Imperfections of Machine Learning:
Experimental Investigations of Human Advice-Taking Behavior

Der Fakultät für Wirtschaftswissenschaften der
Universität Paderborn
zur Erlangung des akademischen Grades
Doktor der Wirtschaftsinformatik
- Doctor rerum politicarum -
vorgelegte Dissertation
von
Dirk Gerhard Leffrang
geboren am 12.10.1995 in Paderborn

30. April 2025

Dekan:	Prof. Dr. Jens Müller
Gutachter:	Prof. Dr. Oliver Müller
	Prof. Dr. Simon Thanh-Nam Trang
Termin der mündlichen Prüfung:	11.06.2025

Abstract - English

The increasing proliferation of machine learning (ML) algorithms raises concerns about their imperfections. Previous behavioral research has primarily focused on a general human aversion toward imperfect algorithms. In contrast, ML research has discovered different forms of imperfections, such as performance uncertainty, transparency issues, and environmental sustainability. This dissertation experimentally explores the impact of communicating these imperfections to the end-user by providing relevant background information about the limitations or specific characteristics of ML algorithms. This dissertation includes ten online experiments across seven papers with a total of 1,428 participants, which yielded three main findings: First, imperfections of algorithms can reduce algorithm aversion. Secondly, the distribution of advice quality shapes algorithm aversion. Third, AI literacy can be associated with algorithm aversion in non-linear ways. These findings are essential for decision-makers, developers, and ML evaluations. They emphasize the need to incorporate and disclose different algorithmic imperfections, which enables more nuanced advice-taking strategies, especially for users with varying degrees of AI literacy.

Keywords: machine learning, algorithm aversion, advice utilization, human-centered computing, artificial intelligence, human-algorithm interaction

Zusammenfassung - Deutsch

Die zunehmende Verbreitung von Machine-Learning-Algorithmen (ML) wirft Fragen über deren Unvollkommenheiten auf. Während sich frühere verhaltenswissenschaftliche Studien vor allem mit einer allgemeinen menschlichen Abneigung gegenüber fehlerbehafteten Algorithmen beschäftigt haben, hat die ML-Forschung verschiedene Arten solcher Schwächen identifiziert - etwa Unsicherheiten bei der Leistung, mangelnde Transparenz oder ökologische Auswirkungen. Diese Dissertation untersucht experimentell, wie sich die Kommunikation solcher Schwächen auf Nutzende auswirkt. Dazu wurde den Teilnehmenden gezielt Hintergrundwissen über Grenzen und besondere Merkmale von ML-Algorithmen vermittelt. Insgesamt wurden zehn Online-Experimente im Rahmen von sieben Fachartikeln mit insgesamt 1.428 Personen durchgeführt. Die Ergebnisse lassen sich in drei zentrale Erkenntnisse zusammenfassen: Erstens kann das Offenlegen von Schwächen die Ablehnung gegenüber Algorithmen verringern. Zweitens beeinflusst die Verteilung der algorithmischen Empfehlungsqualität die Nutzerakzeptanz. Drittens zeigt sich, dass das Wissen über KI in nichtlinearer Weise mit der Akzeptanz von algorithmischen Ratschlägen zusammenhängt. Diese Erkenntnisse sind besonders relevant für Entscheidungstragende, ML-Entwickelnde und alle, die ML-Systeme bewerten. Sie verdeutlichen, wie wichtig es ist, algorithmische Schwächen transparent zu machen, um differenzierte und informierte Entscheidungen bei der Nutzung solcher Systeme zu ermöglichen – insbesondere bei Nutzenden mit unterschiedlichem Verständnis von KI.

Stichworte: machine learning, Algorithmus-Aversion, Nutzung von Empfehlungen, benutzerzentrierte Informatik, künstliche Intelligenz, Mensch-Algorithmus-Interaktion

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

FTI Feature, Training and Inference

JAS Judge-Advice System

MAE Mean Absolute Error

ML Machine Learning

PI Prediction Interval

RQ Research Question

WOA Weight of Advice

XAI Explainable Artificial Intelligence

Part I.

Synopsis

1. Introduction

1.1. Motivation

In recent years, the proliferation of Machine Learning (ML) algorithms has accelerated across research and industry in various domains (Maslej et al., 2025). ML involves training computers to learn from data and experience, enabling them to improve continuously without being explicitly programmed for each task (Samuel, 1959). Examples of application areas where ML algorithms have surpassed human performance are image classification and some areas of natural language processing (He et al., 2015; Devlin et al., 2019). Currently, models such as GPT-4o and Gemini combine language understanding, multimodality and larger natural language contexts (OpenAI, 2024; Gemini Team, 2024). Generative models, such as DALL-E 3, enable image generation, while AlphaFold has transformed biology through accurate protein predictions (OpenAI, 2024; Jumper et al., 2021). However, ML algorithms have several limitations, including issues with reliability, generalizability beyond the training data and ethical concerns related to bias (Maynez et al., 2020; Kapoor and Narayanan, 2022; Heyder et al., 2023). Therefore, it is crucial to understand end-users reactions to these imperfections.

Although algorithms have outperformed human experts in structured regression and classification tasks for decades (Grove et al., 2000), recent advances in ML algorithms are accelerating their growing adoption across various domains (e.g., Surameery and Shakor, 2023; Dell'Acqua et al., 2023). In 2024, 78% of companies reported using ML algorithms, which suggests a growing adoption compared to 55% in 2023 (Singla et al., 2025). ML algorithms enable faster task completion and higher-quality outcomes by narrowing the skill gap between low- and high-skilled workers (e.g., Cambon et al., 2023; Dell'Acqua et al., 2023).

In a survey on the perception of ML algorithms, 66% of respondents believed ML algorithms would significantly impact their lives within three to five years (Carmichael, 2024). However, half of the respondents felt uneasy about ML algorithms (Carmichael,

2024). Despite the continuous improvements of ML algorithms, they are still imperfect (e.g., Lebovitz et al., 2021). Although governmental regulations concerning ML algorithms have risen in recent years, a primary difficulty in ML algorithms lies in the inherent uncertainty about its correctness (Maslej et al., 2025; Janiesch et al., 2021). Therefore, it is crucial to understand how individuals react to such imperfections.

From a human perspective, prior research observed algorithm aversion, which describes a tendency to favor human decision-makers over algorithms (Jussupow et al., 2024). However, the results are inconclusive as the opposite effect — algorithm appreciation — was also observed (Logg et al., 2019). As algorithm appreciation describes the inverse effect of aversion, we will use algorithm aversion to refer to both research streams.

Multiple terms have been used in research on algorithmic decision-making to label algorithms (Langer and Landers, 2019). For instance, "Artificial Intelligence (AI)" involves the development and study of methods and software that enable machines to perceive their surroundings, employ learning and intelligence and take actions that increase their probability of fulfilling specified objectives (Russell and Norvig, 2016, pp. 19-44). In the context of algorithm aversion, these terms are used interchangeably (Burton et al., 2020). Therefore, we will summarize AI and ML under the term "ML algorithm," which we specify as a computational procedure making predictions under uncertainty.

From an algorithmic perspective, predictive ML algorithms differ from others due to their probabilistic nature as they typically involve predictions under uncertainty (Russell and Norvig, 2016). Contrary to prior technologies, an ML algorithm that perfectly matches historical data would typically be viewed as overfitting. Overfitting arises when an ML algorithm adheres too closely to past data, potentially undermining its ability to generalize to new data (Russell and Norvig, 2016).

Furthermore, errors of ML algorithms can stem from various factors beyond overfitting, including data drift, insufficient high-quality training data and irreducible randomness (e.g., Lu et al., 2018; Lebovitz et al., 2021; Russell and Norvig, 2016). As a result, data scientists use performance metrics to evaluate algorithms relative to each other on specific tasks, as an absolute evaluation seems impossible (Huyen, 2022). However, metrics and human judgment are not always aligned (Lebovitz et al., 2021).

1.2. Research Objective and Research Questions

This dissertation addresses the impact of imperfections on advice-taking along three main research questions. We derive these questions in a combination of the Feature, Training and Inference (FTI) architecture from computer science and the psychological role of the Judge-Advice System (JAS) framework, as visualized in Figure 1.1. Visualized by solid lines, the FTI architecture contextualizes the imperfections along the ML pipeline. We used dashed lines to outline the different instances within the JAS framework, with each research question (RQ) targeting one instance.

