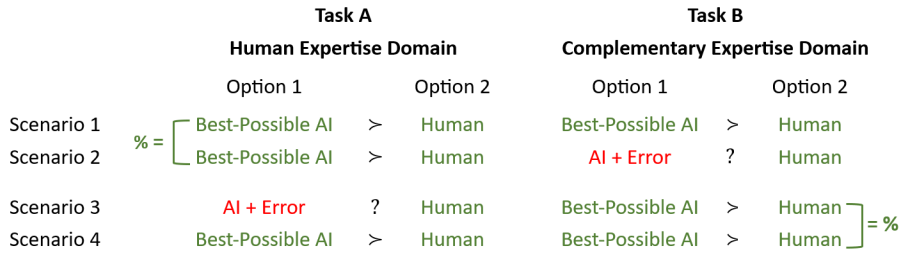

Figure 1: Illustration of the basic IA in our human-AI collaboration framework.

2 BACKGROUND AND RELATED WORK

2.1 Choice Independence

The independence axiom (IA) is an integral part of decision theory across various social sciences. Rational choice theory, for instance, builds on expected utility theory [101], which postulates choice independence as one of four central axioms. The IA is therefore foundational to neoclassical microeconomics and modern mathematical theories of decisions under uncertainty. Following von Neumann and Morgenstern, it states that human preferences between uncertain gambles should not change with the introduction of an additional, common gamble. Thus, if a decision maker prefers gamble A over gamble B, the introduction of a third gamble C should not change the decision maker's preference order over gambles A and B. Since its inception, the assumption has been subject to continuous debate. For decades, experiments have shown that in certain situations, humans fail to comply with the axiom [6, 57, 70, 79]. They often do not evaluate options in isolation, but in reference to, sometimes one, sometimes several other options [70, 72, 95]. One prominent example is the attraction or decoy effect, where the strategic addition of an asymmetrically dominated inferior alternative increases the attractiveness of the dominating option [52, 88]. Recent studies, however, have found it difficult to replicate these violations across a large number of choice environments [38, 105]. Indeed, there is evidence that a significant proportion of people do adhere to choice independence [44, 50, 68, 76] and that previously documented violations of IA can be empirically fragile [13, 26]. Still, several behavioral regularities that contradict the IA, such as the certainty effect or subjective probability weighting, largely remain empirically robust [91].

Overall, it is difficult to ascertain the "true" validity of the IA. There are undoubtedly many everyday decisions where many humans act in accordance with the axiom. Beyond very specific experimental gambling environments, we have little consistent evidence that would allow researchers to make generalizable predictions about which factors determine behavioral violations of IA. There is no one model that can simultaneously account for all choice patterns documented in the literature [58, 81]. Furthermore, to the best of our knowledge, choice independence has not yet been analyzed in forecasting, delegation, or advice-taking contexts. Instead, most of the literature on choice independence focuses on a decision maker's choices between uncertain, risky, or ambiguous alternatives, and how adjustments of existing options, or the introduction of novel options, change the decision maker's revealed preference ordering.

In this paper, we argue that the decision of a human forecaster between their own and an AI system's prediction is comparable to a decision between two uncertain gambles.[2] While the forecaster may have some information about the average performance level of either alternative, the accuracy of each individual prediction is always uncertain. This may be due to imperfect information and limited computational capabilities, or simply environmental randomness. A rational forecaster should evaluate the two options (themselves vs. AI system) for a given task, and, all else equal, choose the one with the highest subjectively expected accuracy. Furthermore, their preference order should not change in the presence of a distinct second task. A rational agent will evaluate both delegation decisions in isolation, implying that across different variations of any Task B (e.g., different levels of human and AI-system prediction accuracy, variance, or error type), preference ratios for any Task A remain constant (see Figure 1). This relationship holds as long as the variations in Task B have no informative value for Task A, meaning the two tasks are independent of one another.

2.2 Delegation in Human-AI-Collaboration

This article relates to the growing literature on reliance and delegation within human-AI collaboration [45, 46]. In their seminal paper, Dietvorst et al. [30] show that human forecasters strongly overweigh errors by superior algorithmic decision systems and therefore tend to rely on inferior human alternatives, resulting in substantial efficiency losses. This remarkably resilient pattern has been replicated in many contexts [16, 21, 22, 29, 31, 51, 59, 83, 85, 89, 90], although humans have also exhibited preferences for algorithms in task domains that are perceived as "objective" [21, 59, 69]. Research on perceptions of and information about algorithms suggests only small to ambiguous effects of AI knowledge on delegation [59, 84]. Similarly, there is mixed evidence on algorithms that demonstrate an ability to learn, although most research points towards increases in utilization [12, 25, 89]. Endowing human decision makers with agency over an algorithm's output substantially improves model

[2] Of course, there are also important differences. The forecaster has agency over their own performance, which can significantly determine choice outcomes. Depending on the human's knowledge about their own and the AI system's performance, there can be asymmetric information, which translates into asymmetric uncertainty. Furthermore, literature on forecasting and advice-taking heavily suggests a broad prevalence of overconfidence and egocentric discounting among human forecasters [14, 24, 98]. These differences underline the importance of our research because previous results on choice independence cannot be readily applied to human-AI collaboration.

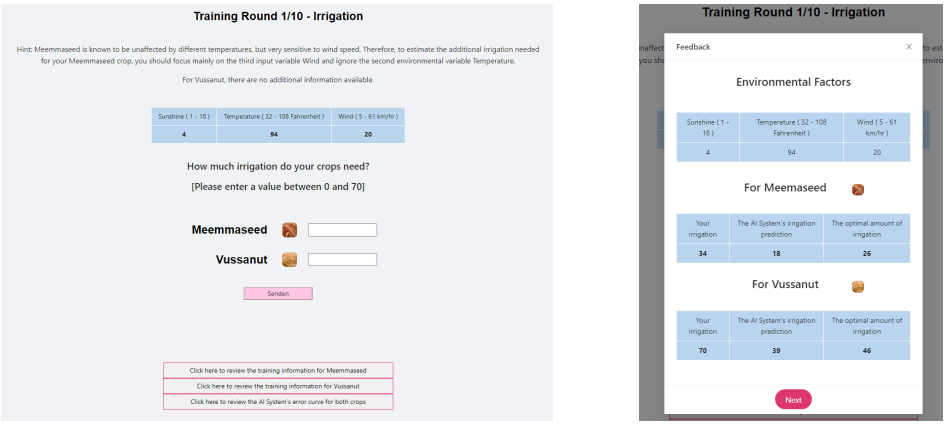

Figure 2: Training round prediction screen (left) and feedback screen (right).

evaluations and delegation choices [16, 22, 31, 55, 56, 60]. Finally, some studies propose that human delegation to superior algorithmic and AI systems is mediated by biased self-assessments, which may manifest in a lack of "metacognition" [39], overconfidence [27], or self-protection [78]. People fail to adequately judge their own performance level in relation to the task's difficulty, complexity [92], or uncertainty [93], and therefore do not implement effective delegation rules. Allowing AI systems to delegate tasks to human decision makers may alleviate these inefficiencies [39, 49].

2.3 Complementary Expertise in Human-AI-Collaboration

AI systems that provide humans with complementary expertise and thereby improve joint outcomes are one of the most promising fields of HCI research [40, 42, 74, 87, 102–104]. Several papers show that human-AI combinations can in principle exceed singular decision makers within the same task, e.g. by avoiding bad predictions or choices [11, 23, 28, 37, 48, 64, 73, 109]. Often, AI systems improve joint performance by giving human decision makers additional information or providing a useful baseline reference [48, 109]. Furthermore, fine-tuning an AI to compensate for specific human weaknesses like identifying false-negatives can also support user performance [53].

Most research analyzes complementary human-AI expertise strictly within the same task. Yet, often and similar to traditional teamwork, human-AI collaboration must be organized across tasks. In such a case, human decision makers decide which kind of task to delegate and which kind of task to complete themselves. Our main contribution to the expertise literature lies in highlighting previously under-explored interdependencies between different human-AI error profiles across different prediction tasks. If choice independence holds, these interdependencies do not exist. It would be, for instance, efficient to optimize a model's performance for tasks where humans have comparative disadvantages, even if it comes at the expense of tasks where humans perform well. However, if people fail to judge an AI system's performance in isolation, optimizing for specific tasks may have unintended consequences.

2.4 Human-AI-Collaboration and Error Types

Research on the influence of error type on human-AI delegation is scarce. Dietvorst and Bharti [29] find that higher uncertainty leads to stronger algorithm aversion because people have a diminishing sensitivity to forecasting errors and exhibit preferences for near-perfect predictions. Recent studies also point to the importance of first impressions in human-AI collaboration, showing that people react significantly stronger to relatively early errors [60, 82, 100]. Furthermore, humans may differentiate between algorithmic false-negatives and false-positives, although evidence for that is mixed and ambiguous [43, 62]. This article extends the exploration of different error types in human-AI collaboration by differentiating between continuous but moderate and large but rare errors. In addition, we look at errors that originate from environmental randomness and those that are systemic to the AI's predictions.

3 EXPERIMENTAL DESIGN

We employ six treatments of a pre-registered online prediction experiment in which participants take on the role of a farmer who predicts the irrigation need of two fictional crops, **Meemmaseed** (*human expertise domain*) and **Vussanut** (*complementary expertise domain*), each consuming one hectare of land. Participants learn that under ideal conditions, both crops require at least 40 thousand gallons of water. Their task is to predict the additional irrigation need, as determined by three observable environmental variables: Sunshine in hours/day (I_S), Average Day Temperature in Fahrenheit (I_T), and Wind Speed in km/h (I_W). Irrigation for Meemmaseed follows: $Y_M = 40 + 0.1 * I_S + 0 * I_T + 0.9 * I_W + \varepsilon$, and irrigation for Vussanut follows: $Y_V = 40 + 0.15 * I_S + 0.55 * I_T - 0.3 * I_W + \varepsilon$, where ε is a treatment-sensitive random error. The environmental input factors are randomly drawn from the following uniform distributions: $I_S \in [1, 18]$, $I_T \in [32, 108]$ and $I_W \in [5, 61]$.[3]

[3]We use farming as an example from real-world contexts where AI systems are increasingly being used, and as a scenario that participants can loosely comprehend. The task design is based on a rich body of literature in psychology, economics, and more recently, Human-AI interaction, where similar forecasting environments have been used to study a broad range of decision phenomena, including e.g., the interaction of humans and algorithms [29], rationality [41, 66], advice-taking and forecasting [24, 75], or overconfidence [47, 86]. Our setup mimics many real-life scenarios in which people use a set of attributes to generate forecasts, e.g., investments, evaluating

Table 1: The different experimental conditions in our study.

