

RESEARCH ARTICLE

# Algorithm, expert, or both? Evaluating the role of feature selection methods on user preferences and reliance

Jaroslav Kornowicz<sup>ID</sup>\*, Kirsten Thommes<sup>ID</sup>

Faculty of Business Administration and Economics, Paderborn University, Paderborn, Germany

\* [jaroslav.kornowicz@uni-paderborn.de](mailto:jaroslav.kornowicz@uni-paderborn.de)



## Abstract

The integration of users and experts in machine learning is a widely studied topic in artificial intelligence literature. Similarly, human-computer interaction research extensively explores the factors that influence the acceptance of AI as a decision support system. In this experimental study, we investigate users' preferences regarding the integration of experts in the development of such systems and how this affects their reliance on these systems. Specifically, we focus on the process of feature selection—an element that is gaining importance due to the growing demand for transparency in machine learning models. We differentiate between three feature selection methods: algorithm-based, expert-based, and a combined approach. In the first treatment, we analyze users' preferences for these methods. In the second treatment, we randomly assign users to one of the three methods and analyze whether the method affects advice reliance. Users prefer the combined method, followed by the expert-based and algorithm-based methods. However, the users in the second treatment rely equally on all methods. Thus, we find a remarkable difference between stated preferences and actual usage, revealing a significant attitude-behavior-gap. Moreover, allowing the users to choose their preferred method had no effect, and the preferences and the extent of reliance were domain-specific. The findings underscore the importance of understanding cognitive processes in AI-supported decisions and the need for behavioral experiments in human-AI interactions.

## OPEN ACCESS

**Citation:** Kornowicz J, Thommes K (2025) Algorithm, expert, or both? Evaluating the role of feature selection methods on user preferences and reliance. PLoS ONE 20(3): e0318874. <https://doi.org/10.1371/journal.pone.0318874>

**Editor:** Sohaib Mustafa, Beijing University of Technology, CHINA

**Received:** August 5, 2024

**Accepted:** January 22, 2025

**Published:** March 7, 2025

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pone.0318874>

**Copyright:** © 2025 Kornowicz, Thommes. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data availability statement:** Experimental data and analysis scripts can be found at [https://osf.io/z2xpy/?view\\_only=90607651bed949d29593c4a176d6c96d](https://osf.io/z2xpy/?view_only=90607651bed949d29593c4a176d6c96d).

## Introduction

As artificial intelligence (AI) becomes increasingly powerful through advances in computing power, improved algorithms, and the availability of more data, its prevalence expands across a wide array of fields and life situations [1–5]. In response to this growing ubiquity, recent research efforts have shifted from solely focusing on improving the accuracy of AI models to addressing the interaction with a more diverse and heterogeneous user base, exploring the potential consequences of AI adoption and understanding users' preferences and concerns [6].

**Funding:** We gratefully acknowledge funding by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG): TRR 318/1 2021 – 438445824.

**Competing interests:** NO authors have competing interests.

One strand of research focuses on the human user and has observed that user reliance on algorithmic decision aids is not uniform and is influenced by various factors [7,8] such as the user's personality, algorithm design, task factors, and high-level factors as organizational and societal aspects. The literature surrounding "algorithm aversion" has documented a stated preference among users for human decision-making over algorithmic advice and has noted that individual aspects of AI systems can impact trustworthiness and reliance [7–10]. However, these results encounter resistance, often described as "algorithm appreciation" that observes the converse—a stated preference in favor of algorithms [11,12].

Another stream of research has concentrated on the system, enhancing transparency and explainability as methods to make AI more accessible, comprehensible, and reliable [13]. Legal institutions also drive this research landscape. The increasing presence of AI in society has prompted governments to establish requirements for greater transparency [14,15]. These regulations have led to "black box" models becoming more informative to end users, with implications for AI reliance among all stakeholders. In addition, interdisciplinary efforts between computer scientists, social scientists, and ethicists are increasingly encouraged to tackle the complex challenges posed by AI integration in society [16,17].

Instead of explaining the model or the outcome, recent research discusses other means of quality control during the development of the AI system, e.g., adding human agency. The basic idea here is that not every user must be able to understand the system, but that experts, e.g., domain experts, are involved in the process of machine learning (ML) development, supervise the system, and add human expert knowledge—resulting in a more trustworthy ML models for every end user [18–20].

Previous research has highlighted the significance of human involvement and its effect on users' perceptions, preferences, and reliance. It can be categorized in two ways: involvement in the development and training (typically beyond the scope of the user) and the degree to which humans can apply AI, giving the user options on how to utilize recommendations for their decisions [7]. Limited research has been directed towards the former. Ashoori and Weisz [9] and Jago [21] demonstrated that users tend to favor models trained by data scientists or experts instead of those trained autonomously, without explicitly specifying the nature of the involvement. In a recent study that inspired our work, Cheng and Chouldechova [22] involved users at various stages. They discovered that permitting users to select the training algorithm can mitigate aversion, whereas modifying the inputs does not. While a detailed description of human involvement may not be necessary in many cases, it can be essential in highly transparent models, where features are readily visible, such as in scoring systems [23]. The literature review by Jussupow et al. [7] reveals that it is important to note that human responses differ between the stated preferences and the chosen behavioral response, i.e. their actual reliance. While many studies find a strong preference for human oversight, the revealed preferences in terms of actual behavior as less clear. In our study, we set out to analyze whether stated and revealed preferences are aligned.

Although there are many areas for human involvement, in this paper we focus on the role of human involvement within feature selection. Feature selection is a pivotal step in the machine learning pipeline. It involves identifying the most relevant variables from the input data, which can significantly impact the predictive performance and interpretability of the resulting model [24,25]. Algorithmic feature selection methods are often criticized for lacking theoretical or expert knowledge. Consequently, many scholars argue for human-based feature selection methods or a collaboration of algorithms and humans for feature selection and other machine learning processes [25–27]. We contribute to answering this call.

In our study, we distinguish three methods of feature selection: algorithm-based feature selection (*Algorithm*), expert-based feature selection (*Expert*), and a combined approach (*Combination*). We seek to answer three research questions:

- 1) What kind of feature selection method do users prefer?
- 2) Does the feature selection method affect reliance?
- 3) Does allowing the user to choose their preferred method affect reliance?

Yet, as far as we know, the question of how feature selection modes contribute to AI reliance has not been systematically analyzed. Nonetheless, feature selection and human preferences for feature selection mechanisms are crucial to understanding a model. The novelty of our study lies in addressing the gap in the literature by examining the effects of different levels of human integration in feature selection on user preferences and reliance.

To answer our questions, we conducted an online study involving 216 participants. Our results reveal that *Combination* was the most preferred, followed by *Expert* and *Algorithm*. However, these relationships vary depending on the task domain. Interestingly, stated preferences do not correlate with behavioral reliance, similar to previous studies [28,29]. In a second treatment, we randomly allocate a new group of users to models whose features are either selected by *Expert*, *Algorithm*, or a *Combination*. We observe no significant effect of the underlying feature selection methods on advice reliance. Moreover, the involvement of participants in choosing their preferred feature selection method does not affect the reliance. Reliance is also different across domains. We find a significantly higher probability of reliance in the medical domain compared to a sports-related domain. Concerning individual differences, we observe that participants displaying higher risk-taking tendencies prefer *Algorithm* and *Combination* over *Expert*.

Our study underscores the value of behavioral experiments with incentivized tasks in understanding human-AI collaboration. It points to the importance of further examining cognitive processes in decision-making with AI assistance and stresses the challenge and importance of considering domain-specific effects.

## Related work

### Feature selection

A critical process in developing ML models is feature selection [30]. Features, also called predictors, variables, dimensions, or inputs, can be defined as measurable properties or characteristics of observed procedures or entities [31,32]. Selecting an appropriate subset of features for an ML model can significantly impact its performance, interpretability, computation time, and overfitting risk [33]. This is especially relevant for high-dimensional datasets, which may contain irrelevant and redundant features that negatively affect the quality of the learned models for stakeholders [34]. Feature selection can be used for simple tabular datasets, but also for image data, for example, to improve super-resolution algorithms [35] or computer-aided diagnosis for glaucoma identification [36] and cancer prediction [37].

The domain of feature selection is extensively studied, with the development of various automated algorithms that aim to select relevant feature subsets from datasets [38]. Feature selection techniques driven by data can be generally divided into three categories: filter methods that assess features solely based on the data; wrapper methods that select features through the predictive capability of a machine learning algorithm; and embedded approaches such as LASSO regression that come with inherent feature selection processes [24]. There are also

hybrid methods that show great promise, indicating that research in this area continues to grow [39].

Equally relevant to our research is incorporating human knowledge in feature selection, sourced directly from domain specialists or literature. For instance, Naher et al. [40] demonstrated that features based on a literature review significantly improved the accuracy of a heart disease classifier. Human knowledge-driven feature selection can involve researching relevant scholarly literature [40–42] or consulting domain experts [43,44]. These approaches are particularly important for model explainability, ensuring that the selected features do not contradict human knowledge [45].

It is also feasible to combine various approaches. Multiple feature sets, potentially sourced from different origins, can be aggregated into a singular final set [46,47]. Additionally, there are interactive methodologies wherein humans and algorithms collaborate iteratively [48,49]. Determining the superior approach among data-driven, knowledge-driven, aggregated, or interactive methods is challenging due to the variety of data sets and the vast array of potential combinations [41].

## Human-AI collaboration

Human decision-makers receiving advice from algorithmic systems is not new and has been studied for many decades [50]. With AI systems' increasing power and practicality, it has found their way into more and more domains, often surpassing human judgment, even with simple methods [51,52]. While they are not infallible, relying solely on them might yield better results when human decision-making is generally less accurate. Yet, this approach will still fall short of the optimal scenario where human and AI decision-making are complementary [53,54].

