Hochschule Neu-Ulm
University of Applied Sciences

Masterarbeit gemäß § 10 der Studien- und Prüfungsordnung vom 17.08.2009
im Masterstudiengang International Enterprise Information Management
an der Hochschule für angewandte Wissenschaften Neu-Ulm

Review of the Conflicting Literature on Algorithm Aversion

Erstkorrektor: Prof. Andy Weeger

Zweitkorrektor: Prof. Heinz-Theo Wagner

Verfasser: Arthur Pfeffer (Matrikel-Nr.: 308579)

Thema erhalten: 30.06.2025

Arbeit abgeliefert: 31.08.2025

Abstract

Algorithms and AI systems increasingly influence decision-making, yet research on human responses remains fragmented across three constructs: Algorithm Aversion, Algorithm Appreciation, and Automation Bias. This thesis conducts a grounded theory-based literature review of 54 studies to clarify their boundaries, operationalizations, and interrelations. The analysis shows that Algorithm Aversion and Automation Bias are structurally symmetrical phenomena, defined as under- and over-reliance relative to rational performance benchmarks, while Algorithm Appreciation constitutes a distinct positive stance that often reflects rational adaptation rather than systematic bias. To reconcile these insights, the thesis develops a cycle model that conceptualizes evaluation, behavior, and consequences as dynamically linked, with mental models shaping how outcomes feed back into future evaluations. This model aligns subjective measures (expectations, risk perceptions, identity relevance, affective reactions) and objective measures (reliance, accuracy, efficiency, safety) with their respective stages of the cycle. The thesis contributes to Information Systems research by refining construct definitions, integrating Algorithm Aversion, Algorithm Appreciation, and Automation Bias into a shared process logic, and emphasizing the practical significance of Algorithm Aversion and Automation Bias while questioning whether Algorithm Appreciation warrants treatment as a standalone phenomenon.

Keywords: Algorithm Aversion, Algorithm Appreciation, Automation Bias, Human-Computer-Interaction, Grounded Theory Review

Table of Contents

List of Figures

List of Tables

1 Introduction

Despite the growing practical adoption of algorithms, academic research consistently shows that decision makers often fail to fully benefit from algorithmic advice, a phenomenon referred to as Algorithm Aversion. Dietvorst et al. (2015) define Algorithm Aversion as “a general tendency for people to more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake”. Early studies primarily investigated the phenomenon in decision-making contexts such as forecasting (e.g., Logg et al. 2019; Prahl and Van Swol 2017; You et al. 2022) and diagnostics (e.g., Cabitza et al. 2023; Liu et al. 2023). Research has since expanded across disciplines and tasks, including Economics (e.g., Horowitz and Kahn 2023; Tse et al. 2024), Human-Computer Interaction (e.g., Nourani et al. 2021; Schaffer et al. 2019), Management (e.g., Dennis et al. 2023; Turel and Kalhan 2023), and Marketing (e.g., Brüns and Meißner 2024; Castelo et al. 2019). This expansion highlights both the phenomenon’s ubiquity and its growing complexity, underscoring the need for greater conceptual clarity.

Empirical findings, however, are far from conclusive. For instance, while some studies confirm that humans tend to discount algorithmic advice and algorithmic errors prompt stronger rejection than human errors (e.g., Commerford et al. 2022; Dietvorst et al. 2018), others reveal the contrasting effect that people sometimes prefer algorithmic advice over human alternatives, a phenomenon termed Algorithm Appreciation (Logg et al. 2019). Adding further complexity, individuals may respond both positively and negatively depending on the measure applied (Dennis et al. 2023), the task employed (Castelo et al. 2019), or even alternate between Algorithm Aversion and Appreciation (Turel and Kalhan 2023). Recent work seeks to reconcile these opposing patterns. Scholars such as Hou and Jung (2021) and Cheng and Chouldechova (2023) have explored their interrelation, and Horowitz and Kahn (2023) propose a conceptual symmetry between Algorithm Aversion and the well-established concept of Automation Bias, which they define as “the tendency of humans to rely on AI-enabled decision aids above and beyond the extent to which they should, given the reliability of the algorithms". This fragmented empirical landscape makes cumulative theory-building difficult, which is further amplified by differences in how the phenomena are conceptualized across studies. Some definitions use generic terms, such as “algorithm” (e.g., Liu et al. 2023; Rix et al. 2025), while others adopt narrower terms, such as “evidence-based algorithms” (e.g., Dietvorst et al. 2018; Keppeler 2024) or “AI-driven decision-making tools” (e.g., Cabitza et al. 2023) as targets of Algorithm Aversion. Similarly, the phenomenon is at times framed as “preference for humans” (e.g., Castelo 2024; Reich et al. 2023) and at other times as “reluctance to use algorithmic advice” (e.g., Cabiddu et al. 2022; Wu et al. 2024). Such inconsistencies make it difficult to compare findings and highlight the absence of an integrated theoretical foundation. As a result, clarifying how these concepts are defined and how they relate to one another becomes a critical first step of my thesis. On this basis, my first research question is:

RQ1: How have scholars conceptualized Algorithm Aversion, Algorithm Appreciation, and Automation Bias and how can these phenomena be theoretically related to one another?

While conceptual clarity is essential, it is equally important to recognize that these phenomena do not manifest uniformly but are shaped by different factors. For instance, Mahmud et al. (2022) catalogued a wide range of situational and design-related variables that influence responses to algorithmic advice in decision-making, suggesting that Algorithm Aversion and Algorithm Appreciation may arise under different but similar conditions, producing divergent outcomes across studies. Empirical records demonstrate that at other times people tend to over-rely on automation even when it harms their performance (Cabitza et al. 2023; Horowitz and Kahn 2023; Skitka et al. 1999). Taken together, this

fragmentation signals a theoretical gap. Current research lacks a unifying account that explains when and why these phenomena occur. Consequently, my second research question is:

RQ2: What overarching themes can be identified across the literature on Algorithm Aversion, Algorithm Appreciation, and Automation Bias, and how can these phenomena be integrated into a unified theoretical framework?

My thesis makes two main contributions. First, it develops a coherent conceptualization of Algorithm Aversion, Algorithm Appreciation, and Automation Bias through the construction of refined definitions based on a systematic semantic decomposition of extant definitions. This step addresses the need for conceptual clarity and provides a basis for comparing and relating the three phenomena. Second, it advances an integrative understanding of the mechanisms underlying these constructs by identifying emerging themes in current research and synthesizing them into a cyclical framework of behavior, consequences, evaluation, and mental models. This framework captures how reliance on algorithmic systems is shaped by dynamic feedback loops rather than isolated responses, offering a theoretical lens for interpreting fragmented empirical findings. To achieve this, my study follows a grounded theory approach to systematically review and synthesize the literature. The results are presented in four stages: (1) a semantic decomposition of extant definitions from which I construct formal definitions of each construct, (2) a cross-phenomenon comparison that relates these constructs to one another, (3) an analysis of operationalization patterns, and (4) an inductive synthesis of emerging themes across the corpus, leading to the development of an integrative framework that links behavior, consequences, evaluation, and mental models. Together, these steps build the basis for a more comprehensive theoretical understanding of the relationship between the phenomena. The next chapter outlines the methodology of my review.

2 Method

My literature review follows the methodological principles of a broad theorizing review as articulated by Leidner (2018). Rather than enforcing rigid boundaries between review types, Leidner encourages scholars to embrace discomfort and novelty in research, advocating for hybrid forms and methodological flexibility. In this spirit, my review synthesizes literature from multiple disciplines to build an emergent theoretical framework that explains the fragmented and often inconsistent empirical findings surrounding Algorithm Aversion.

To systematically construct this framework, I adopt the grounded theory-based approach to structured literature reviews as proposed by Wolfswinkel et al. (2013). Building on the foundational principles of Webster and Watson (2002), this method “enables the researcher to come up with a theory-based or concept-centric yet accurate review”, grounded in transparent procedures. It is particularly well suited for reviews aiming at theoretical integration and conceptual development, especially in fields where the empirical landscape is fragmented, and established theory is lacking or underdeveloped, as demonstrated by recent literature reviews on digital innovation (Hund et al. 2021) and digital transformation (Vial 2019).

The literature on Algorithm Aversion, for instance, originated in decision support systems research (Dietvorst et al. 2015), but has since been applied to vastly different domains such as art creation (Magni et al. 2024) or brand content evaluation (Brüns and Meißner 2024). On the other hand, contrasting studies of Algorithm Appreciation suggest that users may sometimes prefer algorithmic advice, further complicating the conceptual boundaries of the phenomenon (Logg et al. 2019). As a result, the concept

lacks definitional clarity and consistent empirical grounding, justifying the need for a structured, theory-generating review.

To address this, the review follows the five-stage process outlined by Wolfswinkel et al. (2013): (1) Define, (2) Search, (3) Select, (4) Analyze, and (5) Present. Each stage is designed to ensure transparency, replicability, and conceptual rigor. A simplified representation and breakdown of the process sequence is shown in Figure 1.

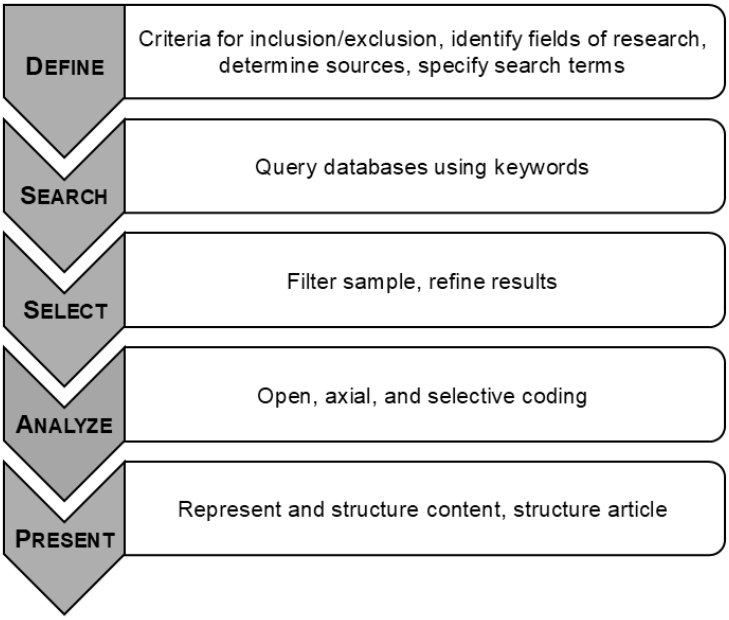

Figure 1: Review Process (Wolfswinkel et al. 2013)

The first four stages are subsequently introduced in this chapter, while the presentation of the results and emergent theoretical contribution is provided in the next chapter (for a more detailed account of the method employed, please refer to Appendix B).

2.1 Definition of Scope and Selection Criteria

In the initial phase of the review, I established the quality threshold as well as clear inclusion and exclusion criteria to ensure both conceptual relevance and methodological rigor.

Quality threshold: To evaluate the quality of journal articles, I relied on the 71st edition of Harzing's Journal Quality List (JQL, 2024) (Harzing 2024), a consolidated meta-ranking that synthesizes insights from eleven internationally recognized journal ranking systems. For conference proceedings, I relied on the CORE conference database maintained by the Computing Research and Education Association of Australasia (CORE 2023).

Inclusion Criteria: Publications were included if they are peer-reviewed, either empirical or conceptual in nature, and address Algorithm Aversion, Algorithm Appreciation, or Automation Bias. In addition, included studies needed to be situated within human–computer interaction settings.

Exclusion Criteria: Studies that lack a behavioral, cognitive, or motivational focus were excluded from the review. Likewise, technical or implementation-focused publications were excluded unless they are explicitly engaged with user acceptance or evaluation. Finally, studies concerned solely with algorithmic bias or fairness, where no human judgment or interaction is involved, were excluded, as were working papers and non-peer-reviewed publications lacking empirical evidence or methodological transparency.

In the second step, I defined the relevant fields of research by conducting exploratory keyword searches using terms such as “Algorithm Aversion”, across a range of databases, including ACM Digital Library, AIS electronic Library, and APA PsycNet. This scoping phase was intended to map the disciplinary landscape and clarify where and how the phenomenon has been addressed in academic research. Although Algorithm Aversion was initially conceptualized within the domain of behavioral decision-making (e.g., Dietvorst et al. 2015; Prahl and Van Swol 2017), it has since been taken up in a variety of scholarly contexts. Through a combination of exploratory searching and citation tracing of foundational publications (e.g., Dietvorst et al. 2015; Dietvorst et al. 2018; Logg et al. 2019), I identified a set of academic fields that have contributed substantially to the development and diversification of the discourse on Algorithm Aversion: psychology; information systems; management and organizational behavior; human-computer interaction; marketing and consumer behavior; medicine and health informatics; business and finance. These fields form the disciplinary foundation for this review and informed the subsequent selection of academic sources.

In the third step, I determined the appropriate sources for the literature search. Given the interdisciplinary character of the phenomenon under study, the selection of academic databases was guided by the need to capture research across both technical and behavioral domains. Particular emphasis was placed on ensuring comprehensive coverage of the IS literature, as my thesis is situated within the IS discipline and aims to contribute to its theoretical development. At the same time, interdisciplinary sources were included to account for conceptual and empirical contributions from adjacent fields. I selected six electronic databases based on their disciplinary relevance and ability to provide peer-reviewed coverage of the fields identified in the previous step: ACM Digital Library; AIS Electronic Library; APA PsycNet; EBSCOhost (Business Source Complete); IEEE Xplore; JSTOR. Together, these databases ensure conceptual coverage of the phenomenon from both behavioral and technical perspectives. This approach aligns with recommendations for rigorous and transparent source selection in interdisciplinary literature reviews (Templier and Pare 2018; Wolfswinkel et al. 2013).

In the fourth step, I defined the search terms. Searches focused on identifying publications with the terms “algorithm aversion” and “algorithm appreciation”. After scoping, I added “automation bias” to capture theoretically adjacent work and explore conceptual symmetry. The next section gives an overview of my search strategy.

2.2 Search Strategy

I subsequently performed searches for both keywords in either the title, abstract, or keywords across all selected databases, identifying 167 articles. The term “automation bias” was included, after my initial findings suggested a conceptual symmetry towards Algorithm Aversion (Horowitz and Kahn 2023). Table 1 shows the exact results for each query. Results retained for Algorithm Appreciation are low due to considerable overlapping with results for Algorithm Aversion and subsequent removal of duplicate articles.

Search Syntax	***Initial Search***	***Final Sample***
“Algorithm Aversion”	103	43
“Algorithm Appreciation”	32	3
“Automation Bias”	32	4

Table 1: Search Syntax & Sample

The following section provides an overview of the study selection strategy employed.

2.3 Study Selection Strategy

To refine the initial result of 167 articles, I applied several filtering steps followed by a forward and backward search. First, I removed duplicates (35), cover pages (2), and a call for papers (1). Next, I excluded publications from outlets that did not meet my predefined quality standards (73). I then reviewed the remaining 56 articles in full and applied the inclusion and exclusion criteria (6). Following the recommendations of Webster and Watson (2002), I conducted a forward and backward search, which led to the identification and inclusion of 4 additional articles that met the criteria. In total, this process yielded 54 articles relevant to my analysis. Figure 2 provides a detailed overview of the search and selection process.

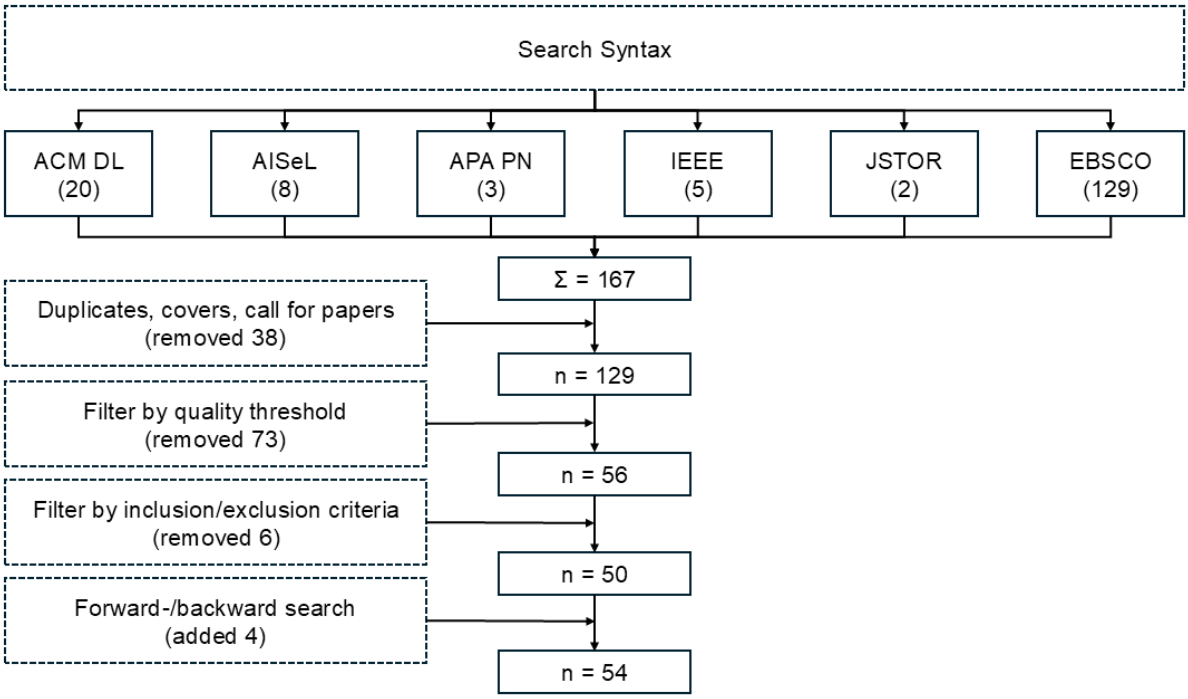

Figure 2: Filter & Selection Process

The following section outlines the data analysis approach.