Treatment	Random Error	AI System Easy Problem		AI System Complex Problem		*N*
		best-possible	systematic error	best-possible	systematic error	
BP_Cont	Continuous	✓	None	✓	None	102
Subst_Cont	Continuous	✓	None	x	Rare & Large	103
Compl_Cont	Continuous	x	Rare & Large	✓	None	100
BP_Rare	Rare, Large	✓	None	✓	None	101
Subst_Rare	Rare, Large	✓	None	x	Continuous & Moderate	100
Compl_Rare	Rare, Large	x	Continuous & Moderate	✓	None	105

Thus, in order to make the best possible predictions, subjects need to learn the relationship between the three environmental inputs and the respective crop's irrigation needs. To achieve that, they complete two training periods, which are described below.

Instead of relying on their own prediction, subjects learn that they can also delegate their irrigation predictions to an AI system. At the beginning, subjects do not know anything about the system's performance. They only know that it does not receive additional information beyond the three environmental inputs.

During the first of the two training periods, subjects then see descriptive information from 20 simulated prediction rounds. Specifically, they first observe a table that shows each input factor in columns 1 – 3, and the actual irrigation requirement for Meemmaseed in column 4. Furthermore, subjects receive the information that Meemmaseed is "*known to be unaffected by different temperatures, but very sensitive to wind speed.*" Therefore, subjects are instructed to focus mainly on the third input variable and ignore the second one. Finally, columns 5 and 6 show the AI system's irrigation prediction, as well as the respective prediction error. For Vussanut, subjects observe the same table with the same environmental inputs, but different actual irrigation requirements, and different AI system predictions. They also receive no additional information about how the inputs relate to irrigation needs. Using all this information, subjects can learn (1) about the relationship between the environment and each crop's irrigation needs, as well as (2) the performance of the AI system. To help subjects evaluate the AI system's accuracy, we also show them a figure that illustrates the system's error curve for both Meemmaseed and Vussanut. We keep the axes constant across all treatments.

In the second training period, subjects complete 10 non-incentivized training predictions (see Figure 2). In each round, subjects observe three environmental input numbers and make two predictions, one for each crop. At the bottom of the page, subjects can always access the descriptive information from the 20 simulated prediction rounds as well as the AI system's error curves by clicking on one of three buttons. This opens a pop-up with the respective information. After submitting their predictions, subjects see a feedback screen that shows for both crops (1) the subject's irrigation prediction, (2) the AI system's irrigation prediction, and (3) the optimal amount of irrigation. The feedback screen also shows the environmental inputs to allow further learning.

After the 10 training predictions, subjects complete 10 incentivized official predictions. They earn 35 Coins for a perfect prediction, and each point that their implemented prediction is off reduces that income by 1 Coin. Coins are converted into pounds at the end of the task where 14 Coins = £1. To determine the final bonus payoff, we randomly select one of the 10 official predictions. Thus, subjects learn that every single official prediction could be the one deciding their income. In contrast to the training predictions, participants do not receive feedback after submitting their predictions. Instead, they decide whether to delegate the predictions for the current round to the AI system. Here, subjects must rely on their previously acquired knowledge, because the AI system's predictions are not observable. Subjects make two delegation decisions, one for each crop. They can for example decide to delegate the irrigation prediction for Vussanut to the AI system but use their own prediction for Meemmaseed.

Finally, upon completing the official predictions, subjects fill out a post-experimental questionnaire. They answer a battery of questions about their confidence in themselves and the AI system, state their risk attitudes [32], complete the subjective numeracy scale [36] as well as the trust in automation questionnaire [63], and share some demographic data.

We share all the data, the original instructions, the preregistration, and this project's code via an online repository (https://osf.io/kh9x6/?view_only=bcc35724db794cc698a6306d9dc6a237) for the benefit of the community and in the spirit of open science.

3.1 Experimental Conditions

We use a 2 (continuous environmental random error vs. rare environmental random error) x 3 (best-possible AI system vs. complementary AI system vs. substitute AI system) between-subject design (see Table 1).

Our first intervention concerns the random error in the environment. Remember that the irrigation need for each crop is determined by the three environmental input factors and a random error ε. Randomness is ubiquitous in real-life environments and is one reason why consistent perfect predictions are almost never possible. We use two different environmental random errors: a relatively

job applicants, diagnosing illnesses, or assessing consumer products. We rely on a linear relationship between input and output factors because (1) it has already been used to analyze human-algorithm interactions [29], (2) is relatively intuitive to human subjects, and (3) provided good results (high accuracy for the "easy" task, low accuracy for the "complex" task) in our pilot. The intervals for the input factors reflect realistic real-life boundaries.

Table 2: A comparison of the main experimental conditions.

Comparison	Outcome of Interest	Research Question	Hypothesis Under IA	Real-Life Example
BP_Cont vs. Subst_Cont BP_Rare vs. Subst_Rare	Share of subjects delegating their Meemmaseed prediction to the AI system	Do AI errors in the complementary expertise domain affect human-AI delegation in the human expertise domain?	There are no differences in delegation behavior between BP_ and Subt_	Experienced Investor using a stock forecasting model; consumer recommender systems for experience goods; physicians diagnosing mainstream illnesses
BP_Cont vs. Compl_Cont BP_Rare vs. Compl_Rare	Share of subjects delegating their Vussanut prediction to the AI system	Do AI errors in the human expertise domain affect human-AI delegation in the complementary expertise domain?	There are no differences in delegation behavior between BP_ and Compl_	Inexperienced investor using a stock forecasting model; consumer recommender systems for unknown products or credence goods; lay people using e.g. GPT-4 to self-diagnose or review professional literature
BP_Cont vs. BP_Rare	Share of subjects delegating their Meemmaseed prediction to the AI system	Does human delegation to the best-possible AI system in their own expertise domain depend on the distribution of randomness in the prediction environment?	N/A, exploratory analysis	Firm using a commercial pricing algorithm vs. physician using a medical expert system
BP_Cont vs. BP_Rare	Share of subjects delegating their Vussanut prediction to the AI system	Does human delegation to the best-possible AI system in a complementary expertise domain depend on the distribution of randomness in the prediction environment?	N/A, exploratory analysis	Layperson using a LLM to forecast stock prices vs. layperson using a LLM to medically self-diagnose

small continuous error,[4] and a larger, rare error. The continuous error is randomly drawn from the uniform distribution $\varepsilon \in [-5, 5]$. The rare error becomes 0 with a probability of 80% and is otherwise randomly drawn from the uniform distribution $\varepsilon \in [-27, 27]$. In both cases, the expected value is 0, and the mean error is virtually the same.

Our second intervention concerns the AI system's performance. The best-possible model makes the best possible prediction by using the correct formula and weights for the three input factors. The only prediction error that remains is caused by the random environmental error ε, which always has an expected value of 0 and is not predictable. It is never possible to beat the best-possible AI system in the long run. Therefore, subjects should always delegate their prediction.

In addition to the best-possible AI system, we introduce two models that exhibit systematic errors. The systematic error depends on the random error in the environment. If there is continuous but small randomness, i.e. $\varepsilon \in [-5, 5]$, the AI system with the systematic error makes the best-possible prediction with a probability of 50%, but has an additional prediction error $\sigma = 24$ for the relatively easy problem Meemmaseed and $\sigma = 30$ for Vussanut with a probability of 50%. Thus, in some cases, the model makes large mistakes. On the other hand, if environmental randomness is rare but large, i.e. $\varepsilon \in [-27, 27]$, the imperfect AI system has a continuous additional prediction error $\sigma \in \{10, 11, 12, 13, 14\}$ for Meemmaseed and $\sigma \in \{13, 14, 15, 16, 17\}$ for Vussanut. Again, both systematic errors have the same mean error.[5]

Combining the two interventions, there are six treatments: **BP_** always refers to an AI system that makes the best-possible prediction for both crops. In **Subst_**, the AI system makes the best-possible prediction for the easy task (Meemmaseed), where humans receive more information and make good predictions, but makes systematically biased predictions for the complex task (Vussanut). Hence, the model is a substitute for the human. The AI system in **Compl_** makes good predictions for complex tasks but systematically biased predictions for easy tasks and is, therefore, a complementary prediction tool. Here, human forecasters can on average optimize their

[4]Note that by a continuous error, we mean a steady and recurring error.

[5]The difference between the systematic errors for both crops is based on a pilot. Because subject predictions for Vussanut are, on average, 2.5 to 4.5 points worse than for Meemmaseed, we choose a systematic AI error for Vussanut that is 3 points larger. Furthermore, we set the size of the systematic error such that human forecasters should, on average, make better predictions than the erring model. The frequency of the rare systematic error is 50% because it allows us to stay within the incentive structure of a pilot. Generally, the difference in error size between continuous and rare inaccuracies is designed to allow for salient disparities in human perception.

