When It Matters to You: The Impact of Personal Relevance on the Evaluation of Algorithmic versus Human Predictions

by

Nalan Akın

A Thesis Submitted to the
Graduate School of Social Sciences and Humanities
in Partial Fulfillment of the Requirements for
the Degree of

Master of Arts

in

Psychology

October 1, 2024

When It Matters to You: The Impact of Personal Relevance on the Evaluation of Algorithmic versus Human Predictions

Koç University

Graduate School of Social Sciences and Humanities

This is to certify that I have examined this copy of a master's thesis by

Nalan Akın

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Zeynep Cemalcılar (Advisor)

_____

Asst. Prof. Terry Eskenazi

_____

Assoc. Prof. Lemi Baruh

Date: 13/10/2024

To my grandfather

ABSTRACT

When It Matters to You: The Impact of Personal Relevance on the Evaluation of Algorithmic versus Human Predictions

Nalan Akın

Master of Arts in Psychology

October 1, 2024

Although AI algorithms that surround our personal and professional lives are improving even further, the use of algorithmic predictions and recommendations in decision-making tasks is not solely guaranteed by their superior capabilities compared to humans. Previous research examining a range of algorithms, varying in complexity from simple decision rules to advanced AI systems, has long discussed the phenomena of both algorithm aversion and appreciation. Findings reveal that people's expectations and reactions to algorithms substantially vary under certain circumstances. Studies have primarily focused on the interaction between algorithm design and individual factors and lacked research on the role of task-related variables. Applying the Theory of Felt Involvement (Celsi & Olson, 1988), this study examines how the level of interest in a task affects the perceived accuracy of predictions that are presented as originating from an AI algorithm or a group of individuals. Participants (n = 78) were randomly assigned to one of three groups and asked to rate the accuracy of predictions for 30 trivia questions provided by one of the following sources: (a) the AI algorithm, (b) previous participants, and (c) the ambiguous source (control group). Analysis with a linear mixed-effects model showed that the accuracy ratings for the AI algorithm were higher than those for both the group of previous participants and the control condition, especially when predictions were made for subject categories participants indicated a higher felt interest. This finding suggests that the influence of the source on perceived prediction accuracy was more pronounced for questions with higher interest levels. We discuss the generalizability of our findings and provide possible future directions for human algorithm interaction research.

ÖZETÇE

İnsanların Tahminlerine Kıyasla Algoritmik Tahminlerin Değerlendirilmesinde Kişisel İlginin Etkisi

Nalan Akın

Psikoloji, Yüksek Lisans

1 Ekim 2024

Kişisel ve profesyonel yaşamlarımızı çevreleyen yapay zeka algoritmaları hızla gelişiyor olsa da, bu algoritmaların insan yargısına kıyasla üstünlüğü, onların karar verme süreçlerinde kullanılacağını garanti etmemektedir. Basit karar kurallarından yapay zeka sistemlerine kadar farklı gelişmişlik düzeyine sahip algoritmaların kullanımını inceleyen önceki araştırmalar, hem algoritmalardan kaçınılması hem de takdir edilmesi olgularını uzun süredir tartışmaktadır. İnsanların algoritmalara yönelik beklenti ve tepkilerinin farklı koşullar altında önemli ölçüde değiştiği görülmüştür. Çalışmalar daha çok algoritma tasarımı ile bireysel faktörler arasındaki etkileşime odaklanmış, kararın verildiği görevle ilgili değişkenlerin rolü üzerine yeterince durulmamıştır. Hissedilen İlişkililik Teorisini (Celsi & Olson, 1988) temel alan bu çalışma, bir göreve duyulan ilgi düzeyinin, bir yapay zeka algoritmasına veya bir insan grubuna ait olduğu söylenen tahminlerin algılanan doğruluğunu nasıl etkilediğini incelemektedir. Katılımcılar (n = 78) üç gruptan birine atanmış ve yapay zeka algoritması, önceki katılımcılar ya da belirsiz bir kaynak (kontrol grubu) tarafından 30 trivia sorusu için yapılan tahminlerin doğruluğunu değerlendirmeleri istenmiştir. Lineer karma model ile yapılan analiz, yapay zeka algoritmasının doğruluk değerlendirmelerinin, özellikle katılımcıların daha fazla ilgi duyduğunu belirttiği konu kategorileri için, hem önceki katılımcılar grubundan hem de kontrol koşulundan daha yüksek olduğunu göstermiştir. Bu sonuç, insanların yapmakta olduğu göreve daha fazla ilgi duyduğunda tahminlerin kaynağından daha fazla etkilendiklerini ve bunun da yapay zeka algoritmaları ve diğer kaynaklara dair yargılarını şekillendirdiğini göstermektedir. Son olarak, ulaşılan bulguaların genellenebilirliği tartışılmış ve insan-algoritma etkileşimi araştırmaları için gelecekteki olası yönler sunulmuştur.

ACKNOWLEDGEMENTS

Writing this thesis has given me something just as valuable as the academic skills I've gained: the chance to witness the depth and unwavering nature of the support and trust from the people in my life.

I am deeply grateful to my advisor, Zeynep Cemalcılar, who has supported me every step of the way, from the day we first met when I shared my vision, to the day this work took its final shape. Thank you for always stepping in and making every task easier when I was on the verge of turning it into something insurmountable. Hearing and feeling your trust in me has ended many of the days I spent doubting myself and my work. I am excited about the new things we will create together, and I cannot thank you enough for helping me make the right decisions for my future.

I would like to thank Lemi Baruh, whose invaluable insights and feedback in the Social Interaction and Media Lab helped me learn by doing and find my way even in uncharted territories. His passion and dedication to his work have left a lasting impression, making him someone I will always look up to. I am also thankful for his constant involvement throughout my thesis and his participation in my defense committee. I extend my gratitude to the other member of my committee, Terry Eskenazi, who, During my five months at Utrecht University, I had the privilege of discussing my thesis ideas, from the design phase to analysis, with Baptist Liefooghe, who listened and improved them with his insightful feedback. I am also grateful to all the members of SCAI for widening my perspective in ways I had not imagined, thanks to being surrounded by people who shared my interests.

A heartfelt thank you to my dear friends at Koç University, Betül, Gamze, Zeynep Betül, Zeynep, Hande, Khaled, and Mehmet. It has been a joy to have them by my side both in school and in life. Special thanks to my close friend Melek, who has walked similar paths with me since our undergraduate days and who has never left me alone during this journey.

I owe endless thanks to my mother, father, brother, and grandfather for always trying to do their best for me and for striving to ensure that, sooner or later, everything turns out as I had hoped.