Despite the potential benefits of incorporating algorithmic advice in decision-making processes, many individuals reject such recommendations [10,55], leading to an under-reliance on the advice and, therefore, often to a decreased decision-making performance [56]. The phenomenon of advice aversion has been extensively studied in human-to-human interactions [57] and, more recently, between humans and AI [7,8]. Algorithm aversion, as defined by Mahmud et al. [8], refers to neglecting algorithmic decisions in favor of one's own decisions or those of others, consciously or unconsciously. The antithesis of algorithm aversion is algorithm appreciation and automation bias [11], potentially causing decision-makers to over-rely on algorithmic advice. This divergence between aversion and appreciation could be partly attributed to the task's nature. Factors such as whether the task appears more objective or subjective from a human perspective [10], or if the employment of algorithms aligns with prevailing social norms [58], may play significant roles. Recent studies have explored methods to mitigate over- and under-reliance, such as employing cognitive-forcing functions [59] and providing XAI explanations [54] with mixed results. For an overview of empirical work on human-AI decision-making, we recommend a recent review by Lai et al. [60].

In this regard, we adopt the definition of reliance provided by Scharowski et al. [61], which describe it as “*a user's behavior that follows from the advice of the system*”. We emphasize that we are not concerned with whether the reliance is *appropriate* or not: In contexts where humans receive advice from AI, decision-making performance can surpass that of individuals only when the human accurately discerns and adheres to correct advice while disregarding erroneous suggestions [53]. Our study's objective is not to enhance the performance of AI-assisted decision-making by optimizing or calibrating the decision makers' reliance or trust [62]. Instead, we view feature selection as a potential factor influencing reliance that could be considered in optimizing advice-giving systems.

To better understand the factors influencing advice-taking interactions between humans and AI, numerous studies have investigated the effects of different AI aspects and advice-taker characteristics. Sundar [19], in his framework for studying human-AI interactions, argues that AI elements can serve as cues that trigger cognitive heuristics during an interaction. These heuristics, which he refers to as “machine heuristics,” can be perceived positively or negatively and depend on individual differences [63]. In their review, Mahmud et al. [8] group influencing factors into four categories: task factors (e.g., subjectivity and morality), high-level factors (e.g., social norms), individual factors (e.g., fear of change, expertise, and demographics), and algorithmic factors (e.g., explainability, accuracy, and integration). Jussupow et al. [7] similarly categorize factors into algorithm characteristics (agency, performance, capabilities, and human involvement) and human agent characteristics (social distance and expertise). Our study focuses explicitly on the feature selection method as a factor. This process is categorized under algorithmic factors and characteristics. It is also related to the category of human involvement in AI systems. In our case, this involves integrating humans as experts and decision-makers in the feature selection process and also the later interaction between decision-maker and AI.

Jussupow et al. [7] emphasize distinguishing who is involved in the machine learning pipeline, whether it is the later end-user or a human developer (e.g., a data scientist) integrated into the development process. Experiments by Jago [21] demonstrate that expert involvement in the training process can enhance algorithm authenticity. Interestingly, participants tend to prefer models trained by data scientists over purely automated methods, as observed by Ashoori and Weisz [9], and they do not even differentiate between prestigious and non-prestigious institutional affiliations [64]. Palmeira and Spassova [65] found that people prefer a combination of expert judgment and decision aid over expert judgment alone. Their results are similar to Waddell's [20], who investigated the differences in the perception of human and algorithmic authors of journalistic articles and found that biases are attenuated when humans and algorithms work in tandem. Lastly, Cheng and Chouldechova [22] investigate three ways in which humans can control AI decisions: altering the input, controlling the process (e.g., the learning algorithm), and adjusting the output for the final decision (the most common type of control in the literature). They found that process and output control reduce algorithm aversion while input modification does not.

Literature exploring algorithm appreciation and aversion suggests that decision-makers favor human involvement in the machine learning process and that human involvement decreases algorithm aversion. Consequently, we hypothesize that when given a choice, users of machine learning models are more inclined to prefer an machine learning model that uses features selected by experts rather than by an algorithm.

**H1a:** *A expert feature selection method is chosen more frequently than a algorithmic feature selection method.*

A machine learning model that uses a combination of an expert and algorithm feature selection method can be perceived as a “tandem,” similar to what Waddell's study showed about the joint effort of algorithms and humans [20]. The involvement of two parties in this process may lead to a cumulative [18] or a “double-dose” effect [66]. Echoing Palmeira's and Spassova's [65] findings, which suggest a preference for combined efforts over sole expert judgment, we hypothesize that the model utilizing a combined method will be more favored than the expert method. Furthermore, we believe that its advice will likely garner the highest level of reliance.

**H1b:** *A combination of expert and algorithmic feature selection methods is chosen more frequently than an expert feature selection method alone.*

We also think that these preferences can be transferred to reliance, allowing us to formulate hypotheses accordingly:

**H2a:** *Advice generated using an expert feature selection method exhibits higher reliance rates than those generated with an algorithmic feature selection method.*

**H2b:** *Advice generated using a combination of expert and algorithmic feature selection methods exhibit higher reliance rates than those generated with an expert feature selection method alone.*

We excluded a variety of feature selection methods here, as we are primarily focused on the different levels of human involvement, and thus concentrate on three distinct stages.

Permitting user to choose their preferred feature selection method introduces a form of control akin to the experiments conducted by Cheng and Chouldechova [22]. Although their results suggest that allowing decision-makers to control the process should increase reliance, feature selection only influences the input, not the processing of information, which may not affect reliance. Kawaguchi [67] found that workers were more receptive to advice when their predictions were considered. An experiment by Köbis and Mossink [68] found that when participants' opinions were incorporated into the decision-making process, it decreased AI aversion. Burton et al. [69] posit that human-in-the-loop decision-making or even an illusion of autonomy can mitigate algorithm aversion. Other factors may explain why the participant's choice might influence reliance positively. For example, the sunk cost fallacy suggests that participants who have invested time and effort in choosing a feature selection method may be more inclined to rely on the model's predictions to justify their initial choice [70].

**H3:** *Giving the users choice to choose their preferred feature selection method positively increases the reliance on the machine learning model's advice.*

## Methods

We employ a behavioral experiment with a between-subject design and two treatments. Our experimental design draws inspiration from prior research on human-AI decision-making processes [60]. It incorporates two distinct decision-making domains: *Cardio*, which focuses on medical diagnoses, and *Football*, which centers around estimating soccer match outcomes. In the first treatment *Choice*, we investigate the decision-maker's preference for these methods when given a choice. Second, we compare this group with another treatment group *No Choice*, which had no option to choose their preferred method. The *No Choice* treatment has three sub-treatments: a human selects features, a data-driven algorithm selects, or feature selection results from a joint effort. We assess the decision-maker's reliance on algorithmic advice in all settings. Do people also prefer ex-ante to what they will rely on ex-post?

Moreover, in an exploratory manner, we examine the correlation between the characteristics of decision-makers and their preferences and reliance on advice. By identifying personality traits related to preference and reliance, we aim to augment the existing literature that has predominantly centered on general trust and reliance rather than specific aspects like feature selection [8,29,71,72]. Hyperlinks to the experimental data can be found in Data within [S1 Data](#).

## Participants and treatments

**Participants.** A total of 265 participants were recruited from Prolific.com between August 2nd and 18th, 2023. The participants were informed about the study and data protection before the start of the experiments and gave their consent digitally; otherwise, they could not participate. The Paderborn University Institutional Review Board approved the study as part of the research project. Each participant provided voluntary and digital consent before the



start of the experiment. Initially, 16 participants were excluded due to failing an initial comprehension check, while another 29 withdrew. Additionally, 4 participants were removed after failing attention checks. Consequently, the final sample comprised 216 participants for analysis. 129 (59.7%) were women, and the average age was 34.2. Participants required, on average, 27.3 minutes to finish the study and earned an average payment of £9.63. We exclusively recruited participants from the United Kingdom to ensure English language proficiency and a higher likelihood of a basic understanding of football, one of the task domains. Upon completing the study, participants received a fixed payment of £5. Additionally, participants received bonus payments contingent upon the accuracy of their decisions.

**Treatments.** 109 participants were randomly assigned to the *Choice* treatment. In this treatment, participants determined who would be responsible for selecting the features upon which the advising AI is trained for both task domains. The remaining 107 participants were assigned to the *No Choice* treatment. Unlike the other treatments, they were not given a choice between methods; instead, they were randomly allocated to one.

## Experimental procedure

The experimental software for this study was developed using oTree [73] and was deployed online. Participants were required to access the study through a desktop client to minimize the risk of distractions and technical issues. The experiment itself is an incentivized behavioral experiment that adheres to design principles found in related literature [60,74,75].

The study began with an explanation of the data protection policy, followed by the general instructions for the study (see Instructions in [S1 Text](#)). Participants were then presented with multiple comprehension questions, with a maximum allowance of two incorrect responses for each question.

The main component of the study is the experiment, including the classification tasks and an advice-giving AI. Screenshots of the classification task and advice-giving can be found in [S1 Fig](#) and [S2 Fig](#). Participants were asked to perform multiple binary classification tasks, wherein they were provided with information on decision problems and required to submit answers. Participants were awarded additionally £0.20 for each correctly solved task. Upon completion, participants completed a survey to collect demographic and personality information.

**Judge-advisor system.** A Judge-Advisor System (JAS), commonly employed in advice-taking research, was utilized in the experiment [57]. Within the JAS, the participant (acting as the decision-maker) is presented with a decision problem. The participant makes an initial decision based on the information provided for the problem. After submitting this initial decision, an advisor (in this case, a machine learning model) offers advice. The participant then makes a subsequent decision, allowing them to reconsider and possibly modify their initial decision by incorporating the advice as they see fit. Moreover, for each initial decision, participants were prompted to rate their confidence on a slider input ranging from 0 (absolutely not confident) to 100 (very confident), with the default value set to 0 [76]. It is central to note that the decision and the advice are presented on the same scale. Screenshots of the decision pages can be found in the [S1 Fig](#) and [S2 Fig](#).

