

Trust in the Machine: How Contextual Factors and Personality Traits Shape Algorithm Aversion and Collaboration

Vinicius Ferraz^{1*}, Leon Houf¹, Thomas Pitz²,
Christiane Schwier¹, Jörn Sickmann²

^{1*} Alfred-Weber Institute for Economics, Heidelberg University,
Bergheimer Str. 58, Heidelberg, 69115, Germany.

² Faculty of Society and Economics, Rhine-Waal University,
Marie-Curie-Str. 1, Kleve, 47533, Germany.

*Corresponding author(s). E-mail(s): visferraz@gmail.com;

Contributing authors: leon.houf@awi.uni-heidelberg.de;

thomas.pitz@hochschule-rhein-waal.de;

christiane.schwieren@awi.uni-heidelberg.de;

joern.sickmann@hochschule-rhein-waal.de;

Abstract

This paper investigates the interplay between contextual factors, personal variables, and algorithm aversion in decision delegation behavior. In an experimental setting with four treatments —baseline, explanation, payment, and automation— subjects chose whether to delegate decisions to an algorithm, considering hidden expected values. Employing Random Forests, Gradient Boosting Machines, and causal analysis with the Uplift Random Forest, we probed key algorithm aversion drivers. In the personal dimension, we assessed Big Five Personality Traits, Locus of Control, Generalized Trust, and demographics. We find that payment reduced delegation, while full automation promoted it. Factors like Age, Extraversion, Openness, Neuroticism, and Locus of Control consistently emerged as significant in shaping delegation decisions. Female participants demonstrated a stronger reaction to algorithmic mistakes. This study offers insights for crafting user-centric AI design to enhance cooperation and minimize aversion.

Keywords: Algorithm Aversion, human-computer interaction, decision behavior, machine learning, causal inference

1 Introduction

Driven by technological advancements, data availability, and computing power, intelligent systems powered by Artificial Intelligence (AI) have become common in our society, largely due to their transformative potential [1]. AI simulates human behaviors like learning and decision-making [2]. AI's ability to efficiently process vast amounts of data, inform decisions, and automate processes has led to its widespread adoption [3].

However, technological shifts can lead to new social phenomena like algorithm aversion, characterized by the reluctance to use algorithms in decision-making, despite their superior ability to undertake certain tasks [4, 5]. The extensive body of literature emerging in a relatively short period reveals an intricate mechanism with various factors that can influence aversion or appreciation of algorithms, demonstrating the complexity of achieving a common understanding of the underlying reasons for this behavior.

As a consensus in the literature, context and personal elements significantly shape an individual's willingness or aversion to delegate decisions to an automated system. Building on this concept, we explore these two impact dimensions in an experimental study, applying a simplified multi-armed bandit problem. In the experiment, subjects repeatedly choose from three options with hidden expected values, aiming to identify the superior option. At each period, they can delegate decisions to a Reinforcement Learning algorithm. For a holistic understanding of this behavior, this study delves into the environmental dimension by investigating the impact of explainability, costs, and full task automation. Concurrently, we assess the personal dimension by examining personality traits commonly associated with algorithm aversion, such as the big five, locus of control, generalized trust, and demographic information. Despite growing awareness of algorithm aversion, there remains a need for more extensive research; therefore, we focused on these psychological and contextual measures inspired by suggestions and recommendations for further research in [6, 7]. This paper aims to contribute to understanding how to design systems that enable fruitful interactions between humans and computers.

Experimental evidence on algorithm aversion and appreciation varies significantly across domains and contexts. Studies have found differing levels of human interaction with automated agents based on factors such as task context, performance expectations, and agent roles [8]. Studies in financial and investment contexts highlight reluctance to fully surrender decision-making authority to automated agents despite their superior performance [9–11]. The presence of human errors and significant decision outcomes seem to exacerbate algorithm aversion [4, 12]. Yet, showcasing an AI-based system's learning ability [13] or exerting time pressure can mitigate this aversion [14]. Notably, the moral implications of decisions also play a role. In morally-charged decisions, people often prefer the discretionary scope of human decision-makers [15], and in situations where discrimination is possible, people prefer

algorithmic evaluation [16]. However, there are instances of preference for algorithmic over human advisory, influenced by factors such as how the expertise of the algorithm is framed against a human [17–19]. In summary, experimental studies on human-machine collaboration and algorithm aversion point to the complexity of these phenomena, influenced by a range of factors from decision consequences and task complexity to decision context framing and perceived algorithm expertise. For comprehensive and interdisciplinary literature collections on algorithm aversion, systematic reviews are provided in [20], [7], and [6].

Given the complexity of the phenomenon, we began our methodological approach with statistical and regression analysis to understand treatment differences and explore variable relationships. We then used machine learning and causal inference techniques, including Logistic Regressions, Random Forest, Gradient Boosting Machines, and Uplift Random Forest classifiers, to probe the nuanced nature of decision delegation behavior. Contextual factors like payment and automation notably affected delegation, with payment reducing and full automation boosting its likelihood. Key personal factors influencing delegation across models were age, extraversion, openness, neuroticism, and locus of control. This paper documents the intricate relationships between individual traits, contextual conditions, and delegation behavior, providing a nuanced understanding of algorithm aversion within the boundaries of an experimental construct.

2 Experimental Design

The experimental setting employed a between-subject design, utilizing a simplified version of the multi-armed bandit problem [21]. Our design finds parallels in previous works, notably by [22], who also examined human interactions with bandit-based decision-making scenarios. The primary task involved participants repeatedly choosing one of three options labeled as "products" over 40 periods. The experiment was conducted online, where participants were instructed to select from three products, each with distinct hidden quality levels that represented their expected values, translated into the probability of receiving a payoff from the chosen option. The three variants of quality were low (50% chance of payoff), medium (70% chance of payoff), and high (90% chance of payoff). These probabilities were randomly assigned to products 1 to 3 at each participant's onset and remained constant throughout the experiment. Through repeated choices, the expected goal was for the participants to identify the high-quality product that would maximize their total payoffs. After each selection, participants received feedback on the outcome of their decision. In each round, participants had the option to delegate the decision to an algorithm. After reading the instructions, we asked participants about their perception of using algorithms for decision-making in regular tasks. The responses were categorized as positive, neutral, or negative. This response was used as a variable in the study, referred to as *perception*.

The basic framework described above is established as the "baseline" treatment. We further introduce three treatments with different contexts — Explanation, Payment, and Automation — to investigate the impact of explainability and transparency, willingness to pay, and complete task automation on delegation behavior. We aim to

better understand user preferences and friction points in algorithmic decision-making by examining these factors. In all treatments, we employ an attention check in a given round by displaying an animal picture below the task, which participants had to identify by the end of the task. Information about the design and the actual experiment screens are documented in appendix D.

2.1 Explanation Treatment

As discussed in numerous studies, transparency and explainability are key factors affecting the acceptance of algorithmic decision support. Algorithm complexity often presents these tools as "black boxes," undermining their acceptance due to the lack of understanding [23–28].

The inherent complexity in high-performing computational models poses a dilemma between accuracy and transparency, as the intricacy of these models could challenge the public's comprehension [29–31]. This complexity underscores the ongoing challenge practitioners face in maintaining explainability [32], necessitating accessible explanations irrespective of the chosen approach. Institutions and regulators also emphasize the need for transparent algorithmic decisions [33].

We tested the information-sharing impact on delegation in this *explanation* treatment, in which participants had access to a description of the algorithm used in the product selection task. The description was supposed to be non-technical and to transmit the essence of the method behind reinforcement learning to the subjects. In the primary experiment page, the following text is displayed in a text box with a prominent design: "*Reinforcement Learning: the algorithm calculates probabilities and chooses an alternative based on the success of choices in previous rounds*". The description text remained visible during the experiment.

2.2 Payment Treatment

Exploring the less examined aspect of financial incentives in algorithm aversion, people might hesitate to pay for transparent AI if costs surpass perceived benefits [34]. During crises, the appeal for robo-advisors—and hence the willingness to pay—escalates due to the need for financial advice [35]. Similarly, radiologists are ready to pay for AI tools that expedite diagnostics [36].

We investigate payment's role in algorithm aversion by assigning a payment requirement to algorithmic support, termed *payment* treatment. Here, participants were informed that while they can delegate decisions to an algorithm, each delegation carries a cost of 0.10 points (one-tenth of a point), aiming to introduce the psychological aspect of payment in a way that participants easily understand. The goal was to simply introduce payment as a contextual variable to gauge its impact, not to explore the complexities of differential willingness to pay. The cost incurred for a decision effectively restricts algorithm support to a pay-per-use basis. The points deduction reduces the expected values of the products by the same amount, introducing a "loss" for rounds where payoffs do not materialize, as the amount is subtracted from the participant's total points.

2.3 Automation Treatment

The task complexity may induce people towards higher acceptance of algorithmic decisions [37]. [38] argue that from a human standpoint, full, compared to partial, automation of decision-making processes can be very desirable in terms of efficiency, such as improving productivity, and effectiveness, for better resource allocation. In essence, the action of delegating the decision is already a form of automation, as the algorithm calculates and selects the best option based on past data. We advance this process by further automating it, thereby reducing the overall task burden. In this way, one can analyze the subjects' behavior toward the delegation of discrete decisions compared to the delegation of the complete task.

In the *automation* treatment, the algorithm takes over the repetitive task of product selection for 40 periods, easing the participants' effort. Unlike previous treatments requiring round-by-round delegation decisions, this feature allows continuous selection without active involvement. Participants could toggle automation on or off at any stage. If they opted for delegation, they had a 5-second window to override the decision, redirecting them to the primary selection interface. Feedback remained available post each round.

2.4 Personal Dimension

Algorithm aversion can be significantly impacted by personal factors such as psychological aspects, personality traits, demographic features, and algorithm/task familiarity [6]. For instance, individuals with an internal locus of control tend to resist human and AI suggestions [39], and neuroticism correlates with lower trust ratings. Delegation to algorithms increases when information scarcity is present and among extroverted individuals [40]. Trust in algorithms is not static but can evolve with personal experiences [41], which similarly impacts attitudes toward autonomous transport [42].

Broadening our research to encompass both contextual and personal aspects of algorithm aversion, we incorporate demographic data, the Big Five Personality traits, Locus of Control, and trust levels into our analysis. The Big Five Personality Traits offer an encompassing view of human personality [43], while Locus of Control illustrates an individual's belief in their power over life events [44]. Generalized trust signifies an individual's confidence in the reliability and benevolence of others [45]. After completing the selection task, participants proceeded to this series of personality questionnaires, which also included control questions (see appendix D).

3 The Algorithm: Reinforcement Learning Implementation Framework

The term "algorithm" has various definitions across different fields. Computer science typically defines it as a step-by-step procedure or set of rules used to perform tasks [46]. In the context of algorithm aversion, it often refers to decision-making tools that assist humans in making choices or predictions [4].