The FTI architecture is a framework that maps the ML pipeline into three stages — feature extraction, model training, and inference — highlighting how data is transformed into predictions. However, several imperfections of ML algorithms can affect each stage. Therefore, developers of ML algorithms have to take these imperfections into account, for instance, when selecting features and configuring model training (Huyen, 2022).

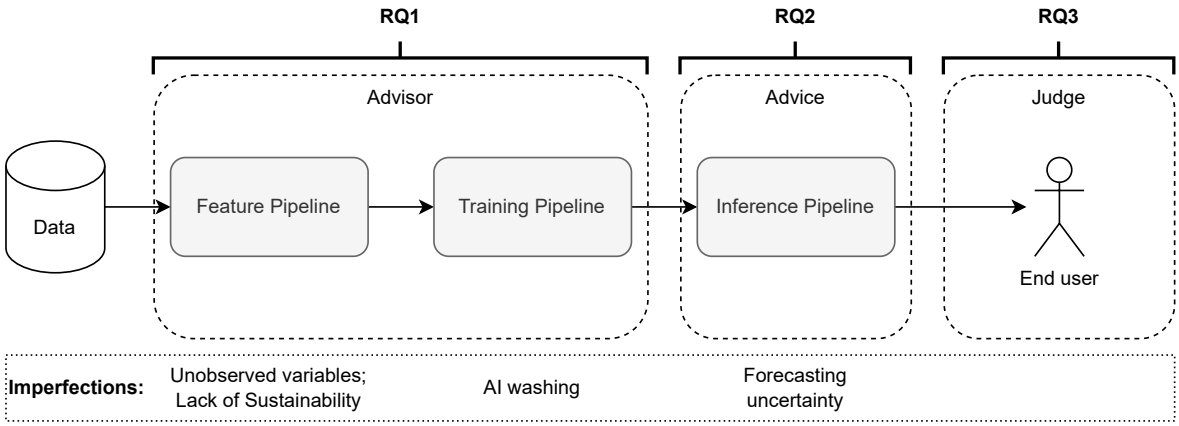

Figure 1.1.: Overview of the research questions (RQ) along the FTI-JAS framework.

The JAS involves a human judge facing a prediction problem, for which they can consult an advisor, which can be an algorithm or a human. The judge weighs their initial prediction and an advisor's advice. We developed one RQ for each instance within the JAS. The first RQ focuses on the algorithm as an advisor, the second on the advice provided by the algorithm and the third on the end user.

Exemplary technical imperfections of ML advisors include unobserved variables, a lack of environmental sustainability due to high energy consumption, and doing "AI" washing by referring to statistical models as "AI" (Patterson et al., 2021; Moore, 2017). Contextualizing these imperfections by highlighting unobserved variables, clarifying

the sustainability-performance trade-off, and using more accurate labels may help users better utilize algorithmic advice. This motivates the following research question:

RQ1: What is the impact of *advisor imperfections* on algorithm aversion?

An example of advice-related imperfection from ML algorithms is forecasting uncertainty. This can appear as performance outliers in hindsight or as distributions of predictive uncertainty in advance (Hyndman and Koehler, 2005; Spiegelhalter et al., 2011). Contextualizing such imperfections — by embedding outliers in repeated advice-taking tasks or visualizing uncertainty distributions — may help users make more informed judgments about algorithmic behavior. This leads to the following research question:

RQ2: What is the impact of *advice imperfections* on algorithm aversion?

Beyond algorithm-specific factors, evaluating ML algorithms can be particularly difficult without sufficient understanding of how they function (Lebovitz et al., 2021). The related concept of AI literacy is therefore essential. It encompasses the ability to critically evaluate ML algorithms, communicate and collaborate effectively with ML systems, and apply them across different settings—whether online, at home, or in the workplace (Long and Magerko, 2020). However, despite the widespread use of ML, empirical research on the role of AI literacy remains limited. This gives rise to the following research question:

RQ3: What is the impact of imperfections on algorithm aversion among individuals *with varying levels of AI literacy*?

1.3. Thesis Structure

This dissertation consists of two parts. Part I includes Chapters 1 to 6. Chapter 2 provides an overview of algorithm aversion from a contextual and methodological view. Afterward, it describes imperfections of ML algorithms along the ML pipeline. Chapter 3 examines the research methodology, while Chapter 4 summarizes the results. Chapter 5 discusses the findings concerning the research questions and concludes with their implications. Chapter 6 concludes the dissertation by outlining its limitations and offering an outlook.

Part II consists of Chapters 7 to 13, each dedicated to a specific publication. Table 1.1 lists these publications. It contains three journal articles and four conference papers. Five papers are published; one is accepted in-principle via a registered report (currently

in Stage 2 review) and one is under review. All papers are ranked in the VHB Rating 2024 (VHB, 2024), with rankings ranging from B to A+. Notably, we are the first to achieve an in-principal acceptance in the Business & Information Systems Engineering journal[1].

[1] `https://www.bise-journal.com/?page_id=2258`

No.	Authors & Title	Status	Outlet	Type	VHB Rating 2024
1	**Leffrang, D.**, Bösch, K., Müller, O. (2023). Do People Recover From Algorithm Aversion? An Experimental Study of Algorithm Aversion Over Time.	P	Hawaii International Conference on System Sciences	C	B
2	**Leffrang, D.** (2023). The Broken Leg of Algorithm Appreciation: An Experimental Study on the Effect of Unobserved Variables on Advice Utilization.	P	Wirtschafts-informatik Conference	C	B
3	**Leffrang, D.**, Müller, O. (2023). AI Washing: The Framing Effect of Labels on Algorithmic Advice Utilization.	P	International Conference on Information Systems	C	A
4	**Leffrang, D.**, Müller, O. (2024). Algorithmic Advice-Taking Beyond MAE: The Role of Negative Prediction Outliers and Statistical Literacy in Algorithmic Advice-Taking.	P	European Conference on Information Systems	C	A
5	**Leffrang, D.**, Müller, O. (2024). Visualizing Uncertainty in Time Series Forecasts: The Impact of Uncertainty Visualization on Users' Confidence, Algorithmic Advice Utilization and Forecasting Performance.	P	Journal of Forecasting	J	B
6	**Leffrang, D.**, Passlack, N., Müller, O., Posegga, O. (2025). Beneficial Mistrust in Generative AI? The Role of AI Literacy in Handling Bad Advice.	IPA; U	Business & Information Systems Engineering	J	B
7	**Leffrang, D.**, Müller, O. (2025). The Sustainability-Performance Trade-Off in AI: The Role of Sustainability Information and Unmet Performance Goals in Sustainable AI Decisions.	U	Information Systems Research	J	A+

Status: IPA: In-principal acceptance; P: Published; U: Under review
Type: C: Conference; J: Journal

Table 1.1.: Publications included in this dissertation.

2. Background

This dissertation builds on several key concepts. First, we examine research on algorithm aversion from a contextual point of view. Next, we analyze the inherent imperfections of ML algorithms throughout the ML pipeline.

2.1. Algorithm Aversion From a Contextual View

In which situations people reject or appreciate algorithms remains an ongoing discussion (Jussupow et al., 2024). Algorithm aversion refers to "a behavior of discounting algorithmic decisions concerning one's own decisions or other's decisions, either consciously or unconsciously" (Mahmud et al., 2022). Such an aversion appeared to be especially prevalent after an error, according to a study in a medical context (Prahl and Van Swol, 2017).

Although the term algorithm aversion was established in the last decade, the phenomenon dates back several decades to a discussion of clinical (human) versus statistical (algorithmic) decision-making. For instance, according to a meta-analysis conducted in 1954, some clinicians prefer human diagnoses over algorithmic diagnoses despite their superiority on average. Clinicians attributed humans with superior perceptive capabilities, experience and attention to improbable events (Meehl, 1954).

However, other studies observed the opposite phenomenon, namely algorithm appreciation. Algorithm appreciation describes a preference for algorithmic over human predictions (Logg et al., 2019). Such an appreciation appears reasonable, with algorithms performing as well as or better than humans in approximately 94% of cases in a literature review of 136 studies (Grove et al., 2000). For instance, individuals put more weight on algorithmic predictions than human predictions when predicting romantic attraction or the popularity of songs (Logg et al., 2019).