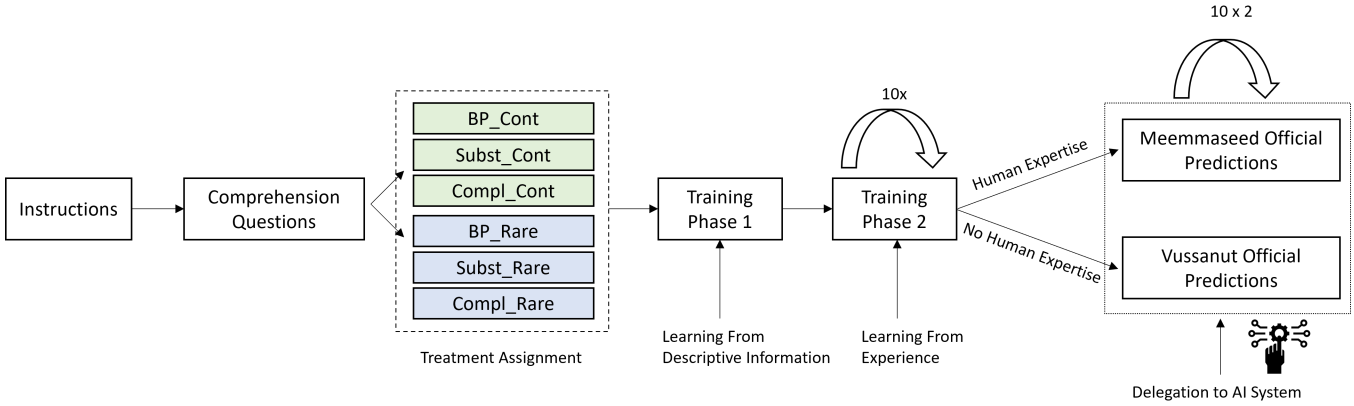

Figure 3: An illustration of our experimental procedure.

accuracy by relying on themselves for Meemmaseed and delegating the prediction for Vussanut to the AI system. Finally, **_Cont** refers to the environment with continuous randomness ε, and **_Rare** to the environment with large but rare random outliers.

3.1.1 Treatment Comparisons. Table 2 provides an overview of our four main treatment comparisons. The Results section comprises the full statistical analysis, as well as additional auxiliary results. Our treatment composition generates four tests of the choice independence hypothesis, conditional on environmental randomness and the respective expertise domain. If the IA holds, then human forecasters evaluate the AI's performance in both tasks independently. Therefore, there can be no differences in delegation between **BP_** and **Subst_** for Meemmaseed (human expertise domain), because the model provides the best-possible Meemmaseed prediction in both treatments. Similarly, there can be no differences in delegation between **BP_** and **Compl_** for Vussanut (complementary expertise domain), because the model provides the best-possible Vussanut prediction in both treatments. If, for example, the share of subjects delegating their irrigation prediction to Vussanut differs between **BP_** and **Compl_**, then this difference is solely driven by the AI system's Meemmaseed prediction errors in **Compl_**.

In addition, we offer a pre-registered exploration of error types on human-AI delegation. The order of the documented treatment comparisons replicates the order in the Results section.

3.2 Procedure

Figure 3 illustrates the experimental procedure. All subjects read the same basic instructions and then proceeded to answer four comprehension questions. Those who correctly answered all four within three trials were allowed to participate in the study.

Participants were then randomly assigned to one of six treatments. The treatments only differ in the AI system's performance across the two problems, and the random environmental error. Otherwise, everything is identical. For each treatment, we selected 5 different 20-round simulations before the experiment and randomly chose between them. This increases the robustness of our results and allows for some exploratory analysis regarding subjects' reactions toward different kinds of large errors (e.g., negative additional irrigation predictions by the AI system). Similarly, we randomly draw the 10 training predictions from a pool of 50 priorly selected rounds to balance variance and between-subject consistency. Participants complete all official predictions in randomized order.

3.3 Participants

We collected data until 100 independent observations per treatment using Prolific. All participants are native English-speakers who reside either in the USA or the UK, have an approval rating of at least 90%, and completed at least 50 prior tasks on the platform. Those who failed to answer four comprehension questions correctly within three trials were not allowed to participate in our experiment. We do not exclude any subject post-data collection. This results in a total of 611 subjects (41% female). Participants earned a base payment of £1.5 and an average bonus of £2.03, resulting in an hourly wage of roughly £10.5.

4 RESULTS

We first analyze choice independence and subjects' general delegation behavior conditional on their and the model's prediction performance. Then, we consider the effects of error type on human-AI collaboration. Throughout, we mainly rely on a panel logistic regression with individual-level random effects and clustered standard errors for delegation hypothesis testing (see Tables 4, 5, 6). P-values are adjusted for multiple hypothesis testing using the Westfall and Young free step-down resampling method [54]. The significance stars in the figures correspond to the following cut-offs: * indicates $p < 0.05$, ** indicates $p < 0.01$, and *** indicates $p < 0.001$. For attitudes and perceptions, we use two-sided t-tests with the same cut-offs.

4.1 Prediction Performance and Manipulation Check

Table 3 shows average human and AI prediction errors across treatments and problems. In all **BP_** conditions, the model clearly outperforms human forecasters. The difference is larger for the complex

Table 3: Average Human and AI Prediction Errors Across Treatments and Problems (SD in parentheses). Bold cells signify instances where humans outperformed the model on average.

	Training Rounds Humans		Training Rounds Model		Official Rounds Humans		Official Rounds Model	
	Meemmaseed	Vussanut	Meemmaseed	Vussanut	Meemmaseed	Vussanut	Meemmaseed	Vussanut
BP_Cont	9.58 (6.5)	12.35 (4.89)	2.51 (0.49)	2.62 (0.46)	7.8 (6.1)	12.03 (5.12)	2.51 (0.48)	2.46 (0.43)
Subst_Cont	9.02 (6.06)	12.89 (5.32)	2.49 (0.6)	**17.03 (4.66)**	8.25 (6.49)	12.79 (5.7)	2.51 (0.46)	**16.58 (1.61)**
Compl_Cont	10.51 (6.8)	13.71 (5.16)	**13.68 (3.8)**	2.58 (0.47)	8.88 (6.27)	12.68 (5.02)	**13.86 (1.25)**	2.54 (0.51)
BP_Rare	10.77 (6.8)	13.66 (5.52)	3.68 (1.14)	3.56 (1.13)	9.11 (6.25)	13.79 (6.26)	2.49 (1.14)	2.63 (1.19)
Subst_Rare	10.71 (7.5)	14.54 (5.69)	3.59 (1.08)	**17.7 (1.66)**	9.34 (7.98)	13.93 (5.92)	3.06 (1.26)	**17.77 (1.23)**
Compl_Rare	9.73 (6.2)	14.18 (4.4)	**15.28 (1.12)**	3.6 (1.18)	8.8 (6.74)	13.59 (5.36)	**14.49 (1.25)**	2.48 (1.08)

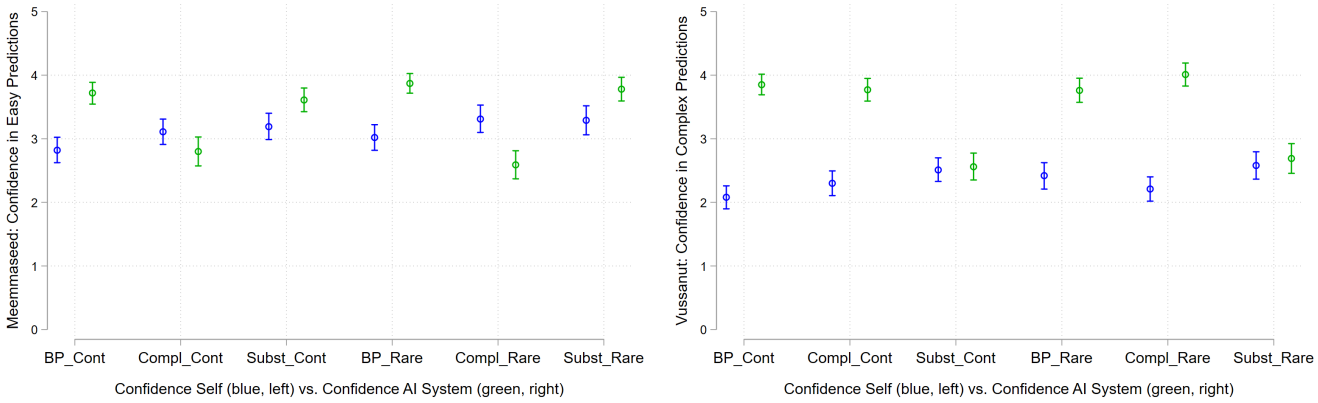

Figure 4: Average subject confidence levels in their own (blue) and the AI system's (green) predictions per treatment and task. Left: human expertise domain. Right: complementary expertise domain. Subjects state their confidence in their own and the model's predictions for each crop on a 5-point Likert scale with the prompt *"How much confidence do you have in the AI system's (your) predictions for optimal [crop name] irrigation?"*.

task, whereas for the easy task, many humans achieve at least comparable accuracy. As expected. humans make better predictions for the complex task in all **Subst_** conditions and better predictions for the easy task in all **Compl_** conditions. This confirms the success of our intervention. The model has little complementary expertise in **Subst_** but is still useful in the easy task domain. The model is highly complementary in **Compl_**, and most humans should only use it for the complex task.

In line with prediction performance, subjects state much higher confidence in their Meemmaseed than their Vussanut predictions (Figure 4). To illustrate, whereas only 5 % have "no" confidence in the human expertise domain, 25% have no confidence in the complementary expertise domain. Similarly, 40% have either "a