A variety of algorithms could be applied to the task of repeatedly selecting alternatives that maximize one’s payoffs. In our design, we aimed to allow participants to observe the algorithm’s training and improvement process throughout the task while keeping it simple enough for participants in the explanation treatment to understand its core mechanism in just a sentence or two. As a result, we chose the Reinforcement Learning (RL) model, a class of solution methods well-suited for learning-based and sequential problems.

Reinforcement learning is typically framed as an optimization problem, with the goal of identifying optimal actions based on defined criteria [47]. The model’s framework is designed to map situations to actions in a way that maximizes rewards, as defined by [48]. Key components of reinforcement-based models include a set of choices or actions, a mechanism for receiving feedback associated with each choice, an updating rule that adjusts previous beliefs or estimates of each choice’s expected value based on the feedback, and a decision rule that determines the probability of selecting each choice based on current beliefs. Our model is based on [49]’s implementation, which incorporates the concept of attractions, or weights attached to strategies that represent the perceived value associated with specific choices [50]. Our implementation assigns an attraction value to each product, which is updated after a decision is made using a learning rule. The attractions are transformed into probabilities of choice using a softmax function. A formalization of the algorithm is presented in appendix A.

The embedding of this algorithm in the experiment generates one instance of reinforcement learning for each participant, which starts with no pre-training or bias. The attraction values are initialized at 0, and the algorithm learns from participant choices and its own choices over time, making the learning process for humans and algorithms comparable.

4 Results

In this section, we conduct a comprehensive six-stage analysis of decision delegation to an algorithm, exploring its contextual, behavioral, and personal dimensions. We begin with an overview of our sample information and attention analysis, followed by an examination of delegation behavior across different treatments. We then use regression methods to identify significant predictors of delegation behavior and machine learning methods for a nuanced understanding of algorithm aversion. We incorporate causal inference methods to clarify causal relationships, analyze participants’ reactions to algorithmic failures, and measure the algorithm’s performance under varying conditions. This multifaceted approach provides a detailed understanding of the complex phenomenon of algorithm aversion¹.

4.1 Sample Information and Attention Analysis

A total of 358 participants took part in our online experiment. Subjects were evenly distributed across the four treatments, with approximately 89 to 91 participants per treatment. On average, the experiment took 11 minutes to complete, and participants earned between 4 and 10 euros, with an average of 6.13 euros. Demographically, the

¹This research project was pre-registered in AsPredicted.org, with the ID 119401.

sample was 52.7% female. Participants were primarily from Germany (51%), with the remaining individuals representing various nationalities. Most participants (73.2%) were from the Rhine-Waal University of Applied Sciences, while 26.8% were from Heidelberg University (both in Germany), aged between 18 and 47 years old; the mean age was 25. Among the subjects, 19% were economics students; the rest were from various other academic disciplines, of which 21% came from STEM majors. The self-reported perception values were 46.6% positive, 43.9% neutral, and 9.5% negative.

We analyzed participants’ attention, particularly focusing on the automation treatment, to determine if active supervision of the algorithms’ decisions persisted in a fully automated task. To measure this, we calculated the total time the web page was active in the subjects’ browsers. Additionally, we implemented attention-check questions in both the experimental task and the personality questionnaires. The results are summarized in the table 1; these values do not account for the first round, which includes the time of reading the instructions.

Table 1 Attention metrics for all treatments

Treatment	Average Active Time (s)	Animal Question (frequency correct)
Baseline	9.6	0.88
Explanation	10.3	0.89
Payment	9.0	0.85
Automation	11.2	0.55

The active time analysis showed consistent results across all treatments, with participants spending an average of 9 to 11 seconds per round. A second attention check involved identifying an animal that appeared during the final rounds, revealing decreased attention in the automated treatment. Even though the screen was active, fewer people in the automated treatment seemed to monitor the task closely. We included an attention self-report question in the automated treatment especially asking if the subject had supervised the algorithm’s decisions during the task. 76% of them answered yes, which deviates from the 55% of participants that got the animal question correct. 15% answered no, and 9% answered not applicable. The delta suggests an overreporting of the attention and supervision levels in the automated treatment. Four control questions were embedded in the personality tests, with 78% of participants answering all four correctly and 93% answering at least three correctly, indicating attentive reading.

4.2 Delegation Behavior and Treatment Effects

We measured the frequency of delegating decisions to the algorithm in each treatment. The absolute frequency of delegation in each treatment is documented in table 2.

In the baseline treatment, we observed a balanced split, where about half of the decisions were delegated across participants and rounds. The information shared in the explanation treatment only slightly increased the number of delegation decisions. The introduction of payment sharply decreases, and the possibility for automation increases the willingness to allow the algorithm to decide.

Table 2 Absolute frequencies of delegation across the four treatments

Treatment	Frequency of Delegation
Baseline	53.02%
Explanation	58.37%
Payment	27.87%
Automation	66.07%

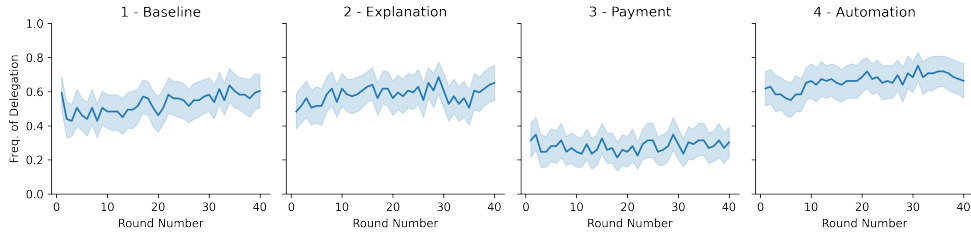


Fig. 1 Mean Frequencies of Delegation Over Time

Figure 1 displays the overall delegation frequencies over time, where the distributions are consistent across treatments and relatively constant, without any large variations in the decision behavior between rounds. We aggregated the experimental data on a participant level to test these findings for statistical significance. Each participant’s cumulative delegation frequency over 40 periods is treated as an independent observation. The distributions of these relative frequencies of delegation are displayed in the histogram in figure 2.

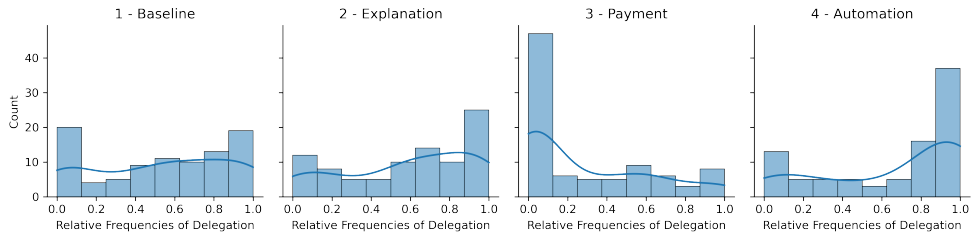


Fig. 2 Histogram of Participants Cumulative Delegation Frequencies

As anticipated, the highest delegation frequencies occur in automation and the lowest in payment treatments. The baseline and explanation treatments exhibit a more even distribution of subjects’ delegation behavior. We employed a Kruskal-Wallis test [51], a non-parametric statistical test comparing the medians of several independent samples. With a test statistic of 52.67 and a p-value < 0.001 , the results indicate a significant difference between the medians of the four independent treatment samples.

While the Kruskal-Wallis test reveals significant differences, it does not provide detailed insights into these differences between the samples. Consequently, we employed a Dunn posthoc test [52] to identify significant pairwise differences between samples. The p-values for these comparisons are in table 3.

Table 3 Dunn posthoc test results, p-values for pairwise treatment comparisons

	Baseline	Explanation	Payment	Automation
Baseline	1	0.373	<0.001	0.009
Explanation	0.373	1	<0.001	0.090
Payment	<0.001	<0.001	1	<0.001
Automation	0.009	0.090	<0.001	1

In summary, these results suggest significant differences between the medians of baseline, payment, and automation, as well as between explanation and payment. There is no significant difference between the medians of baseline and explanation or between explanation and automation. The payment feature was the most influential regarding the willingness to delegate.

The contextual findings highlight the influence of different treatment conditions on the delegation behavior of participants. The baseline and explanation treatments led to a more even distribution of delegation behavior. On the other hand, the payment treatment had a considerable negative impact on the willingness to delegate. The automation treatment led to the highest frequency of delegation among the four treatments, demonstrating the importance of reducing the involved workload in a task in encouraging algorithm-based decision-making. Overall, these results underscore the significance of understanding and addressing the factors that affect delegation behavior to design more effective human-algorithm collaborations and decision-making processes.

4.3 Incorporating the Personal Dimension - Regression Analysis

The design of our treatments provides insights into how exogenous factors influence delegation behavior. However, individual factors also play a significant role in algorithm aversion, as widely discussed in the literature. In this section, we examine the binary action of delegating a decision in relation to treatment conditions and personal factors, including personality test scores, gender, education, and self-reported perception (as explained in chapter 2). Categorical values were encoded as binary dummy variables.

Although correlations between the variables under investigation and delegation are primarily weak, they are highly significant (full correlation results are reported in B1, appendix B). To further explore and quantify these relationships, we constructed a logistic regression model including demographic and personal information as independent variables. The model results are summarized in table 4. A critical remark in the regression modeling is that we use the entire experiment’s dataset: every decision from

each participant at each round. Due to repeated choices made by the same individuals across 40 periods, we clustered the standard errors on the participant level. This approach accounts for intra-participant correlation, considering potential influences from unobserved individual factors or shared experiences, as per [53]’s reasoning.

Table 4 Logistic Regression Results - Delegation

Variable	Coefficient	Standard Error	p-value
Constant	-0.525	1.013	0.605
Explanation	0.252	0.207	0.223
Payment	-1.012	0.235	*** < 0.001
Automation	0.515	0.234	*0.027
Female	-0.144	0.179	0.421
Age	-0.009	0.018	0.596
STEM	0.267	0.227	0.238
Business & Economics	-0.181	0.201	0.37
Extraversion	0.04	0.059	0.497
Agreeableness	0.036	0.073	0.627
Conscientiousness	0.137	0.085	0.106
Neuroticism	-0.047	0.072	0.513
Openness	-0.008	0.087	0.926
Internal LoC	0.057	0.102	0.578
External LoC	0.054	0.106	0.614
Generalized Trust	0.067	0.066	0.307
Perception	-0.368	0.14	**0.009

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

The logistic regression model provides several insights into the effects of treatments and personality traits on delegation and also reinforces the findings in chapter 4.2. Initially, the automation treatment exhibits a positive and statistically significant impact on delegation ($p = 0.027$), suggesting that automating tasks encourages individuals to delegate. Conversely, the payment treatment displays a negative and statistically significant influence ($p < 0.001$), implying that requiring payment could discourage delegation. The explanation treatment, although positive, is not statistically significant ($p = 0.223$). Regarding personal variables, the only statistically significant effects are observed for perception ($p = 0.009$), which negatively impacts delegation, suggesting that an increase in negative perception about algorithms is correlated with a lower likelihood of delegation. Other variables, including gender, age, field of study, and personality traits, do not exhibit statistically significant effects on delegation in this model. A second regression model, including interaction terms, is reported in appendix B, in which payment loses its significance, and Internal Locus of Control becomes significant. Quantile regression models applied to cumulative delegation frequencies (shown in figure 2) showed similar significance and coefficients to logistic regression, despite a marginally better fit. See appendix B for full details.