2.4 Data analysis

In the fourth step, I conducted the data analysis through a sequential coding procedure drawing on grounded theory techniques (Wolfswinkel et al. 2013) and similar applications in IS reviews (Hund et al. 2021; Vial 2019). The process consisted of three phases: (1) open, (2) axial, and (3) selective coding, carried out iteratively to reflect grounded theory's principle of gradual discovery. First, I documented for each work a set of data points, including publication outlet, type of study (empirical, conceptual, or review), context of application, focal concepts, methods, and any construct definitions. Second, I split the sample into random batches of 15 studies and performed open coding, documenting relevant arguments, findings, and conceptual tensions. Excerpts and references were systematically transcribed into an Excel spreadsheet, which allowed me to preserve the contributions of individual studies while maintaining a comparative overview. After each batch, I conducted axial coding to refine the scheme and consolidate first-order codes into higher-order categories. This iterative process yielded 31

categories in total. Finally, I applied selective coding to integrate these categories into a coherent framework. The results of my analysis are presented in the following chapter.

3 Presentation of Findings

This chapter presents the empirical findings of my review in four stages, moving from definitional clarity and comparison of phenomena to operationalization patterns and finally to emerging themes. The objective is to transform a fragmented body of work into a coherent basis for theoretical integration.

First, I conduct a semantic decomposition of definitions of Algorithm Aversion, Algorithm Appreciation, and Automation Bias. Drawing on linguistic methods of decomposition (Akmajian et al. 2017), I identify six recurring primitives, that underpin existing conceptualizations. This step surfaces both shared foundations and heterogeneous elements, providing the basis for constructing refined and comparable definitions of each phenomenon.

Second, I extend this analysis into a cross-phenomenon comparison. By systematically aligning the primitives across constructs, I identify areas of overlap, conceptual boundaries, and points of divergence. This includes an ontological inversion exercise that evaluates whether the phenomena represent symmetrical opposites or instead occupy distinct conceptual spaces. The results reveal that while Algorithm Aversion and Automation Bias form reactive counterparts anchored in rationality benchmarks, Algorithm Appreciation functions more as a baseline evaluative orientation without conditionality.

Third, I examine patterns of operationalization across empirical studies. Here, I distinguish between subjective measures, capturing attitudes, expectations, emotions, and preferences, and objective measures, capturing behavioral reliance and its consequences for accuracy, efficiency, and safety. This analysis shows that reliance is not only a matter of reported attitudes and expectations but also of measurable behavioral choices and performance outcomes, with each approach highlighting different facets of the underlying constructs.

Finally, I synthesize these strands into a set of emerging themes in current research. The analysis identifies four interdependent categories (behavior, consequences, evaluation, and mental models) that together form a dynamic cycle of human–algorithm interaction. This framework demonstrates that Algorithm Aversion, Algorithm Appreciation, and Automation Bias are not isolated reactions but recursive processes shaped by feedback loops between user actions, contextual outcomes, interpretive evaluations, and evolving mental models.

3.1 Semantic Decomposition of Definitions

By applying semantic decomposition (Akmajian et al. 2017), as demonstrated by Vial (2019) and Hund et al. (2021), I inductively identified six primitives that form the basis of definitions of Algorithm Aversion, Algorithm Appreciation, and Automation Bias: (1) Actor, (2) Conduct, (3) Target, (4) Comparator, (5) Condition, and (6) Scope. These primitives serve as the structural building blocks for subsequently constructing and comparing definitions across phenomena (see Appendix D for details on the process of semantic decomposition).

3.1.1 Overview of Semantic Decomposition Results

Across the final sample, I identified a total of 69 formal definitions distributed over 45 sources. Of these, 44 definitions concern Algorithm Aversion (39 sources), 15 Algorithm Appreciation (15 sources), and 10

Automation Bias (4 sources). Several sources provide more than one formal definition for the same phenomenon or define multiple phenomena within the same article (e.g., Bankuoru Egala and Liang 2024; Turel and Kalhan 2023). This overlap is particularly evident for Algorithm Appreciation, which is frequently introduced in relation to Aversion. Indeed, 10 of the 15 articles on Algorithm Appreciation formally define both Algorithm Aversion and Appreciation, reflecting the fact that much of the Algorithm Appreciation literature positions itself as a counterpoint to or extension of Algorithm Aversion research (e.g., Dennis et al. 2023; Reich et al. 2023). By contrast, Algorithm Aversion is often investigated independently, with the majority of articles defining only Aversion (e.g., Burton et al. 2020; Mahmud et al. 2022). Automation Bias appears less frequently overall, but one article stands out for its concentration of definitional material. Cabitza et al. (2023) contributes six separate definitions of Automation Bias and one definition each of Algorithm Aversion and Appreciation. This makes it the most prolific single source within my sample in terms of definitional coverage.

The distribution of definitions underscores the central role of Algorithm Aversion as the most frequently defined construct in the sample, while Algorithm Appreciation emerges mainly in tandem with Aversion, and Automation Bias is less common but still represented. Figure 3 illustrates the overlap in definitional coverage across phenomena in the sample.

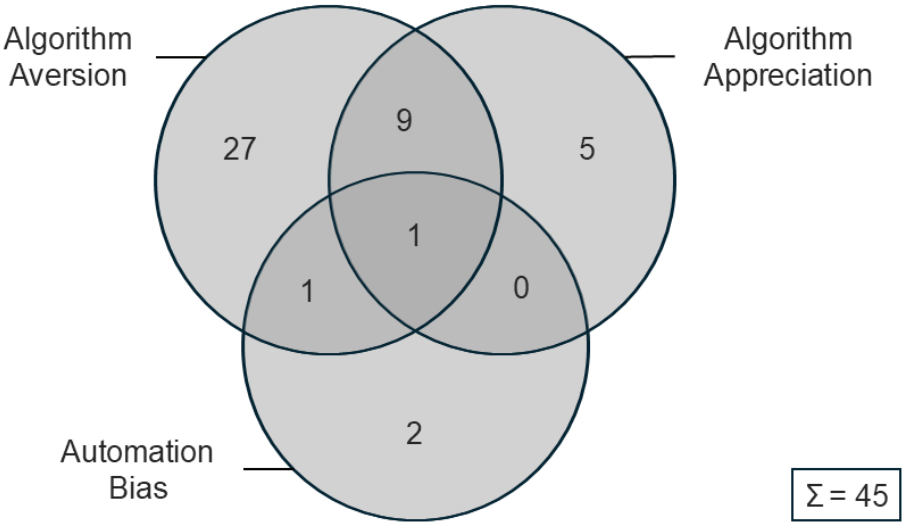

Figure 3: Distribution of sources across definitional themes

Beyond their distribution, it is also important to examine how these definitions are structured. To this end, I summarize the results of the semantic decomposition for each phenomenon. The following tables present, for each of the six primitives (Actor, Conduct, Target, Comparator, Condition, Scope), the extent of coverage across the sample and the range of values identified.

Table 2 summarizes the semantic decomposition of Algorithm Aversion definitions (n = 44). Aversion is the most frequently defined phenomenon in the sample and also the most heterogeneous in its definitional structure. The table shows how each primitive is represented, including the range of actors, the diversity of conduct combinations, and the variation in targets, comparators, conditions, and scope.

Primitive	***Count***	***Identified Values***
Actor	35/44	People (13); Humans (5); Decision-makers (4 explicit, 8 implicit); Users (1 explicit, 2 implicit); Forecasters (1); Clinicians (1)
Conduct	44/44	+A (5); -A (2); -B (11); +AB (1); -AB (17); -AC (3); -ABC (3); +A-B (1); +B-B (1);

Target	44/44	Human (1); Human output (3); Own output (1); Algorithmic/AI output (34); Algorithm (2); Human & algorithm output (2)
Comparator	39/44	Algorithm/AI output (4); Algorithm/AI (1); Human (1); Human output (18); Own output (3); Alternative (12)
Condition	27/44	Algorithm superiority (11); Algorithm imperfection (7) Algorithm superiority & imperfection (9)
Scope	44/44	General scope (12); Decision-making (24); Forecasting (4); Decision-making & forecasting (1); Evaluation (2); Not specified (1)

Table 2: Semantic decomposition of Algorithm Aversion definitions

Definitions of Algorithm Aversion typically describe the phenomenon in general terms, most often referring to "*people"* or "*humans" (e.g., Commerford et al. 2022; Dietvorst et al. 2015)*, with some formulations specifying contextual roles such as *"decision-makers"*, *"forecasters"*, or *"clinicians"* (e.g., Bankuoru Egala and Liang 2024; Burton et al. 2020; Cabiddu et al. 2022). Conduct is most often framed negatively toward algorithmic systems (e.g., Dietvorst et al. 2015; Prahl and Van Swol 2017), though some definitions emphasize a positive stance toward humans (e.g., Castelo 2024; Reich et al. 2023), and a few include both directions simultaneously (e.g., Commerford et al. 2022). These conceptualizations span attitudinal, behavioral, and calibration-related forms of conduct, with roughly equal frequency of attitudinal and behavioral framings, and a smaller but notable share explicitly including calibration asymmetry (see Table 2). While early studies referred narrowly to *"algorithmic"* or *"automation advice"* as the target of aversion (Dietvorst et al. 2015; Prahl and Van Swol 2017), more recent literature positions the construct to span both classic algorithms and contemporary AI applications (e.g., Dennis et al. 2023). A minority shifts the target to human-generated output or both algorithmic and human-generated output (e.g., Commerford et al. 2022; Feng and Gao 2020). Comparators are frequently included, usually contrasting the target output with either human or algorithmic alternatives (e.g., Castelo 2024; Dietvorst et al. 2018). Though some definitions instead invoke one's own judgment (e.g., Chávez et al. 2024; Jain et al. 2025) or unspecified "alternatives (e.g., Burton et al. 2020; Prahl and Van Swol 2017). Conditions are common and often central. Many definitions emphasize that Algorithm Aversion occurs despite algorithms being demonstrably superior (e.g., Koo 2024; Mahmud et al. 2022), while others situate it in the presence of imperfection or error (e.g., Ganbold et al. 2022; Rebholz et al. 2024), and some combine both (e.g., Burton et al. 2020; Dietvorst et al. 2018). Scope is rarely stated explicitly but can often be inferred from context. It spans decision-making, forecasting, evaluation contexts, and domain-specific applications such as medical diagnostics (e.g., Bankuoru Egala and Liang 2024; Dennis et al. 2023; Dietvorst et al. 2018; Prahl and Van Swol 2017). Taken together, these patterns show that Algorithm Aversion is conceptualized as a multifaceted phenomenon defined by reluctance and critical evaluation of algorithmic input, with boundaries shaped by both the direction of comparison and the conditions of performance.

Table 3 presents the decomposition of Algorithm Appreciation definitions (n = 15). Algorithm Appreciation is usually defined in close connection with Aversion. Its definitional structure, however, is more uniform, particularly in terms of target (always algorithmic output) and comparator (almost always human output), though there is still some variation in actor terms, conduct, and scope.

Primitive	***Count***	***Identified Values***
Actor	11/15	People (5); Individuals (3); Human (1); Decision-makers (1); Consumers (1)
Conduct	15/15	+A (5); +B (7); +C (1); +AB (1); +ABC (1)

Target	15/15	Algorithmic/AI output (15)
Comparator	13/15	Human output (11); Alternative output (2)
Condition	4/15	Equivalent output (3); Before AI makes a mistake (1)
Scope	15/15	Decision-making (10 implicit); Forecasting (1); Utilization (2); General (1); Not specified (1)

Table 3: Semantic decomposition of Algorithm Appreciation definitions

Algorithm Appreciation exhibits a consistent orientation toward algorithmic systems. The actor is usually expressed in generic terms such as “individuals” or “people” (e.g., Logg et al. 2019; You et al. 2022), with occasional role-specific formulations such as “decision-makers” or “consumers” (e.g., Koo 2024; Sachin and Schecter 2024). Conduct is uniformly positive, described as a preference for or greater reliance on algorithmic systems. While most definitions emphasize behavioral formulations (e.g., Turel and Kalhan 2023), several employ attitudinal expressions (e.g., Reich et al. 2023), and a small number extends to calibration asymmetry, suggesting stronger positive reactions to algorithmic than to human input under certain conditions (e.g., Cabitza et al. 2023). The target is most often specified as algorithmic or AI advice, other formulations describe information, predictions, or algorithmic conduct more generally, indicating that the phenomenon is not confined to advice alone (e.g., Downen et al. 2024; Rix et al. 2025). Comparators are almost always other human alternatives (e.g., Logg et al. 2019), with occasionally implied unspecified alternatives (e..g., Sachin and Schecter 2024), while conditions are rarely included, limited to occasional constraints such as identical advice or the temporal caveat that Algorithm Appreciation occurs before error cues (e..g., Dennis et al. 2023; You et al. 2022). Scope is usually not explicitly stated but can be inferred from the context. It is mostly situated in decision-making, sometimes forecasting (e.g., Logg et al. 2019; Reich et al. 2023), or more general utilization contexts (e.g., Dennis et al. 2023; Rix et al. 2025). In sum, Algorithm Appreciation captures a positive evaluative stance toward algorithmic systems, typically expressed as both unconditional preference and reliance in contrast to human alternatives.

Table 4 summarizes the decomposition of Automation Bias definitions (n = 10). While less frequently addressed in the sample, Automation Bias shows a distinctive pattern: comparators are typically framed not as alternative agents but as thresholds of appropriate reliance or vigilance.

Primitive	***Count***	***Identified Values***
Actor	8/10	Humans (2); Users (2); People (1); Physician (1); Human operators (1); Human decision-makers (1);
Conduct	10/10	+A (3); +B (3); +AB (2); +B-B (2)
Target	10/10	Algorithmic/AI/automation output (6); Automation (1); Technology advice (1); System advice (1); Computer-generated solution (1)
Comparator	10/10	Threshold of appropriate reliance/vigilance (8); Correct decision (1); Contradictory information (1)
Condition	10/10	Imperfection or erroneous system (2 explicit, 8 implicit)
Scope	10/10	Decision-making (4 implicit); System use (1 implicit); Automated system use (1 implicit); General automated operation (1); Technology use (1 implicit); Human-computer interaction (1); Physician tasks (1)

Table 4: Semantic decomposition of Automation Bias definitions

Automation Bias shows a distinctive definitional pattern. Definitions consistently describe "*humans"* or "*users" (e.g., Nourani et al. 2021)*, sometimes specified as *"human operators"*, *"physicians"*, or *"decision-makers"* (e.g., Cabitza et al. 2023; Horowitz and Kahn 2023). The conduct is uniformly positive, reflecting over-trust or over-reliance on automation, and is phrased as *"complacency"*, *"acceptance"*, or *"a propensity to defer to automated output"* (e.g., Cabitza et al. 2023; Schaffer et al. 2019). Targets differ in wording but generally refer to some form of automated or AI advice (e.g., Cabitza et al. 2023; Horowitz and Kahn 2023). Unlike Algorithm Aversion or Appreciation, comparators are rarely human alternatives, rather, over-reliance is judged against a normative standard of appropriate vigilance or correct decisions (e.g., Cabitza et al. 2023; Schaffer et al. 2019). Conditions are inherent to the construct, since Automation Bias occurs when reliance persists even though systems are imperfect or erroneous (e.g., Cabitza et al. 2023; Horowitz and Kahn 2023). Scope is typically implied as decision-making by framing the target as advice (e.g., Cabitza et al. 2023). In sum, Automation Bias captures a positive evaluative stance toward algorithmic systems, typically expressed as over-reliance on automated advice relative to appropriate standards of vigilance, and defined by persistence of such reliance even when systems are imperfect or erroneous.