Finally, I thank The Scientific and Technological Research Council of Türkiye (TÜBİTAK) 2210-A Graduate Scholarship Program for their support during my master’s education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1:
INTRODUCTION

Recommendations and predictions from artificial intelligence algorithms are increasingly being used in a growing number of decision-making tasks and domains. New tools are added daily that can be used for a wide range of purposes, from finding personal style (Davenport & Miller, 2022) to medical diagnosis (Cadario et al., 2021). Thus, when you have a question about a recipe while cooking, now you have an option instead of calling your mom or looking at online forums to find an answer: models adorned with advanced machine-learning algorithms (Leleko & Holoborodko, 2024). Could these be the first source people turn to when they have problems to solve or decisions to make? Recent research shows that this does not depend solely on the success of the algorithm (Burton et al., 2020). In investigating which features of the algorithms might affect this possibility, researchers have asked various questions about the nature of the algorithms, tasks, users, and the possibilities of interaction among the three (for a review, see Mahmud et al., 2022). However, the role of task-related variables remains relatively unexplored, with many unanswered questions regarding the interaction between users and tasks. For example, would the level of interest the person cooking the dish have in this task affect how they evaluate the outputs of different sources, especially an AI algorithm? This study aims to examine how participants' approaches to predictions made by AI algorithms and other people change depending on their level of interest in the task.

1.1 Perceived Accuracy of Algorithms

The interdisciplinary nature of human-algorithm interaction research brings the advantage of connecting problems in adopting algorithms that are detected empirically and insight and solutions derived from experimental research. Developers who recognized the value of the latter also started to observe how individuals use new AI systems to make decisions rather than solely testing the performance of AI systems (Lai et al., 2021). Green and Chen (2019) demonstrated that broadening assessments of algorithmic tools beyond models' statistical properties to consider sociotechnical contexts is critical since people were unable to reliably evaluate their own predictions or

the risk-assessment algorithm and calibrate their decisions accordingly. Similarly, Tsai et al. (2021) pointed out the challenge that research on online symptom checkers (OSC) compares their diagnosis and advice to medical experts while neglecting users' evaluations of OSCs' validity. Thus, they observed whether providing explanations for the OSC's results impacts health consumers' interaction with it. The results showed that the explanation and justification of OSC's outcomes, both by giving an example and summarizing the user answers, significantly increased perceived diagnostic quality and trust in diagnosis (Tsai et al., 2021).

In another attempt to incorporate users, Kocielnik et al. (2019) sought a technique that decreases the discrepancy between the true and perceived accuracy of the AI-based scheduling assistant by adjusting inflated user expectations. Explicitly indicating system accuracy, explaining how the system works based on an example, and giving a chance to interact with and control the system were all effective mechanisms to prepare users for the system's imperfections, which increased satisfaction and acceptance of the scheduling assistant. Such research contributes to the field, especially in the form of practical implications. However, their primary interest is usually improving the model/system operated in the study. More collaboration between practitioners and researchers from other disciplines, such as psychology and decision sciences, could accelerate understanding human-algorithm interaction mechanisms.

Management, information technologies, consumer research, and social sciences are more inclined to put the human component of human-algorithm interaction at the center of their research (e.g., André et al., 2018; Saragih & Morrison, 2022). The majority of studies have aimed to explore antecedents of preference for algorithms versus humans' predictions and attitudes toward compared sources. Even though it is repeatedly evidenced that algorithmic predictions have superior accuracy than people (for a meta-analysis, see Grove et al., 2000), Dietvorst et al. (2015), who coined the term "algorithm aversion," documented that when people experience algorithms and inevitably witness them making mistakes, an aversion towards algorithms is prone to happen. Reluctance to tolerate algorithmic errors (Sutherland et al., 2016; Rebitschek et al., 2021), an expectation of perfect algorithms (Dzindolet et al., 2001), lack of emotional trust (Zhang et al., 2021), and anxiety about one's ability to handle algorithms (Van Esch et al., 2021) were among factors of algorithm aversion. On the other hand, it was possible to overcome algorithm aversion by giving people the control they need by letting them

modify decisions (Dietvorst et al., 2018; Schaap et al., 2023; Sinclair-Desgagné, 2024), showing algorithms' ability to learn from mistakes (Berger et al., 2021; Reich et al., 2023), and sharing accuracy performance of algorithms (You et al., 2022).

More recent studies demonstrated the phenomenon of algorithm appreciation under both opposing and parallel conditions to previous research (Araujo, 2020; Shin et al., 2020). Algorithmic advice was more effective in changing people's predictions than advice from other people, regardless of whether they saw either one of them or both (Logg et al., 2019). That was applicable to an objective task like weight estimation and subjective tasks such as predicting song popularity and romantic attraction. Moreover, unlike previous studies that contained statistical models and formulas in the definition of algorithms (Önkal et al., 2009; Prahl & Van Swol, 2017), research from the last decade has dealt specifically with reliance on AI algorithms as the use of AI has become rapidly widespread. This new stream of research has signaled that interaction with algorithms may move beyond appreciation. In one such study, just being aware that an AI generated the particular advice led to overreliance on it in an uncertain situation, which led to undesired outcomes for partners in a trust game (Klingbeil et al., 2024).

In an attempt to test a potential mechanism to explain contradicting results on the effects of algorithmic advice, manipulating only the framing of sources for the same advice altered its influence on participants' final judgments since descriptions of sources affected levels of perceived source competence (Hou & Jung, 2021). The most influential source was human experts, followed by an AI system with significantly higher influence than a person and a Mturker. Crucial to the current research, the influence factor of an AI system and an algorithm did not differ significantly. In contrast, Candrian & Scherer (2024) showed that calling a system "AI" after a system error occurs causes less negative reactions and less damage in trust and satisfaction than when the same error is suggested to be "algorithmic," which indicates terminology makes a difference in users' level of trust to AI vs. algorithm as well. In this study, we use the term algorithm as an umbrella term encompassing AI algorithms and any mechanical step-by-step procedures for making predictions.

Despite its limitations, previous research looked into several factors associated with the perceived accuracy of algorithms versus humans. To begin with, earlier studies included a measure of trust, although their primary interest was behavioral outcomes in different domains. For medical and legal decision scenarios where decisions were made

about participants, human experience and intuition were perceived as more accurate and fair than statistical formulas (Eastwood et al., 2012). In another scenario-based study, trust in the recommendation of human professionals was higher than in computers despite the superior performance of the latter (Promberger & Baron, 2006). Task domain was a major factor for perceived self and algorithm accuracy, while neither prior experience with algorithms nor self-confidence in a task had a significant impact on perceived accuracy or reliance on algorithms (Jessup et al., 2024). In addition, users perceived algorithms with slow response times as less accurate but judged people's speed oppositely because slowness was related to higher effort and quality for people but not for algorithms that are expected to perform any task easily (Efendić et al., 2020). Notably, Schmitt et al. (2021) defined a paradoxical gap where people follow the AI system's incorrect recommendations despite a decrease in their trust. This indicates that stated perceptions may not predict actual behaviors equally under all circumstances. Therefore, more studies are needed to incorporate perceptions of accuracy, trust, and confidence in algorithms and to discern nuances in their effects on human behavior.

As it appears, previous studies mainly consisted of behavioral outcomes, such as whether people follow or discount algorithms' predictions. Challenges encountered in interacting with real AI algorithms reveal that we need to investigate user perceptions of such algorithms more closely to understand how perceptions change users' behaviors and choices. Yet the mediatory role of confidence in the study of Dietvorst (2015) has received less attention than the behavioral outcome which demonstrated people's reluctance to bet on algorithms' forecasts of student success. They found that the decline in confidence in the algorithmic estimates due to witnessing it making errors was the main cause of betting on human judgment over algorithms. More studies to expand the exploration of confidence in algorithmic predictions are needed, and task-related factors, which have received relatively less attention in this interaction, should also be considered as part of this endeavor.