In conclusion, examining personality traits and algorithm aversion uncovers the influence of individual factors and treatment conditions on delegation behavior. A critical insight from this analysis is the existence of intricate relationships between various traits. Interaction terms offer a more comprehensive understanding of the relationships

between variables and delegation behavior by accounting for the dependence of some variables' effects on the values of other variables. Gaining insights into these relationships can aid in comprehending how diverse behavioral profiles respond to algorithmic systems.

4.4 Machine Learning for Delegation Behavior Analysis and Causal Inference

To understand whether the personal and contextual pieces of information are helpful in predicting the delegation behavior in such a case, we tested a few prediction techniques using the same variables scheme, that is, predicting the binary outcome of the delegation decision possibility using the treatments, personality, and demographic data.

[54] highlights the benefits of applying machine learning to model behavior, emphasizing its potential for improved predictive accuracy, handling large datasets, capturing non-linear relationships, and adaptability. Additionally, machine learning enables personalization and fosters cross-disciplinary insights, contributing to a better understanding of human decision-making and facilitating more effective interventions across various domains.

The logistic regression model, as detailed in chapter 4.3, offers limited insights into the complex interplay of our variables, accounting for only about 9% (pseudo R-squared) of the variation in delegation decisions. Given the absence of clear linear relationships and the complexity of the data, we turn to more sophisticated methods. We employ machine-learning models to examine the overall impact of variables on predicting delegation, followed by causal machine learning models to separate treatment effects from the personal covariates. In the subsequent models, we refer to within-sample predictions, using 80% of the sample for model training and the other 20% to generate and test predictions. Methodological formalizations for the adopted methods can be found in appendix A, and technical model implementation remarks in appendix C.

4.4.1 Predicting Delegation Behavior

If we use our logistic regression coefficients to generate predictions, the model yields an accuracy score of 0.62, meaning 62% of the delegation decisions were classified correctly, not far from a random baseline. This relatively low accuracy might be due to several factors influencing the results that have yet to be accounted for or the failure of the model to capture complex relationships between the variables. To deepen the understanding of these variables' relationships and the possibility of generating predictions for algorithm aversion behavior using contextual and personal information, we resort to the machine learning techniques Random Forest and Gradient Boosting Machines.

Research shows successful predictions of behavioral elements using personality traits, characteristics, and environmental data. [55] used psychometric test data, including Big 5 and Dark Triad, and Twitter features to predict cyberbullying accurately. [56] employed machine learning to predict mental health status based on

social media and personality data. Similarly, [57] used personality traits to predict smartphone usage behavior. [58] combined reaction time, psychological attributes, and personality traits to predict Loss Aversion Bias, supporting Kahneman’s ”Thinking Fast and Slow” theory [59]. These studies demonstrate the potential of machine learning models in similar prediction tasks.

[60] introduced the Random Forest model, an ensemble learning method designed for classification and regression problems. The algorithm works by creating multiple decision trees, each of which ’votes’ on an answer. In a classification problem such as ours, the Random Forest chooses the class that gets the most votes from all the trees. The key idea behind Random Forest is to create a ”forest” of diverse decision trees constructed from random subsets of training data and features. This approach helps increase the model’s robustness, reduce overfitting, and improve overall predictive accuracy. The Random Forest algorithm is particularly useful for binary classification problems because it can handle non-linear relationships between the input features and the output variable. It can also handle missing values and outliers in the input data and estimate the importance of each input feature in the prediction [61].

In a similar manner, Gradient Boosting Machines (GBMs) are a class of ensemble learning algorithms that build a robust model by iteratively adding weak learners, typically decision trees, to minimize a loss function. The algorithm focuses on correcting the errors of the previous tree by training on the residuals, effectively improving the overall model’s performance, as defined in [62].

As per definitions in [60] and [62], Random Forest and GBMs are ensemble learning methods for similar purposes. The main difference lies in their approach to building the ensemble of decision trees. Random Forest constructs multiple trees independently and in parallel, combining their predictions through averaging or majority voting. It uses bagging (Bootstrap Aggregating) to create diverse trees by resampling the dataset with replacement. In contrast, GBM constructs trees sequentially, with each new tree trying to correct the errors made by the previous tree. It utilizes a technique called boosting, where trees are combined through a weighted majority vote, and the weights are determined by minimizing a loss function during the training process. We apply both methods for comparable results but with distinct processes, enabling comparing and validating the findings from the generated predictions to assess our findings’ consistency. In each model, feature importances highlight the significance of each feature in predicting the target variable. Figure 3 presents an overview of the feature importances.

Both models have been cross-validated during parameter fitting and training using the KFold method to avoid overfitting (details in appendix C). In this process, we split the training data into a number of subsets or ”folds.” We train the model on the remaining data for each fold and test it on this fold. This process is repeated for each fold, allowing us to assess the model’s performance based on its ability to predict new data [63, 64]. Furthermore, with an equivalent objective as clustering the regression errors on a participant level (chapter 4.3), we aggregated the participant observations here using the GroupKFold variant, which ensures instances from the same participant either in the training set or the test set. This approach safeguards against data leakage and maintains a realistic estimate of the model’s performance,

especially when observations within the same group (in this case, participant) are correlated.

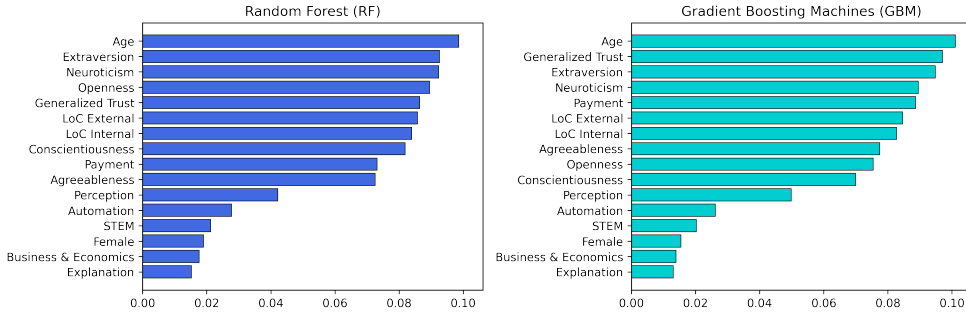


Fig. 3 Machine Learning Models Feature Importances

According to the Random Forest and Gradient Boosting Machine models, the decision to delegate to algorithms is influenced by a complex mix of individual characteristics and contextual factors. Age consistently emerges as the most significant variable in both models, reflecting its significant role in shaping comfort with algorithmic delegation. Similarly, Neuroticism and Extraversion — two Big Five personality traits — feature prominently, signifying their impact on delegation tendencies.

Apart from these, the Locus of Control, both internal and external, appears to influence delegation decisions, although they are more pronounced in the Random Forest model. Contextual factors, like payment and automation, also emerge as crucial determinants across both models. Intriguingly, automation is more influential in the GBM model, suggesting a more substantial bias towards delegation in fully automated scenarios. Gender, education, and the Explanation context appear to have minimal impact in both models.

These findings underscore the intricate dynamics governing decision delegation, with no single factor having a dominating influence. Instead, a nuanced interplay of various individual and contextual elements appears to guide the decision to delegate to algorithms.

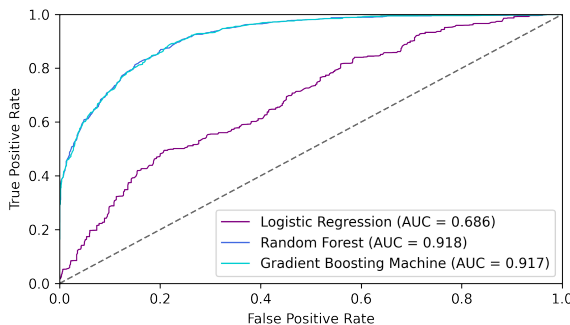
We evaluated the Logistic Regression (LR), Random Forest (RF), and Gradient Boosting Machine (GBM) models using four metrics: Accuracy, Precision, Recall, and F1 score. Accuracy calculates the proportion of correctly classified instances. Precision quantifies how well the model correctly identifies positive instances. Recall gauges the model’s ability to detect positive instances among actual positives. The F1 score, a blend of precision and recall, is the harmonic mean of these two metrics [65, 66]. As summarized in Table 5, both RF and GBM outperformed LR in predictive power, with RF achieving slightly superior performance across all metrics. This outcome highlights the efficacy of tree-based models for our classification problem.

In addition, a Receiver Operating Characteristic (ROC) curve provides a graphical representation of a classifier’s performance across varying decision thresholds (figure

Table 5 Prediction Performance metrics

	LR	RF	GBM
Accuracy	0.6210	0.8332	0.8325
Precision	0.6112	0.8185	0.8120
Recall	0.7018	0.8730	0.8730
F1-score	0.6534	0.8414	0.8415

4). The Area Under the ROC Curve (AUC-ROC) measures the overall performance of a binary classifier. It ranges from 0 to 1, with higher values indicating better performance. A value of 0.5 indicates a random classifier (dashed line), and 1 indicates a perfect classifier. The ROC area quantifies how well the classifier can distinguish between the positive and negative classes, regardless of the choice of classification threshold [67, 68]. In the overall analysis, and in line with previous performance metrics, the LR model is surpassed by the other models, with the RF model showing a slight edge. The high scores achieved by both the RF and GBM models affirm their ability to explain the data, enhancing the reliability of the interpretations documented in our study.

**Fig. 4** ROC Curves for All Models

Although logistic regression provided valuable insights into the direction and significance of individual variables, its ability to handle the complex data relationships in our study was limited. We explored machine learning techniques to capture these relationships better, specifically Random Forest and Gradient Boosting Machines. Both models significantly outperformed logistic regression regarding accuracy, precision, recall, and F1 score, with the Random Forest model having a slight edge in accuracy over the GBM. Both models consistently highlighted the same features, such as payment, extraversion, and neuroticism, as key influencers in delegation decisions.

4.4.2 Causal Inference and Heterogeneous Treatment Effects - Uplift Random Forest

To further understand the factors influencing decision delegation to algorithms, we now focus on disentangling the effects of the treatment conditions from personal data.

While regression and machine learning models have provided insights, they combine all variables, not distinguishing between treatment conditions and personal characteristics effects. Hence, we use causal inference to uncover how treatment effects vary across different subgroups within our sample, focusing on estimating the expected change in the outcome as a result of the intervention. This approach allows us to measure heterogeneous treatment effects and identify the subset of individuals most influenced by the treatment conditions, given their characteristics. To this end, we resort to Uplift Modeling.