Taken together, these overviews show both shared elements and definitional heterogeneity of Algorithm Aversion, Algorithm Appreciation, and Automation Bias. In the next section, I discuss each primitive in detail.

3.1.2 Definition of Primitives

To systematically compare and reconstruct definitions of Algorithm Aversion, Algorithm Appreciation, and Automation Bias, I decomposed them into six recurring semantic primitives: (1) actor, (2) conduct, (3) target, (4) comparator, (5) condition, and (6) scope. These primitives capture who exhibits the phenomenon, how it is expressed, what it is directed toward, against which alternative it is judged, under what circumstances it arises, and in which setting it occurs. The following section summarizes how each primitive appears across extant definitions, highlighting patterns as well as points of divergence.

Actor: The actor is the entity exhibiting the phenomenon. It includes generic terms such as "people", "individuals", "users" (e.g., Jenkin et al. 2024; Nourani et al. 2021; Shariff et al. 2021), or domain-specific roles such as "forecasters", "decision-makers", or "clinicians" (e.g., Bankuoru Egala and Liang 2024; Cabiddu et al. 2022; Xu and Wang 2024). Some studies implicitly specify the actor via the task setup, for example, advice-taking paradigms presuppose a human decision-maker (e.g., Jain et al. 2025; Rebholz et al. 2024). The choice of actor term can signal whether the construct is intended as a universal or as a contextual phenomenon. Taken together, the variation in actor terms shows that most scholars treat the phenomena as generalizable human tendencies, while a smaller subset situates them in role-specific contexts. This suggests a tension between universal conceptualization and domain-bound framing, which has implications for the transferability of findings across settings.

Conduct (A/B/C trichotomy): Conduct captures the nature of the actor's response toward the target of the phenomenon. Across definitions, conduct varies in both form and valence. Valence refers to whether the conduct is expressed in a positive or negative manner toward the target. For example, a positive valence can be expressed as *"biased preference for"* (Talebi et al. 2025), whereas a negative valence can be expressed as *"negative bias against"* (Dennis et al. 2023). My analysis of extant definitions revealed three distinct but related forms of conduct, which I refer to as the A/B/C trichotomy:

- **A – Attitude**: Self-reported evaluations of the target, such as trust (e.g., You et al. 2022), confidence (e.g., Cabitza et al. 2025), perceived reliability (e.g., Skitka et al. 1999), or fairness (e.g., Cheng and Chouldechova 2023).

- **B – Behavior**: Observable actions such as selection between agents (e.g., Rix et al. 2025; Xie et al. 2022) or adjustment toward an advisor's suggestion (e.g., Downen et al. 2024; Feng and Gao 2020).
- **C – Calibration**: Differences in how behavior changes following cues or triggers. This dimension captures asymmetries in participants' reactions to humans versus algorithms. For example, a stronger reduction in reliance on algorithms after observing errors (Dietvorst et al. 2015), a stronger reduction in reliance on humans after conflicting advice is given (Sachin and Schecter 2024), or a stronger increase in reliance on algorithms when framed as capable of learning (Reich et al. 2023).

Each definition in the sample was coded according to both its valence and the combination of these three dimensions. Several articles capture only a single dimension, independent of the phenomenon. For instance, some focus exclusively on attitudes (A; e.g., Dennis et al. 2023; Logg et al. 2019; Schaffer et al. 2019), others on behaviors (B; e.g., Castelo et al. 2019; Commerford et al. 2022; Nourani et al. 2021), and one solely on calibration (C; Cabitza et al. 2023). Some definitions encompass both positive attitude and behavior (+AB; e.g., Cabitza et al. 2023; Feng and Gao 2020; You et al. 2022), others combine negative attitude and calibration (-AC; e.g., Dietvorst et al. 2015), while a few include all three dimensions (ABC; e.g., Liu et al. 2023; Sachin and Schecter 2024). This variation highlights the heterogeneity in how scholars conceptualize the conduct primitive. This heterogeneity indicates that "conduct" is not a single dimension but a composite of attitudes, behaviors, and asymmetries in calibration. The fact that definitions cluster around different combinations (A, B, C) suggests why empirical studies often produce inconsistent findings. They are, in effect, operationalizing different facets of the same construct. Recognizing conduct as a multi-dimensional construct clarifies both the overlaps and the divergences among Algorithm Aversion, Algorithm Appreciation, and Automation Bias.

Target: The target is the object toward which conduct is directed. Across phenomena, this is most often algorithmic or AI output, typically expressed as advice (e.g., Logg et al. 2019; Prahl and Van Swol 2017). In Algorithm Aversion, negative valence conduct is usually directed toward such output, either explicitly (e.g., Turel and Kalhan 2023) or implicitly (e.g., Burton et al. 2020; Dietvorst et al. 2015), with occasional references to the algorithmic source itself (e.g., Liu et al. 2023; Rix et al. 2025). Positive valence definitions, by contrast, sometimes target human alternatives, either other people's output (e.g., Reich et al. 2023) or one's own (e.g., Feng and Gao 2020). A few definitions frame both directions simultaneously, targeting human and algorithmic output together (e.g., Commerford et al. 2022). Algorithm Appreciation consistently targets algorithmic output (e.g., Castelo et al. 2019; Logg et al. 2019), whereas Automation Bias also predominantly focuses on algorithmic or automated output (e.g., Horowitz and Kahn 2023; Nourani et al. 2021), occasionally naming the automation source itself (Cabitza et al. 2023). Despite these nuances, the common denominator is that the target is the output of algorithmic systems, with human output appearing primarily in comparative framings of Algorithm Aversion. This convergence around algorithmic output underscores that the three constructs fundamentally capture human evaluations of algorithmic systems, while differences in human versus algorithmic targets primarily reflect framing choices. Importantly, the consistent centrality of algorithmic output indicates that the phenomena are best understood as responses to system outputs rather than to the technical systems themselves.

Comparator: The comparator is any alternative against which the target is evaluated, most commonly human agents or human-generated output (e.g., Commerford et al. 2022; Dennis et al. 2023). In many definitions, the comparator is implicit. For instance, "preference for humans […] in decision-making" (Jussupow et al. 2024) presupposes comparison to algorithmic output, and vice versa. In Algorithm Aversion, comparators vary considerably. Some definitions frame the comparison against other

humans, as either one's own output (e.g., Jain et al. 2025) or another person's (e.g., Dietvorst et al. 2018), while others contrast algorithmic outputs directly (e.g., Castelo 2024) or invoke unspecified "alternative" options (e.g., Burton et al. 2020; Mahmud et al. 2022). Algorithm Appreciation is more uniform, almost always contrasting algorithmic advice with human advice (e.g., Castelo et al. 2019; Logg et al. 2019). Automation Bias is distinct in that its comparator is rarely another agent but instead a threshold of appropriate reliance (e.g., Horowitz and Kahn 2023; Schaffer et al. 2019) or contradictory information (Cabitza et al. 2023). These differences matter because comparator choice shapes how the construct is understood. Comparing against one's own judgment emphasizes self-other dynamics, comparing against another human emphasizes source preferences, and comparators defined as "appropriate vigilance" shift the construct into the territory of performance norms rather than inter-agent comparison. The divergence in comparator types highlights an important conceptual boundary. Algorithm Aversion and Algorithm Appreciation are relational constructs, defined by preference between agents, while Automation Bias is normative, defined against vigilance thresholds. This structural difference shows that while all three concern reliance on algorithmic systems, they are not symmetrical phenomena but anchored in distinct evaluative logics.

Condition: The condition refers to any facilitating or boundary circumstance under which the phenomenon is defined to occur. In definitions of Algorithm Aversion, conditions are often central, most commonly specifying that algorithms are demonstrably superior yet still avoided (e.g., Mahmud et al. 2022; Turel and Kalhan 2023), or that they are imperfect or error-prone (e.g., Dennis et al. 2023; Dietvorst et al. 2015), with some formulations combining both superiority and imperfection (e.g., Burton et al. 2020; Dietvorst et al. 2018). Algorithm Appreciation, by contrast, rarely includes explicit conditions; when it does, these usually involve equivalence manipulations (e.g., Cabitza et al. 2023; You et al. 2022) or in the case of Dennis et al. (2023) as temporal caveat before an error is observed. Automation Bias consistently carries an implicit condition of system imperfection, since the phenomenon presupposes over-reliance despite erroneous or unreliable outputs (e.g., Nourani et al. 2021; Schaffer et al. 2019). Taken together, these patterns suggest that conditions are definitional for Algorithm Aversion and Automation Bias but largely absent from Algorithm Appreciation. This asymmetry reinforces that Algorithm Aversion and Automation Bias are conceptualized as responses to violated expectations of performance, while Algorithm Appreciation functions more as a default positive orientation without requiring explicit boundary circumstances. This asymmetry in conditionality points to a fundamental difference in how the constructs are theorized. Algorithm Aversion and Automation Bias are reactive, triggered by expectation violations, while Algorithm Appreciation is non-conditional. This helps explain why Algorithm Appreciation often appears as a temporary or fragile effect in experiments, whereas Algorithm Aversion and Automation Bias persist in the presence of performance cues.

Scope: Scope refers to the domain or setting in which the phenomenon occurs. Across all three constructs, scope is most often left implicit, assuming a decision-making or forecasting contexts (e.g., Dietvorst et al. 2015; Logg et al. 2019). Where scope is specified, different patterns emerge. Algorithm Aversion shows the most variety. Some definitions anchor it in decision-making (Jussupow et al. 2024), others in forecasting (e.g., Reich et al. 2023), and a few in evaluative scenarios where algorithms are judged more harshly than humans (e.g., Dennis et al. 2023). Algorithm Appreciation definitions are similarly anchored in decision-making (e.g., Castelo et al. 2019), occasionally extending to forecasting (Reich et al. 2023) or utilization contexts (e.g., Downen et al. 2024). Automation Bias definitions are somewhat broader. While also rooted in decision-making (e.g., Cabitza et al. 2023), they extend into domains of general technology or system use (e.g., Nourani et al. 2021; Schaffer et al. 2019). This pattern suggests that scope is typically assumed rather than explicitly defined, and where it is specified, it reflects the disciplinary origins of the construct (e.g., forecasting for Algorithm Aversion and Appreciation, monitoring and clinical settings for Automation Bias). The reliance on implicit scope

reveals that these constructs are often treated as context-independent, even though actual applications cluster around particular domains. This lack of explicit scoping may inflate generalizability claims and complicates cross-study comparisons, since phenomena studied in safety-critical monitoring may not align with those in consumer decision-making.

The semantic decomposition shows that definitions of Algorithm Aversion, Algorithm Appreciation, and Automation Bias can be consistently broken down into six primitives. While each primitive recurs across phenomena, their formulation varies. Taken together, these results demonstrate that the three phenomena are conceptually related but heterogeneously defined, building the basis for constructing harmonized definitions and comparing them systematically in the following sections.

3.1.1 Construction of Definitions

Building on the definitional patterns identified in the previous section, I construct refined definitions of Algorithm Aversion, Algorithm Appreciation, and Automation Bias. In doing so, I follow established guidelines for strong conceptual definitions, particularly the criteria articulated by Suddaby (2010), who emphasizes that strong definitions should (1) clearly capture the core properties and characteristics of the concept or phenomenon, (2) avoid tautology or circular reasoning, and (3) be parsimonious while capturing essential complexity. Applying these principles, I evaluate the alternative formulations for each primitive and justify the choices that best balance inclusivity, clarity, and generalizability. The result is a set of consolidated definitions that provide a conceptual basis for subsequent comparison and integration.

Algorithm Aversion

To construct a cohesive definition of Algorithm Aversion, I drew on the commonalities and divergences in the literature. For the actor, terms such as *"people"*, *"humans"*, *"decision-makers"*, and *"forecasters"* appear, but the generic formulation of *"people"* is most frequent and avoids contextual narrowing. I therefore adopt *"people"* to preserve generality. For the conduct, definitions vary between attitudinal distrust, behavioral avoidance, and calibration asymmetry. Because the modal pattern emphasizes a negative stance toward algorithms, I use the behavioral phrasing *"reluctance to adopt"*. To acknowledge that some accounts also stress attitudinal evaluation and disproportionate reactions to errors, I extend this with *"critical evaluation of"*, which accommodates attitudinal and calibration elements without overcomplicating the wording. For the target, early studies referred to *"algorithmic advice"*, but later work broadened this to include AI output and related expressions. To generalize across both classic algorithms and contemporary AI systems, I formalize the target as *"output of algorithmic systems"*. Comparators show greater variation, including human output, one's own judgment, other algorithms, or unspecified alternatives but the dominant pattern contrasts algorithmic with human advice. I therefore adopt the inclusive wording *"compared to human alternatives"*. For conditions, most definitions highlight that Algorithm Aversion arises despite evidence of superior performance, while others situate it in reactions to imperfections or errors. I combine these by specifying that aversion occurs for imperfect algorithmic systems, despite evidence of superior performance. Finally, I leave scope implicit to allow applicability across domains. Synthesizing these strands, I define Algorithm Aversion as *people's reluctance to adopt output of imperfect algorithmic systems, and their critical evaluation of such output, despite evidence of the superior performance of those systems compared to human alternatives.*

Algorithm Appreciation

To integrate the definitional variety of Algorithm Appreciation into a cohesive formulation, I adopt *"people"* as the actor, as it represents the most common generic label and avoids narrowing the construct. For conduct, definitions converge on a uniformly positive orientation. Because reliance and

preference appear with similar frequency, and both attitudinal and behavioral framings are evident, I combine them in the phrasing *"preference for and greater reliance on"*. The calibration dimension, although less prominent, is acknowledged in a small number of sources. I do therefore not adopt it in my definition. For the target, the modal expression is *"algorithmic advice"*, but other formulations broaden this to *"predictions"*, *"information"*, or *"algorithmic conduct"*. To encompass this range without overstating the scope, I generalize to *"output from algorithmic systems"*, as I did for Algorithm Aversion. Comparators are almost always human alternatives, which I retain explicitly as *"compared to human alternatives"*. Since most definitions do not specify conditions, and the few that do are limited to operational choices (e.g., identical advice), I omit a condition. Scope is kept implicit to preserve generality. Accordingly, my definition for Algorithm Appreciation refers to *peoples' preference for, and greater reliance on, output from algorithmic systems compared to human alternatives.*

Automation Bias

To construct a cohesive definition of Automation Bias, I adopt the generic label *"humans"* to encompass the variety of actor terms used, such as *"people"*, *"users"*, *"operators"*, or professionals in specific domains. The conduct is uniformly positive and reflects an inclination to accept or defer to automation. Because definitions span attitudinal expressions such as *"over-trust"* and *"complacency"* as well as behavioral terms like *"over-reliance"* or *"a tendency to accept"*, I synthesize these into the phrasing *"tendency to over-rely"*. Targets differ in wording but consistently refer to automated or AI-generated advice, recommendations, or outputs, which I again generalize as *"output from algorithmic systems"*. Unlike Algorithm Aversion or Appreciation, the comparator is not another agent but an implicit normative standard of appropriate vigilance or correctness. For this reason, I leave the comparator unspecified in the definition, as it is implied by the next primitive. Conditions are inherent to the construct, since over-reliance constitutes bias only when systems are imperfect or erroneous. Scope is generally not defined explicitly, and I therefore keep it implicit to preserve generality. Synthesizing these aspects, I define Automation Bias as *the human tendency to over-rely on output from algorithmic systems, even when these systems are imperfect or erroneous.*

These constructed definitions provide a coherent and comparable conceptualization of Algorithm Aversion, Algorithm Appreciation, and Automation Bias. Each definition considers the diversity of formulations found in prior literature and the need for clarity, parsimony, and generalizability. By aligning them around a consistent set of primitives, I establish the basis for systematic comparison. The next section builds on this foundation by examining how the three phenomena relate to one another, identifying their commonalities and boundaries. The phenomena are then compared to one another by ontologically inverting the constructed definitions.

3.2 Cross-Phenomenon Comparison

The next step in understanding the fragmentation of the literature is to examine how the constructs Algorithm Aversion, Algorithm Appreciation, and Automation Bias are positioned relative to one another. While Algorithm Aversion is commonly defined as the rejection of algorithmic advice despite superior performance (Dietvorst et al. 2015), more recent work has proposed Algorithm Appreciation as its conceptual opposite, referring to the preference for algorithmic advice over human judgment (Logg et al. 2019). This binary framing is explicit in several studies, which position Algorithm Appreciation as a counterweight to Algorithm Aversion (e.g., Hou and Jung 2021). A similar contrast has been made between Algorithm Aversion and Automation Bias (Horowitz and Kahn 2023). However, not all scholars support a dichotomous view. Dennis et al. (2023), for example, argue that Algorithm Appreciation and Algorithm Aversion are not mutually exclusive but instead represent temporally situated reactions.