A subtle but important distinction between our study and many prior studies in the JAS literature is that advice was provided only when they deviated from the initial decision. In other JAS experiments, the decision problems often involve regression tasks with cardinal answers, making it more likely for discrepancies between the participant's decision and the advice. However, since our study focuses on binary decisions, offering advice that aligns with

the initial decision seems redundant and offers little to no insight [53]. In a pre-study involving ten students, we observed that when their initial decision matched the advice, an alternation of the participants' decisions did not happen. This appears quite logical: typically, one would only diverge from the advice (that mirrors their own belief) if there's a firm conviction of its inaccuracy. Omitting advice when the advice would only confirm the respondents' initial choice was more efficient. Participants learned they would only revive advice when their initial choice and that AI recommendations would diverge. Participants were briefed about this approach in the instructions.

## Classification domains and machine learning models

**Domains and tasks.** To guarantee the generalizability of our study and reduce the influence of domain-specific effects, we utilized two distinct domains for the decision problem tasks that participants performed during the experiment. These two problems, labeled as and are derived from publicly available datasets.

The *Cardio* problem is a classification task that involves predicting the presence of cardiovascular disease using patient characteristics and symptoms. The dataset for this problem consists of 70,000 patients. The second classification problem, *Football*, focuses on determining whether the home team in a football match won or not, based on match statistics. The original dataset contains 4,070 matches.

These datasets were selected carefully to ensure comprehensibility for the experiment's participants regarding the decision problem and the incorporated features. Furthermore, we sought a diverse set of domains to avoid domain-specific results, as the domain can influence advice reliance due to different task-related factors. For instance, humans exhibit higher aversion for tasks perceived as more subjective than objective [10,77] or when facing morally relevant decisions, particularly in legal or medical fields [78].

We opted for 20 tasks for each domain to allow participants to become more familiar with the decision problem and experience multiple advice-receiving instances. Previous studies have observed that algorithm aversion tends to weaken over time [79]; thus, incorporating multiple tasks should enhance the reliability of our results. Participants were neither provided with feedback about the correctness of their decisions between rounds nor the accuracy of the ML models. This was an intentional choice to focus on the immediate effects of feature selection methods on user preferences and reliance without introducing additional variables that could influence behavior. Providing immediate feedback could lead participants to adjust their strategies based on performance outcomes, potentially introducing noise and confounding the specific effects we aimed to measure. Instead, they received information about their overall payment only at the end of the study.

**Feature subsets.** To maintain comparability between domains, it was necessary to standardize the number of features employed in both the tasks and the models across all three decision problems. Moreover, we needed to provide the models and the participants with sufficient information to make useful predictions. A vital design aspect of the experiment was to explain to participants that a selection of features had occurred and that a selection could impact the quality of the advice. Participants were given 12 features for solving the classification tasks in each decision problem. Still, only 6 of the 12 features were used for the ML models, which were shown and highlighted to the participants. We believe using a subset of the features renders the selection process more intelligible and pertinent. Although supplying participants with more information than the models might adversely affect advice reliance, we also contend that decision-makers in many real-life situations possess a different set of information that could contain more detail.



During the experiment, to ensure that all treatments were equal in all aspects except the feature selection method, it was also vital that the features used for predictions remained consistent in all selection methods, guaranteeing that the advice was uniform across all treatments. We carefully selected the final feature sets employed in the task using multiple feature selection algorithms. For the two domains, we selected the following features, with the first 6 in the list being used for the machine learning models:

*Cardio*: Age, Weight in kg, Body Mass Index, Systolic blood pressure, Diastolic blood pressure, Cholesterol level, Gender, Height in cm, Glucose level, Smoking status, Alcoholism, Physical activity.

*Football*: Offsides away team, Passes away team, Passes home team, Possession home team in %, Shots away team, Shots home team, Corners away team, Corners home team, Fouls conceded home team, Offsides home team, Yellow cards away team, Yellow cards home team.

**Machine learning model.** To train the ML models responsible for the advice, we employed the XGBoost algorithm, a widely used and highly effective algorithm for classification and regression tasks [80]. To ensure the optimal performance of our models, we performed model tuning using the grid search method in conjunction with 5-fold cross-validation. We divided each dataset into a training and a test set. The training set was utilized for hyperparameter tuning and learning, while the test set was employed for evaluating the model's performance. We evaluated the final models using balanced accuracy. The *Cardio* model scored 0.74, while the *Football* model scored 0.64. Although these scores are not exceptionally high and might be considered insufficient for practical applications, their impact on the experiment is likely minimal, as the participants were not briefed on the models' performance. For the tasks, we selected observations, ensuring that the model's accuracy for these specific observations was roughly equivalent to its performance on the test dataset. The sequence of the two domains and the order of tasks were randomized for each participant.

## Evaluation measures

**Advice reliance measurement.** In our study, we primarily aim to explore participants' preferences for the feature selection method and how these methods influence their reliance on the advice. Hereto, we adopt the approach used in two recent studies [53,75]. As the judgments and advice in these tasks are binary (e.g., no disease/disease, home team won/home team did not win), we are particularly interested in instances where the participant's initial decision is unequal to the model's advice. Observing how the participant reconciles the conflicting answers is interesting in such cases. If the participant alters their belief in the subsequent decision to align with the advice rather than maintaining their initial decision, we consider this a reliance on advice. Consequently, the dependent variable is referred to as *Switch to Advice*.

**Explanatory variables.** We draw upon established scales from various social science disciplines to measure individual characteristics. The Big Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) are measured using ten items on a 5-point Likert scale [81]. The lottery choice task by Gächter et al. [82] measures loss aversion. For risk-taking, we rely on the Global Preference Survey (GPS) by Falk et al. [83], which uses a scale and multiple preference-related questions. We adopt two scales to measure affinity for technology (ATI) [84] and artificial intelligence (GAAIS) [72]. ATI consists of 9 items on a 6-point Likert scale. At the same time, GAAIS is divided into two dimensions—positive affinity, measured with 12 items, and negative affinity, assessed through 8 items—both using a 5-point Likert scale.

## Results

The analysis is segmented into two main sections. In the first section, we initially examine the feature selection methods chosen by participants in the *Choice* treatment. The primary aim is to test the first two hypotheses: Do individuals prefer *Expert* over *Algorithm*, and is *Combination* the most favored? Additionally, we seek to determine if distinctions exist between the two domains. In the explanatory segment of this section, we delve into the participant characteristics associated with their choices.

In the second section, we address three hypotheses concerning advice reliance—do individuals' ex-ante preferences align with what they end up relying on ex-post? The dependent variable in this section is *Switch to Advice*, which denotes instances when participants amend their subsequent decisions to the AI's prediction when the advice diverges from their initial decision. We will consider both the participants of the *No Choice* and the *Choice* treatments. This will allow us to determine if choosing the methods influences advice reliance for the third hypothesis. In the explanatory segment of this section, we explore the participant characteristics associated with reliance.

### Feature selection preferences

**General preferences.** During the *Choice* treatment ( $N = 109$  participants with two decisions resulting in  $n = 218$ ) the feature selection method *Algorithm* was chosen 44 times (20.2%), *Expert* 70 times (32.1%), and *Combination* 104 times (47.7%). The chi-squared test indicates that this distribution significantly deviates from what would be expected in a random sample ( $\chi^2 = 24.917, P < 0.001$ ). Pairwise comparisons reveal significant distinctions among all three methods: *Algorithm* vs. *Combination* ( $\chi^2 = 23.324, P < 0.001$ ), *Algorithm* vs. *Expert* ( $\chi^2 = 5.93, P = 0.015$ ), and *Combination* vs. *Expert* ( $\chi^2 = 6.644, P = 0.001$ ). Fig 1 illustrates the distribution of the selections.

**Preferences between domains.** Based on these findings, one might accept hypotheses 1a and 1b, which posit that *Expert* is preferred over *Algorithm* and that *Combination* is favored over *Expert*. However, when examining the data segregated by domains, it becomes evident that participants' preferences are more nuanced and not as straightforward. In *Cardio*, *Algorithm* was chosen 18 times (16.5%), *Combination* 51 times (46.8%), and *Expert* 40 times (36.7%). Once more, we note that the distribution significantly deviates from that of a random sample ( $\chi^2 = 15.541, P < 0.001$ ). Unlike in the analyses conducted on the entire dataset, the pairwise comparison reveals that the difference between *Combination* and *Expert* is no longer significant ( $\chi^2 = 1.33, P = 0.25$ ). Still, the differences between *Algorithm* and both *Combination* ( $\chi^2 = 15.783, P < 0.001$ ) and *Expert* ( $\chi^2 = 8.345, P = 0.004$ ) are statistically significant. In *Football*, a distinct pattern is observed: *Algorithm* was chosen 26 times (23.9%), *Combination* 53 times (48.6%), and *Expert* 30 times (27.5%). Once again, the distribution significantly diverges from that of a random sample ( $\chi^2 = 11.688, P = 0.003$ ). *Combination* was significantly more favored compared to both *Algorithm* ( $\chi^2 = 9.228, P = 0.002$ ) and *Expert* ( $\chi^2 = 6.373, P = 0.003$ ), but no significant difference is found between *Algorithm* and *Expert* ( $\chi^2 = 0.285, P = 0.593$ ). Fig 2 illustrates the selection distributions for both domains. To determine if participants' first and second choices were independent, we examined the distribution of preferences for these choices. Our comparison showed no significant differences ( $\chi^2 = 2.138, P = 0.343$ ). This independence in preferences was observed irrespective of whether *Cardio* ( $\chi^2 = 4.092, P = 0.129$ ) or *Football* ( $\chi^2 = 1.561, P = 0.458$ ) was the first domain in the experiment. While the general analysis allows us to accept both hypotheses H1a and H1b, we point to domain-specific differences that influence the relationships.