In their systematic review, Mahmud et al. (2022) pointed out that relatively limited research exists that focuses on task-related factors of algorithm aversion. There have been efforts to understand the nature of the task, such as their objectivity, need for mechanical skills, and level of difficulty. In terms of task objectivity, a recurring finding has been that objective tasks like predicting the weather and giving directions breed less aversion and more trust towards algorithms in comparison to subjective tasks such as

predicting personality traits and recommending music (Castelo et al., 2019). Likewise, people perceived algorithms and humans as equally fair and trustworthy when they performed mechanical tasks (e.g., work assignment), whereas algorithms were judged as less fair and trustworthy due to their perceived inability to apply intuition and subjective judgment in a task that is assumed to require human skills (e.g., hiring) (Lee, 2018). As for the task difficulty, findings have suggested that it was easier to ignore algorithmic output for simpler tasks (Önkal et al., 2019), but people tended to follow algorithmic suggestions more as task difficulty increased (Bogert et al., 2021).

The only task-related interaction factor investigated in detail so far is the individuals' expertise in the task. Expertise (i.e., substantial domain knowledge) is found to be positively associated with algorithm aversion, even though this decreases the achieved accuracy level in human algorithm collaboration (Arkes et al., 1986; Logg et al., 2019). Hence, the ideal expertise level for algorithm-augmented works was intermediate, suggesting opposing roles of users' skills and their level of algorithmic aversion (Allen & Choudhury, 2022). In the current study, we will explore another task-related factor by examining the impact of interest in the task on evaluations of prediction accuracy of the AI algorithms and other people.

1.2 The Role of Task Interest in Perceptions toward AI-powered technologies

Although the role of interest in the task has not been sufficiently investigated in the context of algorithms, consumer behavior researchers examined the role of interest in the adoption of automation and robots (Leung et al., 2018). For daily and recreational activities such as cooking, driving, and fishing, identity-based consumption, where consumers seek products associated with self, differed from adopting automation for convenience (Leung et al., 2018). In this case, the need for internal attribution made automation of baking and fishing less desirable, especially if the task required high skills to achieve since automation could undermine the credit taken for outcomes. Crucially, controlling expertise did not eliminate the effect of identification with the task. Likewise, consumers resisted robotic labor when higher symbolic value was attached to the products and services, such as artwork and a tattoo (Granulo et al., 2020). The authors suggested this was due to the consumers' desire for uniqueness, which is believed to be more closely linked with human labor. Both studies signaled

resistance to technologies if the context in which they are consumed involves personal motives.

When it comes to AI algorithms that make recommendations or predictions, the Theory of Felt Involvement (Celsi & Olson, 1988) may be crucial in informing the motivational role of interest in the processing and evaluation of informational stimuli generated by algorithms and alternative sources. Felt involvement is defined as an acute state experienced in the presence of information personally relevant to one's values, needs, and goals (Celsi & Olson, 1988, p. 211). Intrinsic and situational factors are found to determine its magnitude; thus, felt involvement is open to manipulation from outside. Once activated, it enhances attention and comprehension processes; specifically, people attend longer, put in more cognitive effort, create more thoughts, and make more inferences. Additionally, Celsi and Olson's (1988) findings revealed that felt involvement affected motivation to process, whereas domain knowledge influenced the ability to process through elaborating inferences. Thus, despite the positive association between felt involvement and domain knowledge, it is observed that they have independent effects on information processing.

Recently, the Theory of Felt Involvement has been utilized to understand the impact of involvement in various areas that emerged due to new technologies, such as social media advertising (Alalwan, 2018), virtual reality spectatorship (Kim & Ko, 2019), and influencer marketing (Trivedi & Sama, 2020). Regarding AI, Sanusi et al. (2024) surveyed Nigerian pre-service teachers and suggested that personal relevance predicted their intention to use AI for occupational purposes. However, no attempt has been made so far to connect individuals' felt involvement with tasks and their evaluations regarding the predictions made by AI algorithms. This association has the potential to offer a significant contribution to the human-algorithm interaction field, providing a basis to examine the differential effects of attention to algorithms and alternative sources such as a group of people, the cognitive effort devoted to interacting with them, comprehension of their outputs, and the level of elaboration in the analysis and evaluation of these sources.

1.3 The Present Study

In this research, we investigate the role of interest in the task in the evaluation of predictions' accuracy presented as coming from an AI algorithm versus a group of people (i.e., previous participants) by testing the hypothesized interaction between interest in the task and source of prediction. The evaluation metric of focus is the perceived accuracy of predictions. It is chosen owing to practitioners' empirical observations that people struggle to evaluate accuracy (Green & Chen, 2019) and the need for more research endeavors to understand its role in algorithm aversion and appreciation. The present study also aims to replicate three previous findings regarding preference for algorithms in decision-making: First, algorithm appreciation was observed for a) objective tasks that do not require human intuition (Castelo et al., 2019; Lee, 2018), b) difficult tasks (Bogert et al., 2021) and when sources had similar procedures and competences (Sunstein & Reisch, 2023).

In order to achieve these goals, a new task was designed that proceeded in a trivia quiz style, where people were asked questions on various subjects. Questions were composed of six different subjects. In this task, participants are not expected to predict answers themselves but to assess the accuracy of predictions of a given source, an AI algorithm, or other people who previously participated in a similar trivia quiz. Choices and descriptions of the sources were built on prior research. Based on insight from Bogert et al. (2021), the AI algorithm condition was alternated with an average of a group of people to compare its influence to social influence or the wisdom of the crowd. Information on how predictions were made was the same across conditions so as not to influence the perceived competence of sources (Hou & Jung, 2021; Logg et al., 2019).

A more favorable evaluation of the AI algorithm compared to humans would indicate the perceived superiority of AI algorithms. Central to the aims of the present study, the differences across sources in their perceived accuracy are expected to be significant only when questions in the trivia quiz came from subjects in which participants had a higher interest. This expectation is grounded in the idea that higher interest is associated with the allocation of greater cognitive effort to comprehend stimuli. Consequently, this can enhance the salience of information sources, making differences in perceived accuracy between sources more pronounced. We referred to *interest* in the current study to capture participants' engagement with different trivia

subjects, as using 'personal relevance' or ‘involvement’ (Celsi & Olson, 1988) seemed less appropriate in this context. To measure interest, before starting the trivia quiz, participants rate their level of interest in 6 subjects: Art, Animals, Entertainment, Geography, History, Human body and health, Science, and Sports. Then, all participants are shown the same set of questions spanning these six subjects in a random order, irrespective of their level of interest.

Particularly, in the present study, we test the following hypotheses:

H1- The perceived accuracy of predictions made by the AI algorithm will be higher than those made by previous participants and the control group (where predictions do not refer to a specific source).