Uplift Modeling, a branch of causal inference, models the impact of incremental treatment effects on individuals' behavior [69]. Early applications of similar methods can be seen in [70]. For a comprehensive definition and literature review on machine learning problems and applications, see [69, 71].

We employ the Uplift Random Forest Algorithm, an ensemble learning method that uses the random forest algorithm to estimate the causal effect of a treatment or intervention on individual outcomes [72, 73]. The uplift random forest classifier [74] incorporates the treatment indicator as a covariate to capture differential effects and uses other covariates to estimate individual treatment effects. The model is tuned using the same cross-validation technique described in 4.4.1, with details in appendix C.

Treatment effects can be evaluated at an individual level by computing uplift scores. These scores represent the predicted likelihood of delegation for each observation under each treatment scenario, essentially providing a probabilistic estimate of how a participant would behave if they were subjected to a specific treatment. The distributions of these predicted likelihoods are plotted in figure 5. The trend observed in this analysis follows the initial assessment of the treatment effects (chapter 4.2) in reference to the baseline. Payment negatively impacts the likelihood of delegation, whereas explanation has a slight positive effect, and automation has a more pronounced positive effect. Each treatment's computed average treatment effects are payment = -0.26 , explanation = 0.05 , and automation = 0.12 .

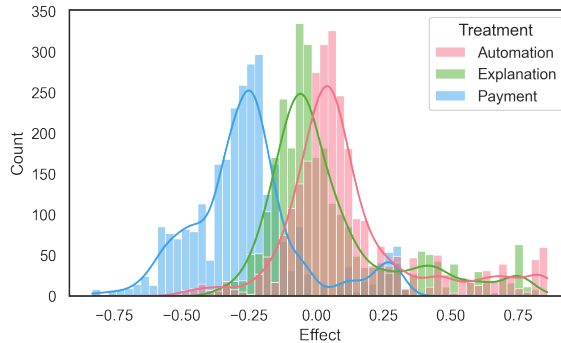


Fig. 5 Distribution of Predicted Treatment Effects (Uplifts)

Feature importance can also be extracted from this model, with a slightly different meaning. Unlike traditional classification models, in Uplift models, feature importance

does not directly equate to the effect of a feature on the outcome but rather its influence on the treatment effects. In other words, an essential feature in the model translates to the influence on the change in the likelihood of delegation mediated by the treatment. These values are presented in figure 6.

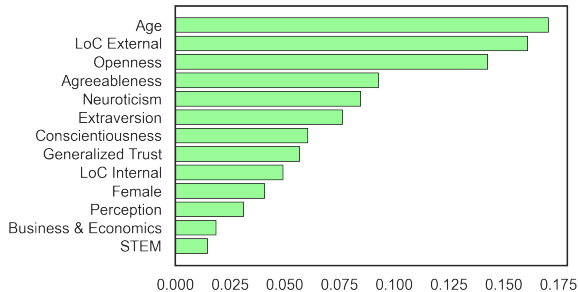


Fig. 6 Feature Importances - Causal Model

In the Uplift Random Forest model, Age, Openness, External Locus of Control, Extraversion, and Agreeableness significantly influence the treatment effectiveness on delegation behavior. Other variables like Internal Locus of Control and Conscientiousness also play a role, but their influence is moderate. On the other hand, Gender, Perception, and Education have the least effect on treatment assignment.

Contrasting with the Random Forest and Gradient Boosting Machine models, the Uplift model emphasizes the impact of these variables on the treatment effects rather than the outcome itself. While age and certain personality traits like Extraversion and Openness are influential across all models, the Uplift model uniquely demonstrates their role in optimizing treatments for delegation.

Evaluating causal inference models, like uplift random forests, is intricate due to counterfactual outcomes. We can only observe a given individual’s delegation decision under one treatment. Unlike traditional classification, where predicted outcomes are compared to observed labels (as in table 5), uplift modeling predicts the difference between observed and unobserved counterfactual outcomes. This lack of observed outcomes for both scenarios for an individual restricts using standard classification metrics. Instead, metrics specific to uplift models, such as uplift curves, assess their performance. The uplift curve, similar in interpretation to the ROC curve, plots cumulative gain from targeting individuals by predicted uplift. Derived from it, the Area Under the Uplift Curve (AUUC) mirrors the AUC-ROC, gauging the model’s ability to prioritize effective interventions. Figure 7 shows our model’s Uplift Curve.

We have computed the AUUC using a synthetic control group consisting of individuals whose predicted optimal treatment matches the actual treatment they received or those in the actual control group, following the method in [75]. The uplift score for each individual in the synthetic control was computed, and individuals were ranked based on these scores. The AUUC was then calculated as the area under the curve plotting the cumulative proportion of actual outcomes against the proportion of the

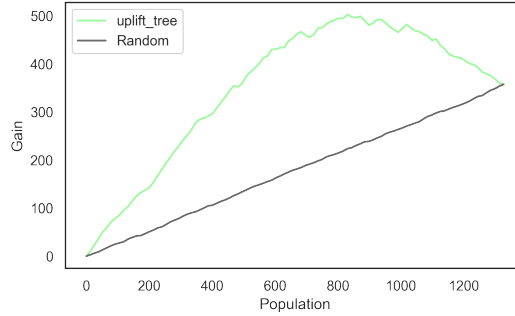


Fig. 7 Uplift Curve

population targeted. The result is 0.977, which indicates relatively high performance in the prediction task and in explanation power.

Applying Uplift Random Forest to our study has offered valuable insights into the factors that influence the impact of treatments in delegation decisions. The model identified age, openness, and certain personality traits as significant determinants. It provided an additional perspective by focusing on the influence of these variables on treatment effectiveness rather than on the outcome itself.

4.5 How Subjects React to Non-Profitable Algorithmic Decisions

Numerous studies show that people initially trust algorithms, but trust may plummet after a mistake occurs [76]. [4] found that people avoid algorithms or computerized decision-making systems even if they make fewer errors than humans due to high expectations for algorithms and attributing errors solely to the algorithm. [77] showed that people are less likely to follow advice from a computer algorithm immediately after receiving incorrect advice. Complementarily, [78] reveals that poor algorithmic performance harms human confidence in the algorithm and self-confidence. [37] complements the idea of adverse reactions by outlining that bad decisions generated by algorithms are more severely punished than those of humans. To investigate this further, we analyzed participants' reactions after delegating a decision to the algorithm and receiving no payoff.

Delving into the impact of the algorithms' performance on the subjects, we calculated the frequency of participants changing their strategies from "delegate" to "not delegate" relative to the number of times the algorithm's decision resulted in a zero payoff, which does not necessarily mean a "wrong" choice but can also indicate a non-realized payoff from the "correct" choice. We extended this analysis to explore potential gender effects. Table 6 presents the absolute proportions of reaction results categorized by gender and treatment.

On average, participants in the payment treatment group exhibited the highest reaction frequency (0.35), suggesting that individuals are more likely to change their decision when a financial incentive is involved. Conversely, the automation treatment group had the lowest frequency of reaction (0.09), indicating that participants are

Table 6 Relative frequencies of changing strategies (reaction) following algorithmic failures

	Baseline	Explanation	Payment	Automation
General (aggregated)	0.30	0.25	0.35	0.09
Males	0.26	0.15	0.27	0.07
Females	0.34	0.31	0.40	0.10

less likely to change their decision when the task is automated, possibly due to the complete handover process or also satisfaction with the algorithm performance, which was overall higher in the automation treatment (further details on the algorithm’s performance are documented in chapter 4.6).

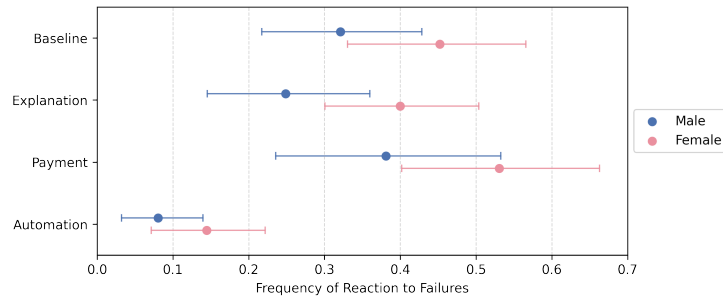


Fig. 8 Frequencies of reaction to algorithmic failures by treatment and gender

Comparing reaction frequencies between males and females reveals that females have a higher reaction frequency across all treatments, suggesting they might be more sensitive to algorithm mistakes (figure 8). To further examine the gender gap in reaction, given that gender differences were not observed elsewhere in the experiment, we conducted statistical tests on both samples.

Similar to the statistical tests performed on the relative frequencies of delegation, we calculated the relative frequencies of reaction for each participant over 40 periods, treating each participant’s decision path as an independent observation and separating the samples by gender. We then applied a Mann-Whitney U test [79] to measure the difference between the two independent samples. The results show a value of 7751.51 and a p-value of 0.0028, outlining a statistically significant difference between the means of the frequency of strategy reactions for males and females. To deepen our understanding of participant reactions, we further analyzed whether contextual or personal factors influenced their behavior. Similar to the methodology used in the delegation behavior analysis (Chapter 4.3), we employed a logistic regression with standard errors clustered at the participant level. The results of this analysis are compiled in Table 7.

The analysis indicates that task automation, gender, and internal locus of control are key factors in strategy changes following unprofitable algorithm decisions. Full task automation and a high internal locus of control reduce the likelihood of strategy shifts,

Table 7 Logistic Regression Results - Reactions

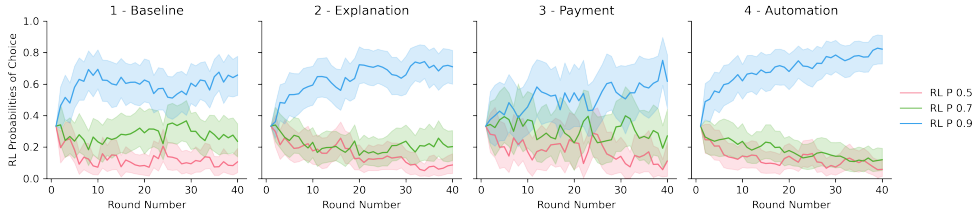
Variable	Coefficient	Standard Error	p-value
Constant	-2.188	0.868	*0.012
Explanation	-0.088	0.181	0.626
Payment	-0.32	0.214	0.134
Automation	-1.042	0.24	*** < 0.001
Female	0.453	0.154	**0.003
Age	0.005	0.016	0.775
STEM	-0.345	0.213	0.106
Business & Economics	0.021	0.182	0.908
Extraversion	0.031	0.053	0.567
Agreeableness	0.121	0.081	0.138
Conscientiousness	0.027	0.082	0.743
Neuroticism	-0.044	0.061	0.465
Openness	0.04	0.09	0.657
Internal LoC	-0.292	0.087	*** < 0.001
External LoC	-0.107	0.104	0.304
Generalized Trust	-0.034	0.07	0.63
Perception	-0.229	0.122	0.06

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

suggesting trust in the process and personal control beliefs. Conversely, female participants are more prone to strategy changes, hinting at potential gender differences in reactions to algorithmic failures. Other factors, including algorithm explanation, payment requirement, and various personality traits, don't significantly influence strategy changes, suggesting their impact may be less direct.