Similar to the phenomenon of algorithm aversion, research on algorithm appreciation also dates back several decades. For instance, clinicians associated algorithms with

objectivity, precision and verifiability (Meehl, 1954). Additionally, when tasked with judging the suitability of solutions across various domains such as education, healthcare and civil law, participants viewed algorithms as more objective and rational than the human advisor (Dijkstra et al., 1998).

Other studies suggest a mixture of algorithm aversion and appreciation. Although individuals initially showed aversion toward algorithms, their aversion decreased after using algorithmic advice for weight estimation tasks (Turel and Kalhan, 2023). Similarly, in a call center scenario, participants unfamiliar with the advisor's quality initially exhibited algorithm aversion. However, those who observed the advisor's consistent improvement developed algorithm appreciation (Berger et al., 2021).

Multiple reviews of the existing literature note that the results remain inconclusive and scattered among disciplines (e.g., Jussupow et al., 2020; Burton et al., 2020; Mahmud et al., 2022; Jussupow et al., 2024). As visualized in Figure 2.1, algorithm aversion can vary based on multiple factors such as algorithmic, high-level, individual, and task factors according to the framework of Mahmud et al. (2022).

Algorithm factors include design, decision and delivery aspects. Design factors refer to how algorithms are developed and structured, including complexity, feedback availability, accessibility, human integration, explainability, understandability, feedback calibration, response speed, learning capabilities and interface anthropomorphism (Mahmud et al., 2022). For example, features such as anthropomorphic design or demonstrating an algorithm's learning ability have been associated with reduced algorithm aversion in classification and resource allocation tasks (Madhavan and Wiegmann, 2007; Berger et al., 2021). While explainability through Explainable Artificial Intelligence (XAI) can influence users' decision-making, it is possible to evaluate algorithm behavior without fully opening the black box of a ML model (Wachter et al., 2018). Therefore, this dissertation focuses on decisions made along the ML pipeline without centering on XAI.

Decision factors focus on the nature of the decision itself, including its accuracy, cost, future implications and role within the broader decision-making process (Mahmud et al., 2022). For instance, aversion to imperfect advice in a human resource setting falls under this category (see Dietvorst et al., 2015). Another example is the relationship between algorithmic and human decisions, which can shape algorithm aversion (Jussupow et al., 2024).

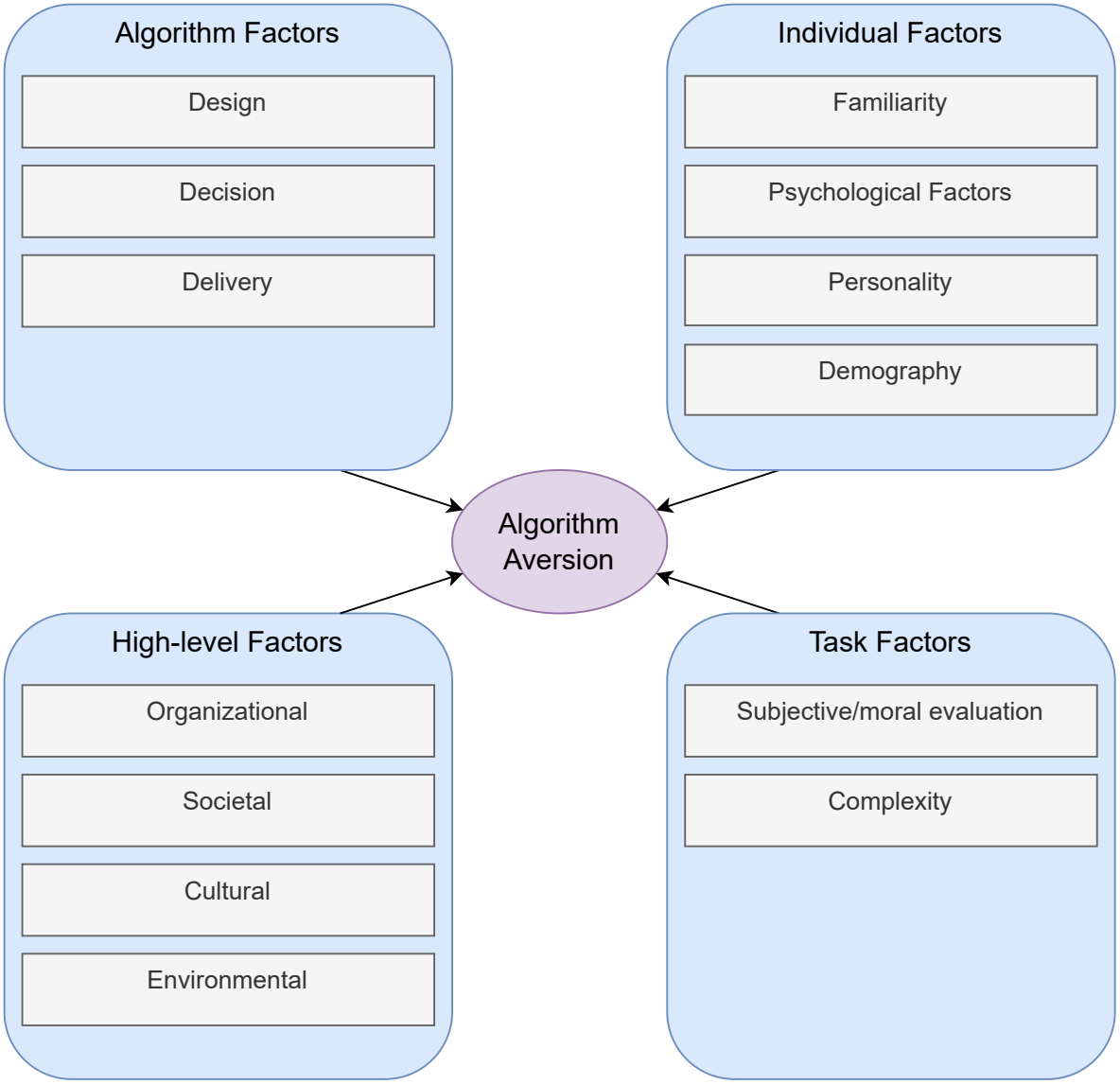

Figure 2.1.: Factors influencing algorithm aversion (adapted from Mahmud et al., 2022).

Factors associated with decision delivery relate to the method and style in which decisions are presented (Mahmud et al., 2022). For example, while humans typically offer the option of oral communication, algorithms generally communicate in written form unless specifically designed otherwise (Önkal et al., 2009). Additionally, human involvement in human-algorithm collaboration affected users' perceived understanding of an algorithm's functionality in a machine maintenance setting (Lebedeva et al., 2024). Consequently, we concentrate on advisors providing written advice without further interaction.

High-level factors include organizational, societal, cultural and environmental influences

(Mahmud et al., 2022). To minimize heterogeneity, this dissertation focuses on incentivized contexts within European and American societies represented by crowdworkers and students. Although incentives do not universally enhance performance across all experimental contexts, research suggests they effectively improve effort-sensitive tasks, such as predictions (Camerer and Hogarth, 1999).

Individual factors cover familiarity, psychological characteristics, personality traits and demographic variables (Mahmud et al., 2022). Familiarity factors capture individuals' prior experiences with algorithms, specific tasks and human advisors. For example, negative experiences with algorithms may lead to increased algorithm aversion toward a forecasting algorithm in a healthcare setting (Prahl and Van Swol, 2017).

Psychological factors refer to reasoning, logic, thinking and emotion related to algorithmic decisions. These factors refer to a general aversion toward algorithms (Mahmud et al., 2022). For instance, individuals' perceived expectations of algorithms influenced their likelihood of algorithm aversion toward a forecasting algorithm in a human resource setting (Dietvorst et al., 2015).

Personality trait factors include self-evaluation and the Big Five personality dimensions (Mahmud et al., 2022). For example, egocentric advice discounting refers to favoring one's own decisions over external advice (Yaniv and Kleinberger, 2000). Thus, high self-esteem can reinforce egocentric bias, increasing resistance to algorithmic recommendations (see Logg et al., 2019).