H2- The effect of source on the perceived accuracy will differ depending on the participants’ level of interest in the trivia quiz subject. Specifically, when the participants’ interest is high, the perceived accuracy of predictions made by the AI algorithm will be higher than those of both previous participants and the control group, whereas when their interest is low, there will be no significant differences among the AI algorithm, previous participants, and control group.

H3- In the AI algorithm condition, the perceived accuracy of the prediction will differ depending on the participants’ level of interest in the trivia quiz subject, with high interest leading to greater perceived accuracy than a subject category with low interest.

Chapter 2:
METHOD

2.1 Pilot

The pilot study was conducted to ensure that the questions were sufficiently challenging such that the majority could not give the true answers. It also served to select 30 appropriate questions from the initial 120, with 5 chosen for each subject category. The pilot task was introduced as a trivia quiz, and participants were asked to evaluate the subject categories, the difficulty of questions, and their confidence in the predictions they made. The sample was reached via the Utrecht University Sona System. Each of the 19 participants answered 32 questions randomly selected from a pool of 120 numerical questions. The pool consisted of an equal number of questions

from 8 different subject categories: Art, Animals, Entertainment, Geography, History, Human body and health, Science, and Sports. A previous study (Nelson, 1980) and trivia quiz websites were searched while preparing questions, but most were created using ChatGPT. All answers were checked from different sources to make sure they were accurate.

The pilot study was crucial to avoid the level of knowledge intervening in the evaluation of accuracy because if participants knew the true answers, then they could objectively evaluate the source's predictions. This would have undermined the study design, as the goal was to measure the perceived accuracy of predictions provided by the AI algorithm and previous participants, which is a subjective metric. For instance, we observed that questions like “How many times is the word ‘blood’ written in the Shakespeare play “Macbeth?” or “How many copies of Minecraft have been sold from its launch to 2021?” were rated more difficult than questions like “What is the speed of the fastest land animal, the cheetah, in kilometers per hour?” or “How many States does the United States consist of?” for which one and three participants already knew the correct answers, respectively. The former questions were included in this case, whereas the latter was excluded.

Based on the results of the pilot, science and history categories were excluded. The design required questions that were challenging to pinpoint the exact answer but open to making the best guess simultaneously, such as “What is the longest recorded time a human has stayed awake without sleep, in hours?”. It was harder than other categories to find enough science and history questions that met these criteria. For example, someone might know, “How long did the Hundred Years' War last?”, but without any information on this historic event, it is hard to predict or judge whether a prediction is reasonable. This would have complicated accuracy ratings.

The choice of 30 questions was guided by insight from the pilot study, and they exhibited the following characteristics. First, an equal number of questions came from 6 subject categories: Art, Animals, Entertainment, Geography, Human body and health, and Sports. Second, all questions were numerical to be convenient to assess predictions’ closeness to being accurate on a scale of 1 to 10. Third, the difficulty of the questions ensured that the ratings of accuracy were based on subjective evaluation rather than a comparison to the true answers available to participants. The mean difficulty rating of the questions was 5.97 (SD = 2.53) out of a maximum of 10.

2.2 Participants

Eighty-two participants were recruited via Prolific (app.prolific.com). The only inclusion criterion was being fluent in English. Four respondents who failed the manipulation check were excluded from the study. When asked who the source of predictions was, two people in the previous participant condition who said it was themselves and Artificial Intelligence were excluded. Two more people from the AI Algorithm condition who could not answer both this question and the source of the algorithm's training data were excluded. As a result, the analysis was conducted with a sample of 78 participants. There was an equal number of females and males; their ages ranged from 20 to 48, with a mean of 26.35 (SD = 6.05). A significant portion of the sample was from South Africa, Poland, and Portugal, with 17, 16, and 14 participants, respectively. 30.8% of participants were college graduates, followed by 29.5% high school graduates, 23.1% with post-graduate or professional degrees, and 16.7% with some college or technical training beyond high school. 48.7% of participants reported being employed, whereas 34.6% were students. More detailed sample characteristics can be found in Table 1. All participants were compensated with £3.00. Data collection took place on May 30, 2023.

Table 2.1: Demographic characteristics of the sample.

Characteristics	Categories	N	%
Gender	Male	39	50
	Female	39	50
Education Level	High school graduate	23	29.5
	Some college or technical training	13	16.7
	College graduate	24	30.8
	Postgraduate or professional degree	18	23.1
Employment Status	Employed or self-employed	38	48.7
	Student	27	34.6

2.3 *Experimental Task*

The task was built on Gorilla software (https://gorilla.sc). Participants were asked to evaluate the accuracy of 30 predictions, chosen based on the pilot study, that served as answers to trivia questions. Firstly, only the question was displayed on the screen. On the next page, the prediction was presented under the question box. It was the true answer and the same across three conditions: the AI algorithm, previous participants, and control group. However, the given source of prediction changed depending on the assigned condition of the participants (for details, see section Procedure). Once they clicked *Next*, a slider appeared on the screen from 1 (not accurate at all) to 10 (extremely accurate). At the end of the trial, they rated the difficulty of the trivia question. Boxes, colors, and a minimal illustration were used in the design of the displays to make it look more like a trivia game.

2.4 *Procedure*

Suitable participants on Prolific who speak English fluently were automatically directed to Gorilla. Only laptops were allowed to take part in the experiment. Participants started by confirming that they understood all the information about the experiment. They were randomly assigned to one of the three conditions: the AI algorithm, previous participants, and the ambiguous source (control group). On the welcoming page, the goal of the research was conveyed as "understanding how people from different backgrounds perform on a general knowledge test in a series of studies." Participants were asked for their participation in evaluating the difficulty of trivia questions and the accuracy of the 30 estimates. On this page, nothing about the sources of estimates was revealed. Next, participants answered eight demographic questions. Then, they sorted six subject categories based on their level of interest and rated their level of knowledge for each using sliders. Regardless of interest level, each participant was presented with the same set of questions covering all six subject categories in a random order. The chosen order of subject categories was used in the data analysis stage to create a variable of interest in the task with two levels.

A practice round was used to explain the task and implement manipulation. The practice question was, "What is the speed of sound in the air under normal conditions, in meters per second?". Between displaying the question and the prediction, participants

in the control condition were shown the message: “After you read the question, we will present an estimate for you to evaluate its accuracy.” On the contrary, at this stage, a source was introduced in other conditions. In the AI Algorithm condition, it was stated, “We will present the prediction of an AI algorithm trained on estimates of participants from a past experiment.” In the previous participants condition, the statement “We will present the average estimate of participants from a past experiment for you to evaluate its accuracy” was inserted.

Next, they engaged in 30 trials without time limitations. Source was a between-subjects condition, which means participants saw predictions from only one source throughout all 30 questions. Except for the control group, predictions were displayed with their source to strengthen the manipulation, so the sentences began with “The prediction of an AI algorithm trained on the estimates of past participants was…” or “The average estimate of participants from a past experiment was…”. Next, they evaluated the difficulty of the question. The order of the subject categories and the five questions within them were randomized. Participants in the AI algorithm and previous participants conditions answered manipulation questions. Both groups answered who came up with the predictions made for trivia questions. Additionally, only the AI algorithm group was asked about the source of the data on which the AI algorithm in the study was trained. The average time of completion was 18.13 minutes (*SD* = 10).