Algorithm Appreciation may emerge prior to algorithmic error, while Algorithm Aversion manifests post-error due to harsher judgment compared to humans. These inconsistencies highlight that the constructs are often used without shared boundaries or clear causal logic. This definitional ambiguity complicates empirical interpretation and motivates the need for a more integrated conceptual structure.

The comparative patterns across Algorithm Aversion, Algorithm Appreciation, and Automation Bias are summarized in Table 5. This table aligns the six primitives across all three phenomena to highlight both commonalities and points of divergence.

Primitive	***Algorithm Aversion***	***Algorithm Appreciation***	***Automation Bias***
Actor	Generic: people (13), humans (5); Role-specific: decision-makers (12 incl. implicit), forecasters (1), users (3), clinicians (1)	Generic: people (5), individuals (3); Role-specific: human, decision-makers, consumers (1 each)	Generic: humans (2), people (1). Role-specific: users (2), physician (1), human operators (1), decision-makers (1). 2 missing
Conduct	Positive valence: +A (5), +AB (1); Negative valence: –A (2), –B (11), –AB (17), –AC (3), –ABC (3); Mixed valence: +A–B (1), +B–B (1); Roughly equal frequency of attitudinal (A) and behavioral (B) framings, some with calibration (C)	Uniformly positive valence. +A (5), +B (7), +C (1), +AB (1), +ABC (1); Mostly attitudinal and behavioral; calibration less frequent	Uniformly positive valence. +A (3), +B (5), +AB (2). Strong emphasis on over-trust/over-reliance
Target	Positive valence: human (1), human output (3), own output (1); Negative valence: algorithmic/AI output (34 explicit/implicit), algorithm (2); Both: algorithm & human (2)	Always algorithmic/AI output (14 explicit, 1 implicit)	Algorithmic/AI/automation output (8 explicit/implicit), automation (1), computer-generated solution (1)
Comparator	Algorithm output (4), algorithmic agent (1), human agent (1), human output (18), own output (3), alternative (12)	Almost always human output (11); occasionally alternative output (2)	Normative standards: threshold of vigilance (8), correct decision (1), contradictory information (1)
Condition	Superior algorithm/output (11), imperfection/error (7), both (9)	Identical advice (3), before error (1)	Always imperfection or erroneous system (2 explicit, 8 implicit)
Scope	General (12), decision-making (24), forecasting (4), decision-making & forecasting (1), evaluation (2), unspecified (1)	Decision-making (10), forecasting (1), general (1), utilization (2), unspecified (1)	Decision-making (4), general automated operation (1), automated system use (1), technology use (1), system use (1), HCI (1), physician tasks (1)

Table 5: Comparison of Primitives across Concepts

The cross-phenomenon comparison reveals both shared foundations and important distinctions in how Algorithm Aversion, Algorithm Appreciation, and Automation Bias have been conceptualized. All three constructs share a broadly similar actor framing, with definitions typically referring to generic terms. This pattern indicates that the constructs are generally intended as universal tendencies but can be adapted to particular contexts.

Conduct is where differences become most salient. While Algorithm Aversion is characterized by definitional heterogeneity, spanning attitudinal, behavioral, and calibration elements with both positive and negative valence, Algorithm Appreciation and Automation Bias exhibit far more uniformity. Algorithm Appreciation is consistently positive toward algorithms, framed as preference or reliance, while Automation Bias is likewise positive but with a distinctive over-reliant quality.

Targets converge more closely, with nearly all definitions across phenomena centering on algorithmic or AI-generated outputs, often phrased as *"advice"*, *"information"*, or *"predictions"*. However, Algorithm Aversion occasionally redirects its focus toward human output (or both algorithmic and human simultaneously), reflecting both its comparative nature and lack of clarity in directionality.

Clearer contrasts emerge in comparators. Algorithm Aversion and Appreciation are typically defined relative to human alternatives, either one's own judgment, another person's advice, or human-generated output, underscoring their character as inter-agent preference phenomena. Automation Bias diverges by positioning reliance against implicit normative standards of appropriate vigilance or correctness. This distinction highlights a conceptual boundary. Algorithm Aversion and Appreciation are comparative constructs, while Bias is a normative one.

Conditions further differentiate the phenomena. Algorithm Aversion is commonly tied to conditions of algorithmic superiority or imperfection, situating the phenomenon as a response to violated expectations. Automation Bias also presupposes conditions, but in the opposite direction. It occurs specifically when systems are imperfect or erroneous yet are still followed, making imperfection inherent to its definition. Algorithm Appreciation, in contrast, rarely includes conditions, functioning more as a default positive orientation. This asymmetry reinforces the interpretation that Algorithm Aversion and Automation Bias are reactive constructs defined by boundary conditions, while Appreciation captures a baseline evaluative stance.

Finally, scope is usually left implicit across all three, but when specified, disciplinary differences become visible. Algorithm Aversion, Algorithm Appreciation, and Automation Bias definitions appear across decision-making. Algorithm Aversion and Appreciation extend towards forecasting, while Algorithm Aversion and Automation Bias appear across more general settings.

In the following subsections, I first discuss areas of overlap, then highlight conceptual distinctions and boundaries before developing an integrative interpretation.

3.2.1 Commonalities across Phenomena

Across all three constructs, actors are generally framed in generic terms (e.g., humans, people, individuals), with role-specific formulations appearing only occasionally. This suggests that the phenomena are broadly conceived as general human tendencies. Targets also converge, with nearly all definitions centering on algorithmic or AI-generated outputs such as advice, recommendations, or predictions. Human output appears less frequently and typically only in comparative framings. Similarly, scope is most often implicit, assumed to involve decision-making contexts. In some instances, it reflects disciplinary origins but still aligns around decision-related tasks. Taken together, these commonalities highlight that all three phenomena are conceptually related. They are defined as human responses to algorithmic systems, centered on their outputs, and frequently situated in contexts of decision-making.

3.2.2 Distinctions and Boundaries

Despite these shared components, clear differences emerge across the three constructs. Conduct diverges most strongly. Algorithm Aversion is heterogeneous and ambivalent, encompassing negative stances toward algorithms, positive stances toward humans, and calibration asymmetries. Algorithm Appreciation is consistently positive toward algorithms, framed mainly as either attitudinal preference or greater behavioral reliance. Automation Bias is also positive but framed as over-reliance. Comparators further delineate boundaries. Algorithm Aversion and Appreciation are comparative constructs, typically defined against human alternatives, whereas Automation Bias relies on implicit normative standards of one's own appropriate vigilance rather than another agent. Conditions also separate the phenomena. Algorithm Aversion and Automation Bias are tied to expectation violations. Algorithm Aversion arises despite algorithmic superiority or in response to imperfections, Automation Bias persists despite errors or unreliability. In contrast, Appreciation largely lacks explicit conditions and functions as an unconditional positive orientation towards algorithmic output. These distinctions delineate conceptual boundaries and underscore that the three phenomena, while structurally related, capture different evaluative orientations toward algorithmic systems.

3.2.3 Ontological Inversion and Concept Symmetry

While the previous subsections outlined commonalities and boundaries across Algorithm Aversion, Algorithm Appreciation, and Automation Bias, a further step is needed to test whether these constructs are merely distinct or whether they represent symmetrical opposites. Following Suddaby (2010), who insists on clear conceptual demarcation, one way to evaluate construct clarity is to invert each definition and examine whether the resulting "ontological opposite" corresponds to an existing phenomenon or reveals a conceptual gap.

Inverting the definition of Algorithm Aversion (*"people's reluctance to adopt output of imperfect algorithmic systems, and their critical evaluation of such output, despite evidence of superior performance"*) yields a construct that would describe *"people's preference for and adoption of algorithmic systems even when these systems are known to be inferior to human alternatives"*. This hypothetical opposite does not map onto Algorithm Appreciation as currently defined, since Algorithm Appreciation emphasizes preference for algorithms in general, often without specific conditions, rather than reliance in the face of demonstrable inferiority. The inversion thus reveals that Aversion and Appreciation are not symmetrical opposites. Algorithm Appreciation is not the logical opposite of Algorithm Aversion, but a related construct of inverted orientation with different boundary conditions.

By contrast, inverting the definition of Automation Bias *("the human tendency to over-rely on output from algorithmic systems, even when these systems are imperfect or erroneous"*) produces a phenomenon that would capture "people's tendency to under-rely on algorithmic systems even when they are correct or appropriate to trust". This is conceptually similar to Algorithm Aversion, suggesting a closer symmetry between Algorithm Aversion and Automation Bias than between Aversion and Appreciation. Both are reactive constructs defined by violations of rational benchmarks, but in opposite directions. Algorithm Aversion is defined as *"reluctance to adopt output […] despite evidence of superior performance"*, Automation Bias as *"over-reliance […], even when these systems are imperfect or erroneous"*.

Finally, the inversion of Algorithm Appreciation *("individuals' preference for, and greater reliance on, output from algorithmic systems compared to human alternatives"*) would yield a construct describing *"individuals' preference for and greater reliance on human output compared to algorithmic systems"*. While this overlaps with the positive directionality embedded in some definitions of Algorithm Aversion, it does not correspond neatly to any single construct. Instead, it highlights how appreciation, unlike the

other two phenomena, functions less as a condition-bound bias and more as a default positive orientation.

These inversions clarify that the constructs are not strictly symmetrical. While Algorithm Aversion and Algorithm Appreciation appear superficially opposed in valence, the definitional structure of Algorithm Appreciation lacks the explicit performance condition of Algorithm Aversion. Similarly, Automation Bias aligns with Algorithm Aversion in being condition-dependent, but it diverges by grounding its comparator in normative standards of one's own vigilance rather than human alternatives, that are not limited to ones own judgment. This analysis shows that the three phenomena overlap in targeting algorithmic output but differ in their conditionality and reference frames. While this suggests that Algorithm Aversion and Automation Bias are conceptually symmetric as under-reliance versus over-reliance with negative consequences, the scope of their definitions diverges. Automation Bias is typically operationalized in monitoring or diagnostic contexts, where reliance is judged against normative standards of appropriate vigilance (Cabitza et al. 2025; Schaffer et al. 2019; Skitka et al. 1999). Algorithm Aversion, by contrast, extends beyond vigilance norms to comparisons with other human advisors, incorporating a broader set of participant roles and scenarios. This asymmetry indicates that Algorithm Aversion encompasses a wider boundary space than Automation Bias, which complicates their functional symmetry. They are thus conceptually opposite in logic, but not equally broad in scope.

3.3 Operationalization Patterns

In this section I examine how Algorithm Aversion, Algorithm Appreciation, and Automation Bias have been operationalized in empirical research. Its purpose is to show how these abstract constructs have been translated into measurable indicators and to highlight the main patterns of measurement across studies. Two broad types of measures recur: (1) subjective measures, which capture attitudinal evaluations through self-reports, and (2) objective measures, which capture reliance and its consequences through observable behavior and performance outcomes. These can partially be mapped onto the A/B/C trichotomy of conduct identified in the chapter on Semantic Decomposition of Definitions. Attitudes (A) are reflected in subjective measures, behaviors (B) in objective reliance indicators such as choice or adjustment, and calibration (C) emerges from comparative shifts in reliance following cues or errors. The following subsections detail how these measures have been implemented, illustrating common operationalization patterns (see Table D. 1 in Appendix D for an overview of all empirical studies).

3.3.1 Subjective Measures

Subjective indicators are the most common operationalization of Algorithm Aversion, Algorithm Appreciation, and Automation Bias. These measures rely on self-reported evaluations, often through Likert-type scales or psychometric instruments, and are designed to capture how participants perceive and evaluate algorithmic or human agents. While widely used across experiments and surveys, the constructs vary in focus and can be clustered into six thematic categories, as shown in Table 6.

First, performance-related expectations include items such as trust (e.g., Horowitz and Kahn 2023; Reich et al. 2023), confidence (e.g., Prahl and Van Swol 2017; Skitka et al. 1999), perceived competence (e.g., Ganbold et al. 2022), perceived learning capabilities (e.g., Chacon et al. 2025; Reich et al. 2023), familiarity with algorithmic systems, or the task itself (e.g., Schaffer et al. 2019). These measures investigate whether participants expect algorithms to perform successfully (e.g., Dietvorst et al. 2015; Nourani et al. 2021).

Second, risk and failure-related perceptions address concerns about possible downsides of algorithmic reliance. These include perceptions of fairness, transparency, risk, complexity, or error-proneness (e.g., Cheng and Chouldechova 2023; Skitka et al. 1999; Wu et al. 2024).

Third, identity and social relevance capture whether interaction with algorithmic systems threatens or reinforces social identity. Measures include identity threat, embarrassment, exploitation, potential loss of standing, uniqueness neglect, collectivist orientation, locus of control, or even xenophobic tendencies (e.g., Jain et al. 2025; Talebi et al. 2025).

Fourth, emotional and affective reactions comprise reported feelings such as anger, pride, guilt, frustration, or positive states like pride and hedonism. They also include affective (dis)trust or discomfort in response to algorithmic advice (e.g., Dennis et al. 2023; Prahl and Van Swol 2017).

Fifth, evaluative intentions and preferences encompass general stances such as preferences for human versus algorithmic advice, willingness to work with algorithms, continuance intentions, purchase or donation intentions, and perceptions of brand-related authenticity or loyalty (e.g., Baek et al. 2024; Brüns and Meißner 2024). In addition, some studies capture evaluations of algorithmically generated artifacts, such as judgments of news quality and authenticity, perceived effort, or impressiveness of an art piece (Reich et al. 2023; Rix et al. 2025). These measures extend evaluative preferences beyond decision advice to include algorithmic outputs in domains such as digital news or creative content.

Finally, personality and individual differences capture traits, including need for cognition, cognitive reflection, competitiveness, or neuroticism (e.g., Commerford et al. 2024; Gill et al. 2024; Jain et al. 2025; You et al. 2022).

Category	*Measured Constructs*
Performance-related expectations	Trust, confidence, competence, effectiveness, familiarity, perceived learning capabilities, utility, self-efficacy
Risk and failure-related perceptions	Fairness, transparency, perceived risk, equity, complexity, average error, consequentialness, difficulty, judgment superiority, risk-aversion
Identity and social relevance	Identity threat, embarrassment, exploitation, uniqueness neglect, collectivism, locus of control, xenophobia, social power
Emotional and affective reactions	Emotional reactions (anger, guilt, pride, shame, etc.), affective trust/distrust, hedonism, discomfort, passion, similarity
Evaluative intentions and preferences	Preference for human vs. algorithm, willingness to use, continuance intention, purchase/donation intention, brand loyalty/authenticity, EWOM
Personality and individual differences	Need for cognition, cognitive reflection, competitiveness, neuroticism, maximizing mindset, high standards

Table 6: Categories of Subjective Measurement

Taken together, subjective measures provide insight into the cognitive and affective orientations that underpin behavioral reliance. Subjective measures thus highlight how individuals think and feel about algorithms, but they do not show how these orientations translate into behavior or performance. To address this gap, studies commonly employ objective measures of reliance.

3.3.2 Objective Measures

Objective measures capture reliance on algorithmic systems through observable behavior and its consequences, providing a complement to self-reported attitudes. They can be grouped into five broad

categories as shown in Table 7. One of the most common operationalizations is the judge–advisor system, in which participants first provide an independent estimate, then receive advice from a human or an algorithmic source and finally revise their estimate. Reliance is quantified by the extent to which participants move toward the advice. The standard metric is Weight of Advice (WOA) (e.g., Castelo et al. 2019; Logg et al. 2019; Sachin and Schecter 2024), while variations such as SHIFT (Prahl and Van Swol 2017), MSHIFT (Tse et al. 2024), and Average Absolute Deviation (AAD) (Cheng and Chouldechova 2023) capture related forms of adjustment, such as absolute shifts, sequential changes, or average deviations from optimal benchmarks. Together, these measures allow researchers to quantify reliance in a continuous manner and to compare responses to human and algorithmic advisors.

Reliance is also operationalized through adherence measures, which capture the proportion of times algorithmic advice is followed across repeated trials (e.g., Jenkin et al. 2024; Schaffer et al. 2019; Xu and Wang 2024). Such measures are particularly useful for detecting calibration effects, for instance, reductions in adherence after observed errors. In addition, studies often employ task-specific reliance indicators embedded in applied contexts. Examples include following algorithmic recommendations in donation tasks (Talebi et al. 2025), investment decisions (Germann and Merkle 2023), adjusting safety stocks in inventory management (Wang et al. 2024b), or clicking on algorithmic suggestions in digital environments (Keppeler 2024). These approaches situate reliance in realistic settings and move beyond abstract decision paradigms.