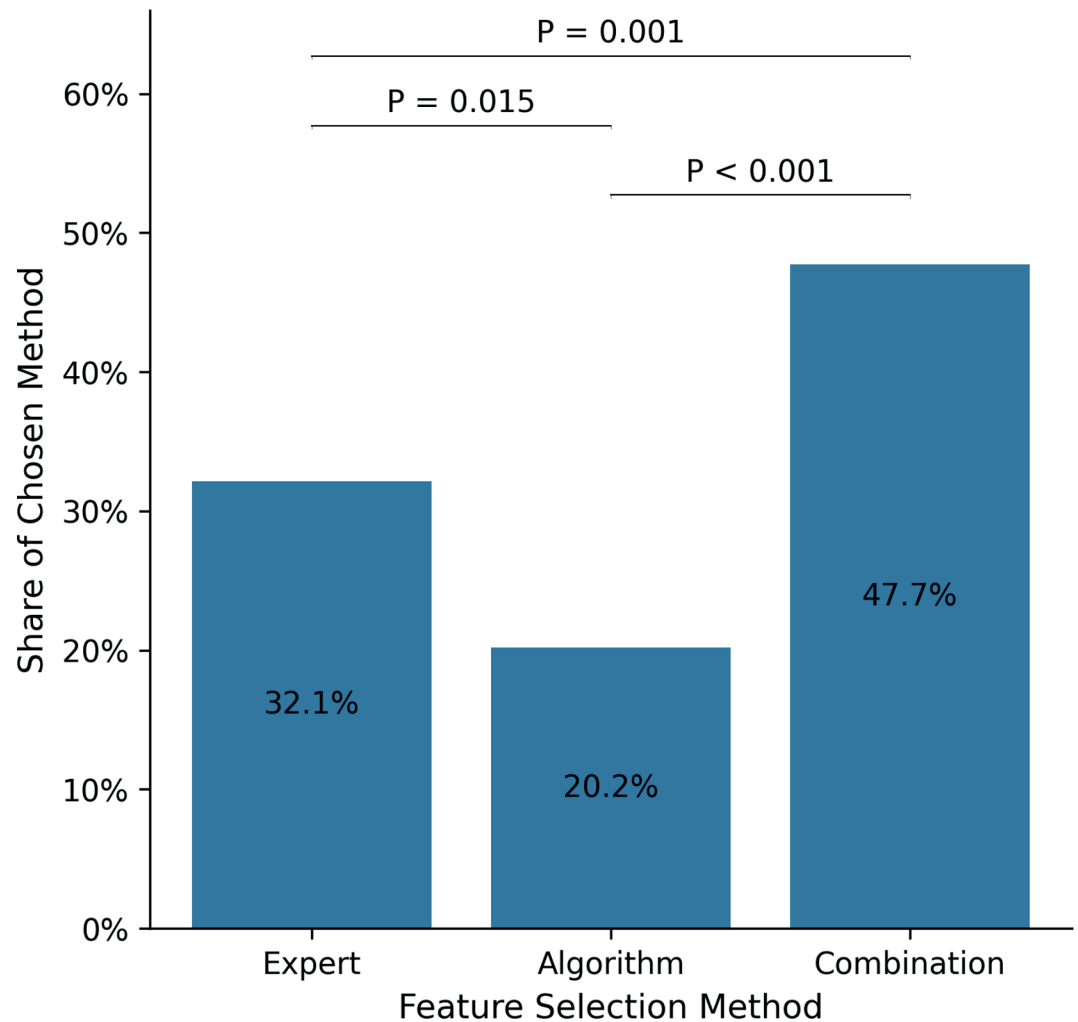


Fig 1. Distribution of the chosen feature selection methods.

<https://doi.org/10.1371/journal.pone.0318874.g001>

**Exploration of characteristics.** Regarding personality characteristics, we found using two multinomial logistic regression models (Table 1) that age is negatively associated with a preference for *Expert* when compared to *Algorithm* ( $\beta = 0.038$ ,  $SE = 0.02$ ,  $P = 0.06$ ) and *Combination*. ( $\beta = 0.032$ ,  $SE = 0.017$ ,  $P = 0.06$ ). *Neuroticism* is positively associated with an increased preference for *Combination* when compared to *Expert* ( $\beta = 0.469$ ,  $SE = 0.233$ ,  $P = 0.045$ ) and *Combination* to *Algorithm* ( $\beta = 0.754$ ,  $SE = 0.264$ ,  $P = 0.004$ ). *Risk-taking* is positively linked with an augmented preference for both *Algorithm* ( $\beta = 1.616$ ,  $SE = 0.687$ ,  $P = 0.018$ ) and *Combination* ( $\beta = 1.458$ ,  $SE = 0.557$ ,  $P = 0.009$ ) over *Expert*.

### Advice reliance

**Descriptive statistics.** In contrast to the previous section, we now utilize data from both treatments, so we observe 216 participants from *Choice* and *No Choice* together. The machine learning models outperformed the participants in the classification tasks. Their predictions were correct in 65% of the *Cardio* and in 60% in *Football* tasks. Participants initially decided correctly in 54.69% of cases (*Cardio*: 63.40%, *Football*: 46.37%). The initial decision aligned

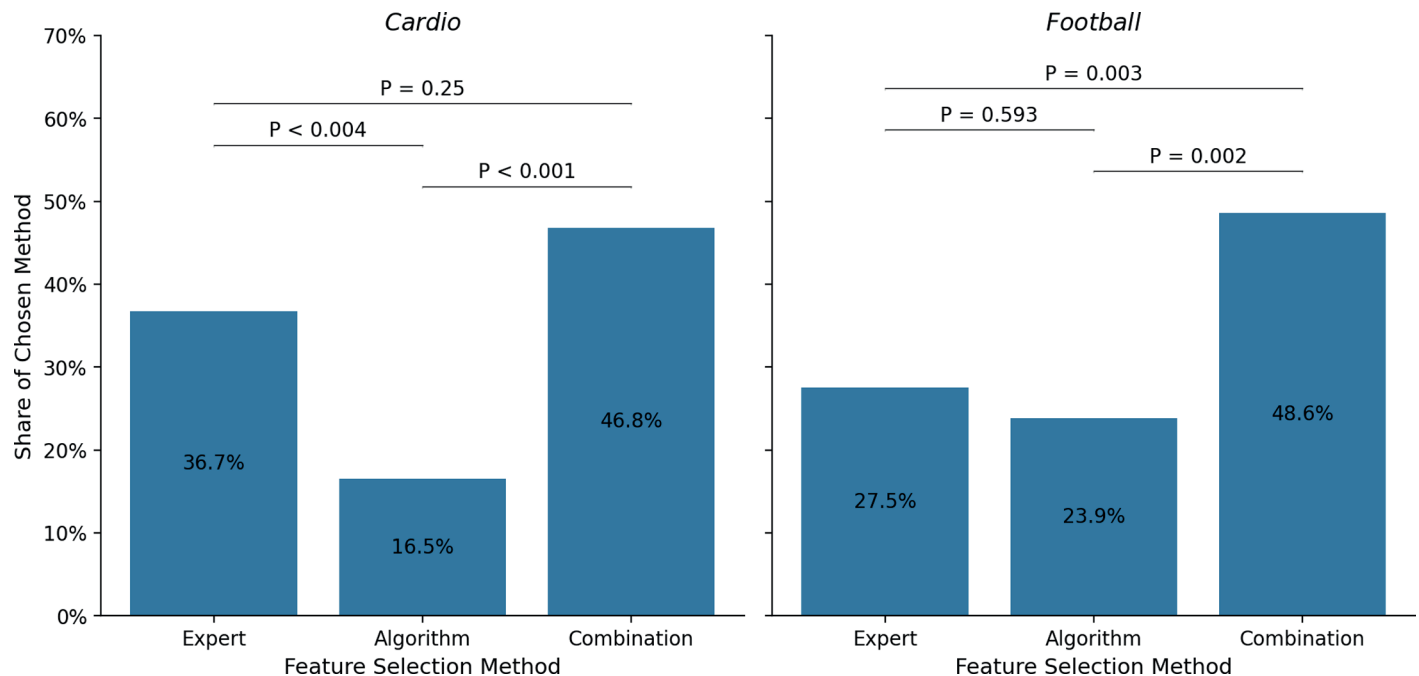


Fig 2. Distribution of the chosen feature selection methods for both domains.

<https://doi.org/10.1371/journal.pone.0318874.g002>

Table 1. Multinomial Logistic Regression results for the feature selection method preferences.

| Base Category                | <i>Expert</i>               |                            | <i>Algorithm</i>   |                             |
|------------------------------|-----------------------------|----------------------------|--------------------|-----------------------------|
|                              | <i>Algorithm</i>            | <i>Combination</i>         | <i>Combination</i> | <i>Expert</i>               |
| <i>Cardio</i>                | -0.714 <sup>†</sup> (0.408) | -0.358 (0.325)             | 0.356 (0.378)      | 0.714 <sup>†</sup> (0.408)  |
| Male                         | -1.051* (0.500)             | -0.485 (0.396)             | 0.566 (0.456)      | 1.051* (0.500)              |
| Age                          | 0.038 <sup>†</sup> (0.020)  | 0.032 <sup>†</sup> (0.017) | -0.006 (0.018)     | -0.038 <sup>†</sup> (0.020) |
| Big 5 Extraversion           | -0.004 (0.245)              | -0.043 (0.189)             | -0.038 (0.232)     | 0.004 (0.245)               |
| Big 5 Agreeableness          | -0.105 (0.316)              | -0.221 (0.249)             | -0.116 (0.279)     | 0.105 (0.316)               |
| Big 5 Conscientiousness      | -0.334 (0.293)              | 0.039 (0.227)              | 0.373 (0.274)      | 0.334 (0.293)               |
| Big 5 Neuroticism            | -0.288 (0.288)              | 0.466* (0.233)             | 0.754** (0.264)    | 0.288 (0.288)               |
| Big 5 Openness               | 0.032 (0.248)               | -0.103 (0.199)             | -0.135 (0.225)     | -0.032 (0.248)              |
| Loss Aversion                | -0.137 (0.150)              | -0.095 (0.124)             | 0.042 (0.137)      | 0.137 (0.150)               |
| Risk Taking                  | 1.619* (0.687)              | 1.458 (0.557)**            | -0.161 (0.620)     | -1.619* (0.687)             |
| ATI                          | 0.221 (0.267)               | 0.085 (0.204)              | -0.137 (0.243)     | -0.221 (0.267)              |
| GAAIS Positive               | 0.278 (0.371)               | 0.357 (0.296)              | 0.079 (0.353)      | -0.278 (0.371)              |
| GAAIS Negative               | 0.391 (0.338)               | 0.053 (0.264)              | -0.338 (0.308)     | -0.391 (0.338)              |
| <i>n</i> (Choices)           | 218                         |                            |                    |                             |
| <i>N</i> (Participants)      | 109                         |                            |                    |                             |
| Pseudo <i>R</i> <sup>2</sup> | 0.0812                      |                            |                    |                             |

The first two models use *Expert* as their base category, while the third and fourth use *Algorithm*. Standard errors in parentheses. <sup>†</sup>  $P < 0.1$ , \*  $P < 0.05$ , \*\*  $P < 0.1$ , \*\*\*  $P < 0.01$ .