2.5 Data Preparation

The data was restructured in Excel to create an unaggregated long-format version, which was desired for the linear mixed-effects modeling (Brown, 2021).

The order in which participants sorted their interests in subject categories was recorded as numbers between 1 and 6. The first three and last three categories were recoded into a different variable: high interest and low interest. Three participants chose the same subject category more than once while sorting them based on their interest in them. That caused 35 accuracy ratings to be lost for four different subjects (5 animals, 15 art, 10 geography, and 5 entertainment predictions). After their exclusion, the final data comprised 2305 rows of individual observations.

Chapter 3:
RESULT

3.1 Analysis Strategy

A linear mixed-effects model (LMM) with crossed random effects was run to analyze data. The study had a repeated-measures design in which all participants were engaged in making predictions for all questions. LMM was preferred due to its advantage of modeling participant and question variability simultaneously, unlike ANOVA (DeBruine & Barr, 2021). It also deals with missing data better by treating each observation individually rather than as a part of a participant's response (Brown, 2021). Yet, a potential drawback of LMMs is the abundance of analytic decisions to be made and the lack of common standards (Meteyard & Davies, 2020). This study followed the guidance of Meteyard and Davies (2020).

The R packages lme4 (Bates et al., 2020) and lmerTest (Kuznetsova et al., 2016) were employed to run the models. It uses restricted maximum likelihood (REML) in this process. Following Arend & Schäfer (2019), a post hoc power analysis based on Monte Carlo simulation was implemented in the R package SIMR (Green & MacLeod, 2016). Other helpful R packages were emmeans (Lenth, 2023) to conduct custom contrasts and MuMIn (Barton, 2024) to calculate R-squared values.

3.2 Model Building

The modeling process started with decisions regarding random effects. Random intercepts were included to model random deviations from the mean accuracy rating and thus solve the nonindependence problem (Brown, 2021). Model comparison went from simple to complex, and the criteria were the results of the Likelihood Ratio Test (LRT) and the Akaike Information Criterion (AIC). LRT revealed that including by-items random intercepts improved the model fit significantly, $\chi 2\ (1) = 43.07$, $p < .001$. Considering moderate intraclass correlation and the variance values of items, this was a necessary output for the decision to warrant the inclusion of this random effect. No convergence issues occurred throughout the process, the details of which can be found in Table 1.

Then, three fixed effects were added to the model all at once. Interest was a within-participants, between-items factor that has two levels (low and high). Source was a between-participants, within-items factor that has three levels (AI Algorithm, Previous Participants, and Control). The third factor was the interaction between source and interest. The final model with an AIC of 10231 showed an improvement relative to the preceding models.

Table 3.1: Model comparisons for random effects.

Model specification	Model fit		LRT Test		
	AIC	LL	df	$\chi 2$	*p*
Random intercept for participants only	10280	-5137.1			
Random intercepts for both participants and items	10240	-5116	1	42.12	<.001
Random intercepts + Fixed effects	10231	-5106.6			

Note. Number of total observations = 2305; Number of subjects = 78; Number of items = 30.

3.3 H1: Source Impacts Perceived Accuracy of Predictions

To analyze whether participants evaluated the accuracy of sources in the study differently, a separate model without the interaction term was run to assess the main effect of source, controlling for interest. It revealed that the effect of source was significant, $F(2,74.9) = 4.48$, $p = .01$, $\eta^2 = 0.11$. Post hoc comparisons with Bonferroni adjusted alpha levels demonstrated that accuracy ratings for the AI algorithm condition ($M = 7.04$, $SD = 2.11$) were higher than in both previous participants ($M = 6.20$, $SD = 2.17$), $t(75) = 2.73$, $p = .02$ and the control condition ($M = 6.32$, $SD = 2.72$), $t(74.7) = 2.78$, $p = .04$. The comparison of the previous participants condition with the control condition was non-significant, $t(75.2) = .30$, $p = 1.0$. According to post hoc power analysis based on 1000 simulations with a significance criterion of $\alpha = .05$, the observed power for the effect of source is 78%. In conclusion, the perceived accuracy of algorithmic predictions was significantly higher than those of previous participants and the control group.

3.4 H2 and H3: The Effect of Source Depends on the Interest in the Task

The interaction between source and interest was found significant, indicating that the effect of source on the perceived accuracy differed depending on the interest levels of the participants, $F(2, 2212.7) = 3.90$, $p = .02$. Based on 1000 simulations and a significance threshold of $\alpha = .05$, post hoc power analysis indicated that the observed power for this effect was 67.3%. Custom contrasts were performed to test our hypotheses using Bonferroni-adjusted alpha levels of .007 per test (.05/7). When the interest was high, a pattern similar to the main effect of source was observed. The accuracy ratings for the AI algorithm ($M = 7.19$, $SD = 2.02$) were higher than in both previous participants ($M = 6.25$, $SD = 2.21$), $t(97.7) = 2.94$, $p = .03$ and the control condition ($M = 6.18$, $SD = 2.76$), $t(97.1) = 3.30$, $p = .01$. However, no significant difference existed between any pair of conditions when the interest was low. Neither previous participants ($M = 6.15$, $SD = 2.13$), $t(98.1) = 2.16$, $p = .23$ nor control condition ($M = 6.46$, $SD = 2.69$), $t(97.1) = 1.36$, $p = 1.0$ differed from AI algorithm condition (see Figure 1). To conclude, higher task interest increases the perceived accuracy of AI predictions in comparison to previous participants and control groups, whereas low task interest reduces this effect.

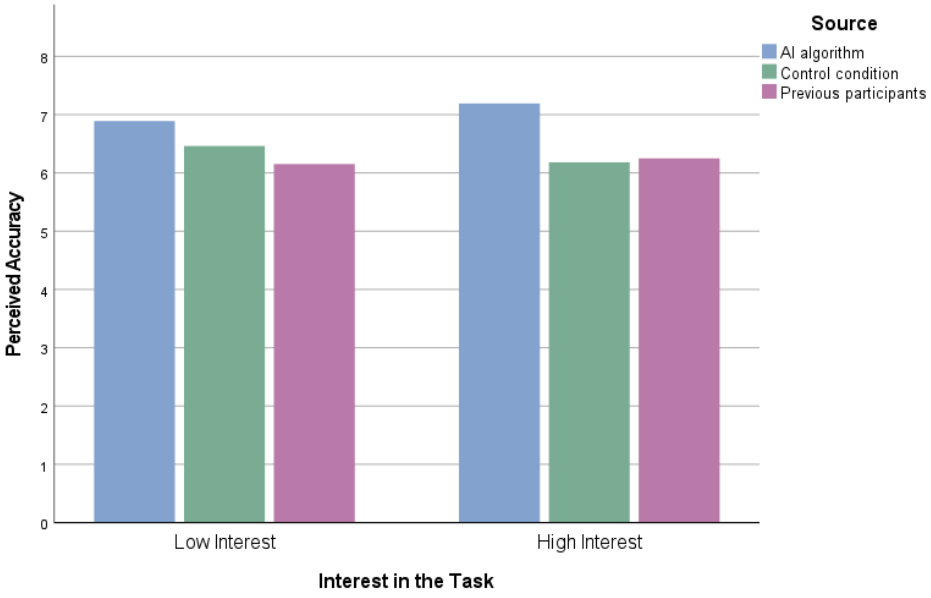

Figure 3.1: Perceived accuracy of predictions by source across levels of task interest

For the last hypothesis, whether interest impacts perceived accuracy within the same source was tested. In the AI algorithm condition, there was no significant difference between the accuracy ratings for subject categories with high (M = 7.19, SD

= 2.02) versus low interest (M = 6.89, SD = 2.20), t(2205.5) = 1.54, p = .87. Similarly, comparison within previous participants condition, t(2186.1) = 0.04, p = 1.0 or control condition t(2221.3) = -2.31, p = .14 did not reach significance level (see Figure 2). Hence, the results did not confirm the third hypothesis.