Demographic factors address how characteristics such as gender, age, and education level affect algorithm aversion (Mahmud et al., 2022). However, findings on variables like age are mixed (e.g., Logg et al., 2019). Nevertheless, we controlled for demographic variables in this research.

Task factors concern the contextual nature of the ML problem, including whether tasks require subjective evaluation, involve moral judgment, or vary in complexity—all of which can influence attitudes toward algorithm acceptance (Mahmud et al., 2022). Research suggests that algorithm aversion is typically lower for more objective tasks (Castelo et al., 2019). Additionally, individuals tend to value actions benefiting others more highly for personal growth than those driven solely by financial motives, often prioritizing empathy and autonomy when decisions impact others (Heßler et al., 2022). To minimize bias and heterogeneity from task factors, this dissertation focuses on objective tasks.

2.2. Imperfections of ML Algorithms Along the ML Pipeline

Figure 2.2 visualizes the FTI architecture, which provides a framework for structuring the ML pipeline (Dowling, 2023). The framework consists of three pipelines. The arrows highlight the primary and most common interconnections among the pipelines.

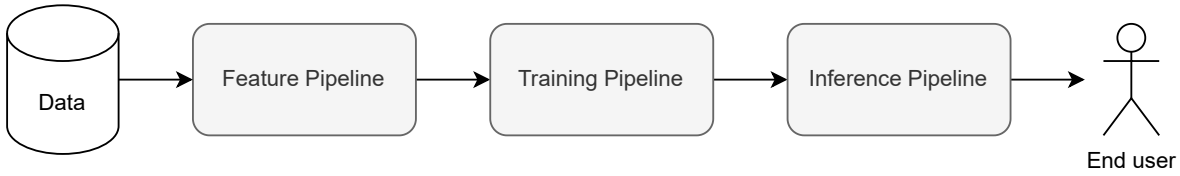

Figure 2.2.: Extended Feature, Training and Inference (FTI) architecture (adapted from Dowling, 2023).

A feature pipeline receives raw data as input, converts it into features and eventually labels (Dowling, 2023). A feature represents a single quantifiable attribute or characteristic within a dataset used as input to ML models to solve the ML problem (Bishop, 2006). At this stage, the data scientist cleans and transforms the data, selecting and engineering relevant features. The extracted features and labels serve as training data (Dowling, 2023).

A training pipeline retrieves the training data, trains the model and stores the trained model. Typically, the data scientist splits the data and chooses multiple algorithms. Afterward, the model learns patterns by minimizing the prediction error. The data scientist can influence the learning process by tuning hyperparameters. The output of this stage is a trained model, which serves as advisor (Dowling, 2023).

An inference pipeline retrieves the trained model, obtains new feature data and generates predictions that the end user can utilize through the ML-enabled product. The model receives raw data through the feature pipeline. Based on the resulting features, the model outputs predictions, which serve as advice (Dowling, 2023).

An end user is not part of the original FTI architecture, which presents the ML-enabled product as the final stage (Dowling, 2023). However, end users frequently use ML algorithms as advisors (Golinelli et al., 2020). The end user serves as a judge who weighs their prediction based on the advice provided by the advisor.

3. Methodology

This dissertation employed online user experiments as the primary research model to investigate human advice-taking behavior in response to algorithmic imperfections. Online experiments offer a controlled yet scalable environment, enabling the systematic manipulation of key independent variables while maintaining ecological validity (Shadish et al., 2003). We designed most experiments using the JAS framework, a well-established paradigm for measuring advice utilization, particularly relevant in contexts of algorithmic advice utilization (Logg et al., 2019). It evaluates how much individuals integrate external advice into their decision-making processes. Figure 3.1 visualizes the three steps of the JAS (Bonaccio and Dalal, 2006):

1. Initially, the judge — who can be a participant in an experiment or a general decision-maker — makes an initial decision under uncertainty (e.g., a regression or programming task).
2. Subsequently, an advisor — a ML algorithm or a person — provides advice.
3. The judge then combines their initial prediction with the advice from the advisor, applying a weighted approach, to arrive at a final decision.

The experimental manipulation varied the nature of the advisor (e.g., human vs. algorithm), advice quality (e.g., error distributions, outliers), advice presentation (e.g., uncertainty visualizations), and contextual information (e.g., sustainability information, AI framing). Additionally, we examined individuals' level of AI literacy. The primary dependent variable in such studies was the Weight of Advice (Weight of Advice (WOA)), defined as the proportional adjustment toward the advisor's recommendation relative to the initial prediction:

$$\text{WOA} = \frac{|\text{final prediction} - \text{initial prediction}|}{|\text{advisor's prediction} - \text{initial prediction}|} \tag{3.1}$$

WOA captures the extent to which individuals incorporate external advice into their decision-making processes. A value of zero indicates complete rejection of the advice,

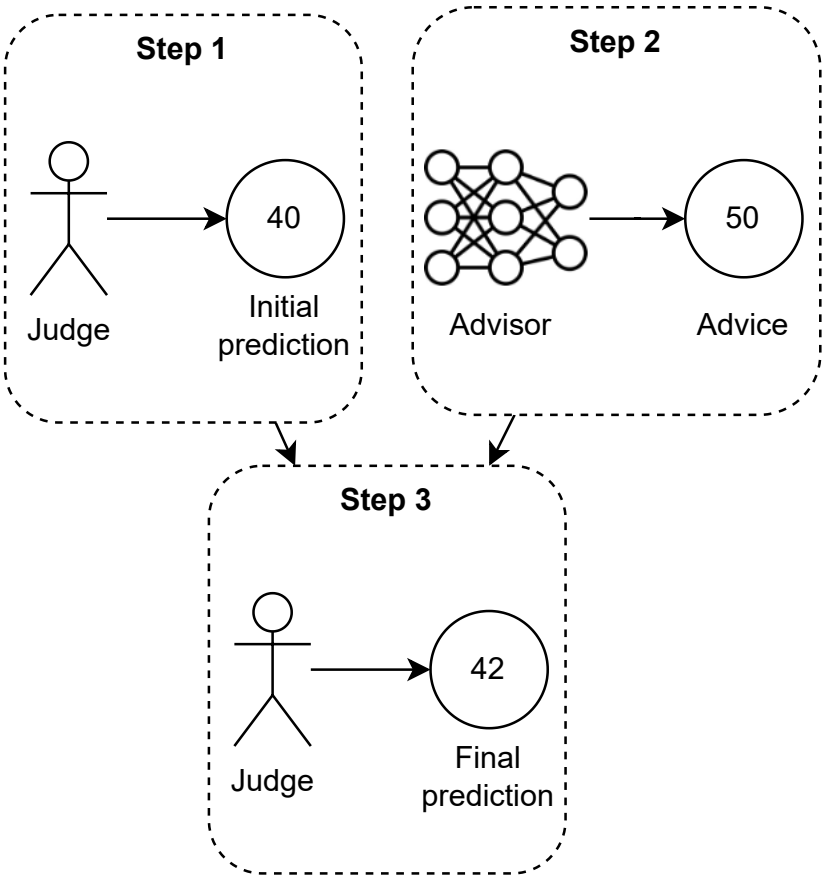

Figure 3.1.: Judge incorporating their initial prediction and the advice in the Judge-Advisor System (JAS).

while higher WOA values reflect greater incorporation of the advice. In line with prior research, WOA is winsorized to values of 1 if there is overshooting (e.g., Logg et al., 2019). The independent variables included the experimental manipulations, their interaction and several control variables such as age and gender. We recruited participants from online platforms (such as Prolific) and university pools to ensure diverse, English-speaking samples that were homogeneous with respect to the pre-defined selection criteria. Participation was incentivized financially or with bonus points to foster serious engagement with the tasks.

The primary statistical analyses involved linear mixed-effects models, linear regressions, logistic regressions and structural equation modeling to account for the variations of the experimental design. All analyses were pre-registered or conducted following established methodological standards to minimize researcher degrees of freedom. Moreover, we consistently applied robustness checks to ensure validity.

The need for high experimental control, random assignment, and access to a diverse participant pool guided the choice of online experiments over alternative research designs such as observational or field studies. Given the focus on algorithmic advice-taking — a phenomenon frequently encountered in online decision environments — the setting aligned well with the contextual demands of the research questions. Furthermore, the online format allowed for the efficient replication of decision tasks at scale, supporting the robustness and generalizability of the findings.