<https://doi.org/10.1371/journal.pone.0318874.t001>

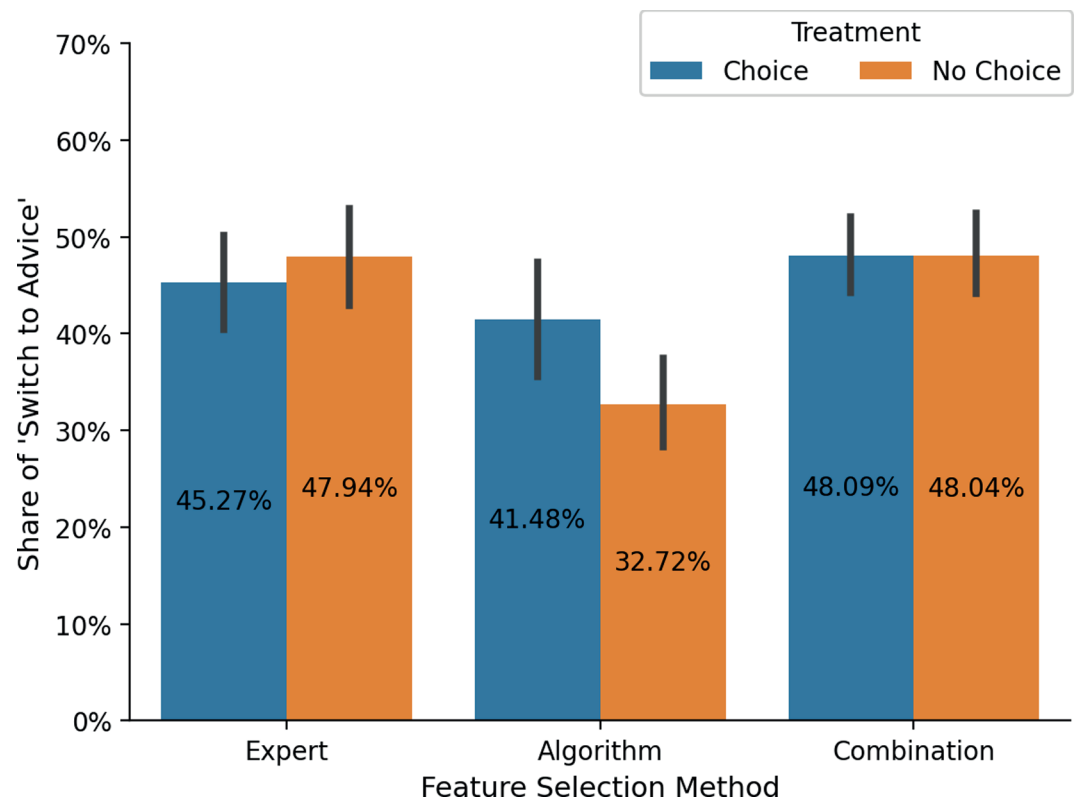
with the models's prediction in 69.11% of instances (*Cardio*: 73.22%, *Football*: 65.00%). In scenarios where the initial decision did not align with the models's advice, participants were correct 37.69% of the time (*Cardio*: 47.02%, *Football*: 30.55%). Conversely, the models's advice was accurate 62.31% of the time in these situations (*Cardio*: 52.98%, *Football*: 69.44%). Participants chose to switch their decisions to follow the models's advice in 44.77% of these cases

(*Cardio*: 53.93%, *Football*: 37.77%). As a result, the overall accuracy rate in advice-receiving situations amounted to 47.47% (*Cardio*: 49.96%, *Football*: 45.57%).

**Reliance between methods and treatments.** While these results indicate that participants partially rejected the advice and, therefore, exhibited an aversion, it's necessary for our research question to examine how reliance depends on the underlying feature selection method and the participant's choice. Fig 3 shows the distribution of *Switch to Advice* across the three methods, distinguishing between both treatments, *Choice* and *No Choice*. Additionally, Fig 4 segregates the data further, delineating the results for both domains.

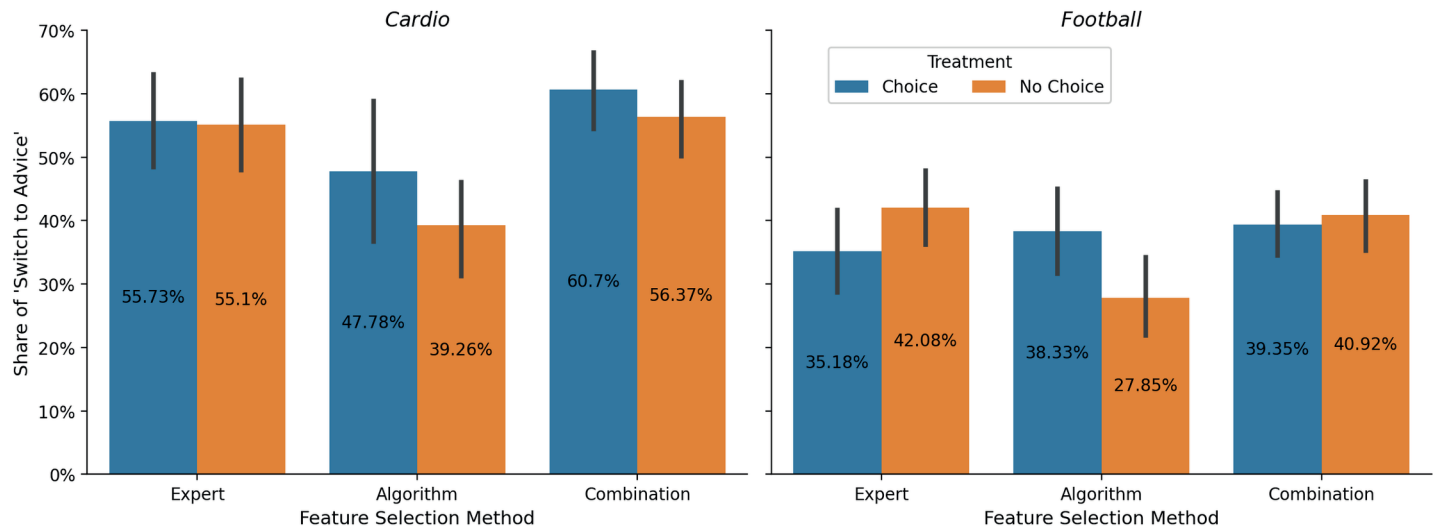
We employ mixed-effects logistic regression models (Table 2) to analyze whether the methods influence reliance. The regressions incorporate a random intercept for each participant, accounting for the multiple observations per individual. For the pairwise comparisons, we alternately set *Expert* and *Algorithm* as the reference categories. We include a dummy variable for the *Choice* treatment and the *Cardio* domain, the number of rounds, the self-reported confidence in the initial decision, and variables representing participant characteristics.

We note 2,669 instances where participants received advice from the AI, as advice was provided only when they deviated from the initial decision of the participants. Both models demonstrate that the respective methods do not have a significant effect on reliance. Furthermore, the option to choose a method also has no influence. Therefore, we reject the hypotheses H2a, H2b, and H3.



**Fig 3. Distribution of *Switch to Advice* by feature selection methods.** Error bars represent 95% confidence intervals.

<https://doi.org/10.1371/journal.pone.0318874.g003>



**Fig 4. Distribution of Switch to Advice by feature selection methods and domains.** Error bars represent 95% confidence intervals.

<https://doi.org/10.1371/journal.pone.0318874.g004>

A significant domain effect is evident through a significant positive coefficient for *Cardio* ( $\beta = 1.008$ ,  $SE = 0.099$ ,  $P < 0.001$ ), a pattern also reflected in our descriptive analysis. This corresponds to a marginal effect of 17.98 percentage points.

**Analysis of covariates.** As the coefficient for the number of tasks is also insignificant, we don't observe any time trends. This was expected as the participants had no feedback during the task. A notable association exists between participants' self-reported confidence in their initial decision and advice reliance ( $\beta = -0.028$ ,  $SE = 0.004$ ,  $P = 0.000$ ). As confidence in one's decision diminishes, the reliance on the AI's advice grows—for each unit (on a scale from 0 to 100), the likelihood of change in the subsequent decision falls by 0.49 percentage points. Regarding personality and demographic attributes, we do not observe any gender-specific effects. However, a significant negative relationship emerges between age and advice reliance ( $\beta = -0.020$ ,  $SE = 0.008$ ,  $P = 0.017$ ). Each year, the likelihood of advice reliance decreases by 0.36 percentage points. Among the Big 5 personality traits, *Openness* is a negative association ( $\beta = -0.225$ ,  $SE = 0.107$ ,  $P = 0.035$ ).