A summary of fixed effects and random effects is presented in Table 2.

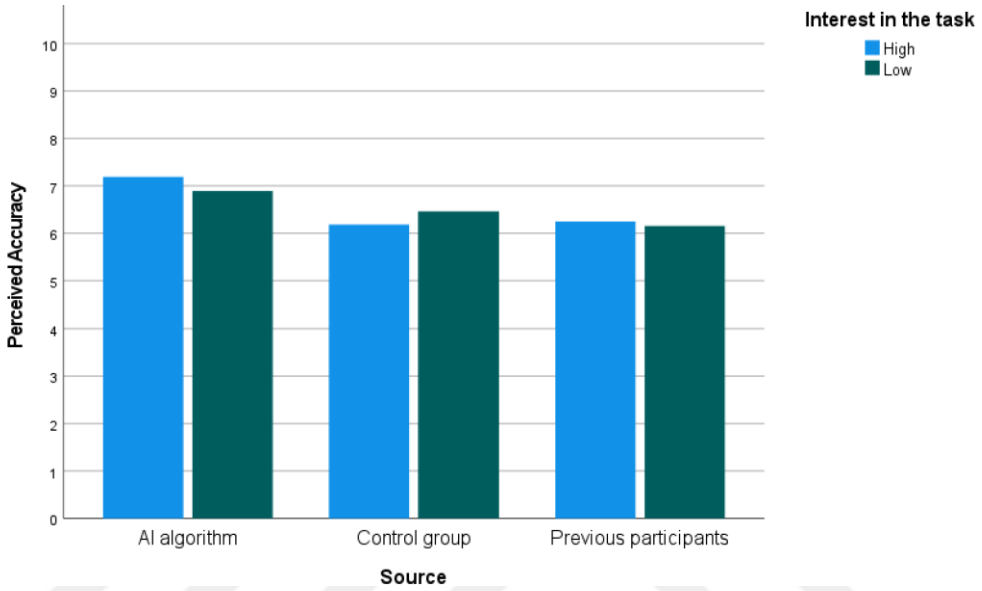

Figure 3.2: Perceived accuracy of predictions by task interest across source types

Table 3.2: Results from linear mixed-effects model.

Fixed Effects	*F*	*Num df*	*Den df*	*p*
Source	0.102	1	1958.1	0.75
Interest	4.483	2	74.9	0.014
Source*Interest	3.898	2	2212.7	0.02

Random Effects	*Variance*	*SD*
Participant (Intercept)	0.94	0.97
Items (Intercept)	0.17	0.41

Model fit		
R^2	*Marginal*	*Conditional*
	0.02	0.21

Note. p-values for fixed effects calculated using Satterthwaites approximations.

Model equation: Accuracy ~ 1 + Interest + Source + Interest:Source +(1 | question_item) + (1 | PartID)

Chapter 4:
DISCUSSION

The present study explored the influence of interest in the task on the perceived accuracy of algorithmic versus human predictions. We designed an experiment in which participants saw predictions of only one source for all numerical trivia questions and evaluated their accuracy. Building on the Theory of Felt Involvement, we hypothesized that participants' accuracy ratings for the AI algorithm would diverge from other sources particularly for predictions related to subjects with higher interest. We also expected to observe algorithm appreciation under the conditions in our design, which had previously been identified as contributing to this phenomenon.

Confirming our first hypothesis, the main effect of source was significant. When predictions were presented as coming from an AI algorithm trained on the estimates of past participants, participants perceived them as more accurate than those attributed to the average estimate of previous participants or predictions provided without any source information. There was no significant difference between the perceived accuracy of previous participants and predictions without any source information in the control condition. Algorithm appreciation has previously been associated with objective (Castelo et al., 2019), mechanical (Lee, 2018), and difficult tasks (Bogert et al., 2021), and our first finding is consistent with these established conditions. Considering that predictions participants saw were the same regardless of the source demonstrates that the difference framing one source as an "AI algorithm" creates is substantial enough to affect people's perceptions of accuracy significantly.

In alignment with our second hypothesis, a significant interaction effect was found between interest in the task and source. This means that the effect of source on the perceived accuracy of predictions was stronger when the question subjects were of high interest compared to when they were of low interest. In particular, participants reported higher accuracy ratings for the AI algorithm than the previous participants and control condition at high interest, whereas no significant difference was found at low interest. These findings can be discussed with respect to the Theory of Felt Involvement (Celsi & Olson, 1988). Accordingly, one possible explanation is that participants paid greater attention to source information when their interest in the subjects was higher,

which enhanced their ability to process and evaluate the source and its output. Then, the prominence of the AI algorithm suggests that it was evaluated more favorably than previous participants in terms of accuracy. This provides initial evidence that the Theory of Felt Involvement can help explain user perceptions and experiences in their collaboration with algorithms for tasks in which they have varying levels of attachment.

Nevertheless, we did not find a significant difference when higher interest and lower interest conditions were compared within the AI algorithm group, unlike the significant differences found between the AI algorithm, previous participants, and control conditions in previous comparisons. The failure to confirm the third hypothesis contradicts the assumptions underlying the Theory of Felt Involvement. One reason might be the simplicity of the faux AI algorithm, which is what people may notice when they attend to the source description (i.e., an AI algorithm trained on the estimates of past participants). Comprehension of this detail may not have added any further impact beyond the initial effect of the 'AI algorithm' label. In addition, this finding might imply that the source can play a more pivotal role in shaping perceptions than participants' level of interest in subject categories within the AI algorithm context. Another possibility might be a limitation in our methodological choices, which led to failure to reflect a broad spectrum of participants' interest in the subjects. Indeed, we used the order in which people ranked the subject categories of questions based on their level of interest to determine the questions corresponding to two categories (i.e., higher interest and lower interest) for each participant. As a more nuanced approach, Celsi and Olson (1988) categorized participants into four levels of personal relevance based on scores they got from their responses to bipolar adjectives.