Other studies assess reliance through its consequences for accuracy and performance. Here, outcomes are benchmarked against optimal or correct standards, such as forecasting accuracy (Dietvorst et al. 2015; Dietvorst et al. 2018), diagnostic accuracy (Cabitza et al. 2025), or prediction error (Cheng and Chouldechova 2023). Economic consequences, such as investment performance (Germann and Merkle 2023), profit deviation (Feng and Gao 2020), or inventory waste (Wang et al. 2024b), are also treated as outcome measures, depending on the context of the study. These indicators are particularly important for evaluating whether reliance on algorithms improves decision quality, thereby testing the conditional assumptions underlying the constructs of Algorithm Aversion and Automation Bias.

A further category of operationalizations captures process efficiency and cognitive load. Such measures focus on the ease or difficulty of interacting with algorithmic systems, including task completion times, error rates, and helpfulness ratings (Nourani et al. 2021), adjusted response times as indicators of cognitive load (You et al. 2022), or ERP measures to detect neural conflict during advice integration (Xie et al. 2022). Such measures highlight that reliance involves not only outcomes but also the cognitive effort required to engage with algorithmic systems.

Finally, several operationalizations emphasize safety and domain-specific risks, particularly in Automation Bias research. Reliance is measured through omission errors, where participants fail to act when automation provides no signal, or commission errors, where they act incorrectly due to faulty automation cues (Skitka et al. 1999). Other indicators include verification checks or tracking performance in dynamic tasks such as driving or aviation (Shariff et al. 2021; Skitka et al. 1999). These measures illustrate that reliance on algorithms can carry risks that extend beyond efficiency or accuracy, especially in applied domains.

Category	***Measured Constructs***
Choice and adjustment paradigms	Binary choice (self vs. algorithm, human vs. algorithm); Weight of Advice (WOA); SHIFT; MSHIFT; AAD; limited adjustment; influence factor; switching rate
Adherence and task-specific reliance	Advice adherence rates; behavioral trust after error; donation behavior; stock adjustments; job responses; clicks on recommendations

Accuracy and performance outcomes	Forecasting accuracy; diagnostic accuracy; prediction error (AAE); situation awareness; investment performance
Process efficiency and cognitive load	Task time; error rates; helpfulness; performance scores/rankings; cognitive load (response time); ERP conflict measures; AI knowledge/familiarity
Safety and domain-specific risks	Omission/commission errors; minimum safety thresholds; tracking performance; perceptions of digital artifact quality or effort

Table 7: Categories of Objective Measurement

Objective measures capture how reliance manifests in practice and what consequences it carries for accuracy, efficiency, and safety. Taken together with subjective measures, they offer a more comprehensive picture of Algorithm Aversion, Algorithm Appreciation, and Automation Bias.

3.4 Emerging Themes in Current Research

Beyond definitional clarity, conceptual relationships, and patterns of operationalization, my analysis reveals a set of recurring themes that describe how individuals interact with algorithmic systems and why Algorithm Aversion, Algorithm Appreciation, or Automation Bias emerge in practice. Across the coded literature, I identified four broad categories: (1) behavior, (2) consequences, (3) evaluation, and (4) mental models. These themes represent different analytical levels that together form a dynamic cycle. User behavior leads to consequences, which in turn shape how systems are evaluated. Evaluations form mental models that structure future expectations and ultimately determine subsequent behavior. This cyclical framework highlights that Algorithm Aversion, Algorithm Appreciation, and Automation Bias are not isolated reactions but situated processes shaped by interdependent cognitive, affective, and contextual mechanisms. Figure 4 illustrates the cyclical structure of emerging themes identified in the literature. This cycle captures the recursive nature of human–algorithm interaction and provides a lens through which to interpret the fragmented empirical findings.

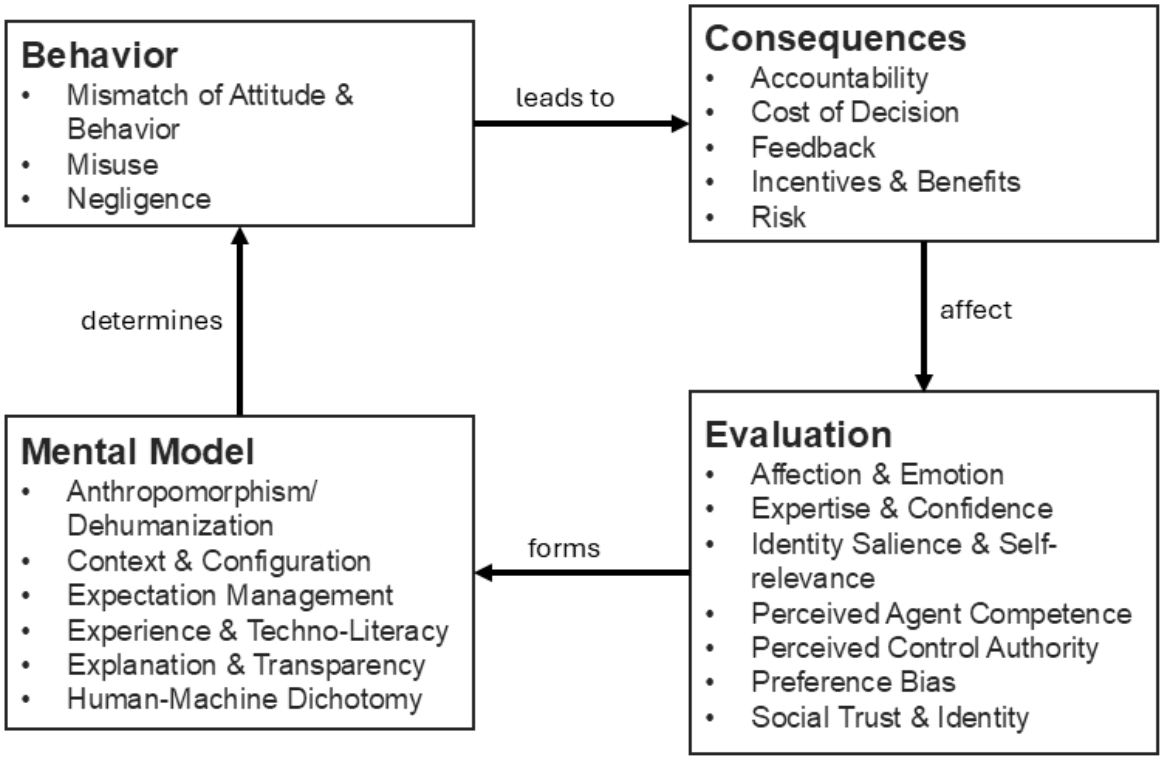

Figure 4: Framework of Emerging Themes in Current Research

The following sections unpack each theme in turn, outlining how they have been conceptualized in current research and how they contribute to the dynamics of Algorithm Aversion, Algorithm Appreciation, and Automation Bias.

3.4.1 Behavior

Research highlights that observable behavior in human-algorithm interaction often diverges from reported attitudes, creating mismatches that complicate interpretation. For instance, individuals may express distrust in algorithms yet still follow their recommendations or conversely report high trust while failing to act on algorithmic advice (Castelo et al. 2019). Beyond these mismatches, behavioral patterns reveal different forms of misuse. Some studies point to excessive reliance on weak algorithms, where individuals continue to follow recommendations despite evidence of poor performance (Tse et al. 2024). Others show the opposite tendency of systematic under-reliance, especially among experts who discount algorithmic input even when doing so reduces accuracy (Logg et al. 2019). Finally, negligence emerges when people disregard reliable cues in favor of flawed automation or persist with low-performing systems despite recognizing their shortcomings (Skitka et al. 1999). Together, these behaviors illustrate that algorithm use is rarely a straightforward matter of rational reliance but is shaped by complex tensions between attitudes, judgments of system reliability, and the situational framing of choice.

3.4.2 Consequences

The behavioral reliance on algorithms has downstream effects that shape both individual outcomes and patterns of trust. A central theme is accountability, which has been shown to reduce Automation Bias by reducing errors when participants are held responsible for their performance or accuracy. Accountability shapes verification behaviors, increases vigilance, and can recalibrate reliance, though

its effects depend on whether accountability is tied to relevant aspects of task performance (Skitka et al. 1999).

Another recurring consequence concerns the cost of decision-making. People weigh the economic or cognitive effort associated with algorithmic reliance, and costs can shift reliance patterns. For instance, users tend to prefer algorithms when human advisors become costly or inefficient, even if initial Algorithm Aversion persists (Germann and Merkle 2023). Similarly, effectiveness beliefs often outweigh discomfort, with individuals tolerating unease if algorithmic advice clearly improves outcomes (Castelo et al. 2019). This cost-to-benefit evaluation highlights that reliance is not only an attitudinal stance but also a pragmatic trade-off shaped by perceived efficiency (Mahmud et al. 2022).

Feedback and interpretation mechanisms play a critical role in sustaining or eroding algorithm reliance. A lack of feedback limits accurate evaluations of algorithm performance, leading to equal trust in both high- and low-performing systems (Tse et al. 2024). Early impressions strongly shape mental models. Exposure to strengths first can inflate confidence and foster Automation Bias, while early exposure to weaknesses can reduce reliance but also underestimate system competence (Nourani et al. 2021). Over time, repeated successes reduce initial resistance (Germann and Merkle 2023), while observed errors recalibrate beliefs (Jussupow et al. 2024). Yet feedback can also bias perception. Primacy effects make early errors disproportionately damaging, and negative feedback generally weighs more heavily than positive outcomes, lowering algorithm use (Dietvorst et al. 2015; Mahmud et al. 2022). These findings suggest that reliance is not static but evolves iteratively as feedback loops reinforce or weaken trust.

A fourth theme concerns incentives and benefits. Algorithm reliance increases when users see clear returns or social value in adoption, whether through higher accuracy, economic efficiency, or fairness (Wu et al. 2024). Incentives can motivate adoption by boosting confidence in algorithmic competence (Mahmud et al. 2022). Perceptions of societal benefits, such as corruption reduction or impartiality, further reduce Algorithm Aversion and encourage reliance (Castelo 2024). Thus, incentives act as a counterweight to discomfort or uncertainty, shifting attention toward perceived utility.

Finally, reliance is shaped by risk assessments. Trust in algorithms is lower for high-stakes tasks where errors carry greater costs (Castelo et al. 2019), and higher decision stakes also reduce perceived fairness of algorithmic models (Cheng and Chouldechova 2023). Importantly, trust and distrust are not mirror opposites. Distrust can exert stronger direct effects by deterring algorithm use in risky contexts (Wu et al. 2024). The framing of outcomes further modulates reliance, with gain- and loss-dynamics influencing whether individuals view algorithmic reliance as an acceptable risk (Mahmud et al. 2022).

Taken together, these findings show that consequences influence the feedback loop between behavior and evaluation. Accountability, cost, feedback, incentives, and risk are integral mechanisms that shape how reliance is calibrated across situations.

3.4.1 Evaluation

Evaluations of algorithms represent the interpretive layer through which behavior and consequences are filtered. They reveal how individuals cognitively and affectively position algorithmic agents in relation to themselves and to human alternatives. These evaluations are multidimensional, spanning emotions, cognitive capacities, expertise, identity, competence, control, personality, and social trust.

A first theme concerns affection and emotional responses. Emotions are not always primary drivers of reliance. Some studies show little difference between algorithmic and human advisors in terms of positive or negative affect (Downen et al. 2024; Prahl and Van Swol 2017). Yet negative emotions such as discomfort, distrust, or perceived coldness can impair decision outcomes and reinforce Algorithm Aversion, particularly in contexts where algorithms are perceived as cost-cutting substitutes for human

effort (Brüns and Meißner 2024). Conversely, positive affect such as hedonism or enjoyment strengthens intention to use algorithms (Chávez et al. 2024), underscoring that emotional framing can either amplify or attenuate reliance.

The second theme involves Experience and confidence. Perceptions of one's own cognitive ability shape algorithm reliance. Experience and perceived self-competence often reduce algorithm reliance, even when objective performance is inferior (Tse et al. 2024). Overconfidence can bias users toward human judgment (Burton et al. 2020), while low numeracy correlates with reduced openness to algorithmic advice (Logg et al. 2019). Paradoxically, laypeople sometimes rely more strongly on algorithms than experts, whose domain confidence makes them less receptive to algorithmic advice (Logg et al. 2019; Rebholz et al. 2024). AI systems can empower low-ability users by providing externalized cognitive resources (Schaffer et al. 2019), but cognitive load can constrain novices in complex tasks, limiting the benefits of algorithmic support (Cabitza et al. 2025). Explanations and feedback sometimes fail to override these biases, particularly when first impressions are strong (Nourani et al. 2021). The result is a paradoxical pattern, that those who might benefit most from algorithmic support are often least inclined to use it.

Another thematic cluster relates to identity salience and self-relevance. Resistance is heightened when algorithms are perceived as threatening self-concept, professional roles, or moral domains (Castelo 2024; Dietvorst et al. 2015; Mahmud et al. 2022). Job candidates judged by AI often report lower self-esteem and reduced willingness to engage (Keppeler 2024), while disclosures of algorithmic use reduce the appeal of offers and products (Brüns and Meißner 2024). Conversely, identity framings such as green identity can increase algorithmic adoption when aligned with personal values (Li et al. 2025). These dynamics suggest that identity relevance often surpasses technical performance in shaping evaluation.

Perceptions of agent competence also play a central role. Algorithms are often doubted in tasks requiring empathy, intuition, or personalization, while humans are credited with "human nature" capacities that machines are presumed to lack (Brüns and Meißner 2024; Castelo et al. 2019). At the same time, labeling algorithms as capable of learning can increase adoption, even if their absolute performance remains lower (Chacon et al. 2025; Reich et al. 2023). These findings highlight that competence judgments are deeply tied to inferred mental traits and assumed capabilities rather than objective accuracy alone.

A further determinant is perceived control authority. Providing users with the ability to adjust, oversee, or partially modify algorithmic outputs consistently reduces Algorithm Aversion and increases satisfaction (Commerford et al. 2024; Dietvorst et al. 2018). Even minimal opportunities for influence can sustain trust and encourage adoption, while fully autonomous systems increase Algorithm Aversion (Chávez et al. 2024). Judicial AI systems that frame humans as the ultimate authority exemplify this by maintaining user agency while still improving decision quality (Cabitza et al. 2025).

Finally, evaluations are shaped by individual differences and social trust. Personality traits such as neuroticism, extroversion, maximizing mindset, or curiosity systematically moderate reliance (Mahmud et al. 2022; Rix et al. 2025; Silber et al. 2025). Preference biases such as general distrust or implicit anti-algorithmic bias can persist even in the face of objective benefits (Turel and Kalhan 2023). Social trust and identity further structure evaluation, with individuals drawing on cues from peers, institutions, and cultural contexts to decide whether to engage with algorithms (Castelo 2024; Wu et al. 2024). Importantly, trust and distrust are not opposites. Distrust can exert a stronger deterrent effect on algorithm use than low trust alone (Wu et al. 2024).

Taken together, evaluation processes reveal that reliance is not merely a function of rational cost–benefit assessments but is deeply embedded in cognitive biases, identity dynamics, and trust relationships. These evaluations form the mental scaffolding through which users interpret algorithmic behavior, creating the conditions under which reliance either solidifies or erodes.

3.4.2 Mental Model

Mental models capture the interpretive frameworks that shape how individuals understand, anticipate, and respond to algorithmic systems. They are not static but formed and reshaped through experience, framing, and contextual cues. The literature points to six recurring dimensions: anthropomorphism and dehumanization, context and configuration, expectation management, experience and techno-literacy, explanation and transparency, and the human-machine dichotomy.

The first dimension is anthropomorphism and dehumanization. Human-likeness in form or interface can buffer negative reactions and increase acceptance, especially for subjective tasks or when disclosure might trigger skepticism (Baek et al. 2024; Castelo et al. 2019). Physical presence and interactivity, such as embodied robots or avatars, foster trust and willingness to collaborate (Jussupow et al. 2024). Competent-looking avatars can even preserve perceptions of system competence after errors, mitigating Algorithm Aversion (Ganbold et al. 2022). Conversely, machine-like framings or lack of humanness intensify skepticism and reduce credibility (Baek et al. 2024).

Second, context and configuration strongly moderate algorithm acceptance. Trust and reliance are higher in objective domains and performance-oriented tasks but diminish for subjective, hedonic, or moral decisions (Castelo et al. 2019; Reich et al. 2023). For example, AI is more accepted for search products than for experience products, where consumers perceive greater conflict (Xie et al. 2022). Similarly, investment tasks provoke more skepticism than trading, given their higher stakes and deliberative character (Koo 2024). Organizational and cultural contexts also matter. In corrupt environments, algorithms are valued more highly for resource allocation, reflecting expectations of impartiality (Castelo 2024). These findings highlight that algorithm use cannot be separated from the task and social setting in which it is embedded.