4. Results

This thesis includes seven research papers—five published, one accepted in principle via a registered report (currently in Stage 2 review) and one under review. Figure 4.1 provides an overview of the FTI-JAS framework developed in this thesis, combining the JAS framework from psychological research and FTI architecture from computer science. For clarity, the figure classifies the primary focus of each research paper along the framework.

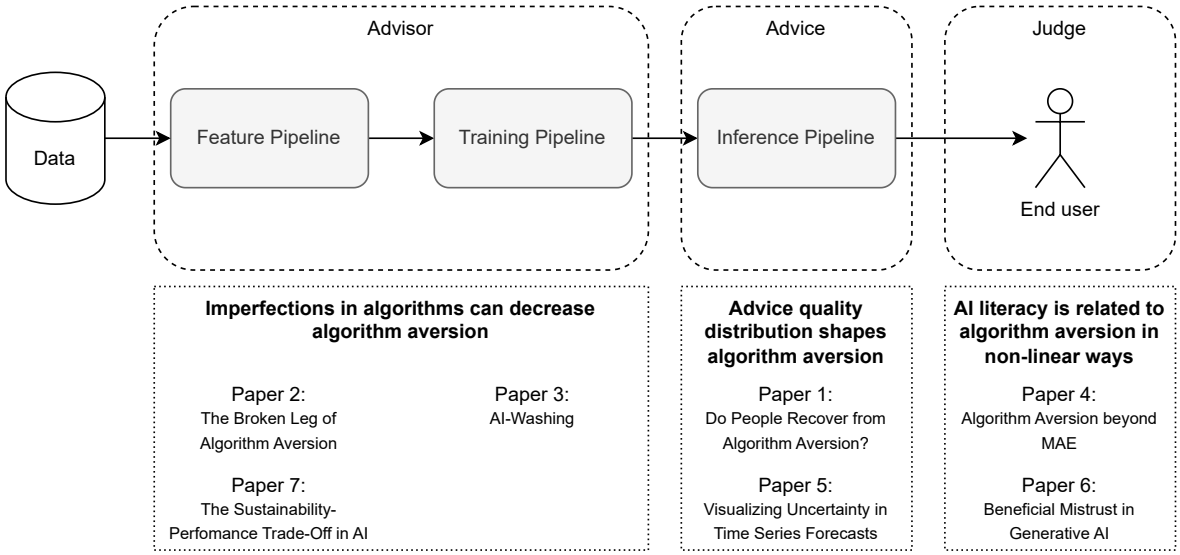

Figure 4.1.: Overview of the primary contributions along the FTI-JAS framework.

Papers 2, 3 and 7 address **RQ1:** What is the impact of *advisor imperfections* on algorithm aversion?
Papers 1 and 5 explore **RQ2:** What is the impact of *advice imperfections* on algorithm aversion?
Finally, papers 4 and 6 examine **RQ3:** What is the impact of imperfections on algorithm aversion among individuals *with varying levels of AI literacy*?
We explicitly incorporated a human baseline in the first three papers. In the subsequent studies, we focused on algorithm-only configuration to explore more nuanced forms of

human-algorithm interaction. The following section summarizes the findings of all seven research papers.

Paper 1: *Do People Recover From Algorithm Aversion?* (Leffrang et al., 2023)
Although algorithmic predictions are, on average, more accurate than human predictions, prior research observed algorithm aversion after bad advice. However, prior research focused on a one-off decision after receiving bad advice. Drawing on expectation-confirmation theory, this paper examines the occurrence of bad advice on advice-taking over time. We conducted an online between-subjects experiment involving 87 participants completing repeated time-series forecasting tasks. Our results show no evidence of immediate algorithm aversion after bad advice. Instead, participants developed a growing appreciation for algorithms over time. This study highlights that the distribution of advice quality (good vs. bad advice) over time can shape algorithm aversion.

Paper 2: *The Broken Leg of Algorithm Appreciation* (Leffrang, 2023)
Despite the capabilities of ML algorithms, individuals showed algorithm aversion when algorithms neglected their unique characteristics. However, the effect of such unobserved variables on algorithm aversion is not yet fully understood, as humans can also suffer from unobserved variables like missing information about events. Based on Meehl's broken leg scenario, this paper examines the impact of an unobserved variable and the advisor's type (algorithm vs. human) on algorithm aversion. We conducted an online within-subjects experiment with 94 participants focused on repeated regression tasks. Our results suggest that an unobserved variable did not reduce advice-taking. Instead, participants showed greater algorithm appreciation, compensating for perceived imperfections. This study contributes to the idea that the imperfection of an unobserved variable can decrease algorithm aversion.

Paper 3: *AI Washing* (Leffrang and Müller, 2023)
Researchers and practitioners often regard ML algorithms as a game changer compared to traditional statistical models. However, some software providers use "AI washing" by rebranding basic statistical solutions as AI systems. Based on attribute framing, this paper examines the impact of framing the advisor's expertise and type on algorithm aversion. We conducted two online within-subjects experiments with 120 participants focused on repeated regression tasks. Our results provide evidence that participants took more advice from human advisors with higher expertise, whereas algorithmic labels did not significantly affect advice-taking. This paper contributes to the idea that presenting imperfections through framing can decrease algorithm aversion.

Paper 4: *Algorithmic Advice-Taking Beyond Mean Absolute Error (MAE)* (Leffrang and Mueller, 2024)
The numerical performance metrics of ML algorithms do not always align with human judgment. However, most prior studies have either concentrated on the statistical evaluation of established metrics or examined how changes in these metrics influence human behavior. Based on the salience bias, this paper examines the impact of individuals' statistical literacy levels and the distribution of the advice quality on advice-taking over time. We conducted an online between-subjects experiment with 115 participants focused on repeated regression tasks. Our results indicate that negative outliers increased advice-taking and statistical literacy had a U-shaped effect on algorithm aversion. Both individuals with low and high levels of statistical literacy were associated with higher advice utilization. This paper contributes to the distribution of advice quality (outlier vs. no outlier) affecting algorithm aversion. It provides evidence for a positive correlation between AI literacy and non-linear relationships between algorithm aversion and individuals' literacy level.

Paper 5: *Visualizing Uncertainty in Time Series Forecasts* (Leffrang and Müller, 2024)
Time series forecasts inherently involve uncertainty. However, experimental research communicating this uncertainty to end users has yielded mixed and inconclusive results. Based on probabilistic and frequency framing, this paper examines the impact of prediction interval (Prediction Interval (PI)) and ensemble plots against a point estimate control on advice-taking. We conducted an online between-subjects experiment with 239 participants focused on a time series forecasting task. Our results indicate a U-shaped relationship between uncertainty visualization and forecasting performance, moderated by confidence, graph literacy and domain knowledge. Both hiding uncertainty and making it overly salient reduced confidence in the forecasting algorithm, led to lower advice utilization and resulted in higher forecasting errors. This paper contributes to the idea that the presentation of advice quality distribution in the form of different uncertainty visualizations influences algorithm aversion.

Paper 6: *Beneficial Mistrust in Generative AI* (Leffrang et al., 2025)
Although generative ML algorithms continue to improve, they can present inaccurate or misleading information as fact — a phenomenon known as "hallucinations." However, with generative ML algorithms becoming increasingly accessible to the public, it is essential to understand how AI literacy affects individuals' responses to bad advice from these systems. Drawing on correspondence bias, this paper examines the impact of individuals with varying levels of AI literacy and the presence of bad advice on advice-

taking. We conducted an online between-subjects experiment with 542 participants focused on a programming task. Our results indicate that high-AI-literacy individuals were less likely to follow advice, especially bad advice. This suggests a form of beneficial mistrust in the case of bad advice but potentially disadvantageous consequences in the case of good advice. This paper supports a positive correlation between AI literacy and algorithm aversion depending on the distribution of advice quality (good vs. bad advice).