4.6 Task Performance and Human-Algorithm Interaction

Finally, to evaluate the performance of Reinforcement Learning in the product selection task, we analyzed the mean probabilities of selecting each product quality level, grouping them based on their probabilities of receiving a payoff. The task was not straightforward due to the possibility of receiving a zero payoff even after identifying the best option, which could alter the weight of correct attractions. This ambiguity challenged human subjects and affected the algorithms' convergence capabilities. Figure 9 illustrates the development of choice probabilities for each product type.

**Fig. 9** RL Choice Probabilities Over Time

In all instances, the algorithm could identify the highest quality product compared to the inferior alternatives. However, performance levels varied across treatments. We observed improved performance in generating optimal choice probabilities in treatments with higher delegation rates, such as explanation and automation, compared to the other groups, with the payment group being the most impacted. In treatments where participants exhibited higher "trust" in the algorithmic decision-making process, the performance in identifying the optimal product was better.

Table 8 Frequency "high" Product Selected

	Algorithm	Human
Baseline	0.592	0.511
Explanation	0.634	0.495
Payment	0.506	0.505
Automation	0.694	0.515

In a complementary analysis, table 8 compares performance between the algorithm and human subjects throughout the task. The values denote the success frequencies, normalized by the number of human or algorithm decisions. As expected, even with a non-trained algorithm that learned on the spot, the algorithm consistently outperformed the human subjects.

5 Conclusion and Discussion

This paper investigated the impact of framing conditions, explainability, willingness to pay, and complete task automation on delegation behavior in the context of algorithmic decision-making. Additionally, the study explored individual differences by examining the Big Five Personality Traits, Locus of Control, Generalized Trust, and other individual characteristics such as gender, age, and education.

We investigated the algorithm aversion phenomenon employing a multi-stage analysis covering hypotheses testing, regressions, machine learning, and causal inference models. Our findings revealed that context conditions significantly influenced participants' delegation choices. The study demonstrated that explaining the algorithm used in the product selection task improved user trust and increased the likelihood of delegation. In contrast, introducing a cost for delegation (Payment treatment) led to a decrease in delegation rates. Finally, the Automation treatment highlighted that participants were likelier to delegate decisions to the algorithm when the task was wholly automated.

In the machine learning application, we adopted a two-pronged approach to decipher the complex dynamics of decision delegation. We utilized traditional machine learning models — Random Forest and Gradient Boosting Machines — and an Uplift Random Forest model, providing complementary perspectives on the influences on delegation behavior. The Random Forest and Gradient Boosting Machine models offered insights into the direct impacts of individual and contextual variables on delegation decisions. Age, personality traits like Neuroticism and Extraversion, and factors

like Payment and Perception consistently emerged as significant influences. These models underscored the intricate interplay of individual traits and contextual conditions, with no single factor dominating the decision to delegate. Complementing this, our Uplift Random Forest model provided direct heterogeneous treatment effects, which confirmed the impacts observed in the statistical analysis: the strong negative influence from the payment context and the moderately strong positive influence of automation. As for the impact of personal variables, this focused on their influence on quantifying the likelihood of delegation. Key variables such as Age, Openness, and certain personality traits significantly shaped the uplifts in treatment assignments. The model highlighted the importance of these factors in optimizing interventions to enhance delegation, adding a unique dimension to our understanding. The machine learning analysis revealed a nuanced understanding of how individual characteristics and contextual factors, alongside their interplay, shape decision delegation to algorithms. The machine learning models identified a set of influential factors with high predictive accuracy, while the Uplift model shed light on optimizing intervention impacts. This complexity and interconnectedness of personal and contextual factors was also reported by [80]. These insights provide valuable guidance for practitioners designing algorithmic decision systems, emphasizing the need for a personalized, context-sensitive approach.

In examining responses to algorithmic errors, we discovered pronounced reactions in scenarios involving payment treatments. Interestingly, these reactions were significantly more frequent among females, indicating the presence of gender effects. Confirmatory statistical analyses reinforced these observations, revealing that factors such as Automation, Payment, and Internal Locus of Control significantly influenced participants' responses to algorithmic mistakes. These findings highlight the influence of both gender and specific situational contexts and confirm previous experiments in the literature pointing to algorithmic failures as a driver of aversion.

Employing a non-biased algorithm and allowing it to learn exclusively from the interaction with participants allowed us to observe how the algorithm's learning process evolved alongside the participants' decision-making behavior. In particular, treatments with lower delegation rates negatively affected the algorithm's performance, generating sub-optimal choice probabilities.

The implications of these findings are manifold. By better understanding the factors influencing delegation behavior in algorithmic decision-making, we can develop more user-friendly systems that facilitate trust and encourage appropriate delegation. These insights can contribute to designing decision support tools tailored to individual preferences and optimize human-algorithm collaboration.

This study has several limitations, including the simplicity of the experimental design, which may not fully capture the complexity of real-world decision-making scenarios, and a potentially non-representative sample. The interconnectedness and multicollinearity of personal traits also present challenges in isolating and interpreting their individual effects on delegation behavior. Further research could employ more realistic product designs and decision-making tasks and investigate the effects of combined treatment conditions, e.g., payment and automation, payment and explanation, among other things, to understand better the interplay between various contextual

factors and their impact on delegation behavior in algorithmic decision-making. Moreover, future studies could also consider evaluating purely economic behaviors and attitudes, such as risk, loss, and ambiguity aversion.

In conclusion, this paper contributes to the growing literature on algorithm aversion and delegation behavior. It highlights the importance of framing conditions, explainability, individual differences, and the complex interaction between variables in shaping user preferences and trust in algorithmic decision-making systems. Future research could delve deeper into the interaction between these factors and explore the impact of different explanation styles, varying costs for delegation, and other contextual factors on delegation behavior. By understanding the nuances of human-algorithm collaboration, we can develop systems that enhance decision-making and contribute to more efficient and effective outcomes in various domains.

Declarations

- **Funding:** The project was supported by funds from the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support program.
- **Conflict of interest/Competing interests:** The authors have no conflicts of interest to declare
- **Ethics approval:** The experiment has been assessed and approved by the Ethics Commission from the Faculty of Society and Economics at the Rhine-Waal University
- **Availability of data and materials:** The datasets used in this paper will be made publicly available.
- **Code availability:** The reproducible code used in the data analysis and processing will be made publicly available

Appendix A Methodological Formalizations

This chapter provides an overview of the machine learning methods used in the project. The following subchapters account for the Random Forest, Gradient Boosting, and Uplift Random Forest methods, providing generalizations of the algorithms' implementations.

A.1 Reinforcement Learning Implementation and Tuning

The underlying problem introduces three options or products, expressed as Q_i , each associated with distinct probabilities of receiving a payoff that can be selected at each period, t . Each product Q_i is associated with an attraction value $A_{Q_i}(t)$, representing the decision weight attached to product Q_i at period t . Following the theoretical frameworks in [49, 50], the attraction values are updated based on the payoffs received

by selecting product Q_i using the following update rule:

$$A_{Q_i}(t) = \phi A_{Q_i}(t-1) + I(Q(t) = Q_i) \pi_{Q_i}(t) \quad (\text{A1})$$

This model features the indicator function, which means that a player’s attraction to a strategy can only increase if they choose it. The attraction increases by the amount of payoff received from it. In the update rule, the indicator functions $I(Q(t) = Q_i)$ equals 1 if a participant chooses product Q_i at period t and 0 otherwise, while $\pi_{Q_i}(t)$ represents the payoff received when choosing product Q_i at period t . The recency parameter ϕ indicates how quickly past payoffs are forgotten, which acts as a form of learning rate. Attractions from the previous period determine choice probabilities in any period. A logistic transformation over the attraction values calculates the probabilities:

$$P_{Q_i}(t+1) = \frac{e^{\lambda A_{Q_i}(t)}}{\sum_{k=1}^m e^{\lambda A_{Q_k}(t)}} \quad (\text{A2})$$

In this equation, $P_{Q_i}(t+1)$ represents the probability of selecting product Q_i at time $t+1$, $A_{Q_i}(t)$ denotes the attraction of product Q_i at time t , and m indicates the number of available product options. The second parameter, λ , reflects the sensitivity of choice probabilities to differences in attractions. The two necessary parameters were tuned using observed data from 1000 simulations, testing for the ranges 0 – 1 for ϕ and 0 – 10 for λ . The tuning resulted in $\phi = 0.47$ and $\lambda = 4.5$, associated with higher payoffs. The experiment parameters were set to these values statically.

A.2 Random Forest

The Random Forest algorithm concept builds a large collection of de-correlated decision trees and then aggregates them through a majority voting system for classification problems. [81] generalized the algorithm as follows:

More details on the Random Forest algorithm can be found in [60].

A.3 Gradient Boosting Machines

Gradient Boosting Machines (GBM) is a machine learning method that builds a sequence of decision trees, each correcting its predecessor’s mistakes, to create a final, robust predictive model [62]. [81] also provides a generalization of this model, with the stepwise algorithm defined as:

Lines 2-6 are repeated K times at each iteration m , once for each class. For a more detailed description of the Gradient Boosting Machines and their derivations, see the comprehensive overview in [81].

A.4 Uplift Modelling

The underlying method is the same as that of the Random Forest. However, For the uplift random forest classifier, the uplift tree consists of a combination of methods based on uplift modeling, with the tree split criterion based on differences in the

Algorithm 1 Random Forest Algorithm

Require: B trees to be grown, N size of bootstrap sample, M total variables, m selected variables, n_{\min} minimum node size

Ensure: Output the ensemble of trees $\{T_b\}_1^B$

- 1: **for** $b = 1$ to B **do**
 - 2: Draw a bootstrap sample of size N from the training data
 - 3: Grow a decision tree T_b on this data by:
 - 4: **while** each terminal node of the tree until the minimum node size n_{\min} is reached **do**
 - 5: Select m variables at random from all M variables
 - 6: Pick the best variable/split-point among the m
 - 7: Split the node into two daughter nodes
 - 8: **end while**
 - 9: **end for**
 - 10: To make a prediction for a new point x , let $\hat{C}_b(x)$ be the class prediction of the b th random forest tree
 - 11: The random forest chooses $\hat{C}_{\text{rf}}(x) = \text{majority vote}\{\hat{C}_b(x)\}_1^B$
-

uplift. In the standard notation [82], we consider $Y_i(1)$ an individual's i being treated and $Y_i(0)$ for being in the control group. In this case, the causal effect τ_i is given by $\tau_i = Y_i(1) - Y_i(0)$. Having $W_i \in 0, 1$ as a binary variable indicating if person i is in the active treatment group, and 0 otherwise (control group), the observed outcome is $Y_i^{\text{obs}} = W_i Y_i(1) + (1 - W_i) Y_i(0)$.