## Discussion

### Main findings

To begin with, we discover that decision-makers in our experiment prefer the *Expert* over *Algorithm* and favor *Combination* over *Expert*. Yet, when separating the data by the two domains, it becomes evident that the specific domains may have affected participants' choices. In the domain where participants classified patients based on symptoms and characteristics into groups with and without cardiovascular disease, we find no significant difference between the popularity of *Combination* and *Expert*. In contrast, in determining a home team win based on match statistics, *Combination* is significantly the most popular, with *Algorithm* and *Expert* being equally favored.

In our analysis regarding the classification tasks, we observe, contrary to our expectations, no significant effect of the underlying feature selection methods on advice reliance and no effect of the opportunity to choose the method by the participants. Significant predictors of reliance are the domain (with a higher reliance in the medical domain), personal confidence



**Table 2. Mixed-effects logistic regression results for *Switch to Advice*.**

|   | (1)<br><i>Switch to Advice</i> | (2)<br><i>Switch to Advice</i> |
|---|--------------------------------|--------------------------------|
| <i>Expert</i>                                   | /                              | 0.236 (0.199)                  |
| <i>Algorithm</i>                                | −0.236 (0.199)                 | /                              |
| <i>Combination</i>                              | −0.017 (0.164)                 | 0.220 (0.189)                  |
| <i>Choice</i>                                   | 0.154 (0.188)                  |                                |
| <i>Cardio</i>                                   | 1.008*** (0.099)               |                                |
| Round Number                                    | −0.003 (.004)                  |                                |
| Own Confidence                                  | −0.028*** (0.003)              |                                |
| Male  | −0.292 (0.222)                 |                                |
| Age   | −0.020* (0.008)                |                                |
| Big 5 Extraversion                              | −0.065 (0.104)                 |                                |
| Big 5 Agreeableness                             | 0.174 (0.103)                  |                                |
| Big 5 Conscientiousness                         | 0.202† (0.122)                 |                                |
| Big 5 Neuroticism                               | −0.028 (0.116)                 |                                |
| Big 5 Openness                                  | −0.225* (0.107)                |                                |
| Loss Aversion                                   | −0.001(0.070)                  |                                |
| Risk Taking                                     | 0.203 (0.28)                   |                                |
| ATI   | −0.042 (0.120)                 |                                |
| GAAIS Positive                                  | 0.232 (0.165)                  |                                |
| GAAIS Negative                                  | −0.028 (0.143)                 |                                |
| Participant Intercept                           | 1.329 (.208)                   |                                |
| Constant  | 0.626 (1.253)                  | 0.556 (1.253)                  |
| Log-likelihood                                  | −1565.109                      |                                |
| Wald $\chi^2$ (23)                              | 185.89                         |                                |
| Prob > $\chi^2$                                 | 0.000                          |                                |
| LR test vs. logistic model: $\bar{\chi}^2$ (01) | 270.53                         |                                |
| Prob $\geq \bar{\chi}^2$                        | 0.000                          |                                |
| Observations                                    | 2,669                          |                                |
| Number of groups                                | 216                            |                                |

The first model uses *Expert* as the base category, and the second *Algorithm*. Standard errors in parentheses. † P<0.1, \* P<0.05, \*\* P<0.01, \*\*\* P<0.001.

<https://doi.org/10.1371/journal.pone.0318874.t002>

in the decision, and age, both showing negative correlations with reliance. From the Big 5 scale *Openness* was negatively associated with reliance.

Together, the findings from our analysis of preferences do not align with those concerning reliance. Given the notable differences in popularity between *Combination* and both *Algorithm* and *Expert* (especially in one domain), one might anticipate greater advice reliance on *Combination* during the classification task. Yet, we observe no effect. While AI users express their preferences regarding AI characteristics, their ultimate behaviors remain largely uninfluenced by these stated preferences. This result is similar to two previous studies: Rabinovitch et al. [28] found that participants explicitly preferred a human advisor over an algorithmic one, but the advice was used equally. Rebitschek et al. [29] discovered a discrepancy between the acceptable, perceived, and actual error rates of algorithms. This can be attributed to various cognitive factors. For instance, according to dual-process theory [85], when asked about their preferences, participants may have engaged in System 2 thinking, carefully evaluating the perceived benefits of the three options. However, during the actual decision-making process, they likely reverted to System 1 thinking due to the complexity of the task and the cognitive load. As a result, they may have paid less attention to the subtle details of the feature selection methods. Another possible explanation is social desirability bias [86], which

could have led participants to perceive the combined feature selection method as the most advanced, and therefore, the most acceptable option.

In conjunction with the unobserved selection effect, these results resonate with the findings of Cheng and Chouldechova [22]. Their research suggested that while choosing the training algorithm can alleviate algorithm aversion, modifications to the information utilized by the algorithm do not offer similar mitigation. Our results partly confirm the framework by Jussupow et al. [7], as in our study, humans state a preference for human involvement in AI development by asking humans to (partly) select the features. However, we find no evidence that this stated preference also unfolds its effects when humans face AI advice. Gogoll and Uhl [87] found a comparable trend: while their participants leaned towards delegating tasks to humans over machines, their trust did not differ.

## Secondary findings

In addition to the relationships of the treatments analyzed, our results indicate that other factors, notably the task domain and the users themselves, play a significant role. Our results indicate caution when analyzing human-AI collaborations, as results may be artifact-specific. Utilizing a self-reported scale for risk-taking behavior [83], a multinomial model shows that participants displaying higher risk-taking tendencies exhibited a preference for *Algorithm* and *Combination* over *Expert*. This inclination might be explained by the “Diffusion of Innovations” theory—historically, early adopters of novel technologies tend to be more risk-prone [88,89]. If *Expert* is perceived as more conservative, then a method incorporating or entirely based on algorithms might be perceived as a more innovative approach.

We observe a significant positive effect of the medical domain on the likelihood of adjusting the decision toward the AI prediction. Notably, our findings do not entirely align with previous research on algorithm aversion in medical settings. For instance, Arkes and Blumer [70] reported that participants favored physicians who did not utilize decision aids. Similarly, Longoni et al. [90] noted a hesitancy towards AI providers compared to human providers in a medical context. While reliance is typically linked to perceived risk, and medical decisions usually carry more risk than sports-related ones, the payoff for both domains is identical, making the risk equivalent. Other factors contributing to the differences in reliance could include perceived AI competence in each domain or participants’ own confidence in their classification abilities. However, in this case, we observed higher confidence among participants in the medical domain. Our analysis indicates a significant negative correlation between the decision-makers’ confidence and their reliance on AI, consistent with prior experimental findings [11,56,91]. The inverse relationship between a participant’s age and reliance diverges from findings by Ho et al. [92], who determined that older adults exhibited a higher trust in decision aids. Similarly, Logg et al. [11] discovered a consistent appreciation for algorithms irrespective of age. Gender was not a significant predictor, as in the study by Logg et al. [11]. The reported inconsistencies may be partially attributed to the rapid integration of AI into society. This is because algorithm aversion and appreciation can be understood through normative processes [58] and long-term learning effects [79].

## Limitations and implications

One potential reason for the missing differences in reliance between the methods might be due to a manipulation that is too subtle. There’s a possibility that the methods’ signals are too faint within the task to detect an effect corresponding to the significant differences observed in preferences. Despite this, the presentation mirrors real-world scenarios where detailed explanations of AI feature selection methods are rarely provided. Participants were able to

review the selected features during the tasks, unlike during the method selection phase. This visibility allowed them to reasonably assess the selection's validity, likely comparing it with their judgment. Consequently, the feature selection method information likely serves as only a minor indicator of the selection's validity, possibly leading to the observed results. Future studies might consider not displaying the features, although this approach could reduce realism.

Another limitation impacting the generalizability of our findings is the recruitment of non-professional decision-makers from an online participant pool instead of domain professionals. While we acknowledge that expertise is crucial in many real-world applications, using lay participants offers important advantages, especially in the context of fundamental research like ours. Lay participants provide an opportunity to study baseline human-AI interactions without the influence of pre-existing domain-specific knowledge, allowing us to isolate general behavioral patterns related to trust, preferences, and reliance on AI systems. Future studies could build on this foundation by replicating the experiment with domain experts to enhance the real-world applicability of our findings.

Nonetheless, it is plausible that domain experts would not yield substantially different outcomes. On the one hand, the literature reveals that the same biases are prevalent among both laypeople and experts [93,94]. On the other hand, a meta-analysis shows that in human-AI collaboration experiments, there are no differences in decision-making performance between professional and non-professional participants [95]. We believe that, in addition to expertise in one's own domain, experience in machine learning and feature selection is also needed to form a strong opinion. With only domain experience, we expect similar results as seen with laymen, both concerning the preference for human oversight and the reliance on AI advice.

Another way to expand the research in this study would be to shift the focus from short-term interactions to long-term time horizons, exploring how preferences and reliance evolve over time. Long-term research has often been avoided in the human-AI literature due to its empirical challenges, but previous studies suggest the presence of temporal effects [79].

By examining algorithm-based, expert-based, and combined feature selection approaches, we offer fresh insights into how human involvement shapes user trust, preferences, and reliance on AI-driven decisions. Our findings highlight the nuanced and complex relationships between human involvement and user behavior, revealing that the degree of human input can significantly influence perceptions of transparency and trustworthiness, yet these perceptions may not always translate into greater reliance on the system. We reveal a significant attitude-behavior gap, known in many disciplines and for many instances: While humans reveal strong stated preference for human oversight *ex ante*, individuals are equally likely to rely on AI advice, independent of human oversight.

Our results have practical implications, especially when transparency is essential in decision support systems and there is a lack of trust towards them. Those overseeing or designing AI systems could communicate that the data the AI uses was selected from a joint effort between human experts and algorithms. However, they also need to consider individual traits. As AI systems are often developed in this way, making this known might align with users' preferences, potentially increasing the likelihood of using these systems and leading to better decision-making outcomes.