Paper 7: *The Sustainability-Performance Trade-Off in AI* (Leffrang and Müller, 2025) Despite the remarkable advancements of ML algorithms, growing concerns persist regarding its environmental footprint. Yet, organizations seldom adopt sustainable ML practices and previous research has primarily concentrated on promoting sustainability for non-ML products or investigating technical solutions within ML applications. Drawing on goal-setting theory, this paper examines the impact of sustainability information (energy consumption) and an unmet performance goal on retraining decisions for ML algorithms. We conducted three online experiments with 343 participants focused on a time series forecasting task. Our results indicate that providing sustainability information increases the likelihood of choosing sustainable ML. However, presenting an unmet performance goal decreases the likelihood and offsets the beneficial impact of sustainability information. This paper provides evidence that the imperfection of sustainability-performance trade-off in ML can influence algorithm aversion.

5. Discussion

5.1. Addressing the Research Questions

In recent years, the use of ML algorithms has expanded rapidly across research and industry in a wide range of fields (Maslej et al., 2025). However, there are mixed findings on whether individuals utilize these algorithms (Mahmud et al., 2022). While prior studies have examined a general aversion to imperfect algorithms (e.g., Dietvorst et al., 2015), the advice-taking may vary depending on the type of imperfection along the ML pipeline. By using seven studies, this study examines the impact of advisor imperfections, advice imperfections and the varying reactions of individuals with different levels of AI literacy:

The first research question examines the impact of advisor imperfections on algorithm aversion. Papers 2, 3 and 7 provide evidence that imperfections can decrease algorithm aversion. Paper 2 demonstrates that revealing an omitted variable bias can increase advice-taking. Paper 3 reveals that framing influences perceptions of algorithms, with algorithm appreciation observed when algorithms were compared to students but not when compared to experts. Additionally, paper 7 indicates that while providing sustainability information increases the likelihood of choosing sustainable ML, presenting an unmet performance goal decreases the likelihood and offsets the beneficial impact of sustainability information.

The second research question investigates the impact of advice imperfections on algorithm aversion. Papers 1, 4 and 5 suggest that advice quality distribution shapes algorithm aversion. In paper 1, the results indicate that algorithm appreciation increases over time after a one-time error. Paper 4 finds that a one-time outlier leads to more advice-taking over time compared to the repeated occurrence of small errors. In paper 5, disclosing predictive uncertainty resulted in a U-shaped relationship with moderate uncertainty boosting confidence, advice-taking and performance but excessive uncertainty diminishing these coefficients.

Finally, the third research question explores the effects of imperfections on algorithm aversion among individuals with different levels of AI literacy. Papers 4 and 6 reveal situations in which AI literacy can be associated with algorithm aversion. Although paper 4 did not confirm a positive impact of statistical literacy, the results suggest that statistical literacy follows a U-shaped pattern concerning advice-taking. In paper 6, more AI-literate individuals exhibited less advice-taking from ML algorithms, which is beneficial in case of bad advice but potentially harmful in case of good advice.

5.2. Research Implications

Each paper contributes to the overarching goal of improving knowledge about imperfections influencing human advice-taking behavior. For the systematic classification of the dissertation, we utilize the contextual classification based on Mahmud et al. (2022), who model algorithm aversion as the result of algorithmic, individual, high-level and task factors. Figure 5.1 illustrates the contextual classification of the papers, introduced in Section 2.1.

RQ1 examines the impact of *advisor imperfections*. From a contextual perspective, we focused on factors related to the design of the algorithm. In paper 2, we used a regression task based on tabular data. The design factor involved an unobserved variable and a variation in advisor type. In paper 3, we used a similar setup and the design factor was the framing of the advisor. Paper 7 we provided participants with graphical data on binary retraining decisions. The design factor in this study was the sustainability-performance trade-off.

We contribute that **imperfections of algorithms can decrease algorithm aversion**. The discussion on algorithm aversion started with the claim of a general aversion toward imperfect algorithms (Dietvorst et al., 2015; Logg et al., 2019). Therefore, prior research on design factors has focused on mitigating advisor imperfections using an anthropomorphic design or demonstrating an algorithm's learning (Madhavan and Wiegmann, 2007; Berger et al., 2021). However, there remain mixed findings on algorithm aversion and appreciation (see Jussupow et al., 2020; Burton et al., 2020; Mahmud et al., 2022; Jussupow et al., 2024, for reviews). Moreover, imperfections in ML algorithms extend beyond prediction accuracy. For instance, the environmental costs of ML algorithms, such as rising energy consumption due to large-scale model training on exponentially growing datasets, are becoming increasingly relevant (Crawford, 2024;

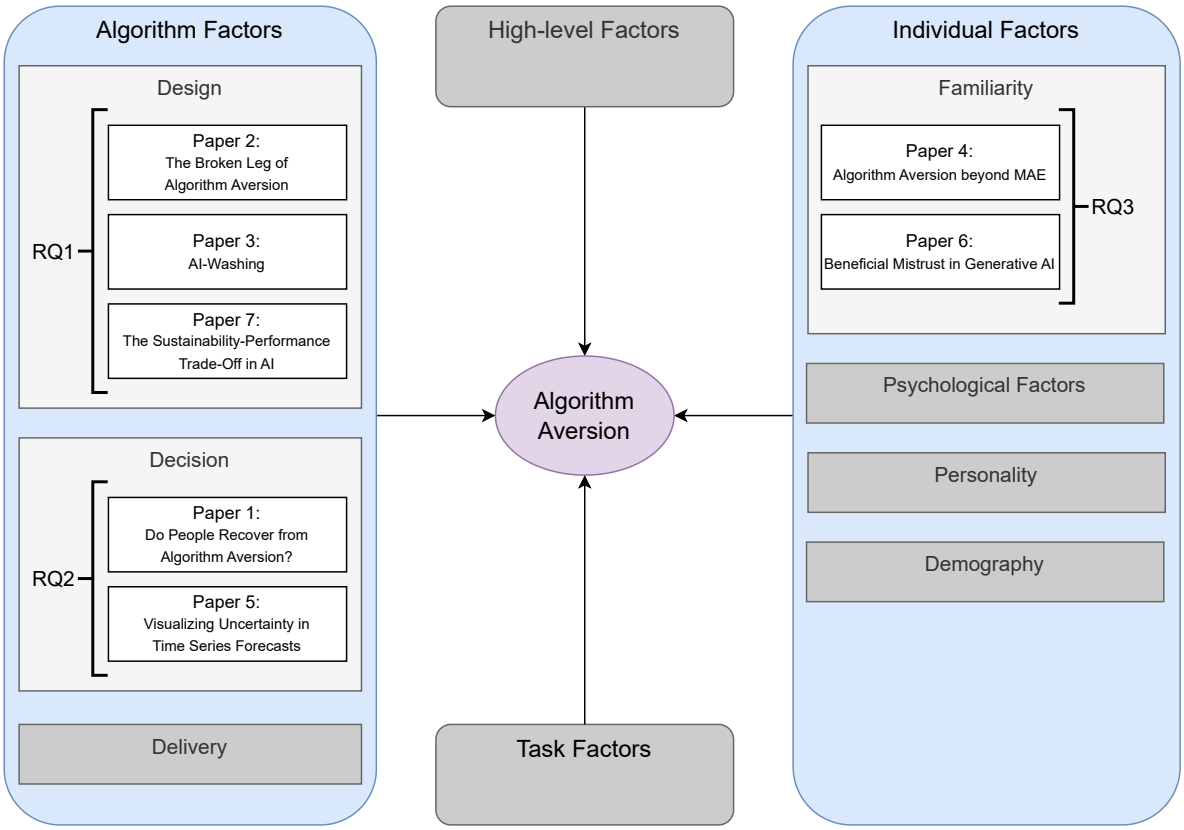

Figure 5.1.: Contextual classification of the papers included in this dissertation (adapted from Mahmud et al., 2022). The dark gray factors lie outside the scope of this dissertation.

Wu et al., 2022). We particularly extend research on algorithmic design-related algorithm aversion and human-algorithm interaction by providing empirical evidence that imperfections linked to the ML pipeline can decrease algorithm aversion as individuals draw conclusions based on the provided design factors.

RQ2 focuses on the impact of *advice imperfections* in the form of forecasting uncertainty. Here, we shifted the contextual focus to factors related to decisions with the algorithm. In paper 1, we used a regression task visualized by time series graphs depicting river water levels. the decision factor was the presence of bad advice and the recovery process in advice-taking over time. In paper 5, time series graphs illustrated a regression task using COVID-19 hospitalization data in the paper. The decision factor was the type of uncertainty visualization.