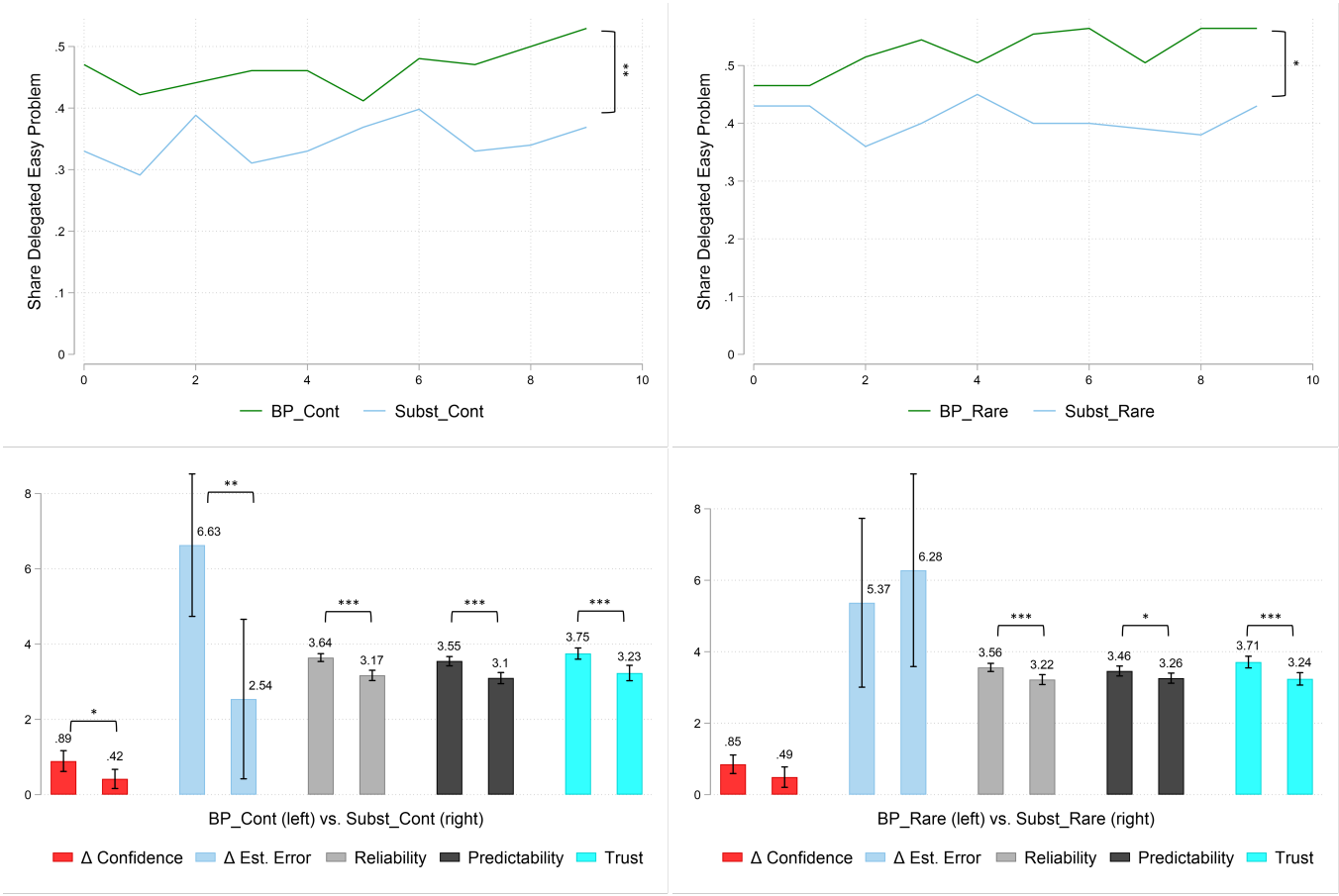

Figure 5: Top: Subject delegation shares for the *easy* problem with *continuous* environmental randomness and *rare* environmental randomness. Bottom: Corresponding treatment differences in relative subject confidence, estimated prediction accuracy, perceived model reliability, predictability, and trust. Subjects (1) state their confidence in their own and the model's predictions for each crop on a 5-point Likert scale with the prompt *"How much confidence do you have in the AI system's (your) predictions for optimal [crop name] irrigation?"* and (2) judge the accuracy of themselves and the AI system by answering the prompt *"How many units do you think the AI system's (your) predictions are off by for [crop name] on average? [Please enter a number 0 – 100)]"*. Differences are calculated by subtracting self-confidence (self-assessment) from model confidence (model assessment). Reliability, Predictability, and Trust are measured using the trust in automation questionnaire.

fair amount" or "a lot of" confidence in their own Meemmaseed predictions, as compared to 13% for Vussanut. Overall, subjects make much better predictions in the human expertise domain and have a lot more confidence in themselves.

4.2 Choice Independence

If choice independence holds, there are no differences in subject delegation for the easy task (Meemmaseed, human expertise domain) between **BP_Cont** vs. **Subst_Cont** as well as **BP_Rare** vs. **Subst_Rare**. For the complex task (Vussanut, complementary expertise domain), there should be no differences between **BP_Cont** vs. **Compl_Cont** and **BP_Rare** vs. **Compl_Rare**.

4.2.1 Meemmaseed Easy Task. Figure 5 depicts delegation shares over the 10 official predictions for the easy task. The data show a significant violation of choice independence (see Tables 4 and 5). Irrespective of the environmental error type, subjects delegate the easy prediction more often to the best-possible model when the AI system also makes the best-possible prediction for the complex prediction. On average, subjects delegate 46% (52) in **BP_Cont** (**BP_Rare**) and 35% (41) in **Subst_Cont** (**Subst_Rare**). The differences are significant both in the panel regressions and using a t-test on average delegation shares (*Cont*: t = 2.13, p = 0.034; *Rare*: t = 2.09, p = 0.038). In line with these results, Figure 5 (bottom panel) shows that the bad performance of the AI system in the complex task domain significantly alters subject perceptions. Note that the answers to the trust in automation questionnaire [63] refer to the AI system in general and not to one specific prediction problem. The questions regarding subjects' confidence in the model and themselves, as well as their estimation of their and the

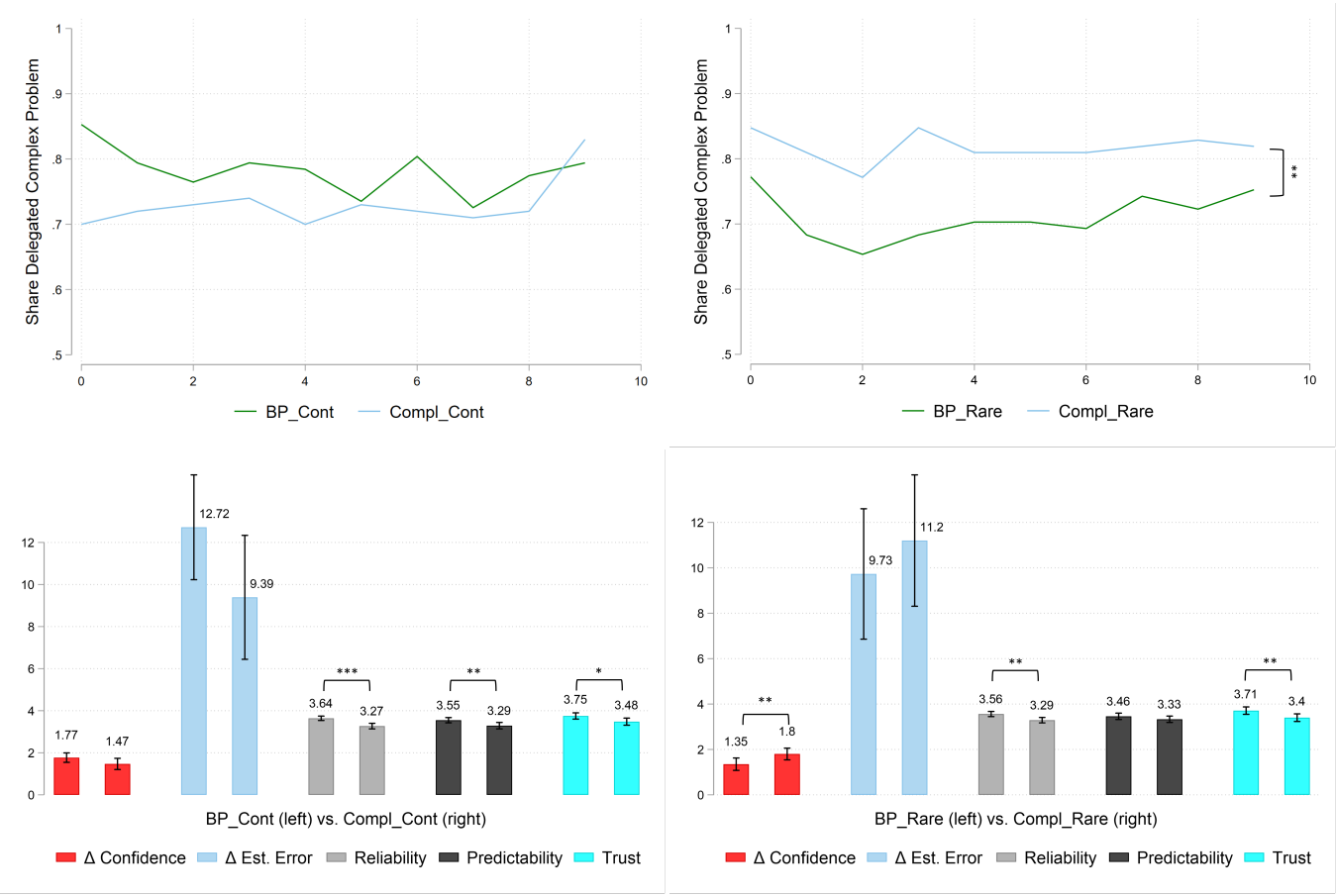

Figure 6: Top: Subject delegation shares for the *complex* problem with *continuous* environmental randomness and *rare* environmental randomness. Bottom: Corresponding treatment differences in subject confidence, estimated prediction accuracy, perceived model reliability, predictability, and trust. Subjects (1) state their confidence in their own and the model's predictions for each crop on a 5-point Likert scale with the prompt *"How much confidence do you have in the AI system's (your) predictions for optimal [crop name] irrigation?"* and (2) judge the accuracy of themselves and the AI system by answering the prompt *"How many units do you think the AI system's (your) predictions are off by for [crop name] on average? [Please enter a number 0 – 100)]"*. Differences are calculated by subtracting self-confidence (self-assessment) from model confidence (model assessment). Reliability, Predictability, and Trust are measured using the trust in automation questionnaire.

model's accuracy, differentiate between Meemaseed and Vussanut. In the *continuous* random error environment, subjects have more confidence in the AI system's Meemmaseed predictions when it also makes the best-possible prediction for Vussanut, estimate a stronger accuracy advantage compared to themselves, and find it overall more reliable, predictable, and trustworthy. Interestingly, when environmental randomness is more erratic, bad performances for the second task do not significantly alter confidence and accuracy estimates. Therefore, rare but high variance randomness may improve peoples' ability to infer accurate performance estimates. Subjects again find the AI system in the **BP_** condition more reliable and trustworthy, replicating the general effect on model perceptions. Thus, subjects in the **_Rare** condition can relatively accurately infer the performance advantage of the AI system for Meemmaseed irrespective of the model's performance in the second task, i.e., choice independent, but still violate choice independence when it comes to actual delegation behavior.

> **Result 1:** Humans violate choice independence in human-AI collaboration when delegating to a substitute-model. Systematic AI errors in the complementary expertise domain reduce delegation to the best-possible model in the human expertise domain.