Based on [71], considering a balanced randomized experiment, the average treatment effects (uplifts) are estimated as:

$$\hat{\tau} = \underbrace{\frac{\sum_i Y_i^{\text{obs}} W_i}{\sum_i W_i}}_p - \underbrace{\frac{\sum_i Y_i^{\text{obs}} (1 - W_i)}{\sum_i (1 - W_i)}}_q, \quad (\text{A3})$$

which represents the difference in the sample average outcome between the treated and untreated observations. For the splitting criterion, the gain difference after splitting is defined as:

$$D_{\text{gain}} = D_{\text{after_split}}(P^T, P^C) - D_{\text{before_split}}(P^T, P^C) \quad (\text{A4})$$

Where D is the difference and P^T and P^C is the probability distribution of the outcome variable in the treatment and control groups [83]. The uplift trees were split using the Chi function, rooted in a statistical test that determines significant associations between two categorical variables. Within uplift modeling, this function aids in prioritizing splits that highlight a significant relationship between the treatment and the outcome. The divergence in this method is represented by X^2 :

$$X^2(P : Q) = \sum_{k=\text{left, right}} \frac{(p_k - q_k)^2}{q_k} \quad (\text{A5})$$

Algorithm 2 Gradient Boosting Machines Algorithm (Generalized)

Require: M iterations, n number of observations, L loss function, y_i observed response, $F(x_i)$ predicted response, $h_m(x)$ base learner at iteration m

Ensure: Output $F_M(x)$ as the final model

1: Initialize the model with a constant value:

$$F_0(x) = \operatorname{argmin}_{\gamma} \sum_{i=1}^n L(y_i, \gamma)$$

2: **for** $m = 1$ to M **do**

3: Compute pseudo-residuals:

$$r_{im} = - \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)}, \quad \text{for } i = 1, \dots, N.$$

4: Fit a base learner $h_m(x)$ to pseudo-residuals, i.e., train it using the training set $\{(x_i, r_{im})\}_{i=1}^n$

5: Compute multiplier:

$$\gamma_j m = \operatorname{argmin}_{\gamma} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

6: Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

7: **end for**

where p indicates the sample mean in the treatment group, q is the sample mean in the control group, and k denotes the leaf in which p and q are calculated.

Appendix B Additional Data and Analyses

This chapter presents additional data analysis elements not included in the main manuscript.

B.1 Correlations

Delegation behavior exhibits weak positive correlations with STEM degrees, extraversion, agreeableness, conscientiousness, internal locus of control, and external locus of control. Conversely, it has weak negative correlations with gender (female), business and economics degrees, and neuroticism. Age and openness display almost no correlation with delegation behavior (figure B1).

Table B1 displays the results of point-biserial correlation coefficients between the personality traits and delegation behavior (binary).

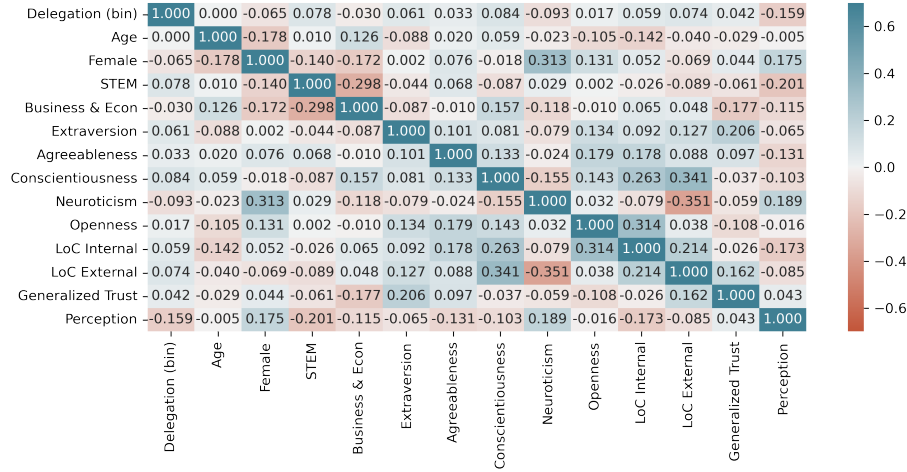


Fig. B1 Spearman correlation coefficients

Table B1 Point-biserial correlation coefficients to binary action of delegation

Variable	Correlation Coefficient	p-value
Age	-0.019	*0.026
Female	-0.065	*** < 0.001
STEM	0.078	*** < 0.001
Business & Economics	-0.03	*** < 0.001
Extraversion	0.06	*** < 0.001
Agreeableness	0.039	*** < 0.001
Conscientiousness	0.089	*** < 0.001
Neuroticism	-0.087	*** < 0.001
Openness	0.024	**0.005
Internal LoC	0.06	*** < 0.001
External LoC	0.068	*** < 0.001
Generalized Trust	0.044	*** < 0.001
Perception	-0.147	*** < 0.001

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

B.2 Regressions

This regression model includes interaction terms to account for the correlation between independent variables (table B2), providing a more nuanced analysis of the relationships between variables and delegation behavior. In this model, the main effects of some variables change, and the added interaction terms help us better understand how the relationships between variables affect the outcome.

The internal locus of control variable becomes significant ($p = 0.041$) in the model with interaction terms, while it was not significant in the model without interactions.

Table B2 Logistic Regression results - delegation, with interaction Terms

Variable	Coefficient	Standard Error	p-value
Constant	5.195	4.046	0.199
Explanation	0.195	0.21	0.354
Payment	-1.053	0.24	*** < 0.001
Automation	0.453	0.235	0.054
Female	0.821	0.567	0.147
Age	-0.012	0.018	0.501
STEM	0.269	0.232	0.246
Business & Economics	-0.193	0.201	0.337
Extraversion	0.018	0.059	0.755
Agreeableness	0.044	0.073	0.552
Conscientiousness	-0.411	0.597	0.491
Neuroticism	0.111	0.361	0.759
Openness	-0.71	0.416	0.088
Internal LoC	-1.266	0.618	*0.041
External LoC	0.251	0.604	0.678
Generalized Trust	0.07	0.065	0.284
Perception	-0.361	0.14	*0.01
Female x Neuroticism	-0.234	0.136	0.084
Internal Loc x Conscientiousness	0.136	0.089	0.129
External Loc x Conscientiousness	-0.043	0.094	0.65
External Loc x Neuroticism	-0.011	0.079	0.894
Internal Loc x Openness	0.139	0.079	0.08

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

This change suggests that the relationship between internal locus of control and delegation behavior might be more complex than initially estimated by the first model. Including interaction terms allow us to capture the combined effects of internal locus of control with other variables, such as openness, which might help explain this shift in statistical significance.

The interaction between female gender and neuroticism is significant at the 10% level ($p = 0.084$). For instance, women generally report higher neuroticism scores than men [84–86], which is also true for our sample. Given that women generally report higher neuroticism scores than men, this term indicates that the relationship between neuroticism and delegation behavior differs for males and females. Specifically, the effect of neuroticism on delegation behavior may be more substantial for one gender than the other. As a result, the positive coefficient for the female gender in the second model suggests that the likelihood of delegation among females might depend more on their neuroticism level than males.

Another noteworthy interaction term is the one between internal locus of control and openness, which is significant at the 10% level ($p = 0.080$). This interaction suggests that the effect of internal locus of control on delegation behavior is more pronounced for individuals with specific levels of openness. For example, participants with a high internal locus of control and high openness might be more likely to delegate tasks than those with a high internal locus of control and low openness. This finding further emphasizes the importance of considering the interaction effects when examining the relationships between variables and delegation behavior.

We also have fit quantile regression models [87] using the cumulative frequency of delegation for each participant across all periods, removing the time dimension. We employed this method due to the varying relationships between the variables across different parts of the outcome distribution and the lack of normality. The results are summarized in table B3.

Table B3 Quantile Regression results - cumulative delegation frequencies

Variable	Coefficient	Standard Error	p-value
Intercept	0.386	0.28	0.169
Explanation	0.088	0.059	0.138
Payment	-0.367	0.06	*** < 0.001
Automation	0.192	0.06	**0.002
Female	-0.027	0.046	0.558
Age	-0.003	0.005	0.526
STEM	0.078	0.057	0.173
Business & Economics	-0.051	0.055	0.348
Extraversion	0.023	0.017	0.179
Agreeableness	0.011	0.021	0.598
Conscientiousness	0.025	0.024	0.285
Neuroticism	0.002	0.018	0.927
Openness	0.005	0.023	0.841
Internal LoC	-0.02	0.027	0.467
External LoC	0.058	0.028	*0.04
Generalized Trust	0.013	0.019	0.49
Perception	-0.143	0.034	*** < 0.001

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

This model explains approximately 18.95% of the sample variance. Similarly to the logistic regression results, these findings show that the condition involving payment significantly reduces the frequency of delegation ($p < 0.001$), while full automation significantly increases it ($p = 0.002$). Among personal characteristics, only External Locus of Control significantly contributes to delegation, indicating that participants who believe outcomes are beyond their control are more likely to delegate decisions ($p = 0.04$). Moreover, a negative perception of algorithms significantly corresponds to a less frequent delegation of decisions ($p < 0.001$). Other actors such as explanation condition, demographics, Big Five personality traits, Internal Locus of Control, and Trust do not significantly affect the delegation frequency. We have also controlled for correlated variables in this model by adding interaction terms; the results are summarized in table B4.

Upon adding interaction terms, the pseudo-R-squared value rose to 21.01%, showing a marginally improved model fit. Payment ($p < 0.001$) and automation ($p = 0.01$) still significantly influence delegation. Notably, individuals with a STEM background ($p = 0.017$) show a significant positive association with delegation. Openness to experience negatively correlates with delegation ($p = 0.034$). A significant interaction

Table B4 Quantile Regression Results - cumulative delegation frequencies, with interaction terms

Variable	Coefficient	Standard Error	p-value
Intercept	1.206	1.043	0.248
Explanation	0.039	0.057	0.497
Payment	-0.397	0.058	*** < 0.001
Automation	0.151	0.058	*0.01
Female	0.105	0.139	0.45
Age	-0.007	0.005	0.14
STEM	0.131	0.055	*0.017
Business & Economics	-0.034	0.052	0.512
Extraversion	0.011	0.017	0.509
Agreeableness	0.015	0.02	0.451
Conscientiousness	0.058	0.166	0.727
Neuroticism	0.02	0.086	0.814
Openness	-0.22	0.103	*0.034
Internal LoC	-0.258	0.144	0.074
External LoC	0.18	0.157	0.253
Generalized Trust	0.008	0.018	0.672
Perception	-0.102	0.033	**0.002
Female x Neuroticism	-0.032	0.033	0.331
Internal Loc x Conscientiousness	0.016	0.026	0.548
External Loc x Conscientiousness	-0.025	0.025	0.307
External Loc x Neuroticism	-0.003	0.018	0.879
Internal Loc x Openness	0.041	0.02	*0.04

Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

emerges between internal locus of control and Openness ($p = 0.04$): those high in internal locus of control and openness tend to delegate more. A negative view of algorithms remains a strong deterrent to delegation ($p = 0.002$).