We find that **the qualitative distribution of advice shapes algorithm aversion**. While prior research on algorithm aversion has examined algorithmic performance as one key factor influencing advice-taking (see Mahmud et al., 2022, for a review), this

work has primarily examined one-off decisions and focused on reactions to single errors (e.g., Dietvorst et al., 2015; Prahl and Van Swol, 2017). However, such decision-making scenarios involve uncertainty, which can alternate future advice-taking (Spiegelhalter et al., 2011). Additionally, individuals can change their advice-taking behavior depending on the advice quality (Yaniv and Kleinberger, 2000). This is in contrast to ML research, which typically uses averaging metrics to evaluate the quality of an advisor (e.g., Hyndman and Koehler, 2005, pp. 77-87). We particularly extend research on algorithm aversion with respect to decision factors and ML by highlighting the role of advice quality distributions instead of focusing on one-time algorithm decision-making scenarios or aggregated measures.

RQ3 explores the impact of imperfections *among individuals with varying levels of AI literacy*. From a contextual perspective, we concentrated on the individual's familiarity factors. In paper 4, we used tabular data for a regression task on the market values of soccer players. Statistical literacy was the key familiarity factor in this study. Paper 6 involved a programming task with the help of a generative ML algorithm. The familiarity factor in this study was individuals' AI literacy.

Our results indicate that **AI literacy can be associated with algorithm aversion**. Most prior work on algorithm aversion has focused on psychological factors such as personality, familiarity and demography (Mahmud et al., 2022). While conceptual work has emphasized the importance of AI literacy—including subliteracies such as statistical literacy—as essential for engaging with ML algorithms (e.g., Gal, 2002; Long and Magerko, 2020), empirical research could not confirm a generally positive effect of literacy on advice-taking. Specifically, individuals with lower AI literacy relied more heavily on algorithmic advice due to a lack of understanding or overtrust in the system (Ehsan et al., 2021; Jacobs et al., 2021). This has prompted calls for companies to market ML algorithms solutions specifically for individuals with low AI literacy (Dell'Acqua et al., 2023; Tully et al., 2025). We particularly extend prior research on familiarity factors in algorithm aversion and AI literacy by warning about nonlinear relationships between AI literacy and advice-taking.

5.3. Practical Implications

Our study provides several practical implications for decision-makers, developers and test managers in algorithmic decision-making:

First, **decision-makers should incorporate different types of algorithmic imperfections into their decision-making process**. This dissertation offers insights that enable managers to develop effective strategies for incorporating algorithms into their decision-making processes. Recognizing and devising ways to overcome algorithms' imperfections is essential for successful human-algorithm cooperation. By identifying the sources of potential imperfections and disclosing end-users' responses to them, this dissertation provides managers with practical tools for successful algorithmic integration:

1. *Persist with algorithmic approaches, even if your initial attempt encounters challenges.* Predictive algorithms are always associated with uncertainty. These imperfections may lead to initial algorithm aversion (Dietvorst et al., 2015; Turel and Kalhan, 2023). However, good performance in repeated decision-making scenarios can offset the initial uncertainty (paper 1, 4).

2. *Embrace genuine ML instead of superficial AI washing.* Even though expectations for ML algorithms can be high during development — with developers believing that ML algorithms hold greater potential than traditional statistical models — our results suggest that users are unlikely to use the advice of ML algorithms until its performance outshines other options (paper 3).

3. *Incorporate imperfections of ML algorithms beyond established metrics as users think about them.* Even when performance remains unchanged, an unobserved variable can boost advice-taking. This insight can help practitioners understand unintended human behavior (paper 2). Decision-makers should not rely solely on performance metrics when evaluating algorithms; they must also consider the impact of other factors such as performance distributions, the sustainability-performance trade-off and the AI literacy of the end users (papers 4, 5, 6, 7).

Second, **developers should disclose imperfections along the ML pipeline**. This dissertation presents a comprehensive overview of various factors along the ML pipeline that can influence algorithm-based decision-making. This work synthesizes findings from previous studies on information systems, information technology, behavioral economics and psychology to provide developers with a knowledge base for algorithm design. This foundational understanding can enable the creation of more transparent algorithms, ultimately boosting users' advice-taking and performance:

1. *Enable end-users to correct for algorithmic imperfections.* For instance, they may correct for an omitted variable or outlier (papers 2, 4).

2. *Communicate the uncertainty associated with algorithms.* Although there seems to be a U-shaped relation between uncertainty and advice-taking, communicating uncertainty through a prediction interval (PI) appears to be an appropriate trade-off (paper 5).

3. *Communicate the trade-offs of algorithms through transparency.* Developers should also disclose the trade-offs to end users, such as the sustainability-performance trade-off, to democratize design decisions concerning ethics (paper 7).

Third, **test managers should incorporate end-users' AI literacy levels and alternative goals into ML evaluations**. This dissertation offers insights that empower test managers to develop a robust AI evaluation strategy for incorporating algorithms into end users' decision-making processes. It emphasizes that understanding both the challenges and benefits of AI literacy is essential for effective human-algorithm collaboration:

1. *Respect individuals with higher AI literacy levels for their mistrust in ML algorithms.* While mistrust can be harmful in the case of good advice, it is beneficial in the case of bad advice (papers 4, 6).

2. *Differentiate between the conjoint performance of human-algorithm collaboration and naive advice-taking from algorithms.* Naive and nuanced advice-taking strategies might lead to similar results in some situations (paper 6).

3. *Ensure that metrics are aligned with conflicting goals.* Selected measures may not lead to the desired goal, for instance, because of communicating omitted variables, outliers or adverse side effects such as higher energy consumption (papers 2, 4, 7).

6. Closure

Given the accelerating proliferation of ML algorithms, exploring the impact of imperfections throughout the ML pipeline offers a promising direction for future research. Future research should further explore the interactions between algorithmic, individual and other factors, such as high-level and task-related factors. For instance, future work could further explore the interaction of uncertainty-related insights with models that predict and communicate their uncertainty, offering users dynamic confidence measures alongside advice.

We focused on arguably objective application tasks. Although we observed consistent phenomena like algorithm appreciation across multiple studies, future work should examine how different application scenarios may influence these findings (see Castelo et al., 2019). Additionally, exploring alternative dependent variables can offer valuable insights. For instance, although WOA is a common dependent variable in the algorithm appreciation literature (e.g., Logg et al., 2019), it has certain limitations (Bonaccio and Dalal, 2006). It is important to note that our work primarily focused on the effects of individual variables rather than modeling comprehensive user behavior. Future work could aim to develop integrative models that better capture the multifaceted nature of advice-taking.

Due to the varying application contexts, we used different framing, which future work can examine further. While studies frequently assess advice-taking among students and crowdworkers using experiments (e.g., Fügener et al., 2022), laboratory experiments tend to offer low external validity, as artificial environments can distort natural behavior (Shadish et al., 2003). In addition, participants can unconsciously adapt their behavior due to observation effects such as social desirability (Shadish et al., 2003). Therefore, future work might adopt alternative research designs, such as qualitative designs with data scientists or examining long-term learning effects in human-algorithm collaboration using field experiments with more specific populations. Additionally, participants who completed the task independently before receiving ML advice might respond differently than those who receive advice immediately (Buçinca et al., 2021).

Given the sample size constraints typical of laboratory experiments, future research should prioritize replicating our findings with larger and more diverse populations. Additionally, the user interfaces in our studies included supplementary task-related information — such as MAE values or historical time series data — which may have influenced participants' responses, potentially mediating or moderating the observed effects. In real-world settings, the contextual information provided by algorithms and human advisors often differs; for example, human advisors typically do not present statistical metrics but may allow for interactive follow-up questions (Önkal et al., 2009).

While this dissertation examined several algorithmic imperfections, future studies should investigate additional forms and their effects on human advice-taking behavior. Future work can apply the FTI-JAS framework developed in this dissertation to facilitate cross-disciplinary communication by integrating technical knowledge of ML algorithm imperfections from computer science with behavioral insights from psychology. Although this work enhances our understanding of individuals' responses to algorithmic advice, the broader challenge of successfully integrating AI into organizational contexts remains critical. Ensuring user acceptance and fostering effective human–algorithm collaboration will be key to unlocking the full economic potential of AI systems.