In studies where people interact with algorithms, the selection and implementation of tasks are closely related to the generalizability of results. The majority of user-centered research starts with describing what a supposed algorithm does and how it works. Then, the algorithms' (faux) outputs, like predictions and advice, are presented, which are, indeed, predetermined by researchers (e.g., Bigman & Gray, 2018; Bogert et al., 2021). This procedure is extensively practiced since researchers need experimental control, and building real AI algorithms is challenging, but as a byproduct, participants rarely truly experience using real algorithms. Eventually, it can undermine external validity (Mahmud et al., 2022), as it implies asking participants to

imagine a decision-making scenario involving algorithms. In the real world, the consequences of decisions may be more consequential.

Although our task substantially takes after the mentioned procedures and task features, it diverges from some of them in significant aspects. First, participants were asked to evaluate the question and prediction properties under the impression that researchers would use their feedback to revise the task for future general knowledge studies. Thus, rather than representing a specific domain, our task is an artificial task. This can make its generalizability questionable because it does not target an actual domain. However, how and why the algorithm in our study makes predictions can be applied to everyday and professional usage, as it simply takes an average of previous predictions to answer factual questions. It also enables targeting laypeople without any inclusion criteria like domain knowledge. Second, algorithms making personally relevant predictions can be challenging to administer in experiments because they typically require a real algorithm trained on personal information and have higher output variability. For example, an algorithm that predicts a match between users and potential romantic partners must consider users' personal preferences to make realistic predictions. We introduced another way to explore the impact of a motivational personal factor (i.e., personal relevance) on the human algorithm interaction with an objective task by asking about the interest in question subjects. Testing an alternative path, the present study represents a step toward discussing more individual-level variables that can simultaneously affect people’s interactions with tasks and AI algorithms.

Risk, required expertise, and subjectivity are other essential task characteristics for evaluating generalizability (Lai et al., 2021). Our task has low stakes considering its artificial nature and the absence of something to gain or lose, like bonus incentives often offered (Dietvorst et al., 2015; Jessup et al., 2024). Rewards are not appropriate for the present study since participants do not perform but evaluate other sources of information, such as AI vs other people. Finally, our task is objective, and a ground truth exists – a factual answer being sought. However, the required expertise to evaluate the accuracy of predictions is relatively high because the questions in the trivia quiz are designed (and pilot-tested) to be difficult enough to prevent the confounding effect of participants knowing or even guessing the correct answer (For details, see section Pilot).

4.1 Limitations and Future Studies

Our study had limitations regarding its methodology and generalizability. First, due to resource constraints, the sample size was determined without fully aligning with the recommendations of a formal power analysis, potentially affecting the sensitivity of the results. Second, a potential limitation of the theory is that it was initially developed and predominantly applied to customer behavior and marketing research. Although algorithmic predictions are integrated into products and services aimed at customers, the users of algorithms encompass a broader audience that extends beyond buyers. Third, the generalizability of findings is limited by the relatively low stakes of the task, making it difficult to extend these results to high-risk contexts. For example, sources can be evaluated differently in a high-pressure situation, such as creating an original recipe for a competition, compared to experimenting casually to explore new flavors.

Future research should fully leverage the explanatory potential of the Theory of Felt Involvement (Celsi & Olson, 1988) by examining attention processes as a mechanism of the confirmed interaction effect between the source of predictions and interest in the task on perceived accuracy. Researchers should consider evaluating the role of expertise in the task alongside interest in shaping perceptions of algorithmic predictions and look for objective and reliable ways to measure these variables. We did not incorporate expertise into our design because ensuring task difficulty was a higher priority, making expertise less relevant. However, Celsi and Olson (1988) measured domain knowledge in tennis with a 15-question tennis quiz. This could serve as a reliable measure, while our attempt to measure the level of knowledge in subject categories was only a subjective evaluation of participants' expertise and could not be used to control the effect of expertise. In terms of interest, future studies could include multiple tasks or domains, categorizing them as high- and low-interest (e.g., solving mathematical problems and fortune-telling) to observe how individuals collaborate with algorithms in tasks of varying personal relevance.

For our research, we designed a task reflecting machine learning principles, a field that involves using algorithms to learn from data and make predictions (Delgadillo, 2020). Today, AI-driven chatbots utilizing natural language processing (NLP) techniques and trained on significantly larger datasets can play an increasingly prominent role in similar tasks. At the time of data collection, the first version of

ChatGPT had only recently been released. Therefore, we did not ask participants whether they had used it or similar tools. However, with the current popularity of generative AI, any AI-generated prediction or advice might now be perceived as a product of such technologies due to participants' familiarity and prior experience. Thus, future studies should control participants' experience with and understanding of artificial intelligence.

Chapter 5:

CONCLUSION

The present study was the first attempt to apply the Theory of Felt Involvement to people's interactions with AI algorithms in decision-making processes. We introduced the interest in the decision task as a moderator in the relationship between the source of predictions (generally algorithms versus humans) and their perceived accuracy. Our analysis showed that higher interest in the task enhances the effect of source on the perceived accuracy of algorithmic versus human predictions. Specifically, the AI algorithm had significantly higher accuracy ratings than previous participants and the control group. We argue that the differences in perceived accuracy between sources could be explained by greater attention and comprehension, driven by the activation of felt involvement for tasks with higher interest. This study contributes to the human algorithm interaction research by investigating perceptions of AI algorithm's accuracy compared to a human crowd's accuracy experimentally. Future research could explore ways to measure and manipulate attention and comprehension levels to better understand the mechanism underlying these findings.

BIBLIOGRAPHY

Alalwan, A. A. (2018). Investigating the impact of social media advertising features on customer purchase intention. *International Journal of Information Management*, *42*, 65–77. https://doi.org/10.1016/j.ijinfomgt.2018.06.001

Allen, R., & Choudhury, P. (2022). Algorithm-Augmented Work and Domain Experience: The Countervailing Forces of Ability and Aversion. *Organization Science*, *33*(1), 149–169. https://doi.org/10.1287/orsc.2021.1554

André, Q., Carmon, Z., Wertenbroch, K., Crum, A., Frank, D., Goldstein, W., Huber, J., van Boven, L., Weber, B., & Yang, H. (2018). Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data. *Customer Needs and Solutions*, *5*(1), 28–37. https://doi.org/10.1007/s40547-017-0085-8

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society.*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on monte carlo simulation. *Psychological Methods*, *24*(1), 1–19. Scopus. https://doi.org/10.1037/met0000195

Arkes, H. R., Dawes, R. M., & Christensen, C. (1986). Factors influencing the use of a decision rule in a probabilistic task. *Organizational Behavior and Human Decision Processes*, *37*(1), 93–110. https://doi.org/10.1016/0749-5978(86)90046-4

Bartoń, K. (2024). *MuMIn: Multi-Model Inference* (Version 1.48.4) [R package]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., & Green, P. (2020). *lme4* (Version 1.1-26) [Computer software]. Comprehensive R Archive Network. https://cran.r-project.org/web/packages/lme4/lme4.pdf

Berger, B., Adam, M., Rühr, A., & Benlian, A. (2021). Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering*, *63*(1), 55–68. https://doi.org/10.1007/s12599-020-00678-5

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920960351. https://doi.org/10.1177/2515245920960351

Burton, J. W., Stein, M., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, *33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*(12), 1636–1642. https://doi.org/10.1038/s41562-021-01146-0