4.2.2 Vussanut Complex Task. Figure 6 depicts delegation shares over the 10 official predictions for the complex task (Vussanut). We compare treatments **BP_**, in which the AI system makes the best-possible prediction for both tasks, and **Cont_**, in which the AI system makes the best-possible prediction only for the complex task

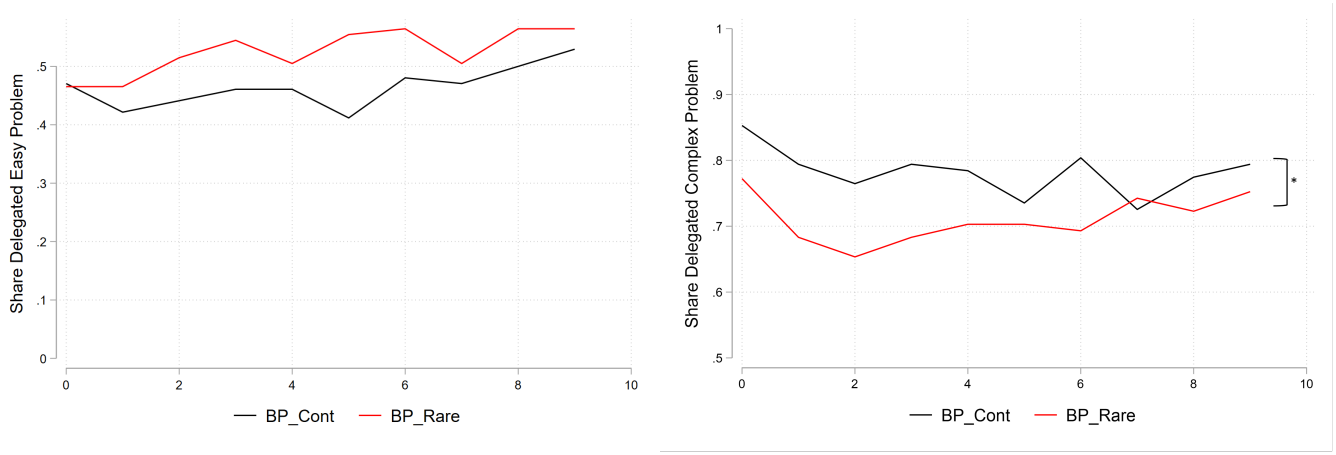

Figure 7: Subject delegation shares to the best-possible AI system for Meemmaseed (*easy*) and Vussanut (*complex*).

and is therefore highly complementary. The results are noticeably different from those above. For the environment with continuous randomness, there is no significant difference in delegation (Table 4). Subjects delegate 78% in **BP_Cont** and 73% in **Compl_Cont** (t = 1.23, p = 0.22). The direction is qualitatively the same as before, in that subjects delegate less when the model has systematic errors for the easy task. Still, the overall difference is smaller and less consistent. Under rare but more impactful randomness, we again document a significant violation of choice independence (Table 5). However, the effect is *reversed* compared to the **Subst_** conditions where the AI system functions as a substitute rather than a complement. Now, subjects delegate the complex task *more* when the AI system makes systematic errors for the easy task (**Compl_Rare:** 82% vs. **BP_Rare:** 71%, t = -2.45, p = 0.015). This surprising and, to us, unexpected result also reflects itself in subject perceptions. In the **_Rare** conditions, subjects state more confidence in the AI system's predictions for Vussanut when it makes systematic errors in the Meemmaseed predictions. Yet, generalized attitudes towards the AI system align with the other scenarios, and the best-possible model garners higher scores for trust and reliability. Thus, participants override their general feelings about the AI system in favor of a task-based approach.

Result 2: Humans violate choice independence in human-AI collaboration when delegating to a complementary model. Systematic AI errors in the human expertise domain increase delegation to the best-possible model in the AI expertise domain. This effect only holds for moderate and continuous, but not large and rare, systematic AI errors.

4.3 Error Type and Algorithm Aversion

The section on choice independence illustrates that human delegation can vary across different environmental and AI error types. This section analyzes the effect of error type on algorithm aversion toward the best-possible AI system.

We compare subject behavior in **BP_Cont** and **BP_Rare**. Here, the AI system makes the best-possible prediction for both tasks, and almost every human should always delegate the prediction to the model. Figure 7 shows delegation shares for the easy and the complex task.

Delegation shares do not differ significantly between treatments in the baseline regressions (Table 6). The same is true for all model perceptions, except the subject's relative confidence levels in the AI's predictions for Vussanut. Compared to themselves, **BP_Cont** subjects have significantly more confidence in the model's Vussanut predictions than those in **BP_Rare**. This effect is driven by lower self-confidence under continuous randomness than rare high-variance randomness (**_Cont:** 2.08 vs. **_Rare:** 2.42, t = 2.43, p = 0.016), despite larger human prediction errors in **BP_Rare** (13.8) than **BP_Cont** (12). Intuitively, continuous randomness impairs the human ability to form useful heuristics (or the perception of) due to high levels of noise, whereas rare randomness allows for a relatively large number (80% in our case) of noise-free observations. This is not consequential for the easy problem because subjects already have a heuristic, i.e., always focus on the third input number and ignore the second. In line with that, we again see a reversal in delegation shares between the two problem types. For Meemmaseed, subjects tend to delegate more with rare environmental errors. For Vussanut, subjects tend to delegate more with continuous errors. While the simple regression model does not show a treatment effect for either problem, incorporating risk attitudes and numeracy reveals a significant effect for Vussanut but not Meemmaseed (Table 6). Hence, there is moderate evidence that error type does play a role in algorithm aversion, but only for complex problems where subjects need to learn and build up their own decision rules.

Result 3: Algorithm aversion does not generally depend on whether the model makes small and continuous or large but rare mistakes. However, there is evidence that continuous randomness can reduce algorithm aversion outside the human expertise domain through lower self-confidence.

4.3.1 Systematic AI Errors. We provide some auxiliary results to parse out two particular effects of systematic AI errors on human behavior. One, how does human delegation change with the introduction of relatively large systematic errors that lead to model predictions that are, on average, worse than human predictions? Two, do humans differentiate between continuous but moderate and rarer but larger systematic errors? The full analysis is detailed in the Appendix (see section 7). Here, we only present the main conclusions.

Our data shows that participants react to the introduction of a systematic error by correcting their delegation behavior downwards. In the human expertise domain, this effect is confined to forecasters who exceed the AI system's performance. Only 10% – 20% rely on the AI system. Those who perform worse still delegate around 55% of tasks to the model. In contrast, when humans are not endowed with a useful decision rule, systematic AI errors lead to substantially less delegation across all participants, irrespective of their own performance level. Hence, a lack of expertise inhibits peoples' ability to judge their own performance level against the AI system properly and therefore limits meta-knowledge [39]. Finally, there is moderate evidence that participants punish continuous but moderate systematic AI errors stronger than rare and large errors in their own expertise domain.

5 DISCUSSION

This paper is the first to systematically analyze how differences in AI performance across distinct prediction tasks influence human utilization of superior AI models. As AI systems are increasingly capable of complementing or supporting human expertise, it is essential to understand which factors may drive or inhibit their adoption. This process is complicated by the fact that many systems do not simply occupy one very specific role but instead provide predictions for a variety of different questions or problems. One relevant example is recommender and expert systems. Spotify, for instance, recommends its diverse and heterogeneous set of users music from different genres and time periods, as well as podcasts and shows. Some customers may be very good at finding new music from their favorite genres on their own but struggle with unfamiliar genres or novel podcasts. Others may know exactly which kind of podcast they like but have not yet developed a good sense of their musical taste. Experts like financial advisors, lawyers, or physicians are often highly specialized and may, therefore, in theory, benefit in particular from systems that complement their expertise. However, in almost all cases, expert systems have a large overlap with the experts they advise. This allows for comparisons, and interdependencies between different kinds of AI predictions may arise. If, for instance, financial advisors refuse trading advice because their algorithm performs relatively worse in capital investing, or cardiologists forego AI heart attack diagnoses because the system may err when identifying myocarditis, there could be a number of inefficiencies that all relevant actors, including not only the experts but regulators and developers, should be aware of.

RQ1: Does the independence axiom of choice hold for delegation decisions in human-AI collaboration?

The independence axiom of choice does not hold for delegation decisions in human-AI collaboration. Systematic AI prediction errors in the complementary domain significantly reduce delegation to the superior best-possible model in the human expertise domain. Systematic AI prediction errors in the human expertise domain significantly increase delegation to the superior best-possible model in the complementary domain, but only as long as environmental randomness allows for a large number of perfect complementary AI predictions. Humans, therefore, do not judge AI predictions task-independently.

RQ2: Do humans condition their delegation choices on objective performance differences between prediction problems?

Beyond a violation of choice independence, we are able to make some more general inferences about human delegation to superior AI systems. When humans have some expertise in a prediction domain, their behavior is sensitive to the model's relative performance advantage. Systematic errors strongly decrease delegation, but only for those who benefit from their own predictions. This illustrates a general ability to properly assess their own accuracy in relation to the AI system. Still, many subjects fail to delegate when appropriate, and the level of algorithm aversion is high.

In the complementary domain without any real human expertise, delegation adjustments are less optimal. Instead of assessing their own accuracy in relation to the AI system's performance, subjects respond to the introduction of systematic errors with a general decline in delegation, irrespective of their ability. This speaks to a lack of meta knowledge as discussed, e.g., in Fügener et al. [39].

RQ 3: How do different prediction error types influence human reliance on a more accurate AI system?

The second important part of this paper investigates whether humans react differently towards two kinds of errors: continuous but moderate and rare but large prediction inaccuracies. Our results show that humans are less likely to delegate their complex predictions to the best-possible AI system when it makes rare but relatively large errors due to randomness. This, however, does not seem to be driven by lower confidence in the model but higher human self-confidence. Thus, while e.g., Dietvorst and Bharti [29] show that people forego algorithms because they prefer (the possibility of) perfect predictions, in our case, more perfect AI predictions leads to less delegation because less frequent environmental noise increases human forecasters' confidence in their own performance. Here, one significant takeaway is the importance of differentiating between systematic prediction errors and random prediction errors that always befall all forecasting agents. Because randomness affects both the model and the human, the two prediction agents may have conflicting behavioral effects. Furthermore, our results also suggest that human forecasters do not differentiate between the two environmental error types in their own expertise domain, possibly due to – as mentioned above – better meta-knowledge.