Third, expectation management shapes mental models even before interaction. Framing algorithms as “AI” rather than “algorithms” increases reliance (Hou and Jung 2021), while highlighting their capacity to learn reduces Algorithm Aversion (Chacon et al. 2025; Reich et al. 2023). Conversely, AI disclosure in advertising contexts often undermines credibility, authenticity, and generosity of donations, fueling skepticism (Baek et al. 2024; Brüns and Meißner 2024). More generally, emphasizing heterogeneity among algorithms or human involvement in AI development can restore trust and soften Algorithm Aversion (Castelo 2024; Koo 2024). These framings demonstrate that expectations are highly malleable and play a decisive role in whether Algorithm Appreciation or Aversion emerges.

Fourth, experience with a system and general techno-literacy reinforce or weaken Algorithm Aversion over time. Familiarity with algorithmic systems correlates with increased trust in algorithmic advice and reduced Algorithm Aversion (Castelo et al. 2019), while lack of understanding exacerbates skepticism (Mahmud et al. 2022). Seeing an algorithm err can reduce reliance initially, but repeated exposure allows for more nuanced mental models that accommodate imperfection (Burton et al. 2020). Importantly, Algorithm Aversion is not static but shifts across time and with repeated experience (Jussupow et al. 2024).

Fifth, explanation and transparency interventions have mixed effects. Rich explanations can increase perceived understanding and trust in recommender systems (Cabitza et al. 2023), but sometimes they create an illusion of understanding without substantive knowledge gains (Cabitza et al. 2023; Nourani

et al. 2021). Simple, performance-focused explanations reduce Algorithm Aversion (Castelo 2024), and persuasive communication styles (e.g., personalization, warmth) enhance the credibility of explanations (Mahmud et al. 2022). Transparency about errors or prediction accuracy can paradoxically reduce Algorithm Appreciation, while learning labels and performance transparency tend to increase trust and reliance on advice (Chacon et al. 2025; Reich et al. 2023). Thus, transparency is a double-edged sword. It can empower novices but also foster overconfidence or misplaced reliance.

Finally, the human-machine dichotomy frames how individuals reason about algorithmic competence. Humans are associated with intuition, flexibility, and moral judgment, while algorithms are seen as predefined, rigid, and reductionist (Burton et al. 2020; Jussupow et al. 2024; Mahmud et al. 2022). People often expect algorithms to be perfect (Mahmud et al. 2022), making their errors less forgivable than human mistakes (Dietvorst et al. 2015). This “perfection schema” creates asymmetries. Users abandon algorithms more quickly than humans after errors, even when human errors are larger (Dietvorst et al. 2015; Prahl and Van Swol 2017). At the same time, reliance on algorithms increases to a greater degree compared to humans when their learning ability is made salient (Reich et al. 2023).

Taken together, mental models reveal that user responses to algorithms are grounded less in raw performance and more in how systems are framed, experienced, and contextualized. Anthropomorphic cues, domain fit, expectations, familiarity, and transparency all shape the interpretive lenses through which algorithms are judged. The human-machine dichotomy persists as a powerful organizing principle, often privileging human flexibility and emotional capacity over algorithmic consistency. These mental models, in turn, determine behavior by setting the stage for either trust, skepticism, or overreliance.

4 Discussion

This review set out to clarify and integrate the fragmented body of research on Algorithm Aversion, Algorithm Appreciation, and Automation Bias. By systematically decomposing definitions, comparing constructs, analyzing operationalizations, and inductively synthesizing themes, the study makes two contributions. First, it demonstrates that Algorithm Aversion and Appreciation, while often treated as opposites on a spectrum, differ in their conditionality. Further comparison of Algorithm Aversion and Automation Bias revealed that both are reactive constructs bound to expectation violations, whereas Algorithm Appreciation functions as a default positive orientation. Second, the thematic synthesis reveals that reliance on algorithms unfolds as a cyclical process. Behavior produces consequences, which shape evaluations, which in turn form mental models guiding future behavior. This perspective provides a higher-order structure for reconciling inconsistent empirical findings.

These findings reposition Algorithm Aversion research within the broader tradition of judgment and decision-making studies, showing that reliance on algorithmic systems cannot be reduced to a single attitudinal or behavioral disposition. Instead, reliance reflects an interplay of situational cues, identity concerns, and evolving mental models. In doing so, the framework addresses the conceptual ambiguity that has characterized prior research and highlights the need for integrated theorizing across domains.

5 Avenues for Future Research

The findings of my thesis not only clarify existing constructs but also reveal areas where further inquiry is needed. In particular, two avenues for future research stand out. The first concerns the updating of construct boundaries and targets to ensure their continued relevance for contemporary algorithmic

systems. The second relates to clarifying the need for conceptual symmetry among Algorithm Aversion, Algorithm Appreciation, and Automation Bias. Together, these directions provide a foundation for refining theory and guiding more cumulative research on human responses to algorithms.

5.1 Updating Construct Boundaries and Targets

A first avenue for future research lies in refining how Algorithm Aversion, Algorithm Appreciation, and Automation Bias are conceptually defined. Much of the existing literature continues to frame “algorithms” as deterministic, rule-based procedures and situates the constructs within advice-taking tasks (e.g., Castelo 2024). This narrow conceptualization no longer reflects contemporary realities, in which algorithms increasingly take the form of adaptive, probabilistic, or generative systems (e.g., Dietvorst et al. 2015; Jussupow et al. 2024; Mahmud et al. 2022). Similarly, restricting the target to advice ignores how individuals now evaluate a much broader range of algorithmic outputs, such as creative artifacts or news content (Reich et al. 2023; Rix et al. 2025). If definitional formulations remain narrow, the phenomena risk being restricted to a subset of contexts and underestimating their relevance for contemporary AI systems. Future work should therefore adopt a more inclusive target formulation that captures both deterministic and adaptive algorithms as well as outputs beyond advice. Doing so would not only improve the conceptual coherence of the constructs but also ensure their ongoing relevance as algorithmic systems become ever more integrated into organizational and societal practices.

5.2 Clarifying the need for conceptual symmetry

A recurring challenge in existing literature is the need for conceptual symmetry among Algorithm Aversion and Algorithm Appreciation. Many studies frame the constructs as conceptual opposites (e.g., Cheng and Chouldechova 2023; Turel and Kalhan 2023). This pairing suggests a symmetrical relationship, with Algorithm Appreciation as the “anti-aversion.” My analysis indicates that this symmetry is overstated and conceptually misleading. Algorithm Aversion is anchored in a normative behavioral benchmark derived from expected utility. When algorithms are demonstrably superior, rational actors should choose them, and failure to do so constitutes Aversion (Dietvorst et al. 2015; Prahl and Van Swol 2017). By contrast, Algorithm Appreciation lacks such a normative anchor. It is generally defined as a positive evaluative stance toward algorithms. It may simply reflect rational adaptation to situational assumptions (e.g., expectations of fairness, utility, or competence) rather than a systematic bias (e.g., Cabitza et al. 2025; Ganbold et al. 2022; Talebi et al. 2025). Future research should build on the distinctions outlined in this thesis to further consolidate the conceptual landscape of Algorithm Aversion, Algorithm Appreciation, and Automation Bias. While this study has clarified where the constructs overlap, diverge, and how they can be positioned relative to one another, more work is needed to test these distinctions empirically, refine their theoretical boundaries, and examine their implications across different contexts. Such efforts can help to reduce conceptual friction and work towards a more.

6 Limitations

As with all research, this thesis has certain limitations. First, while it systematically reviews definitions, operationalizations, and empirical findings, it does not test them empirically. The arguments therefore rest on interpretive synthesis rather than experimental validation. Second, the analysis is constrained by the available sample of literature. Although the review draws on a broad and interdisciplinary corpus, it inevitably underrepresents studies that did not meet the chosen quality and inclusion criteria. Third, the definitions and categorizations developed in this thesis reflect trade-offs between inclusivity and

parsimony. Alternative formulations are possible, and different coding decisions could emphasize other aspects of the phenomena. Finally, while the semantic decomposition offers conceptual clarity, it cannot fully resolve the evolving meaning of “algorithmic systems,” which continues to expand with technological advances. These limitations underscore the need for ongoing conceptual refinement and empirical validation to advance cumulative theorizing in this domain.

7 Conclusion

Algorithm Aversion, Algorithm Appreciation, and Automation Bias constitute a fragmented yet rapidly growing body of research that mirrors the increasing role of algorithmic and AI-based systems in organizational and societal contexts. This review has synthesized insights from a broad interdisciplinary corpus and positioned these constructs within a shared evaluative–behavioral–consequential cycle. By clarifying conceptual boundaries, highlighting the structural symmetry between Algorithm Aversion and Automation Bias, and questioning whether Algorithm Appreciation warrants recognition as a standalone phenomenon, the thesis advances conceptual clarity in a field marked by inconsistent definitions and operationalizations. Beyond theoretical integration, the findings underscore the practical significance of Algorithm Aversion and Automation Bias, both of which can undermine the effective and safe use of algorithmic systems. Taken together, this work contributes to Information Systems research by offering a clearer conceptual foundation for studying human–algorithm interaction and by providing a basis for more cumulative, practice-relevant theorizing as the role of algorithmic systems continues to expand.

Appendix A – Declaration on the Use of GenAI tools

In the preparation of this paper, I have used following tools based on generative artificial intelligence (GenAI):

1. ChatGPT
2. Adobe AI Assistant

I further declare that:

- I have labeled the content taken from the GenAI tools listed above with my details in Table A. 1,
- I have verified that the content generated by the above-mentioned GenAI tools and adapted by me is factually correct,
- I am aware that, as the author of this work, I am responsible for the information and the statements made in it, and
- I am aware that violating the disclosure of the use of generative AI in my work is a deception and leads to an evaluation with an insufficient grade.

I have used the above-mentioned AI systems as indicated below.

Table A. 1: Overview of GenAI Tools across Areas of Contribution

Areas of contribution	AI tool(s) used	Description of the manner of use and compliance with good scientific practice
Development and conception of the research project	-	-
Identification of literature	-	-
Synthesizing of literature	2	-
Structuring the text	1	All sections
Formulation of text	1	All sections
Revision of text	-	-
Creation of visualizations	-	-
Further contributions	-	-

Appendix B – Methods

This thesis adopts a grounded theory approach to literature review, following the methodology outlined by (Wolfswinkel et al. 2013). The review process is structured into five main stages, further divided into eleven specific tasks that systematically guide the review from initiation to completion (see Table B. 1). The following section outlines how each of these tasks was applied.

Table B. 1. Grounded Theory for Literature Reviews (Wolfswinkel et al. 2013)

Step	Task(s)
Define the scope and criteria for the review	• Define the criteria for Inclusion/exclusion • Identify the fields of research • Determine the appropriate sources • Decide on the specific search terms
Search for relevant literature	• Conduct the actual search
Select sources by refining, filtering, and applying criteria to produce the final sample	• Refine the sample
Analyze the sample by applying principles of Grounded Theory	• Open coding to identify concepts • Axial coding to establish relationships • Selective coding to integrate and refine categories
Present by structuring the findings and insights gathered during the analysis	• Represent and structure the content • Structure the article

Define: Foundation for the Review

The first step consisted of defining the scope of the review, in line with established literature review practices (Templier and Paré 2015). This involved setting clear inclusion and exclusion criteria, identifying relevant research domains, selecting appropriate sources, and outlining procedures for locating and retrieving literature. To initiate this process, I performed exploratory searches in online databases (ACM Digital Library, AIS Electronic Library, APA PsycNet) using the keyword “algorithm aversion” to identify peer-reviewed journal articles and conference proceedings.

This preliminary search helped establish an overview of the topic’s coverage across several disciplines, including psychology, information systems (IS), management and organizational behavior, human-computer interaction (HCI), marketing and consumer behavior, medicine and health informatics, as well as business and finance. The results revealed that algorithm aversion is a widely discussed subject across a variety of domains. To maintain conceptual coherence and ensure a manageable scope, I chose to focus on literature within the IS field.

Accordingly, I selected six databases as my primary sources: ACM Digital Library, AIS Electronic Library, APA PsycNet, IEEE, JSTOR, and EBSCO. ACM Digital Library offers extensive access to publications in computer science and HCI, particularly relevant for research on algorithm design, transparency, and user interaction. AIS Electronic Library includes key IS journals and conference proceedings, supporting a socio-technical lens on algorithmic decision-making. APA PsycNet was included due to its comprehensive coverage of psychological research on trust, identity threat, heuristics, and user attitudes toward algorithms, and as the outlet where the foundational study by Dietvorst et al. (2015) was published. EBSCO provided access to business, management, and consumer behavior research not fully covered by IS or psychology databases. IEEE, with its focus on

engineering and applied technical domains, was selected for its relevance to algorithmic systems in contexts such as healthcare and diagnostics. JSTOR was added to capture literature from adjacent disciplines given its broad, interdisciplinary coverage.

With the source landscape defined, I then set a quality threshold and formulated specific inclusion and exclusion criteria to ensure the final sample would meet both conceptual relevance and methodological rigor.

Quality threshold: To evaluate the quality of journal articles, I relied on the 71st edition of Harzing's Journal Quality List (Harzing 2024), a consolidated meta-ranking that synthesizes insights from eleven internationally recognized journal ranking systems, including the ABDC 2022, the EJL 2024, and META 2023. This enabled the identification of consistently high-performing journals across disciplines relevant to the review. For conference proceedings, I applied the classification system maintained by the International Computing Research and Education Association of Australasia (CORE 2023), whose ranking database is jointly governed by an international committee of representatives of CORE, GII, GRIN, and SCIE. Although Information Systems conferences were removed from the ICORE ranking after 2020, earlier versions were retained to ensure coverage of IS-specific venues. Conferences rated A or higher in any available annual ICORE evaluation were considered acceptable.

Inclusion Criteria: Publications were included if they were peer-reviewed, either empirical or conceptual in nature, and addressed algorithm aversion, algorithm appreciation, or automation bias. In addition, included studies needed to be situated within decision-making contexts or human–computer interaction settings and had to present methodologically interpretable research designs, settings, and claims.

Exclusion Criteria: Studies that lacked a behavioral, cognitive, or motivational focus were excluded from the review. Likewise, technical or implementation-focused publications were omitted unless they were explicitly engaged with user acceptance or evaluation. Finally, studies concerned solely with algorithmic bias or fairness, where no human judgment or interaction was involved, were excluded, as were working papers and non-peer-reviewed publications lacking empirical evidence or methodological transparency.

I defined my search queries using the terms "**algorithm aversion**" and "**algorithm appreciation**," in either the title, abstract, or keywords. During my initial review of the literature, I encountered the work of Horowitz and Kahn (2023) who argue that automation bias represents a conceptual counterpart to algorithm aversion. Based on this insight, I decided to include "**automation bias**" as an additional search term and to extend the focus of my research.

Search: Relevant Literature

Each search result was assigned a unique identifier to maintain data integrity throughout the subsequent data collection process. For each selected database, I executed the defined search queries, with the final search iteration completed on July 2, 2025. Where possible, I applied the inclusion and exclusion criteria at this stage, i.e., filtering for peer-reviewed publications and specific article types. Search results were exported in the most convenient format available from each database (e.g., CSV, BibTeX, EndNote), and included titles, abstracts, and all available metadata. These results were then imported into Excel and consolidated into a single spreadsheet to facilitate further analysis.

Select: Finalize Sample

The initial search yielded a sample of 167 works. After removing duplicates and applying the defined inclusion and exclusion criteria, the sample was reduced to 50 publications. I then extracted key citation information from each paper to conduct a preliminary backward and forward search, which resulted in the addition of 5 further sources. The final sample thus consists of 55 works.

Analyze: Gain Insights from Sources

In the fourth step, I applied techniques from grounded theory to develop a deeper understanding of the literature corpus. Following Wolfswinkel et al. (2013), this process relied on three phases: open coding, axial coding, and selective coding. To structure the analysis, I randomly selected batches of 15 sources from the sample and coded them iteratively, consistent with grounded theory's principle of gradual discovery.

I conducted open coding by annotating each source, noting conceptual insights, tensions, and relevant elements. Excerpts and references were systematically transferred into an Excel spreadsheet, enabling me to preserve the contributions of individual studies while maintaining an overarching view of emerging concepts.

I then conducted axial coding to refine the scheme and group first-order categories into higher-level categories. This process was iterative, involving continuous comparison between raw data, open codes, and emerging categories. To ensure transparency, I transferred my coding into the Excel spreadsheet. At the end of axial coding, I retained 31 higher-order categories that provide the basis for subsequent analysis.

Finally, I applied selective coding to integrate the categories developed through axial coding into a coherent framework. This step involved identifying core categories that captured the essence of the phenomena under study and systematically relating all other categories to these. Sequential coding thus provided the bridge from descriptive fragmentation to theoretical integration. It allowed the concepts of Algorithm Aversion, Algorithm Appreciation, and Automation Bias to be positioned within a shared evaluative, behavioral, and consequential cycle. In this way, sequential coding served not only to reduce complexity but also to highlight the relational structure among constructs, ensuring that the emerging framework remained grounded in the reviewed literature.