Candrian, C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, *134*, 107308. https://doi.org/10.1016/j.chb.2022.107308

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Celsi, R. L., & Olson, J. C. (1988). The Role of Involvement in Attention and Comprehension Processes. *Journal of Consumer Research*, *15*(2), 210–224. https://doi.org/10.1086/209158

Davenport, T. H., & Miller, S. M. (2022). Stitch Fix: AI-Assisted Clothing Stylists. In *Working with AI: Real Stories of Human-Machine Collaboration* (pp. 15–19). Working with AI: Real Stories of Human-Machine Collaboration. MIT Press. https://ieeexplore.ieee.org/abstract/document/9855560/authors#authors

DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920965119. https://doi.org/10.1177/2515245920965119

Delgadillo, J. (2020). Machine learning: A primer for psychotherapy researchers. *Psychotherapy Research*, *31*, 1–4. https://doi.org/10.1080/10503307.2020.1859638

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting Misuse and Disuse of Combat Identification Systems. *Military Psychology*, *13*(3), 147–164. https://doi.org/10.1207/S15327876MP1303_2

Eastwood, J., Snook, B., & Luther, K. (2012). What People Want From Their Professionals: Attitudes Toward Decision-making Strategies. *Journal of Behavioral Decision Making*, *25*(5), 458–468. https://doi.org/10.1002/bdm.741

Efendić, E., Van de Calseyde, P. P. F. M., & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, *157*, 103–114. https://doi.org/10.1016/j.obhdp.2020.01.008

Granulo, A., Fuchs, C., & Puntoni, S. (2021). Preference for Human (vs. Robotic) Labor is Stronger in Symbolic Consumption Contexts. *Journal of Consumer Psychology*, *31*(1), 72–80. https://doi.org/10.1002/jcpy.1181

Green, B., & Chen, Y. (2019). The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359152

Green, P., & MacLeod, C. (2016). simr: An R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. https://doi.org/10.1037/1040-3590.12.1.19

Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–25. https://doi.org/10.1145/3479864

Jessup, S. A., Alarcon, G. M., Willis, S. M., & Lee, M. A. (2024). A closer look at how experience, task domain, and self-confidence influence reliance towards algorithms. *Applied Ergonomics*, *121*, 104363. https://doi.org/10.1016/j.apergo.2024.104363

Kim, D., & Ko, Y. J. (2019). The impact of virtual reality (VR) technology on sport spectators' flow experience and satisfaction. *Computers in Human Behavior*, *93*, 346–356. https://doi.org/10.1016/j.chb.2018.12.040

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3290605.3300641

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). *Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies* (arXiv:2112.11471). arXiv. http://arxiv.org/abs/2112.11471

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 205395171875668. https://doi.org/10.1177/2053951718756684

Leleko, S. & Holoborodko, Y. (2024, June 22). *Machine learning and AI in the Food Industry*. SPD Technology. https://spd.tech/machine-learning/machine-learning-and-ai-in-food-industry/

Leung, E., Paolacci, G., & Puntoni, S. (2018). Man Versus Machine: Resisting Automation in Identity-Based Consumer Behavior. *Journal of Marketing Research*, *55*(6), 818–831. https://doi.org/10.1177/0022243718818423

Lenth, R. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.10.4) [R package]. https://rvlenth.github.io/emmeans/

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, *175*, 121390. https://doi.org/10.1016/j.techfore.2021.121390

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. https://doi.org/10.1016/j.jml.2020.104092

Morewedge, C. K. (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, *26*(10), 824–826. Scopus. https://doi.org/10.1016/j.tics.2022.07.007

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 338–368. https://doi.org/10.1016/S0022-5371(80)90266-2

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*(6), 691–702. https://doi.org/10.1002/for.2464

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468. https://doi.org/10.1002/bdm.542

Rebitschek, F. G., Gigerenzer, G., & Wagner, G. G. (2021). People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Scientific Reports*, *11*(1), 20171. https://doi.org/10.1038/s41598-021-99802-y

Reich, T., Kaju, A., & Maglio, S. J. (2023). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, *33*(2), 285–302. https://doi.org/10.1002/jcpy.1313

Sanusi, I. T., Ayanwale, M. A., & Tolorunleke, A. E. (2024). Investigating pre-service teachers' artificial intelligence perception from the perspective of planned behavior theory. *Computers and Education: Artificial Intelligence*, *6*, 100202. https://doi.org/10.1016/j.caeai.2024.100202

Saragih, M., & Morrison, B. W. (2022). The Effect of past Algorithmic Performance and Decision Significance on Algorithmic Advice Acceptance. *International Journal of Human–Computer Interaction*, *38*(13), 1228–1237. https://doi.org/10.1080/10447318.2021.1990518

Schaap, G., Bosse, T., & Hendriks Vettehen, P. (2023). The ABC of algorithmic aversion: Not agent, but benefits and control determine the acceptance of automated decision-making. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01649-6

Schmitt, A., Wambsganss, T., Söllner, M., & Janson, A. (2021). Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. *Forty-Second International Conference on Information Systems (ICIS)*, 1-17.

Shin, D., Zaid, B., & Ibahrine, M. (2020). Algorithm Appreciation: Algorithmic Performance, Developmental Processes, and User Interactions. *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, 1–5. https://doi.org/10.1109/CCCI49893.2020.9256470

Sinclair-Desgagné, B. (2024). On the (non-) reliance on algorithms—A decision-theoretic account. *Journal of Mathematical Psychology*, *119*, 102844. https://doi.org/10.1016/j.jmp.2024.102844

Sunstein, C. R., & Reisch, L. (2023). *Do People Like Algorithms? A Research Strategy* (SSRN Scholarly Paper 4544749). https://doi.org/10.2139/ssrn.4544749

Sutherland, S. C., Harteveld, C., & Young, M. E. (2016). Effects of the Advisor and Environment on Requesting and Complying With Automated Advice. *ACM Transactions on Interactive Intelligent Systems*, *6*(4), 1–36. https://doi.org/10.1145/2905370

Trivedi, J., & Sama, R. (2020). The Effect of Influencer Marketing on Consumers' Brand Admiration and Online Purchase Intentions: An Emerging Market Perspective. *Journal of Internet Commerce*, *19*(1), 103–124. https://doi.org/10.1080/15332861.2019.1700741

Tsai, C., You, Y., Gui, X., Kou, Y., & Carroll, J. (2021). *Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers*. https://doi.org/10.1145/3411764.3445101

Van Esch, P., Black, J. S., & Arli, D. (2021). Job candidates' reactions to AI-Enabled job application processes. *AI and Ethics*, *1*(2), 119–130. https://doi.org/10.1007/s43681-020-00025-0

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation? *Journal of Management Information Systems*, *39*(2), 336–365. https://doi.org/10.1080/07421222.2022.2063553

Zhang, L., Pentina, I., & Fan, Y. (2021). Who do you choose? Comparing perceptions of human vs robo-advisor in the context of financial services. *Journal of Services Marketing*, *35*(5), 634–646. https://doi.org/10.1108/JSM-05-2020-0162