5.1 Practical Implications

Understanding how human decision makers react to imperfect models is essential for applying and deploying current AI systems. One straightforward implication of our results is that optimizing or fine-tuning models for better performances in domains where human

decision makers are relatively inaccurate is not a neutral process. In a world where choice independence holds, developers can largely ignore AI errors for tasks performed by a human, thereby maximizing joint human-AI performance through specialization. Instead, our findings suggest that humans cognitively bracket the AI system's performance across different tasks, translating into changes in attitudes and delegation behavior.

These changes appear to be conditional on the AI's application domain. If the human decision maker should delegate a task in their own expertise domain to a superior system, then observing objectively unrelated AI errors for another task reduces appropriate reliance. However, if humans delegate a task for which they have close to zero expertise, then unrelated AI errors can increase appropriate reliance. This observation implies that different kinds of human decision makers, e.g., experts and laypeople, may react differently to the same AI error and that the design, optimization, and deployment of AI systems should be explicitly stakeholder-driven.

The latter conclusion is also apparent in the context of AI error types. Reactions thoroughly depend on the considered expertise domain and are often reversed. First, our results suggest that baseline randomness matters and appropriate delegation can be lower for tasks where "correct" predictions are likely and possible. This may, for instance, include medical self-diagnoses by laypeople. Randomness often plays little to no role in mainstream diagnoses, which allows for (1) perfect learning observations and (2) perfect predictions. Such a pattern may be relevant for regulators, but also, e.g., in the design of user apps for health services. Second, people who share expertise with the AI system — e.g., almost every expert, such as physicians, financial advisors, or lawyers — react more strongly to moderate and continuous model mistakes. Developers optimizing or fine-tuning applied AI systems may want to consider that to maximize appropriate uptake.

Finally, our results point to a potential benefit of further education for users who regularly confront heterogeneous AI output. For instance, there is good evidence that market experience and domain knowledge can correlate with higher rationality [18, 19, 35, 61, 77, 97], including specifically reductions in violation of choice independence [68]. This suggests that people learn to adjust their behavior autonomously through feedback, which may be provided via additional training.

5.2 Caveats and Limitations

Experimental abstraction. One goal of this paper is to empirically test the assumption of choice independence in human-AI collaboration. We use an abstract forecasting task that gives us control over each model's output and what specifically human forecasters observe. In reality, many contextual factors determine how performance differences across tasks determine human behavior. We abstract from almost all of those, and a valuable direction for future research would be to apply the logic of choice independence to problems that consider commonly used AI systems and models. This includes not only the problem domain but many procedural and environmental factors. For example, human decision makers in our experiment make simultaneous predictions and then simultaneously choose between themselves and the AI system for both problems. A more realistic scenario may include sequential decision tasks or time delays. Moreover, our results are restricted to prediction domains under uncertainty. While these are highly relevant, they are not the only field of application for modern AI systems, and specifically, the degree of uncertainty and risk involved could have large consequences for human behavior.

Artificial expertise. In our experiments, we differentiate between a human expertise domain and a complementary expertise domain. However, expertise is induced artificially through the provision of a decision heuristic. It would be interesting to compare such a scenario with real experts with more entrenched, far-reaching expertise and, thus, presumably, a higher awareness of their strengths and weaknesses. We also do not test the validity of choice independence *within* an expertise domain. Our setup assumes that humans face problems outside their field of expertise and, therefore, always judge the AI system on two different levels with two different reference points. In many situations, this may not be valid. However, we argue that most professionals will experience these situations, even if only because of a novel problem or case for which they have not yet accrued the relevant experience or information.

Error type specificities. We measure choice independence by introducing a systematic AI system error to the second, objectively unrelated problem. This systematic error always differs in type from the baseline error induced by environmental randomness. In that sense, we introduce a second type of error. Therefore, we cannot guarantee that any error induces a violation of choice independence. Here, we see a lot of room for future research to experiment with different types of AI system errors and to gauge how these error types influence human behavior.

Due to experimental restrictions, we rely on a "rare" systematic error that happens 50% of the time. A more pronounced difference to the continuous error in frequency and intensity may produce stronger differences in human behavior. Similarly, having a human expertise domain where a sizeable share of humans actually outperforms the AI system or choosing a more ambiguous systematic error could affect our results. For instance, one may argue that AI systems with large systematic errors will always be judged as not market-worthy and, therefore, never be deployed until a certain performance benchmark has been reached. This results in smaller inaccuracies, which may not induce violations of choice independence. One counter-argument would be the recent deployment of ChatGPT – an AI system accessible to almost anyone and simultaneously very inaccurate in certain domains. Still, the question of how substantial or salient AI errors have to be so that they reduce human utilization in an unrelated problem is a very relevant one.

Task similarity. Finally, reactions to objectively unrelated prediction errors could be mediated by the perceived similarities of the different tasks. In this experiment, subjects observe an AI system's performance across two tasks with a very similar dependent variable: the irrigation need of a crop. Our data shows that subjects differentiate between the two tasks and can over-write their general attitudes towards the AI system in favor of a task-based evaluation approach. This means that even when subjects trust the AI system less overall, they may have more confidence in its predictions for a particular task. Still, such an approach may make it easier for humans to cognitively conflate the AI system's performance across tasks. For example, some participants may have constructed a simple evaluation heuristic that estimates the model's

ability to accurately predict irrigation needs – independent of the target outcome. While this is still in violation of the IA and thus does not contradict our interpretation, it is a potential limitation. Some real-life instances, like Spotify's recommendation algorithm for songs, artists, and podcasts, or certain diagnostic models may allow for similar heuristics.[6] Others, however, will be less comparable, such as self-driving cars, weather apps, or physicians that utilize models across more dissimilar domains, e.g., image classification and mental health diagnoses. We, therefore, highlight the potential mediating role of task similarity for choice independence as an important avenue for future research.

6 CONCLUSIONS

This article analyzes appropriate reliance in human-AI collaboration when decision-makers face multiple tasks. Using two different error types, our experimental design systematically varies the AI system's performance across a human and a complementary expertise domain. We are the first to show that human forecasters consistently violate choice independence by taking the AI's performance in an unrelated second task into account. As a consequence, subjects reduce delegation to the superior best-possible system in their own expertise domain. Interestingly, subjects react to systematic AI errors in the human expertise domain by increasing appropriate reliance on the complementary AI expertise domain. Furthermore, our results suggest that human rejection of superior algorithms is sensitive to the forecasting environment's error type and that humans tend to punish continuous AI errors stronger than large but rare ones. These results enhance our theoretical understanding of human-AI collaboration by considering previously unexplored interdependencies. They also highlight the importance of stakeholders and user expertise for algorithmic design and AI adoption. In particular, human experts with domain-specific knowledge might be especially likely to forego useful systems due to biased evaluations.

ACKNOWLEDGMENTS

We would like to thank the anonymous participants in our study and the reviewers who provided valuable feedback. This work was supported by the Lower Saxony Ministry of Science and Culture under grant number ZN3492 within the Lower Saxony "Vorab" of the Volkswagen Foundation, the Center for Digital Innovations (ZDIN), the TU Delft Design@Scale AI lab within the TU Delft AI Initiative, and the NWO ROBUST-LTP G.E.N.I.U.S. project.

REFERENCES

[1] Daron Acemoglu and Pascual Restrepo. 2018. Artificial intelligence, automation, and work. In *The economics of artificial intelligence: An agenda*. University of Chicago Press, 197–236.

[2] Daron Acemoglu and Pascual Restrepo. 2018. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American economic review* 108, 6 (2018), 1488–1542.

[3] Daron Acemoglu and Pascual Restrepo. 2019. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives* 33, 2 (2019), 3–30.

[6] For example, a model may be reliable in diagnosing infectious or physiological diseases but markedly worse at hereditary or deficiency diseases because of limited data. A similar problem may arise through heterogeneous data availability across different population groups. A related false inference would then be: "the model is bad at diagnosing diseases."

[4] Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. 2019. Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives* 33, 2 (2019), 31–50.

[5] Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access* 6 (2018), 14410–14430.

[6] Maurice Allais. 1953. Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: journal of the Econometric Society* (1953), 503–546.

[7] Victoria Angelova, Will S Dobbie, and Crystal Yang. 2023. *Algorithmic recommendations and human discretion.* Technical Report. National Bureau of Economic Research.

[8] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* (2023).

[9] Aaron Baird and Likoebe M Maruping. 2021. The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS quarterly* 45, 1 (2021).

[10] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[11] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[12] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. 2021. Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering* 63, 1 (2021), 55–68.

[13] Pavlo Blavatskyy, Andreas Ortmann, and Valentyn Panchenko. 2022. On the experimental robustness of the Allais paradox. *American Economic Journal: Microeconomics* 14, 1 (2022), 143–163.

[14] Silvia Bonaccio and Reeshad S Dalal. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes* 101, 2 (2006), 127–151.

[15] Erik Brynjolfsson, Tom Mitchell, and Daniel Rock. 2018. What can machines learn and what does it mean for occupations and the economy?. In *AEA papers and proceedings*, Vol. 108. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 43–47.

[16] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making* 33, 2 (2020), 220–239.

[17] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[18] Colin Camerer. 1992. The rationality of prices and volume in experimental markets. *Organizational Behavior and Human Decision Processes* 51, 2 (1992), 237–272.

[19] Colin F Camerer. 1987. Do biases in probability judgment matter in markets? Experimental evidence. *The American Economic Review* 77, 5 (1987), 981–997.

[20] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.

[21] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[22] Lingwei Cheng and Alexandra Chouldechova. 2023. Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–27.

[23] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248* (2020).

[24] Jeremy Clark and Lana Friesen. 2009. Overconfidence in forecasts of own performance: An experimental study. *The Economic Journal* 119, 534 (2009), 229–251.

[25] Melanie Clegg, Reto Hofstetter, Emanuel de Bellis, and Bernd H Schmitt. 2023. Unveiling the Mind of the Machine. *Available at SSRN 4564832* (2023).

[26] John Conlisk. 1989. Three variants on the Allais example. *The American Economic Review* (1989), 392–407.

[27] Marie-Pierre Dargnies, Rustamdjan Hakimov, and Dorothea Kübler. 2022. Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. (2022).