Appendix C – Semantic Decomposition

To develop definitions for *Algorithm Aversion*, *Algorithm Appreciation*, and *Automation Bias*, I applied semantic decomposition (Akmajian et al. 2017), a method that breaks phrases or words into fundamental semantic units. This approach, illustrated in IS research by Hund et al. (2021) and (Vial 2019), was used here to segment existing definitions into their grammatical components (e.g., verbs, nouns, adjectives) and assign each segment a semantic role (e.g., actor, conduct, condition).

The goal of this process was to identify a consistent set of semantic primitives applicable across all three phenomena, enabling systematic comparison and revealing conceptual connections. Guided by the principles of *construct clarity, distinctiveness, and parsimony* (Suddaby 2010), I arrived at six primitives: Actor, Conduct, Target, Comparator, Condition, and Scope. Conduct

I followed five steps for each phenomenon:

1. Collected and organized all formal definitions
2. Decomposed them into the six primitives
3. Summarized patterns and notable variations
4. Chose the wording for each primitive based on majority trends and inclusivity
5. Synthesized these choices into a constructed definition

The following sections present the findings for each phenomenon in the order of the six primitives, along with the final constructed definition. Detailed decomposition tables follow each section.

Algorithm Aversion

Across the sample, I identified 44 formal definitions of Algorithm Aversion in 39 distinct papers. Table C. 1 lists all excerpts by year (most recent first) and then alphabetically by author(s). Five articles include two formal definitions each. Castelo (2024) contrasts Dietvorst et al. (2015) with Morewedge (2022) to highlight different views on behavioral directionality. Bankuoru Egala and Liang (2024) cite Dietvorst et al. (2015) and later narrow the definition to a medical context, drawing on Longoni et al. (2022). Jain et al. (2025) offer a definition that frames Algorithm Aversion as a general preference for human judgment, followed by a reference to Mahmud et al. (2022). Turel and Kalhan (2023) first reference Burton et al. (2020) and then propose their own, broader definition. Talebi et al. (2025) cite Dietvorst et al. (2015) but distinguish between a positive attitude toward humans and a negative behavior toward algorithms. Most authors point to Dietvorst et al. (2015) as the original source (27 instances), even though their final definitions differ substantially. A cross-article comparison reveals similar ambiguities, underscoring the

fragmented conceptualization of Algorithm Aversion in the current literature. I synthesized the cross-source evidence into a single, generalizable definition. In the following section I summarize the rationale and final choice for each primitive and present the constructed definition for AIgorithm Aversion:

1. The **actor** is most often described generically as *people* (13) or *human/humans* (5). Some definitions specify a contextual role such as *human decision-makers* (4), *human forecasters* (Cabiddu et al. 2022), or *users* (3). A domain-specific exception is *clinicians* (Bankuoru Egala and Liang 2024). In several cases, the actor is implicit and can be inferred from the input target, *advice* (Commerford et al. 2022; Ganbold et al. 2022), *decisions* (Jain et al. 2025), *recommendations (Talebi et al. 2025)*, or *predictions* (Reich et al. 2023), which presuppose either a *human* actor (9) or specifically a *decision-maker* (8). **Rationale:** Most sources keep the actor general; over-specifying (e.g., clinicians) narrows applicability. Many definitions even leave the actor implicit. **Choice:** *people*
2. Definitions vary by valence (positive vs. negative) and by the nature of **conduct** (attitude; behavior; calibration). Valence depends on direction (toward an algorithm vs. toward a human). Some emphasize observable behavior (Cabitza et al. 2023), others underlying attitude (Baek et al. 2024), asymmetry of change in attitude or behavior, which I labelled *calibration* (Dietvorst et al. 2015), or combinations (Talebi et al. 2025). Most frame conduct negatively toward algorithms (36), some positively toward humans (6), and a few include both directions (Commerford et al. 2022; Downen et al. 2024). Operationally, aversion is defined as attitude (7), behavior (13), attitude and behavior (18), attitude and calibration (3), or attitude, behavior, and calibration (3). This spread indicates ambiguity in both direction and nature, even though the label Algorithm Aversion suggests a negatively oriented attitude toward algorithms. **Rationale**: The phenomenon is primarily a negative **attitude** with **behavioral** implications, but operationalizations vary. To stay maximally inclusive and acknowledge asymmetry in **calibration**, conduct should capture reluctance to use and the tendency to scrutinize. **Choice:** "*reluctance to adopt* … and *critical evaluation of* …" (covers attitude, behavior, calibration)
3. The **target** of conduct typically refers to outputs of algorithmic origin such as *advice, information, recommendation* (36, of which 25 implicit), or to the algorithmic source itself (Liu et al. 2023; Rix et al. 2025). When direction is reversed, the target becomes outputs of human origin (5, one implicit) or the human source (Castelo 2024). Terminology varies, *algorithm* (as computational procedure) is most frequent (26), more recent work names *AI* (4). Some studies reference integrated artefacts such as *algorithm-augmented* or *AI-augmented* devices (6). This diversity suggests broad applicability across targets with differing characteristics. **Rationale**: Many definitions refer to algorithmic *advice* or similar outputs, but the phenomenon extends to modern AI beyond classic algorithms. A broader phrasing should include both discrete algorithms and contemporary AI applications. **Choice:** "out*put of algorithmic systems*" (generalizes advice to input and covers algorithms/AI)
4. Most definitions include a **comparator** (16 explicit; 23 implicit), typically the conceptual opposite of the target: human origin (22) or algorithmic origin (5). A few specify the human comparator as *one's own decision* (Chávez et al. 2024; Horowitz and Kahn 2023). Elsewhere, an inferior alternative (7) is implied via evaluative terms like *superior* or inferred from comparison cues such as *preference* (5). A small subset (5) lacks any explicit or implicit comparator, indicating that Algorithm Aversion can also be evident without a clear alternative. **Rationale:** To retain generality and match the modal pattern in prior work, the comparison should be against human options without locking in a specific role. **Choice.** "*compared to human alternatives*"

5. In absence of a **condition**, about one third of the definitions (17) present the phenomenon as unconditional. Among those specifying conditions, two recur: the *superiority* of algorithmic input (12) and *perceived unreliability* as a trigger (7), sometimes via observed error; a few state both (9). The unreliability trigger also adds a temporal aspect, that Algorithm Aversion becomes salient or is reinforced after a perceived flaw in algorithmic performance. **Rationale**: Aversion is theoretically meaningful when it occurs *despite* clear evidence favoring the algorithmic option; this sets the facilitating condition. Algorithm imperfection as additional condition to exclude ignorance (despite perfect results) **Choice:** *"imperfect [algorithmic systems] ... despite evidence of the superior performance of those systems*"
6. **Scope** is often inferred from the actor's context or the specified input target. Most definitions, explicit or implicit, place the actor where algorithmic input is meant to be used: *decision-making* (5 explicit; 20 implicit), *forecasting* (1 explicit; 3 implicit), or *general system interaction* (3 explicit; 9 implicit). More specific scopes include *clinical decision-making* (Bankuoru Egala and Liang 2024) and *managerial decision-making* (Wang et al. 2024b). Rix et al. (2025) and Dennis et al. (2023) do not go beyond an evaluation setting. **Rationale**: Because the phenomenon appears across tasks and settings, keeping scope implicit maximizes applicability wherever the above primitives hold. **Choice:** *implicit*

Constructed definition: Algorithm Aversion refers to "people's reluctance to adopt output of imperfect algorithmic systems, and their critical evaluation of such systems, despite evidence of the superior performance of those systems compared to human alternatives".

Table C. 1: Extracting Primitives for the Phenomenon of Algorithm Aversion

Exerpt & Source	Reference	Primitive #1: Actor	Primitive #2: Conduct	Primitive #3: Target	Primitive #4: Comparator	Primitive #5: Condition	Primitive #6: Scope
"[…] a general tendency to avoid algorithms in favor of human instinct, experience and judgment." (Jain et al. 2025)	-	decision-makers (implicit)	"avoid", preference for subjective or intuitive reasoning (implied)	"algorithms"	"in favor of human instinct, experience and judgment"	-	decision-making (implicit)
"a behavior of neglecting algorithmic decisions in favor of one's own decisions or other's decisions, either consciously or unconsciously" (Jain et al. 2025)	Mahmud et al. (2022)	decision-makers (implicit)	"neglecting […] either consciously or unconsciously"	"algorithmic decisions"	"in favor of one's own decisions or other's decisions"	-	decision-making (implicit)
"biased assessment of an algorithm which manifests in negative behaviours and attitudes toward the algorithm compared to a human agent" (Rix et al. 2025)	Jussupow et al. (2020)	-	"biased assessment ", "negative behaviors and attitudes"	"algorithm"	"compared to human agent"	-	evaluation settings

"[…] the tendency to reject AI-driven decision-making tools in favor of human judgment […]" (Silber et al. 2025)	Dietvorst et al. (2015)	decision-makers (implicit)	"reject"	"AI-driven decision-making tools"	"in favor of human judgment"	-	decision-making (implicit)
“[…] the reluctance of human decision makers to use superior but imperfect algorithms […]” (SimanTov-Nachlieli 2025)	Burton et al. (2020); Dietvorst et al. (2015); Mahmud et al. (2022)	"human decision makers"	"reluctance […] to use"	"superior but imperfect algorithms"	inferior alternatives (implied)	"superior but imperfect algorithms"	decision-making (implicit)
"[…] people's biased preference for human recommendations over those of algorithms." (Talebi et al. 2025)	Dietvorst et al. (2015)	"people"	"biased preference for"	"human recommendations"	"over those [recommendations] of algorithms"	-	decision-making (implicit)
"[…] reluctance toward the use of AI over humans […]" (Talebi et al. 2025)	Dietvorst et al. (2015)	-	"reluctance toward the use of"	"AI"	" over humans"	-	general "use of AI"
"[…] the tendency to distrust information provided by algorithms." (Baek et al. 2024)	Dietvorst et al. (2015)	-	"tendency to disstrust"	"information provided by algorithms"	-	-	decision-making (implicit)
"[…]a tendency to lose confidence in algorithms faster than in humans after seeing them err." (Castelo 2024)	Dietvorst et al. (2015)	-	"lose confidence in […] faster"	"algorithms"	"than in humans"	"after seeing them err"	general (implicit)
"[…] a preference for humans relative to algorithms." (Castelo 2024)	Morewedge (2022)	-	"preference for"	"humans"	"relative to algorithms"	-	general (implicit)
“rejecting models in favor of their own (mis)judgments even when given evidence of the superior performance of models” (Chávez et al. 2024)	(Petropoulos et al. 2016)	decision-makers (implicit)	"rejecting"	"models" (synonymously with algorithms)	"in favor of their own (mis)judgments"	despite "evidence of the superior performance of models", performance salience (implicit)	decision-making (implicit)
"[…] the tendency for people to more rapidly lose faith in an erring decision-making algorithm than in humans, making comparable errors." (Chávez et al. 2024)	Dietvorst et al. (2015)	people	"tendency to more rapidly lose faith in"	"decision-making algorithm"	humans	erring algorithm, comparable human	decision-making (implicit)
"[…] people choose not to rely upon algorithmic decision aids and instead prefer human judgment in guiding their decision-making." (Downen et al. 2024)	Dietvorst et al. (2015)	"people"	"choose not to rely upon [algorithmic decision aids]", "instead prefer [human judgment]"	"algorithmic decision aids"	"human judgment"	-	"guiding their decision-making"

"the reluctance to use superior algorithm-augmented devices known to be unreliable." (Bankuoru Egala and Liang 2024)	Dietvorst et al. (2015)	-	"reluctance to use"	"algorithm-augmented devices"	inferior alternatives (implied)	"superior […] devices", "devices known to be unreliable"	"use of […] devices"
"the phenomenon where clinicians exhibit resistance, scepticism, or hesitation towards the use of algorithmic-augmented aids in clinical decision-making processes." (Bankuoru Egala and Liang 2024)	Longoni et al. (2020)	"clinicians"	"resistance, scepticism, or hesitation towards the use"	"algorithmic-augmented aids"	alternative (implied)	-	"use", "clinical decision-making processes"
"[…] the tendency of humans in some situations to discard algorithms in favor of their own judgment despite evidence in favor of relying on an algorithm." (Horowitz and Kahn 2023)	-	"humans"	"discard"	"algorithms"	"in favor of their own judgment"	"despite evidence in favor of relying on an algorithm", performance salience (implicit)	"in some situations"
"[…] the preference for humans over algorithms in decision-making […]" (Jussupow et al. 2024)	-	decision-makers (implicit)	"preference"	"humans"	"algorithms"	-	"decision-making"
“although evidence-based algorithms consistently outperform human forecasters, people often fail to use them after learning that they are imperfect” (Keppeler 2024)	Dietvorst et al. (2015)	"people"	"fail to use"	"evidence-based algorithms"	"human forecasters"	"algorithms consistently outperform human forecasters", "after learning that they [evidence-based algorithms] are imperfect"	forecasting
“the tendency of humans to shy away from using algorithms even when algorithms observably outperform their human counterparts” (Koo 2024)	Germann and Merkle (2023)	"humans"	"shy away from using"	"algorithms"	"human counterparts"	"even when algorithms observably outperform their human counterparts", performance salience (implicit)	general (implicit)
"a reluctance to integrate advice from (erroneous) algorithms as compared to (erroneous) humans […]" (Rebholz et al. 2024)	Dietvorst et al. (2015);Burton et al. (2020); Jussupow et al. (2020);Mahmud et al. (2022)	decision-makers (implicit)	"reluctance to integrate"	"advice from (erroneous) algorithms"	"as compared to [advice from] (erroneous) humans"	erroneos algorithm, erroneos human	decision-making (implicit)

"[…] people often fail to rely on good algorithms after learning that the algorithms are imperfect." (Tse et al. 2024)	Dietvorst et al. (2018)	"people"	"fail to rely"	"good algorithms"	-	"good algorithms", "after learning that the algorithms are imperfect"	general
"[…] humans tend to trust their intuition more than analytical algorithms when making managerial decisions." (Wang et al. 2024b)	Dietvorst et al. (2015)	"humans"	"trust […] more"	"their intuition"	"analytical algorithms"	preference for own intuitive reasoning (implied)	"managerial decisions"
"People tend to oppose and critically evaluate algorithms, even if they and artificial intelligence are frequently superior to human decision-making." (Wang et al. 2024a)	Dietvorst et al. (2015)	"people"	"oppose and critically evaluate"	"algorithms […] and artificial intelligence"	"to human decision-making"	algorithms and AI "are frequently superior"	decision-making (implicit)
"[…] the reluctance to use superior but imperfect algorithms." (Wu et al. 2024)	Dietvorst et al. (2015)	humans/decision-makers(implied)	"reluctance to use"	"algorithms"	inferior alternatives (implied)	"superior but imperfect algorithms"	general (implicit)
[…] decision-makers can be reluctant to delegate decisions to or accept advice from AI-based systems, especially when they see these systems exhibit errors […]" (Xu and Wang 2024)	Dietvorst et al. (2015)	"decision-makers"	"can be reluctant to delegate […] to or accept"	" advice from AI-based systems"	alternative (implied)	"especially when they see these systems exhibit errors"	decision-making (implicit)
"[…] non-reliance on AI intervention […]" (Cabitza et al. 2023)	-	-	"non-reliance"	"AI-intervention"	-	-	general (implicit)
"human decision-makers are unwilling to use data-driven algorithms despite being presented with evidence that the algorithms consistently outperform expert human judgment" (Cheng and Chouldechova 2023)	Dietvorst et al. (2015)	"human decision makers"	"are unwilling to use"	"data-driven algorithms"	"expert human judgment"	"despite being presented with evidence that the algorithms consistently outperform", performance salience (implicit)	decision-making (implicit)
"[…] a negative bias against AI after AI makes a mistake because AI is judged more harshly than humans who make similar mistakes." (Dennis et al. 2023)	Dietvorst et al. (2018)	-	"negative bias"	"against AI"	"humans who make similar mistakes"	"after AI makes a mistake because AI is judged more harshly"	evaluation settings
"[…] people often resist and critically judge algorithms." (Liu et al. 2023)	Dietvorst et al. (2015)	"people"	"resist and critically judge"	"algorithms"	-	-	general (implicit)
"[…] the general preference for humans' recommendations or predictions." (Reich et al. 2023)	Dietvorst et al. (2015)	decision-makers (implicit)	"preference for"	humans' recommendations or predictions	-	-	decision-making, forecasting