Bibliography

Berger, B., Adam, M., Rühr, A., and Benlian, A. (2021). Watch me improve - Algorithm aversion and demonstrating the ability to learn. *Business and Information Systems Engineering*, 63(1):55–68.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006th edition.

Bonaccio, S. and Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2):127–151.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *ACM on Human-Computer Interaction*, 5(CSCW):1–21.

Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.

Cambon, A., Hecht, B., Edelman, B., Ngwe, D., Jaffe, S., Schwarz, M., and Teevan, J. (2023). Early LLM-based tools for enterprise information workers likely provide meaningful boosts to productivity. Technical report, Microsoft Research. `https://www.microsoft.com/en-us/research/uploads/prod/2023/12/AI-and-Productivity-Report-First-Edition.pdf`, Accessed: 14 Feb, 2025.

Camerer, C. F. and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19:7–42.

Carmichael, M. (2024). The Ipsos AI monitor 2024. Technical report, Ipsos. `https://www.ipsos.com/en-us/ipsos-ai-monitor-2024`, Accessed: 14 Feb, 2025.

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.

Crawford, K. (2024). Generative AI's environmental costs are soaring — and mostly secret. *Nature*, 626(693):693–693.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Management Unit Working Paper*, 24(013):1–54.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. ACL.

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

Dijkstra, J. J., Liebrand, W. B., and Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour and Information Technology*, 17(3):155–163.

Dowling, J. (2023). From MLOps to ML systems with feature/training/inference pipelines. https://www.hopsworks.ai/post/mlops-to-ml-systems-with-fti-pipelines. Accessed: 03 Oct 2024.

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., and Riedl, M. O. (2021). The who in explainable AI: How AI background shapes perceptions of AI explanations. *arXiv*, pages 1–47.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696.

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70(1):1–25.

Gemini Team (2024). Gemini: A family of highly capable multimodal models. *arXiv*, pages 1–90.

Golinelli, D., Boetto, E., Carullo, G., Nuzzolese, A. G., Landini, M. P., and Fantini, M. P. (2020). Adoption of digital technologies in health care during the COVID-19 pandemic: Systematic review of early scientific literature. *Journal of Medical Internet Research*, 117(11):1–23.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., and Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1):19–30.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *IEEE Conference on Computer Vision*, pages 1026–1034. IEEE.

Heßler, P. O., Pfeiffer, J., and Hafenbrädl, S. (2022). When self-humanization leads to algorithm aversion: What users want from decision support systems on prosocial microlending platforms. *Business and Information Systems Engineering*, 64(3):275–292.

Heyder, T., Passlack, N., and Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *Journal of Strategic Information Systems*, 32(3):101772.

Huyen, C. (2022). *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. O'Reilly, 1st edition.

Hyndman, R. J. and Koehler, A. B. (2005). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Jacobs, M., Pradier, M. F., McCoy, T. H., Perlis, R. H., Doshi-Velez, F., and Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: The example of the antidepressant selection. *Translational Psychiatry*, 11(108):1–9.

Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3):685–695.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger,

M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In *European Conference on Information Systems*, pages 1–16. AISeL.

Jussupow, E., Benbasat, I., and Heinzl, A. (2024). An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making. *MIS Quarterly*, 48:1–16.

Kapoor, S. and Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. *arXiv*, pages 1–29.

Langer, M. and Landers, R. N. (2019). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, 106878:9–25.

Lebedeva, A., Protte, M., van Straaten, D., and Fahr, R. (2024). Involvement of domain experts in the AI training does not affect adherence: An AutoML study. In Arai, K., editor, *Advances in Information and Communication*, pages 178–204. Springer Nature.

Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. (2021). Is AI ground truth really true? The dangers of training and evaluating AI tools based on experts' know-what. *MIS Quarterly*, 45(3):1501–1525.

Leffrang, D. (2023). The broken leg of algorithm appreciation: An experimental study on the effect of unobserved variables on advice utilization. In *Wirtschaftsinformatik Conference*, pages 175 – 189. AISeL.

Leffrang, D., Bösch, K., and Müller, O. (2023). Do people recover from algorithm aversion? An experimental study of algorithm aversion over time. In *Hawaii International Conference on System Sciences*, pages 4016–4025. AISeL.

Leffrang, D. and Mueller, O. (2024). Algorithmic advice-taking beyond MAE: The role of negative prediction outliers and statistical literacy in algorithmic advice-taking. In *European Conference on Information Systems*, pages 1–16. AISeL.

Leffrang, D. and Müller, O. (2023). AI washing: The framing effect of labels on algorithmic advice utilization. In *International Conference on Information Systems*, pages 1–17. AISeL.

Leffrang, D. and Müller, O. (2024). Visualizing uncertainty in time series forecasts: The impact of uncertainty visualization on users' confidence, algorithmic advice utilization, and forecasting performance. *Journal of Forecasting*, 44:1–12.

Leffrang, D. and Müller, O. (2025). The sustainability-performance trade-off in AI: The role of sustainability information and unmet performance goals in sustainable AI decisions. *Working Papers Dissertations, Paderborn University*, 135:1–31.

Leffrang, D., Passlack, N., Müller, O., and Posegga, O. (2025). Beneficial mistrust in generative AI? the role of AI literacy in handling bad advice. *Working Papers Dissertations, Paderborn University*, 136:1–35.

Logg, J. M., Minson, J. A., Moore, D. A., and States, U. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Long, D. and Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Conference on Human Factors in Computing Systems*, pages 1–16. ACM.

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., and Zhang, G. (2018). Learning under concept drift: A review. *Transactions on Knowledge and Data Engineering*, 31(12):2346–2363.

Madhavan, P. and Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, 49(5):773–785.

Mahmud, H., Islam, A. K., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:1–26.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J. B., Rome, A., Shi, A., and Oak, S. (2025). The AI index 2025 annual report.

Technical report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University. https://doi.org/10.48550/arXiv.2405.19522, Accessed: 28 Apr, 2025.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Annual Meeting of the ACL*, pages 1906–1919. ACL.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press, 1st edition.

Moore, S. (2017). Gartner says AI technologies will be in almost every new software product by 2020. https://www.gartner.com/en/newsroom/press-releases/2017-07-18-gartner-says-ai-technologies-will-be-in-almost-every-new-software-product-by-2020. Accessed: 08 Sep 2022.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., and 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.

OpenAI (2024). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 23 Apr, 2025.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. (2021). Carbon emissions and large neural network training. *arXiv*, pages 1–22.

Prahl, A. and Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702.

Russell, S. J. and Norvig, P. (2016). *Artificial Intelligence: A Modern Approach.* Pearson, 3rd edition edition.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2003). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin Company, 1st edition.

Singla, A., Sukharevsky, A., Yee, L., Chui, M., and Hall, B. (2025). How organizations are rewiring to capture value. *McKinsey & Company*, pages 1–23.

Spiegelhalter, D., Pearson, M., and Short, I. (2011). Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400.

Surameery, N. M. S. and Shakor, M. Y. (2023). Use Chat GPT to solve programming bugs. *International Journal of Information technology and Computer Engineering*, 3(1):17–22.

Tully, S., Longoni, C., and Appel, G. (2025). Express: Lower artificial intelligence literacy predicts greater ai receptivity. *Journal of Marketing*, 0(0):1–20.

Turel, O. and Kalhan, S. (2023). Prejudiced against the machine? Implicit associations and the transience of algorithm aversion. *Management Information Systems Quarterly*, 47(4):1–26.

VHB (2024). VHB Rating 2024 - Wirtschaftsinformatik/Information Systems. https://www.vhbonline.org/en/verband/wissenschaftliche-kommissionen/wirtschaftsinformatik/vhb-rating-2024-wirtschaftsinformatik/information-systems. Accessed: 03 Mar, 2025.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):1–52.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H., Akyildiz, B., Balandat, M., Spisak, J., Jain, R., Rabbat, M., and Hazelwood, K. (2022). Sustainable AI: Environmental implications, challenges and opportunities. In *Machine Learning and Systems*, pages 795–813. MLSys organization.

Yaniv, I. and Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2):260–281.