[28] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing*

Systems. 1–12.

[29] Berkeley J Dietvorst and Soaham Bharti. 2020. People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science* 31, 10 (2020), 1302–1314.

[30] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[31] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.

[32] Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G Wagner. 2011. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the european economic association* 9, 3 (2011), 522–550.

[33] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with AI systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[34] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 43–52.

[35] Dorla A Evans. 1997. The role of markets in reducing expected utility violations. *Journal of Political Economy* 105, 3 (1997), 622–636.

[36] Angela Fagerlin, Brian J Zikmund-Fisher, Peter A Ubel, Aleksandra Jankovic, Holly A Derry, and Dylan M Smith. 2007. Measuring numeracy without a math test: development of the Subjective Numeracy Scale. *Medical Decision Making* 27, 5 (2007), 672–680.

[37] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.

[38] Shane Frederick, Leonard Lee, and Ernest Baskin. 2014. The limits of attraction. *Journal of Marketing Research* 51, 4 (2014), 487–507.

[39] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2022. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research* 33, 2 (2022), 678–696.

[40] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. 2021. Human-AI collaboration with bandit feedback. *arXiv preprint arXiv:2105.10614* (2021).

[41] C Garner. 1982. Experimental evidence on the rationality of intuitive forecasters. *Research in experimental economics* 2 (1982), 113–128.

[42] Yashesh Gaur, Walter S Lasecki, Florian Metze, and Jeffrey P Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*. 1–8.

[43] Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Ewart de Visser, Poornima Madhavan, Gopikrishna Deshpande, and Frank Krueger. 2017. An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social neuroscience* 12, 5 (2017), 570–581.

[44] David W Harless and Colin F Camerer. 1994. The predictive utility of generalized expected utility theories. *Econometrica: Journal of the Econometric Society* (1994), 1251–1289.

[45] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[46] Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*.

[47] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[48] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2022. On the effect of information asymmetry in human-AI teams. *arXiv preprint arXiv:2205.01467* (2022).

[49] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 453–463.

[50] John D Hey and Chris Orme. 1994. Investigating generalizations of expected utility theory using experimental data. *Econometrica: Journal of the Econometric Society* (1994), 1291–1326.

[51] Yoyo Tsung-Yu Hou and Malte F Jung. 2021. Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[52] Joel Huber, John W Payne, and Christopher Puto. 1982. Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of consumer research* 9, 1 (1982), 90–98.

[53] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, Libuše Hannah Vepřek, and Gabrielle Quinn. 2022. Advancing Human-AI Complementarity: The Impact of User Expertise and Algorithmic Tuning on Joint Decision Making. *ACM Transactions on Computer-Human Interaction* (2022).

[54] Damon Jones, David Molitor, and Julian Reif. 2019. What do workplace wellness programs do? Evidence from the Illinois workplace wellness study. *The Quarterly Journal of Economics* 134, 4 (2019), 1747–1791.

[55] S Mo Jones-Jang and Yong Jin Park. 2023. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication* 28, 1 (2023), zmac029.

[56] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. (2020).

[57] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–292.

[58] Konstantinos V Katsikopoulos and Gerd Gigerenzer. 2008. One-reason decision-making: Modeling violations of expected utility theory. *Journal of Risk and Uncertainty* 37 (2008), 35–56.

[59] Esther Kaufmann, Alvaro Chacon, Edgar E Kausel, Nicolas Herrera, and Tomas Reyes. 2023. Task-specific algorithm advice acceptance: A review and directions for future research. *Data and Information Management* (2023), 100040.

[60] Antino Kim, Mochen Yang, and Jingjing Zhang. 2023. When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Transactions on Computer-Human Interaction* 30, 1 (2023), 1–36.

[61] Peter Knez, Vernon L Smith, and Arlington W Williams. 1985. Individual rationality, market rationality, and value estimation. *The American Economic Review* 75, 2 (1985), 397–402.

[62] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[63] Moritz Körber. 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*. Springer, 13–30.

[64] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[65] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[66] Johannes Leitner and Ulrike Leopold-Wildburger. 2011. Experiments on forecasting behavior with several sources of information–A review of the literature. *European Journal of Operational Research* 213, 3 (2011), 459–469.

[67] Sohye Lim and Byron Reeves. 2010. Computer agents versus avatars: Responses to interactive game characters controlled by a computer or other player. *International Journal of Human-Computer Studies* 68, 1-2 (2010), 57–68.

[68] John A List and Michael S Haigh. 2005. A simple test of expected utility theory using professional traders. *Proceedings of the National Academy of Sciences* 102, 3 (2005), 945–948.

[69] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[70] Graham Loomes and Robert Sugden. 1987. Testing for regret and disappointment in choice under uncertainty. *The Economic Journal* 97, Supplement (1987), 118–129.

[71] Brian Lubars and Chenhao Tan. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *Advances in Neural Information Processing Systems* 32 (2019).

[72] R Duncan Luce and Detlof Von Winterfeldt. 1994. What common ground exists for descriptive, prescriptive, and normative utility theories? *Management Science* 40, 2 (1994), 263–279.

[73] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[74] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems* 31 (2018).

[75] Laureen A Maines. 1996. An experimental examination of subjective forecast combination. *International Journal of Forecasting* 12, 2 (1996), 223–233.

[76] Jacob Marschak. 1950. Rational behavior, uncertain prospects, and measurable utility. *Econometrica: Journal of the Econometric Society* (1950), 111–141.
[77] Kevin A McCabe and Vernon L Smith. 2000. A comparison of naive and sophisticated subject behavior with game theoretic predictions. *Proceedings of the National Academy of Sciences* 97, 7 (2000), 3777–3781.
[78] Carey K Morewedge. 2022. Preference for human, not algorithm aversion. *Trends in Cognitive Sciences* (2022).
[79] Ivan Moscati. 2016. Retrospectives: how economists came to accept expected utility theory: the case of samuelson and savage. *Journal of economic perspectives* 30, 2 (2016), 219–236.
[80] Sendhil Mullainathan and Ziad Obermeyer. 2022. Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics* 137, 2 (2022), 679–727.
[81] William Neilson and Jill Stowe. 2002. A further examination of cumulative prospect theory parameterizations. *Journal of risk and uncertainty* 24 (2002), 31–46.
[82] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.
[83] Dilek Önkal, Paul Goodwin, Mary Thomson, Sinan Gönül, and Andrew Pollock. 2009. The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22, 4 (2009), 390–409.
[84] Marc Pinski, Martin Adam, and Alexander Benlian. 2023. AI Knowledge: Improving AI Delegation through Human Enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
[85] Andrew Prahl and Lyn Van Swol. 2017. Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* 36, 6 (2017), 691–702.
[86] Till Proeger and Lukas Meub. 2014. Overconfidence as a social bias: Experimental evidence. *Economics Letters* 122, 2 (2014), 203–207.
[87] Charvi Rastogi, Liu Leqi, Kenneth Holstein, and Hoda Heidari. 2022. A unifying framework for combining complementary strengths of humans and ML toward better predictive decision-making. *arXiv preprint arXiv:2204.10806* (2022).
[88] Srinivasan Ratneshwar, Allan D Shocker, and David W Stewart. 1987. Toward understanding the attraction effect: The implications of product stimulus meaningfulness and familiarity. *Journal of Consumer Research* 13, 4 (1987), 520–533.
[89] Taly Reich, Alex Kaju, and Sam J Maglio. 2023. How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology* 33, 2 (2023), 285–302.
[90] Laetitia A Renier, Marianne Schmid Mast, and Anely Bekbergenova. 2021. To err is human, not algorithmic–Robust reactions to erring algorithms. *Computers in Human Behavior* 124 (2021), 106879.
[91] Kai Ruggeri, Sonia Alí, Mari Louise Berge, Giulia Bertoldo, Ludvig D Bjørndal, Anna Cortijos-Bernabeu, Clair Davison, Emir Demić, Celia Esteban-Serna, Maja Friedemann, et al. 2020. Replicating patterns of prospect theory for decision under risk. *Nature human behaviour* 4, 6 (2020), 622–633.
[92] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 215–227.
[93] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.
[94] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916* (2022).
[95] Eldar Shafir, Itamar Simonson, and Amos Tversky. 1993. Reason-based choice. *Cognition* 49, 1-2 (1993), 11–36.
[96] Ben Shneiderman. 2022. *Human-centered AI.* Oxford University Press.
[97] Vernon L Smith. 1985. Experimental economics: reply. *The American Economic Review* 75, 1 (1985), 265–272.
[98] Jack B Soll, Asa B Palley, and Christina A Rader. 2022. The bad thing about good advice: Understanding when and how advice exacerbates overconfidence. *Management Science* 68, 4 (2022), 2949–2969.
[99] Lea Strohm, Charisma Hehakaya, Erik R Ranschaert, Wouter PC Boon, and Ellen HM Moors. 2020. Implementation of artificial intelligence (AI) applications in radiology: hindering and facilitating factors. *European radiology* 30 (2020), 5525–5532.
[100] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*. 77–87.
[101] John Von Neumann and Oskar Morgenstern. 2007. *Theory of games and economic behavior (60th Anniversary Commemorative Edition)*. Princeton university press.
[102] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. 2016. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718* (2016).
[103] Guangyu Wang, Xiaohong Liu, Jun Shen, Chengdi Wang, Zhihuan Li, Linsen Ye, Xingwang Wu, Ting Chen, Kai Wang, Xuan Zhang, et al. 2021. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nature biomedical engineering* 5, 6 (2021), 509–521.
[104] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).
[105] Sybil Yang and Michael Lynn. 2014. More evidence challenging the robustness and usefulness of the attraction effect. *Journal of Marketing Research* 51, 4 (2014), 508–513.
[106] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421* 9 (2023).
[107] Mireia Yurrita, Agathe Balayn, and Ujwal Gadiraju. 2023. Generating Process-Centric Explanations to Enable Contestability in Algorithmic Decision-Making: Challenges and Opportunities. *arXiv preprint arXiv:2305.00739* (2023).
[108] Fabio Massimo Zanzotto. 2019. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.
[109] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–28.