## Conclusion

AI-supported decision-making is becoming increasingly relevant in everyday contexts, making it essential to understand the factors that influence human-AI interactions. While researchers advocate for greater transparency and explainability, it raises questions about

how users perceive different elements. In this paper, we focus on two critical aspects: human involvement and feature selection, both central to many ML models. Our findings suggest that decision-makers tend to prefer a combination of human and algorithmic feature selection methods. However, we also discovered that neither the methods themselves nor the decision-makers' involvement in choosing these methods significantly influences reliance. These insights underscore the complexity of human-AI interactions and highlight the importance of behavioral experiments in this field of research.

## Supporting information

**S1 Text.**  
(PDF)

**S1 Fig.**  
(PDF)

**S2 Fig.**  
(PDF)

**S1 Data.**  
(PDF)

## Author contributions

**Conceptualization:** Jaroslaw Kornowicz, Kirsten Thommes.

**Data curation:** Jaroslaw Kornowicz.

**Formal analysis:** Jaroslaw Kornowicz.

**Funding acquisition:** Kirsten Thommes.

**Investigation:** Jaroslaw Kornowicz.

**Methodology:** Jaroslaw Kornowicz, Kirsten Thommes.

**Project administration:** Jaroslaw Kornowicz, Kirsten Thommes.

**Resources:** Jaroslaw Kornowicz.

**Software:** Jaroslaw Kornowicz, Kirsten Thommes.

**Supervision:** Jaroslaw Kornowicz, Kirsten Thommes.

**Validation:** Jaroslaw Kornowicz, Kirsten Thommes.

**Visualization:** Jaroslaw Kornowicz.

**Writing – original draft:** Jaroslaw Kornowicz, Kirsten Thommes.

**Writing – review & editing:** Jaroslaw Kornowicz, Kirsten Thommes.

## References

1. Aoki N. An experimental study of public trust in AI chatbots in the public sector. *Gov Inf Quart.* 2020;37(4):101490. <https://doi.org/10.1016/j.giq.2020.101490>
2. Cetinic E, She J. Understanding and creating art with AI: review and outlook. *ACM Trans Multim Comput Commun Appl.* 2022;18(2):66:1–66:22. <https://doi.org/10.1145/3475799>
3. Deranty JP, Corbin T. Artificial intelligence and work: a critical review of recent research from the social sciences. *AI Soc.* 2022. <https://doi.org/10.1007/s00146-022-01496-x>

4. Hallur GG, Prabhu S, Aslekar A. In: Das S, Gochhait S, editors. Entertainment in era of AI, Big Data & IoT. Singapore: Springer; 2021. p. 87–109. Available from: [https://doi.org/10.1007/978-981-15-9724-4\\_5](https://doi.org/10.1007/978-981-15-9724-4_5)
5. Makridakis S. The forthcoming Artificial Intelligence (AI) revolution: its impact on society and firms. *Futures*. 2017;90:46–60. <https://doi.org/10.1016/j.futures.2017.03.006>
6. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. *arXiv preprint* 2021. <http://arxiv.org/abs/2103.11251> [cs, stat]
7. Jussupow E, Benbasat I, Heinzl A. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *ECIS 2020 Research Papers*. 2020.
8. Mahmud H, Islam AKMN, Ahmed SI, Smolander K. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technol Forecast Soc Change*. 2022;175:121390. <https://doi.org/10.1016/j.techfore.2021.121390>
9. Ashoori M, Weisz JD. In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. *arXiv preprint* 2019. <http://arxiv.org/abs/1912.02675>
10. Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. *J Market Res*. 2019;56(5):809–25. <https://doi.org/10.1177/0022243719851788>
11. Logg JM, Minson JA, Moore DA. Algorithm appreciation: People prefer algorithmic to human judgment. *Organiz Behav Hum Decis Process*. 2019;151:90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
12. You S, Yang CL, Li X. Algorithmic versus human advice: does presenting prediction performance matter for algorithm appreciation? *J Manag Inf Syst*. 2022;39(2):336–65. <https://doi.org/10.1080/07421222.2022.2063553>
13. Barredo Arrieta A, Diaz-Rodriguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
14. Albrecht JP. How the GDPR will change the world. *Eur Data Prot L Rev*. 2016;2:287.
15. McCarthy M. An examination of the Algorithmic Accountability Act of 2019. Available at SSRN 3615731. 2019.
16. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell*. 2019;267:1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
17. Rohlfing KJ, Cimiano P, Scharlau I, Matzner T, Buhl HM, Buschmeier H, et al. Explanation as a social practice: toward a conceptual framework for the social design of AI systems. *IEEE Trans Cognit Develop Syst*. 2020. p. 1. <https://doi.org/10.1109/TCDS.2020.3044366>
18. Sundar SS, Knobloch-Westerwick S, Hastall MR. News cues: information scent and cognitive heuristics. *J Am Soc Inf Sci Technol*. 2007;58(3):366–78. <https://doi.org/10.1002/asi.20511>
19. Sundar SS. Rise of machine agency: a framework for studying the psychology of Human–AI Interaction (HAI). *J Comput-Mediat Commun*. 2020;25(1):74–88. <https://doi.org/10.1093/jcmc/zmz026>
20. Waddell TF. Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism Mass Commun Quart*. 2019;96(1):82–100. <https://doi.org/10.1177/1077699018815891>
21. Jago AS. Algorithms and Authenticity. *Academy of Management Discoveries*. 2019;5(1):38–56. <https://doi.org/10.5465/amd.2017.0002>
22. Cheng L, Chouldechova A. Overcoming algorithm aversion: a comparison between process and outcome control. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg Germany: ACM; 2023. p. 1–27. Available from: <https://dl.acm.org/doi/10.1145/3544548.3581253>
23. Ustun B, Rudin C. Supersparse linear integer models for optimized medical scoring systems. *Mach Learn*. 2016;102(3):349–91 <https://doi.org/10.1007/s10994-015-5528-6>
24. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. *Neurocomputing*. 2018;300:70–9. <https://doi.org/10.1016/j.neucom.2017.11.077>
25. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3(Mar):1157–82.
26. Holzinger A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Inf*. 2016;3(2):119–31. <https://doi.org/10.1007/s40708-016-0042-6> PMID: 27747607
27. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206–15. <https://doi.org/10.1038/s42256-019-0048-x> PMID: 35603010

28. Rabinovitch H, Budescu DV, Meyer YB. Algorithms in selection decisions: Effective, but unappreciated. *J Behav Decis Making*. 2024;37(2):e2368. <https://doi.org/10.1002/bdm.2368>
29. Rebitschek FG, Gigerenzer G, Wagner GG. People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Sci Rep*. 2021;11(1):20171. <https://doi.org/10.1038/s41598-021-99802-y> PMID: 34635779
30. Studer S, Bui TB, Drescher C, Hanuschkin A, Winkler L, Peters S, et al. Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology. *Mach Learn Knowl Extract*. 2021;3(2):392–413. <https://doi.org/10.3390/make3020020>
31. Mera-Gaona M, Lopez DM, Vargas-Canas R, Neumann U. Framework for the ensemble of feature selection methods. *Appl Sci*. 2021;11(1717):8122. <https://doi.org/10.3390/app11178122>
32. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. vol. 112. Springer; 2013.
33. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electric Eng*. 2014;40(1):16–28. <https://doi.org/10.1016/j.compeleceng.2013.11.024>
34. Liu H, Motoda H. Feature selection for knowledge discovery and data mining. Springer Science & Business Media; 2012.
35. Yin L, Wang L, Lu S, Wang R, Ren H, AlSanad A, et al. AFBNet: a lightweight adaptive feature fusion module for super-resolution algorithms. *Comput Model Eng Sci*. 2024;140(3):2315–47. <https://doi.org/10.32604/cmes.2024.050853>
36. Singh LK, Khanna M, Singh R. Feature subset selection through nature inspired computing for efficient glaucoma classification from fundus images. *Multim Tools Appl*. 2024;83(32):77873–944. <https://doi.org/10.1007/s11042-024-18624-y>
37. Khanna M, Singh LK, Shrivastava K, Singh R. An enhanced and efficient approach for feature selection for chronic human disease prediction: a breast cancer study. *Heliyon*. 2024;10(5):e26799. <https://doi.org/10.1016/j.heliyon.2024.e26799>
38. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: a data perspective. *ACM Comput Surv*. 2017;50(6):94:1–94:45. <https://doi.org/10.1145/3136625>
39. Tiwari A, Chaturvedi A. A hybrid feature selection approach based on information theory and dynamic butterfly optimization algorithm for data classification. *Exp Syst Appl*. 2022;196:116621. <https://doi.org/10.1016/j.eswa.2022.116621>
40. Nahar J, Imam T, Tickle KS, Chen YPP. Computational intelligence for heart disease diagnosis: a medical knowledge driven approach. *Exp Syst Appl*. 2013;40(1):96–104. <https://doi.org/10.1016/j.eswa.2012.07.032>
41. Corrales DC, Lasso E, Ledezma A, Corrales JC. Feature selection for classification tasks: expert knowledge or traditional methods? *J Intell Fuzzy Syst*. 2018;34(5):2825–35. <https://doi.org/10.3233/JIFS-169470>
42. Wang J, Oh J, Wang H, Wiens J. Learning credible models. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. KDD'18. New York, NY, USA: Association for Computing Machinery; 2018. p. 2417–26. Available from: <https://doi.org/10.1145/3219819.3220070>
43. Cheng TH, Wei CP, Tseng VS. Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. In: 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06). IEEE; 2006. p. 165–70.
44. Moro S, Cortez P, Rita P. A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Exp Syst*. 2018;35(3):e12253. <https://doi.org/10.1111/exsy.12253>
45. Shin D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int J Hum-Comput Stud*. 2021;146:102551.
46. Bolon-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: a review and future trends. *Inf Fusion*. 2019;52:1–12. <https://doi.org/10.1016/j.inffus.2018.11.008>
47. Wald R, Khoshgoftaar TM, Dittman D, Awada W, Napolitano A. An extensive comparison of feature ranking aggregation techniques in bioinformatics. In: 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI); 2012. p. 377–84.
48. Bianchi F, Piroddi L, Bemporad A, Halasz G, Villani M, Piga D. Active preference-based optimization for human-in-the-loop feature selection. *Eur J Control*. 2022;66:100647. <https://doi.org/10.1016/j.ejcon.2022.100647>
49. Correia AHC, Lecue F. Human-in-the-loop feature selection. *Proc AAAI Conf Artif Intell*. 2019;33(01):2438–45. <https://doi.org/10.1609/aaai.v33i01.33012438>
50. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. *Science*. 1989;243(4899):1668–74. <https://doi.org/10.1126/science.2648573> PMID: 2648573



51. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518(7540):529–33. <https://doi.org/10.1038/nature14236> PMID: 25719670.
52. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint*. 2023. <http://arxiv.org/abs/2303.13375>
53. Schemmer M, Kuhl N, Benz C, Bartos A, Satzger G. Appropriate reliance, explainable AI, human-AI collaboration, human-AI complementarity. *arXiv preprint* 2023. <http://arxiv.org/abs/2302.02187>
54. Vasconcelos H, Jorke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. Explanations can reduce overreliance on AI systems during decision-making. *Proc ACM Hum-Comput Interact*. 2023;7(CSCW1):1–38.
55. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J Exp Psychol: Gen*. 2015;144(1):114–26. <https://doi.org/10.1037/xge0000033> PMID: 25401381
56. He G, Kuiper L, Gadiraju U. Knowing about knowing: an illusion of human competence can hinder appropriate reliance on AI systems. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*; 2023. p. 1–18. <http://arxiv.org/abs/2301.11333>
57. Gino F, Brooks AW, Schweitzer ME. Anxiety, advice, and the ability to discern: feeling anxious motivates individuals to seek and use advice. *J Personal Soc Psychol*. 2012;102(3):497–512. <https://doi.org/10.1037/a0026413> PMID: 22121890
58. Bogard J, Shu S. Algorithm aversion and the aversion to counter-normative decision procedures; 2022. Available from: <https://www.researchsquare.com/article/rs-1466639/v1>
59. Bucinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc ACM Hum-Comput Interact*. 2021;5(CSCW1):188:1–188:21. <https://doi.org/10.1145/3449287>
60. Lai V, Chen C, Smith-Renner A, Liao QV, Tan C. Towards a science of human-AI decision making: an overview of design space in empirical human-subject studies. In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. Chicago IL USA: ACM; 2023. p. 1369–85. Available from: <https://dl.acm.org/doi/10.1145/3593013.3594087>
61. Scharowski N, Perrig SA, von Felten N, Bruhlmann F. Trust and reliance in XAI—distinguishing between attitudinal and behavioral measures. *arXiv preprint* 2022. <http://arxiv.org/abs/2203.12318>
62. Wischniewski M, Kramer N, Muller E. Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI'23. New York, NY, USA: Association for Computing Machinery; 2023. p. 1–16. Available from: <https://dl.acm.org/doi/10.1145/3544548.3581197>
63. Molina MD, Sundar SS. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society*. 2022; p. 14614448221103534. <https://doi.org/10.1177/14614448221103534>
64. Arkes HR, Shaffer VA, Medow MA. Patients derogate physicians who use a computer-assisted diagnostic aid. *Int J Soc Med Decis Making*. 2007;27(2):189–202. <https://doi.org/10.1177/0272989X06297391> PMID: 17409368
65. Palmeira M, Spassova G. Consumer reactions to professionals who use decision aids. *Eur J Market*. 2015;49(3/4):302–26. <https://doi.org/10.1108/EJM-07-2013-0390>
66. Lee S, Kim KJ, Sundar SS. Customization in location-based advertising: effects of tailoring source, locational congruity, and product involvement on ad attitudes. *Comput Hum Behav*. 2015;51:336–43. <https://doi.org/10.1016/j.chb.2015.04.049>
67. Kawaguchi K. When will workers follow an algorithm? A field experiment with a retail business. *Manag Sci*. 2021;67(3):1670–95. <https://doi.org/10.1287/mnsc.2020.3599>
68. Kobis N, Mossink LD. Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput Hum Behav*. 2021;114:106553. <https://doi.org/10.1016/j.chb.2020.106553>
69. Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. *J Behav Decis Making*. 2020;33(2):220–39. <https://doi.org/10.1002/bdm.2155>
70. Arkes HR, Blumer C. The psychology of sunk cost. *Organiz Behav Hum Decis Process*. 1985;35(1):124–40. [https://doi.org/10.1016/0749-5978\(85\)90049-4](https://doi.org/10.1016/0749-5978(85)90049-4)
71. Kaya F, Aydin F, Schepman A, Rodway P, Yetisensoy O, Demir Kaya M. The roles of personality traits, AI anxiety, and demographic factors in attitudes toward artificial intelligence. *Int J Hum-Comput Interact*. 2022; p. 1–18.
72. Schepman A, Rodway P. The General Attitudes towards Artificial Intelligence Scale (GAAIS): confirmatory validation and associations with personality, corporate distrust, and general trust. *Int J Hum-Comput Interact*. 2023;39(13):2724–41. <https://doi.org/10.1080/10447318.2022.2085400>

73. Chen DL, Schonger M, Wickens C. oTree—an open-source platform for laboratory, online, and field experiments. *J Behav Exp Finance*. 2016;9:88–97. <https://doi.org/10.1016/j.jbef.2015.12.001>
74. Hemmer P, Westphal M, Schemmer M, Vetter S, Vossing M, Satzger G. Human-AI collaboration: the effect of AI delegation on human task performance and task satisfaction. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. Sydney, NSW, Australia: ACM; 2023. p. 453–63. Available from: <https://dl.acm.org/doi/10.1145/3581641.3584052>
75. Zhang Y, Liao QV, Bellamy RKE. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. FAT\*20. New York, NY, USA: Association for Computing Machinery; 2020. p. 295–305. <https://doi.org/10.1145/3351095.3372852>
76. Liu M, Conrad FG. Where should i start? On default values for slider questions in web surveys. *Soc Sci Comput Rev*. 2019;37(2):248–69. <https://doi.org/10.1177/0894439318755336>
77. Bonnefon JF, Rahwan I. Machine thinking, fast and slow. *Trends Cognit Sci*. 2020;24(12):1019–27. <https://doi.org/10.1016/j.tics.2020.09.007> PMID: 33129719
78. Bigman YE, Gray K. People are averse to machines making moral decisions. *Cognition*. 2018;181:21–34. <https://doi.org/10.1016/j.cognition.2018.08.003> PMID: 30107256
79. Freisinger E, Unfried M, Schneider S. The adoption of algorithmic decision-making agents over time: algorithm aversion as a temporary effect? *ECIS 2022 Research Papers*. 2022.
80. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016. p. 785–94. Available from: <https://dl.acm.org/doi/10.1145/2939672.2939785>
81. Rammstedt B, Kemper CJ, Klein MC, Beierlein C, Kovaleva A. Big Five Inventory (BFI-10). Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). 2014. <https://doi.org/10.6102/ZIS76>
82. Gächter S, Johnson EJ, Herrmann A. Individual-level loss aversion in riskless and risky choices. *Theory Decis*. 2022;92(3):599–624. <https://doi.org/10.1007/s11238-021-09839-8>
83. Falk A, Becker A, Dohmen T, Huffman D, Sunde U. The preference survey module: a validated instrument for measuring risk, time, and social preferences. *Manag Sci*. 2022;69(4):1935–50. <https://doi.org/10.1287/mnsc.2022.4455>
84. Franke T, Attig C, Wessel D. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *Int J Hum-Comput Interact*. 2019;35(6):456–67. <https://doi.org/10.1080/10447318.2018.1456150>
85. Kahneman D. Thinking, fast and slow. Farrar, Straus and Giroux. 2011.
86. Nederhof AJ. Methods of coping with social desirability bias: a review. *Eur J Soc Psychol*. 1985;15(3):263–80. <https://doi.org/10.1002/ejsp.2420150303>
87. Gogoll J, Uhl M. Rage against the machine: automation in the moral domain. *J Behav Exp Econ*. 2018;74:97–103. <https://doi.org/10.1016/j.socsc.2018.04.003>
88. Dale V, McEwan M, Bohan J. Early adopters versus the majority: characteristics and implications for academic development and institutional change. *J Perspect Appl Acad Pract*. 2021;9(22):54–67. <https://doi.org/10.14297/jpaap.v9i2.483>
89. Wejnert B. Integrating models of diffusion of innovations: a conceptual framework. *Annu Rev Sociol*. 2002;28(1):297–326 <https://doi.org/10.1146/annurev.soc.28.110601.141051>
90. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. *J Consum Res*. 2019;46(4):629–50. <https://doi.org/10.1093/jcr/ucz013>
91. Gino F, Moore DA. Effects of task difficulty on use of advice. *J Behav Decis Mak*. 2007;20(1):21–35. <https://doi.org/10.1002/bdm.539>
92. Ho G, Wheatley D, Scialfa CT. Age differences in trust and reliance of a medication management system. *Interact Comput*. 2005;17(6):690–710. <https://doi.org/10.1016/j.intcom.2005.09.007>
93. Butler D, Butler R, Eakins J. Expert performance and crowd wisdom: evidence from English premier league predictions. *Eur J Oper Res*. 2021;288(1):170–82. <https://doi.org/10.1016/j.ejor.2020.05.034>
94. Kynn M. The ‘Heuristics and Biases’ bias in expert elicitation. *J Roy Statist Soc Ser A: Statist Soc*. 2008;171(1):239–64. <https://doi.org/10.1111/j.1467-985x.2007.00499.x>
95. Vaccaro M, Almaatouq A, Malone T. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat Hum Behav*. 2024:1–11. <https://doi.org/10.1038/s41562-024-02024-1>