Appendix C Technical Remarks

The documented experiment was executed online, programmed with the oTree open-source platform [88]. The data work was performed using Python language. The statistical tests were done using statsmodels [89]. The machine learning models were deployed, tuned, and cross-validated using Scikit-Learn [90]. Both models were tuned using a grid search algorithm with the target to maximize the AUC-ROC. It is important to outline that this is a computationally expensive procedure. The parameter set for the Random Forest model is in table C5.

Parameter	Value	Definition
bootstrap	True	Determines whether or not to use bootstrap samples when building trees
class_weight	balanced_subsample	Adjusts the weights of the classes. balanced_subsample means it computes weights based on the bootstrap sample for every tree
criterion	entropy	Defines the function to measure the quality of a split. entropy is for information gain
max_depth	15	Specifies the maximum depth of the tree
max_features	auto	The number of features to consider when looking for the best split. auto means the square root of the total number of features
min_samples_leaf	1	The minimum number of samples required to be at a leaf node
min_samples_split	min_samples_split	The minimum number of samples required to split an internal node
n_estimators	100	The number of trees in the forest

Table C5 Random Forest Classifier parameters

Similarly, the grid search-generated parameters for the GBM model are described in table C6

Parameter	Value	Definition
learning_rate	0.05	Determines the impact of each tree on the final outcome
max_depth	10	Specifies the maximum depth of the tree
max_features	sqrt	The number of features to consider when looking for the best split. sqrt means the square root of the total number of features
min_samples_leaf	1	The minimum number of samples required to be at a leaf node
min_samples_split	15	The minimum number of samples required to split an internal node
n_estimators	100	The number of boosting stages to perform. Each stage adds a new tree into the ensemble
subsampling	0.7	The fraction of samples to be used for fitting the individual base learners

Table C6 Gradient Boosting Machine Classifier parameters

The cross-validation technique used in both models was the GroupKFold algorithm, which aggregated samples for the same participant. This procedure was performed in both the parameter search and model training steps, using five validation folds.

The uplift random forest classifier was implemented using the causalml library [75]. Since this method, in conjunction with the group cross-validation using synthetic control groups, was performance costly, we implemented a less-exhaustive approach for the parameter-fitting method, using the Optuna library [91]. It employs efficient search algorithms, such as Tree-structured Parzen Estimator (TPE). We ran an optimization study for 150 trials and selected the parameter set that yielded satisfactory AUUC scores. One important remark here is that calculating the AUUC in this way might produce abnormally high results due to the stochastics in place, so practitioners might have to supervise the optimization process. Table C7 describes the parameter values.

Parameter	Value	Definition
n_estimators	850	The number of trees in the forest
max_depth	8	The maximum depth of each decision tree
max_features	9	The number of features to consider when looking for the best split
min_samples_leaf	45	The minimum number of samples required to be at a leaf node
min_samples_treatment	15	The minimum number of samples in a leaf node that come from the treatment group
n_reg	14	The regularization parameter used in the causal tree procedure
evaluationFunction	Chi	The evaluation function used to evaluate splits

Table C7 Uplift Random Forest Classifier parameters

Appendix D Experiment Design Screens

In this appendix session, we added the most important screens for the experiment. Figure D3 contains the main task screens for each treatment. Figure D4 shows the attention questions.

References

- [1] Russell, S.J.: Artificial Intelligence a Modern Approach, 4th edn. Pearson Education, Inc., London (2010)

Instructions

In this experiment, you will be presented with three product choices, which will be named generically.

Product Selection

Each of the products has a **quality level**, which will be randomly allocated between the three alternatives. During the experiment, you will have to select one of the products, and you will receive a payoff according to your selection, as in the example below:

Product Selection

Round 1 out of 5
Total payoff: 0

Below you find three products of different qualities. You must decide for one of them. The product quality is associated with higher probabilities of receiving a payoff. *Please choose your preferred option.*



Product 1



Product 2



Product 3

The product quality levels introduce different probabilities of acquiring points as a payoff, defined as:

- **High quality** – 0.9 probability of receiving 1 point
- **Medium quality** – 0.7 probability of receiving 1 point
- **Low quality** – 0.5 probability of receiving 1 point

You will have the opportunity to perform repeated product selections for 40 rounds. The product qualities will remain unchanged across the rounds after initial random allocation. For example, if your quality allocation was Product 1: High, Product 2: Low, and Product 3: Medium; these values will remain throughout the experiment.

If you learn which product is associated with which quality level, you will have higher chances of better payoffs.

Decision Delegation

In each round, you will have the opportunity to delegate your decision to an algorithm, which will select the product for you. If you choose to delegate, you get feedback on the selection of the algorithm and move to the next round. If you do not delegate, you can make the product selection yourself.

Questionnaire

When the experiment is finished, you will be redirected to a questionnaire with simple multiple choice questions. After that, you will see the conclusion page, where your results will be shown, as well as the final results and your earnings.

To start with the experiment, please click "next".

Fig. D2 General Instructions Screen

- [2] McCarthy, J.: What is artificial intelligence (2007)
- [3] Azucar, D., Marengo, D., Settanni, M.: Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences* **124**, 150–159 (2018)
- [4] Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* **144**(1), 114 (2015)
- [5] Ku, C.-Y.: When ais say yes and i say no: On the tension between ai's decision and human's decision from the epistemological perspectives. *Információs Társadalom* **19**(4), 61–76 (2020)

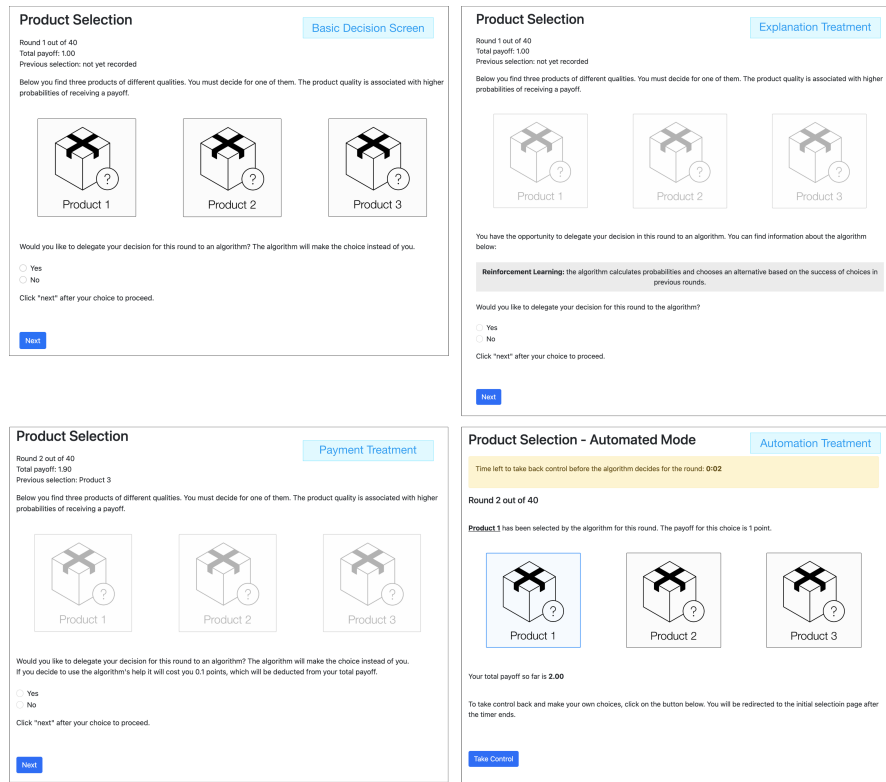


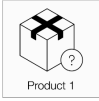
Fig. D3 Main Task Experiment Screens

- [6] Mahmud, H., Islam, A.N., Ahmed, S.I., Smolander, K.: What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change* **175**, 121390 (2022)
- [7] Burton, J.W., Stein, M.-K., Jensen, T.B.: A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* **33**(2), 220–239 (2020)
- [8] Chugunova, M., Sele, D.: We and it: An interdisciplinary review of the experimental evidence on how humans interact with machines. *Journal of Behavioral and Experimental Economics* **99**, 101897 (2022)
- [9] Filiz, I., Judek, J.R., Lorenz, M., Spiwox, M.: Algorithm aversion as an obstacle in the establishment of robo advisors. *Journal of Risk and Financial Management* **15**(8), 353 (2022)
- [10] Logg, J.M.: Theory of machine: When do people rely on algorithms? Harvard Business School working paper series# 17-086 (2017)


Product Selection

Round 3 out of 4
Total payoff: 100
Previous selection: Product 2


Below you find three products of different qualities. You must decide for one of them. The product quality is associated with higher probabilities of receiving a payoff.



Product 1



Product 2




Product 3

Would you like to delegate your decision for this round to an algorithm? The algorithm will make the choice instead of you.

Yes
 No

Click "next" after your choice to proceed.



The image is not relevant for the task.

[Next](#)

Questions

Below you see a number of statements, each of which starts with "I see myself as someone who". For each statement, indicate how much you agree with this.

I see myself as someone who...	Strongly Disagree	Disagree	Somewhat Disagree	Neither agree nor disagree	Somewhat Agree	Agree	Strongly Agree
worries a lot	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
gets nervous easily	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
remains calm in tense situations	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
is talkative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
control question, please select somewhat agree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
is outgoing, sociable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
is reserved	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
is original, comes up with new ideas	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

[Next](#)

Attention Question

Please answer the short question below before proceeding to the final questionnaire. This question will not affect your earnings in any way, so please answer it honestly.

A picture of an animal was shown in one of the rounds, can you name which animal it was?