7 APPENDIX – AUXILIARY RESULTS AND REGRESSION TABLES

Beyond prediction inaccuracies induced by randomness, AI systems may have systematic errors. That is, many AI systems do not deliver the best-possible prediction but are confounded in some way, e.g., due to data constraints. We now look at (1) how human delegation changes by introducing relatively large systematic errors that lead to, on average, model predictions that are worse than human predictions and (2) whether humans differentiate between continuous but moderate and rarer but larger systematic errors. Figure 8 compares delegation shares for the two problems with and without a systematic error, where the AI system always makes the best-possible prediction for the second unrelated problem. This avoids confounding through violations of choice independence.

Subjects react to the introduction of a systematic error by significantly decreasing delegation (Tables 4 and 5. The effect is smaller in the human expertise domain, primarily due to lower baseline delegation. Here, the vast majority of participants perform worse in the **BP_** conditions (95% and 100% respectively), and 48 – 52% of (easy) problems are delegated to the AI system. With the introduction of a systematic error, only 22% in **Compl_Cont** and **Compl_Rare** have a larger average prediction error than the model. That sub-population delegates 55% of their official predictions to the AI system. In contrast, those with higher accuracy on average delegate only 20% (10%) in **Compl_Cont** (**Compl_Rare**). Behavioral patterns in the AI-expertise domain for the complex problem are similar, but not the same. In the **BP_** conditions, no human on average beats the best-possible model, and the vast majority of problems (78% and 71%) are delegated. Introducing a systematic error in **Subst_Cont** and **Subst_Rare** enables 74% and 77% of humans respectively to make more accurate predictions. Those have, again, relatively low delegation rates of 30% in **Subst_Cont** and 25% in **Subst_Rare**. However, in contrast to the human expertise domain, subjects who perform the complex predictions worse than the systematically erring AI system are significantly less likely to delegate a prediction to the flawed model (**BP_Cont:** 78% vs. **Subst_Cont:**

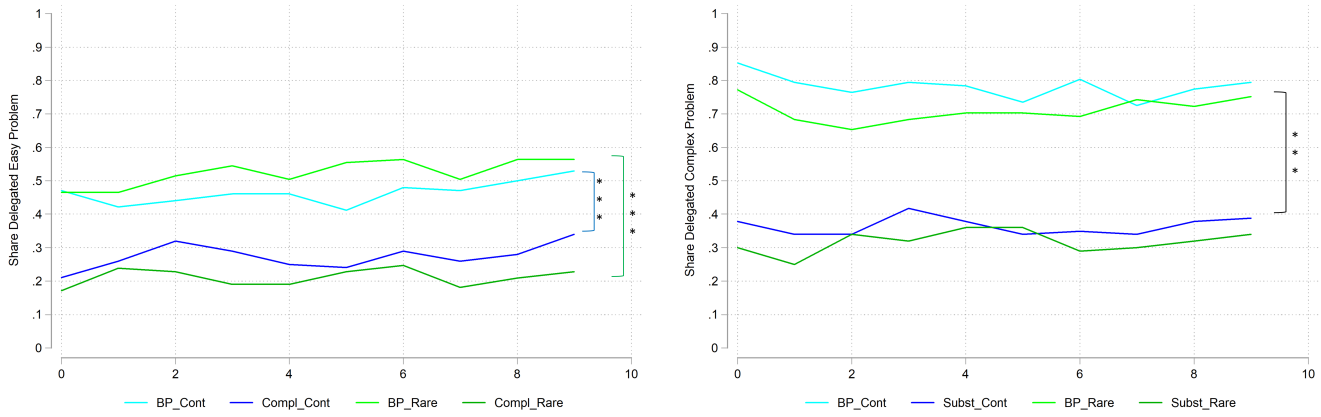

Figure 8: Left: Subject delegation shares to the AI system in the human expertise domain for Meemmaseed (*easy*). Right: Subject delegation shares to the AI system in the AI expertise domain for Vussanut (*complex*). We compare delegation shares to the best-possible AI system with delegation to the AI system that has a systematic error for the problem of interest but is still the best-possible for the second problem.

56%, t = 3.32, p = 0.001; **BP_Rare:** 71% vs. **Subst_Rare:** 55%, t = 2.05, p = 0.04).

> **Result 4:** Subjects react to the introduction of a systematic error by correcting their delegation behavior downwards. In the human expertise domain, this effect is confined to those human forecasters who exceed the AI system's performance. In the complementary expertise domain where humans have no default decision rule, systematic errors exert negative externalities by also reducing the likelihood that bad human forecasters delegate predictions to the system.

Second, we look at the effect of different systematic errors on subject delegation. In **Subst_Cont** and **Compl_Cont**, humans observe a systematic error that is relatively rare (50%), but large (24 and 30 for Meemmaseed and Vussanut respectively). In **Subst_Rare** and **Compl_Rare**, the systematic error is continuously drawn from [10, 11, 12, 13, 14] and [13, 14, 15, 16, 17]. Therefore, the expected average systematic error is always either 12 or 15. To test for a differential impact of systematic error type on human delegation, we run logistic random effects panel regressions interacting a binary systematic error treatment variable with a binary systematic error type variable (see Table 6). This analysis reveals a significant and negative interaction effect of continuous systematic error type on delegation for easy predictions in the human expertise domain (Meemmaseed) but not for complex predictions (Vussanut). Hence, in our sample, subjects react more strongly to continuous and moderate than rare and large systematic errors in their own expertise domain but do not differentiate between them in the complementary expertise domain.

> **Result 5:** Subjects punish continuous but moderate systematic AI errors stronger than rare and large errors in their own expertise domain. For complex problems in the complementary expertise domain, there is no effect of systematic error type on delegation.

Table 4: This table reports marginal effects of panel logistic regressions using individual-level random effects and a cluster–robust VCE estimator. The dependent variable is a binary variable that equals 1 if the participant delegates to the AI system and 0 otherwise. P-values are adjusted by controlling for the family-wise error rate using Westfall and Young [54]. – * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Continuous Environmental Randomness							
	Choice Independence (1)		**Choice Independence (2)**		**Reaction Systematic Error (1)**		**Reaction Systematic Error (2)**	
	Easy Problem	Complex Problem	Easy Problem	Complex Problem	Easy Problem	Complex Problem	Easy Problem	Complex Problem
BP_Cont	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
Subst_Cont	-0.076**		-0.069*			-0.191***		-0.189***
	(0.021)		(0.023)			(0.015)		(0.015)
Compl_Cont		-0.056		-0.072	-0.200***		-0.184***	
		(0.039)		(0.040)	(0.04)		(0.041)	
SNS Ability			-0.005	0.032			-0.001	-0.008
			(0.026)	(0.018)			(0.02)	(0.019)
SNS Preferences			-0.035	0.012			-0.043	-0.026
			(0.031)	(0.022)			(0.028)	(0.023)
Risk			0.019*	-0.018*			0.030**	-0.004
			(0.009)	(0.009)			(0.009)	(0.008)
N	2050	2020	2050	2020	2020	2050	2020	2050

Table 5: Table reports marginal effects of panel logistic regressions using individual-level random effects and a cluster–robust VCE estimator. The dependent variable is a binary variable that equals 1 if the participant delegates to the AI system and 0 otherwise. P-values are adjusted by controlling for the family-wise error rate using Westfall and Young [54]. – * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Rare Environmental Randomness							
	Choice Independence (1)		**Choice Independence (2)**		**Reaction Systematic Error (1)**		**Reaction Systematic Error (2)**	
	Easy Problem	Complex Problem	Easy Problem	Complex Problem	Easy Problem	Complex Problem	Easy Problem	Complex Problem
BP_Rare	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline	Baseline
Subst_Rare	-0.056		-0.058*			-0.188***		-0.184***
	(0.026)		(0.024)			(0.014)		(0.015)
Compl_Rare		0.118*		0.125**	-0.268***		-0.281***	
		(0.037)		(0.036)	(0.04)		(0.037)	
SNS Ability			-0.011	0.017			-0.018	-0.000
			(0.024)	(0.018)			(0.022)	(0.021)
SNS Preferences			0.032	0.033			-0.009	-0.005
			(0.026)	(0.020)			(0.023)	(0.024)
Risk			-0.031**	-0.026***			0.010	-0.024*
			(0.010)	(0.008)			(0.007)	(0.009)
N	2010	2060	2010	2060	2060	2010	2060	2010

Table 6: Table reports marginal effects of panel logistic regressions using individual-level random effects and a cluster–robust VCE estimator. The dependent variable is a binary variable that equals 1 if the participant delegates to the AI system and 0 otherwise. "Systematic Error" is a dummy variable that equals 1 for each treatment where the AI system makes systematic errors. "_Rare" is a dummy variable that equals 1 for all treatments with a random environmental error and consequently a continuous systematic AI error. In the interaction model for the easy problem with the substitute model, we use treatments BP_Cont, Compl_Cont, BP_Rare and Compl_Rare. For the complex problem with the complementary model, we use BP_Cont, Subst_Cont, BP_Rare and Subst_Rare. – * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

	Continuous vs. Rare Environmental Randomness							
	Algorithm Aversion BP (1)		**Algorithm Aversion BP (2)**		**Δ Reaction Systematic Error (1)**		**Δ Reaction Systematic Error (2)**	
	Easy Problem	Complex Problem	Easy Problem	Complex Problem	Easy Problem	Complex Problem	Easy Problem	Complex Problem
BP_Cont	Baseline	Baseline	Baseline	Baseline				
BP_Rare	0.016	-0.024	0.015	-0.026*				
	(0.016)	(0.013)	(0.016)	(0.013)				
Systematic Error					-0.276***	-0.413***	-0.272***	-0.410***
					(0.027)	(0.031)	(0.028)	(0.031)
_Rare					-0.011	-0.062*	-0.008	-0.059*
					(0.027)	(0.031)	(0.028)	(0.030)
Systematic Error × _Rare					-0.066*	-0.052	-0.071*	-0.049
					(0.032)	(0.049)	(0.033)	(0.048)
SNS Ability			0.008	0.006			-0.016	-0.006
			(0.025)	(0.020)			(0.015)	(0.015)
SNS Preferences			0.013	0.019			-0.020	-0.017
			(0.028)	(0.020)			(0.018)	(0.015)
Risk			-0.009	-0.032***			0.008	-0.013*
			(0.011)	(0.009)			(0.006)	(0.006)
N	2030	2030	2030	2030	4080	4060	4080	4060