I don't know
 Dog
 Lion
 Bear
 Snake
 Cat
 Elephant
 Parrot
 Lizard

[Next](#)

Fig. D4 Attention Measures

- [11] Gaudeul, A., Giannetti, C.: Trade-offs in the design of financial algorithms. Available at SSRN 4432707 (2023)
- [12] Filiz, I., Judek, J.R., Lorenz, M., Spiwoкс, M.: The tragedy of algorithm aversion (2021)
- [13] Berger, B., Adam, M., Rühr, A., Benlian, A.: Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering* **63**(1), 55–68 (2021)
- [14] Jung, M., Seiter, M.: Towards a better understanding on mitigating algorithm aversion in forecasting: an experimental study. *Journal of Management Control* **32**(4), 495–516 (2021)
- [15] Jauernig, J., Uhl, M., Walkowitz, G.: People prefer moral discretion to algorithms: algorithm aversion beyond intransparency. *Philosophy & Technology* **35**(1), 1–25 (2022)
- [16] Jago, A.S., Laurin, K.: Assumptions about algorithms' capacity for discrimination. *Personality and Social Psychology Bulletin* **48**(4), 582–595 (2022)
- [17] Logg, J.M., Minson, J.A., Moore, D.A.: Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* **151**, 90–103 (2019)

- [18] Hou, Y.T.-Y., Jung, M.F.: 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), 1–25 (2021)
- [19] Candrian, C., Scherer, A.: Rise of the machines: Delegating decisions to autonomous ai. *Computers in Human Behavior* **134**, 107308 (2022)
- [20] Jussupow, E., Benbasat, I., Heinzl, A.: Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion (2020)
- [21] Robbins, H.: Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58**(5), 527–535 (1952)
- [22] Hoelzemann, J., Klein, N.: Bandits in the lab. *Quantitative Economics* **12**(3), 1021–1051 (2021)
- [23] De Bruyn, A., Viswanathan, V., Beh, Y.S., Brock, J.K.-U., Wangenheim, F.: Artificial intelligence and marketing: Pitfalls and opportunities. *Journal of Interactive Marketing* **51**, 91–105 (2020)
- [24] Zhang, Y., Chen, X., *et al.*: Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* **14**(1), 1–101 (2020)
- [25] Enholm, I.M., Papagiannidis, E., Mikalef, P., Krogstie, J.: Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, 1–26 (2021)
- [26] Trocin, C., Mikalef, P., Papamitsiou, Z., Conboy, K.: Responsible ai for digital health: a synthesis and a research agenda. *Information Systems Frontiers*, 1–19 (2021)
- [27] Vlačić, B., Corbo, L., Silva, S.C., Dabić, M.: The evolving role of artificial intelligence in marketing: A review and research agenda. *Journal of Business Research* **128**, 187–203 (2021)
- [28] Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* **267**, 1–38 (2019)
- [29] Herm, L.-V., Heinrich, K., Wanner, J., Janiesch, C.: Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability. *International Journal of Information Management*, 102538 (2022)
- [30] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA),

pp. 80–89 (2018). IEEE

- [31] Gunning, D.: Explainable artificial intelligence (xai). Defense advanced research projects agency (DARPA), nd Web **2**(2), 1 (2017)
- [32] Castelluccia, C., Le Métayer, D.: Understanding algorithmic decision-making: Opportunities and challenges (2019)
- [33] Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine* **38**(3), 50–57 (2017)
- [34] König, P.D., Wurster, S., Siewert, M.B.: Consumers are willing to pay a price for explainable, but not for green ai. evidence from a choice-based conjoint analysis. *Big Data & Society* **9**(1), 20539517211069632 (2022)
- [35] Ben-David, D., Sade, O.: Robo-advisor adoption, willingness to pay, and trust—before and at the outbreak of the covid-19 pandemic. *Willingness to Pay, and Trust—Before and at the Outbreak of the COVID-19 Pandemic* (May 1, 2001) (2001)
- [36] Wedel, P., Hagist, C.: Physicians’ preferences and willingness to pay for artificial intelligence-based assistance tools: a discrete choice experiment among german radiologists. *BMC Health Services Research* **22**(1), 1–14 (2022)
- [37] Bogert, E., Schechter, A., Watson, R.T.: Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports* **11**(1), 1–9 (2021)
- [38] Bucklin, R., Lehmann, D., Little, J.: From decision support to decision automation: A 2020 vision. *Marketing Letters* **9**, 235–246 (1998)
- [39] Sharan, N.N., Romano, D.M.: The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon* **6**(8), 04572 (2020)
- [40] Goldbach, C., Kayar, D., Pitz, T., Sickmann, J.: Transferring decisions to an algorithm: A simple route choice experiment. *Transportation research part F: traffic psychology and behaviour* **65**, 402–417 (2019)
- [41] Fenneman, A., Sickmann, J., Pitz, T., Sanfey, A.G.: Two distinct and separable processes underlie individual differences in algorithm adherence: Differences in predictions and differences in trust thresholds. *Plos one* **16**(2), 0247084 (2021)
- [42] Goldbach, C., Sickmann, J., Pitz, T., Zimasa, T.: Towards autonomous public transportation: attitudes and intentions of the local population. *Transportation Research Interdisciplinary Perspectives* **13**, 100504 (2022)
- [43] Goldberg, L.R.: An alternative” description of personality”: the big-five factor structure. *Journal of personality and social psychology* **59**(6), 1216 (1990)

- [44] Rotter, J.B.: Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied* **80**(1), 1 (1966)
- [45] Yamagishi, T., Yamagishi, M.: Trust and commitment in the united states and japan. *Motivation and emotion* **18**(2), 129–166 (1994)
- [46] Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: *Introduction to Algorithms*, 2nd edn. The MIT Press, Cambridge, Massachusetts (2001)
- [47] Barto, A.G.: Reinforcement learning. In: *Neural Systems for Control*, pp. 7–30. Elsevier, Hoboken, New Jersey, USA. (1997)
- [48] Sutton, R.S., Barto, A.G.: *Reinforcement Learning: An Introduction*. MIT press, Cambridge Massachussets (2018)
- [49] Erev, I., Roth, A.E.: Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 848–881 (1998)
- [50] Camerer, C., Hua Ho, T.: Experience-weighted attraction learning in normal form games. *Econometrica* **67**(4), 827–874 (1999)
- [51] Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* **47**(260), 583–621 (1952)
- [52] Dunn, O.J.: Multiple comparisons among means. *Journal of the American statistical association* **56**(293), 52–64 (1961)
- [53] Bertrand, M., Duflo, E., Mullainathan, S.: How much should we trust differences-in-differences estimates? *The Quarterly journal of economics* **119**(1), 249–275 (2004)
- [54] Camerer, C.F.: Artificial intelligence and behavioral economics. In: *The Economics of Artificial Intelligence: An Agenda*, pp. 587–608. University of Chicago Press, Chicago, IL 60637 U.S.A. (2018)
- [55] Balakrishnan, V., Khan, S., Fernandez, T., Arabnia, H.R.: Cyberbullying detection on twitter using big five and dark triad features. *Personality and individual differences* **141**, 252–257 (2019)
- [56] Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* **18**, 43–49 (2017)
- [57] Stachl, C., Hilbert, S., Au, J.-Q., Buschek, D., De Luca, A., Bischl, B., Hussmann, H., Bühner, M.: Personality traits predict smartphone usage. *European Journal of Personality* **31**(6), 701–722 (2017)

- [58] Saltık, Ö., Söyü, R., Değirmen, S., Şengönül, A., *et al.*: Predicting loss aversion behavior with machine-learning methods. *Humanities and Social Sciences Communications* **10**(1), 1–14 (2023)
- [59] Kahneman, D.: *Thinking, Fast and Slow*. macmillan, United Kingdom (2011)
- [60] Breiman, L.: Random forests. *Machine learning* **45**, 5–32 (2001)
- [61] Liaw, A., Wiener, M., *et al.*: Classification and regression by randomforest. *R news* **2**(3), 18–22 (2002)
- [62] Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232 (2001)
- [63] Kohavi, R., *et al.*: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai*, vol. 14, pp. 1137–1145 (1995). Montreal, Canada
- [64] Berrar, D.: Cross-validation. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (eds.) *Encyclopedia of Bioinformatics and Computational Biology*, pp. 542–545. Academic Press, Oxford (2019). <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>
- [65] Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. *Information processing & management* **45**(4), 427–437 (2009)
- [66] Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020)
- [67] Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition* **30**(7), 1145–1159 (1997)
- [68] Fawcett, T.: An introduction to roc analysis. *Pattern recognition letters* **27**(8), 861–874 (2006)
- [69] Radcliffe, N.J., Surry, P.D.: Real-world uplift modelling with significance-based uplift trees. *White Paper TR-2011-1, Stochastic Solutions*, 1–33 (2011)
- [70] Radcliffe, N., Surry, P.: Differential response analysis: Modeling true responses by isolating the effect of a single action. *Credit Scoring and Credit Control IV* (1999)
- [71] Gutierrez, P., Gérardy, J.-Y.: Causal inference and uplift modelling: A review of the literature. In: *International Conference on Predictive Applications and APIs*, pp. 1–13 (2017). PMLR
- [72] Guelman, L., Guillén, M., Pérez-Marín, A.M.: Random forests for uplift modeling: an insurance customer retention case. In: *International Conference on Modeling*

and Simulation in Engineering, Economics and Management, pp. 123–133 (2012). Springer

- [73] Guelman, L., Guillén, M., Pérez-Marín, A.M.: Uplift random forests. *Cybernetics and Systems* **46**(3-4), 230–248 (2015)
- [74] Sołtys, M., Jaroszewicz, S., Rzepakowski, P.: Ensemble methods for uplift modeling. *Data mining and knowledge discovery* **29**, 1531–1559 (2015)
- [75] Chen, H., Harinen, T., Lee, J.-Y., Yung, M., Zhao, Z.: Causalml: Python package for causal machine learning. arXiv preprint arXiv:2002.11631 (2020)
- [76] Glikson, E., Woolley, A.W.: Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* **14**(2), 627–660 (2020)
- [77] Prah, A., Van Swol, L.: Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting* **36**(6), 691–702 (2017)
- [78] Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., Cagan, J.: Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior* **127**, 107018 (2022)
- [79] Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60 (1947)
- [80] Snijders, C., Conijn, R., Fouw, E., Berlo, K.: Humans and algorithms detecting fake news: Effects of individual and contextual confidence on trust in algorithmic advice. *International Journal of Human–Computer Interaction* **39**(7), 1483–1494 (2023)
- [81] Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* vol. 2. Springer, New York (2009)
- [82] Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66**(5), 688 (1974)
- [83] Rzepakowski, P., Jaroszewicz, S.: Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems* **32**, 303–327 (2012)
- [84] Costa Jr, P.T., Terracciano, A., McCrae, R.R.: Gender differences in personality traits across cultures: robust and surprising findings. *Journal of personality and social psychology* **81**(2), 322 (2001)
- [85] Schmitt, D.P., Realo, A., Voracek, M., Allik, J.: Why can't a man be more like

- a woman? sex differences in big five personality traits across 55 cultures. *Journal of personality and social psychology* **94**(1), 168 (2008)
- [86] Weisberg, Y.J., DeYoung, C.G., Hirsh, J.B.: Gender differences in personality across the ten aspects of the big five. *Frontiers in psychology*, 178 (2011)
- [87] Koenker, R., Bassett Jr, G.: Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50 (1978)
- [88] Chen, D.L., Schonger, M., Wickens, C.: oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance* **9**(C), 88–97 (2016) <https://doi.org/10.1016/j.jbef.2015.12.00>
- [89] Seabold, S., Perktold, J.: statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)
- [90] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
- [91] Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631 (2019)