"[...] users […] unwillingness to accept and utilize advice from algorithms (or the systems within which the algorithms operate), even though such algorithms can be more accurate than humans yet probabilistically imperfect." (Turel and Kalhan 2023)	Burton et al. (2020)	"users"	"unwillingness to accept and utilize"	"advice from algorithms (or the systems within which the algorithms operate)"	humans / human advice (implied)	"even though such algorithms can be more accurate than humans yet probabilistically imperfect"	decision-making (implicit)
"[…] unwillingness to accept and utilize reasonably good algorithmic advice […]"(Turel and Kalhan 2023)	-	users (implicit)	"unwillingness to accept and utilize"	"algorithmic advice"	-	"reasonably good algorithmic advice"	decision-making (implicit)
“the reluctance of human forecasters to use superior but imperfect algorithms” (Cabiddu et al. 2022)	Burton et al. (2020)	"human forecasters"	"reluctance […] to use"	"superior but imperfect algorithms"	inferior alternatives (implicit)	"superior but imperfect algorithms”	forecasting (implicit)
"The human tendency to discount advice from algorithms and rely more readily on human input, as compared to computer-generated input." (Commerford et al. 2022)	Eastwood et al. (2012); Dietvorst et al. (2015)	"human"	" discount advice from algorithms" and "rely more readily on human input"	"advice from algorithms","computer-generated input"	"human input"	-	decision-making (implicit)
"The tendency to ignore or discount advice from algorithms that have made an error […]." (Ganbold et al. 2022)	Dietvorst et al. (2015)	decision-makers (implicit)	"ignore or discount"	"advice from algorithms"	-	"algorithms that have made an error"	decision-making (implicit)
"The tendency of humans to shy away from using algorithms—even when algorithms observably outperform their human counterpart—has been referred to as algorithm aversion." (Germann and Merkle 2023)	Dietvorst et al. (2015)	"humans"	"tendency […] to shy away from using"	"algorithms"	"human counterpart"	"even when algorithms observably outperform their human counterpart"; performance salience (implicit)	general "use of algorithms"
"[…] people reject algorithms despite being familiar with their superior performance […]" (Mahmud et al. 2022)	Dietvorst et al. (2015)	"people"	"reject"	"algorithms"	inferior alternatives (implicit)	"despite being familiar with their superior performance"; performance salience	general (implicit)
"[…] people tend to dismiss input from algorithms even when given information about the algorithm’s superior performance […]" (Hou and Jung 2021)	Dietvorst et al. (2015)	"people"	"tend to dismiss"	"input from algorithms"	-	"even when given information about the algorithm’s superior performance", performance salience (implicit)	decision-making (implicit)

"[…] people are resistant to using algorithms even if they are equal or superior to a human." (Shariff et al. 2021)	Dietvorst et al. (2015)	"people"	"are resistant to using"	"algorithms"	"to a human"	"even if they [algorithms] are equal or superior"	general (implicit)
"[…] the reluctance of human decision makers to use superior but imperfect algorithms." (Burton et al. 2020)	Dietvorst et al. (2015)	"human decision makers"	"reluctance […] to use"	"superior but imperfect algorithms"	inferior alternatives (implicit)	"superior but imperfect algorithms"	decision-making
"[…] people tend to adopt the inferior decisions made by humans rather than the superior algorithmic recommendations." (Feng and Gao 2020)	Dietvorst et al. (2015)	"people"	"tend to adopt"	"inferior decisions made by humans"	"rather than the superior algorithmic recommendations "	"superior algorithmic recommendations "	decision-making (implicit)
"Although evidence-based algorithms consistently outperform human forecasters, people often fail to use them after learning that they are imperfect […]" (Dietvorst et al. 2018)	-	"people"	"fail to use"	"evidence-based algorithms"	"human forecasters"	"algorithms consistently outperform human forecasters", "after learning that they are imperfect "	forecasting (implicit)
"The irrational discounting of automation advice […]" (Prahl and Van Swol 2017)	Meehl (1954)	-	"irrational discounting"	"of automation advice"	rational alternative	-	decision-making (implicit)
"[…] a general tendency for people to more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake." (Dietvorst et al. 2015)	-	"people"	"to more quickly lose confidence"	"in algorithmic […] forecasters"	"human forecasters"	"after seeing them [algorithmic forecasters] make the same mistake"	forecasting (implicit)

Algorithm Appreciation

I identified 1 distinct formal definitions of Algorithm Appreciation across 14 sources. Table C.2 lists all excerpts by year (most recent first) and then alphabetically by author(s). Most authors cite Logg et al. (2019) as the original source, despite substantial variation in their own definitions, indicating fragmentation similar to Algorithm Aversion. Findings by primitive follow:

1. Most definitions describe a general human **actor** such as “people” (5 counts), “individuals” (3), or “humans” (1). Some specify context, e.g., “decision makers” (Sachin and Schecter 2024) or “consumers” (Koo 2024). Even when unnamed (4), the context typically implies a human engaging with algorithmic output. **Rationale**: To maintain consistency across sources and avoid domain restrictions (e.g., “consumers,” “decision makers”), a generic human label is preferable. **Choice:** *people*

2. The **conduct** of the phenomenon is characterized in attitude terms (4) or behavior terms (8). You et al. (2022) define it as both attitude and behavior; Dennis et al. (2023) as attitude and calibration; Sachin and Schecter (2024) as attitude, behavior, and calibration. Unlike Algorithm Aversion, conduct is consistently positive and directed toward the algorithmic source. **Rationale**: Many studies operationalize appreciation behaviorally via reliance (e.g., judge-advisor setups), but attitude is also evidenced (preference), and some findings imply asymmetry in calibration (e.g., stronger reactions or a general positive bias toward AI). To reflect practice while remaining inclusive, conduct should capture both behavior (reliance) and attitude (preference), with room for calibration effects. **Choice:** *preference for and greater reliance on*
3. All definitions direct conduct toward the **target** of algorithmic output. The source is usually termed "algorithm" (13), with alternatives "AI" (Dennis et al. 2023) or "AI algorithm" (Jenkin et al. 2024). Cabitza et al. (2023) and Liu et al. (2023) note that merely labeling or framing advice as algorithmic can trigger appreciation. **Rationale**: Narrow phrases like "algorithmic advice" are historically and contextually limiting. Contemporary systems (including AI) generate recommendations, predictions, information, creative artefacts, and automated actions. A broader formulation should encompass all such outputs. **Choice:** *output from algorithmic systems*
4. Relative to the algorithmic target, the **comparator** is typically human input (9 explicit; 2 implicit). Exceptions include Cabitza et al. (2023) and Liu et al. (2023), who compare labeling/framing conditions. Jenkin et al. (2024) and Turel and Kalhan (2023) omit a comparator, though the context suggests the individual's own judgment. **Rationale**: Most definitions contrast algorithmic with human input; keeping this generic preserves applicability across tasks and roles. **Choice:** *compared to human alternatives*
5. **Conditions** are rarely explicit. Three sources reference equivalent (Cabitza et al. 2023; Rebholz et al. 2024) or identical advice (You et al. 2022) to establish a comparative basis. Dennis et al. (2023) add a temporal note, that appreciation arises "before the AI makes a mistake" (cf. Algorithm Aversion). Overall, the literature treats Algorithm Appreciation as typically occurring without specific conditions. **Rationale**: Identity/equivalence manipulations (e.g., identical advice) are operationalization choices rather than prerequisites of the phenomenon; most sources do not require specific conditions. **Choice:** *no explicit condition specified*
6. **Scope** is generally inferred from context, most often decision-making (10 counts) or forecasting (Reich et al. 2023). Broader applications also appear: Rix et al. (2025) and Downen et al. (2024) describe general evaluative settings, and Dennis et al. (2023) explicitly frame Algorithm Appreciation as a broad, generalizable phenomenon. **Rationale**: Although often studied in decision-making (and sometimes forecasting), appreciation appears in broader evaluative contexts; specifying scope would unnecessarily narrow applicability. **Choice:** *implicit*

Constructed definition: Algorithm Appreciation refers to "peoples' preference for and greater reliance on output from algorithmic systems compared to human alternatives".

Table C.2: Extracting Primitives for the Phenomenon of Algorithm Appreciation

Exerpt & Source	Reference	Primitive #1: Actor	Primitive #2: Conduct	Primitive #3: Target	Primitive #4: Comparator	Primitive #5: Condition	Primitive #6: Scope

"[…] humans' preference for algorithmic over human conduct." (Rix et al. 2025)	Dennis et al. (2023);You et al. (2022)	"humans"	"preference for"	"algorithmic […] conduct"	"over human conduct"	-	evaluation settings
"[…] people will rely more on information provided by algorithms than on information provided by humans […]" (Downen et al. 2024)	Logg et al. (2019)	"people"	"will rely more on"	"information provided by algorithms"	"than on information provided by humans"	-	evaluation settings
"[…] individuals tend to follow an AI algorithm's advice." (Jenkin et al. 2024)	Logg et al. (2019)	"individuals"	"tend to follow"	"AI algorithm's advice"	-	-	decision-making (implicit)
"[…] consumers may prefer advice from algorithms to advice from humans […]" (Koo 2024)	Logg et al. (2019)	"consumers"	"may prefer"	"advice from algorithms"	"to advice from humans"	-	decision-making (implicit)
"[…] people integrating algorithmic advice more than quantitatively equivalent advice provided by human advisors." (Rebholz et al. 2024)	Logg et al. (2019)	"people"	"integrating […] more"	"algorithmic advice"	"than advice provided by human advisors"	"quantitatively equivalent advice"	decision-making (implicit)
"decision makers will react more strongly to advice from algorithms and will report greater preference for an algorithmic advisor" (Sachin and Schecter 2024)	-	"decision makers"	"will react more strongly to and will report greater preference for"	"advice from algorithms"; "algorithmic advisor"	alternative (implicijt)	-	decision-making (implicit)
"[…] people consistently give more weight to equivalent advice when it is labeled as coming from an algorithmic versus human source." (Cabitza et al. 2023)	(Logg et al. 2019)	"people"	"give more weight to"	" advice when it is labeled as coming from an algorithmic […] source"	versus "when it is labeled as coming from […] a human source"	"equivalend advice"	decision-making (implicit)
"a general positive bias for AI before the AI makes a mistake." (Dennis et al. 2023)	Logg et al. (2019)	-	"positive bias for"	"AI"	-	"before the AI makes a mistake"	"general"
"[…] individuals adhere more to algorithm-framed than human-framed advice […]" (Liu et al. 2023)	Logg et al. (2019)	"individuals"	"adhere more to"	"algorithm-framed […] advice"	"than human-framed advice"	-	decision-making (implicit)
"[…] algorithms are preferred over humans as sources of prediction […]" (Reich et al. 2023)	Dietvorst and Bharti (2020);Dijkstra (1999);Dijkstra et al. (1998); Logg et al. (2019)	-	"prefered"	"algorithms [...] as sources of prediction"	"over humans"	-	forecasting (implicit)
"[…] adherence to algorithmic advice." (Turel and Kalhan 2023)	Logg et al. (2019)	-	"adherence"	"to algorithmic advice"	-	-	decision-making (implicit)

"[…] individuals largely […] follow algorithmic advice to a greater extent than identical human advice due to a higher trust in an algorithmic than human advisor." (You et al. 2022)	Logg et al. (2019)	"individuals"	"follow […] to a greater extent", "due to a higher trust in an algorithmic […] advisor"	"algorithmic advice"	"identical human advice"	"advice identical"	decision-making (implicit)
"[…] individuals largely […] follow algorithmic advice to a greater extent than identical human advice due to a higher trust in an algorithmic than human advisor." (Hou and Jung 2021)	Logg et al. (2019)	"people"	"rely more on"	"algorithmic advice"	"than human advice"	-	"in some situations"
"[…] greater reliance on advice from an algorithm than from a human." (Castelo et al. 2019)	Logg et al. (2019)	-	"greater reliance on"	"advice from an algorithm"	"[advice] than from a human"	-	decision-making (implicit)
"[…] people actually prefer advice from algorithms to advice from people." (Logg et al. 2019)	-	"people"	"actually prefer"	" advice from algorithms"	“to advice from people”	-	decision-making (implicit)

Automation Bias

I identified ten formal definitions of Automation Bias across four distinct papers. Table C.3 lists all excerpts by year (most recent first) and then alphabetically by author(s). Horowitz and Kahn (2023) provide two definitions, Cabitza et al. (2023) references five and contributes one additional definition. As the longest-known phenomenon in this set, cited sources range from Mosier and Skitka (1996) to Lyell and Coiera (2017). Findings by primitive follow:

1. Most definitions reference a general human **actor** such as “humans” (2), “people” (1), or “users” (2). Some specify roles such as “human operators” (Horowitz and Kahn 2023), “physicians,” or “human decision makers” (Cabitza et al. 2023). Even when unnamed (Schaffer et al. 2019), context clearly implies a human engaging with algorithmic input. **Rationale:** Labels vary across sources (“people,” “users,” “human operators”), but a generic term avoids domain lock-in and covers common usages. **Choice:** *human*
2. Definitions frame the **conduct** of the phenomenon in attitude terms (3), behavior terms (4), or both (3). Behaviorally, it appears as over-reliance on automated systems; in attitude terms, as over-trust. A notable outlier is Cabitza et al. (2023) citation of Cummings (2004), defining Automation Bias as a “disregard for contradictory information”. **Rationale:** Definitions span attitude (e.g., “over-trust”, “complacency”), behavior (e.g., “over-reliance”, “tendency to accept”), or both (e.g., “propensity to over rely”). To stay inclusive while keeping a clear line between disposition and action, use wording that captures both. Calibration is less explicit but compatible with an over-reliance pattern. **Choice:** *tendency to over-rely*
3. In nearly all cases, conduct is directed toward output of an automated source (9) as **target**. The exception is the Cummings (2004) definition (Cabitza et al. 2023), where conduct targets “contradictory information,” which is disregarded. Expressions for automated sources range from general “technology” (Schaffer et al. 2019) to specific “AI systems” (Nourani et al. 2021), and “Clinical Decision Support Systems” (Cabitza et al. 2023). **Rationale:** Expressions

range from “automated advice”, “AI-enabled decision aids”, and “computer output” to domain-specific systems (e.g., clinical decision support). A broad term better accommodates classic automation and modern AI. **Choice:** *output from algorithmic systems*

4. As **comparator**, all definitions invoke a more appropriate level of reliance, either explicitly (5) or implicitly (5), thereby implying the existence of a better alternative to the observed reliance level. **Rationale:** The notion of *over-reliance* already implies a superior alternative or more appropriate reliance level; most sources leave this implicit. **Choice:** *no explicit comparator specified*
5. Several accounts imply insufficient reliability (Horowitz and Kahn 2023) or cite an inaccurate computer-generated solution (Cabitza et al. 2023) as a **condition**. Elsewhere, the description of over-reliance implies an imperfect (7) or erroneous (Cabitza et al. 2023) automated system. **Rationale:** the cognitive basis of over-reliance suggests a threshold of appropriate reliance. Explicit mention of imperfection or erroneous quality of systems for clarity **Choice:** *imperfect or erroneous [systems]*
6. **Scope** is usually inferred from the target (7). It spans general technology use (Schaffer et al. 2019), general system or automated system use (Cabitza et al. 2023; Nourani et al. 2021), and more specific decision-making contexts (4). Overall, the phenomenon appears broadly applicable across human-computer interaction settings. **Rationale:** Occurs across varied human-computer interaction contexts; naming a specific domain would unnecessarily narrow applicability. **Choice:** *implicit*

Constructed definition: Automation Bias is the “human tendency to over-rely on output from algorithmic systems, even when those systems are imperfect or erroneous”.

Table C.3: Extracting Primitives for the Phenomenon of Automation Bias

Exerpt & Source	Reference	Primitive #1: Actor	Primitive #2: Conduct	Primitive #3: Target	Primitive #4: Comparator	Primitive #5: Condition	Primitive #6: Scope
“tendency [for human operators] to over-rely on automation.” (Horowitz and Kahn 2023)	(Goddard et al. 2012); Skitka et al. (1999)	"human operators"	"over-rely"	"on automation"	threshold of appropriate reliance (implicit)	imperfect system (implicit)	general automated operation
"[…] the tendency of humans to rely on AI-enabled decision aids above and beyond the extent to which they should, given the reliability of the algorithms." (Horowitz and Kahn 2023)	-	"humans"	"rely"	"AI-enabled decision aids"	"above and beyond the extent to which they should"	"given the reliability of the algorithms"	decision-making (implicit)
"[…] the 'type of human cognitive bias due to over-reliance on the recommendations of an AI system'." (Cabitza et al. 2023)	-	"human"	" cognitive bias due to over-reliance on"	" the recommendations of an AI system"	threshold of appropriate reliance (implicit)	imperfect system (implicit)	decision-making (implicit)
"[…] automation-included complacency" (Cabitza et al. 2023)	Lyell and Coiera (2017)	-	"complacency" (passivity, lack of vigilance)	"automation"	lower vigilance than required (implicit)	imperfect system (implicit)	automated systems use (implicit)

"[…] the propensity of people to over rely on automated advice […]" (Cabitza et al. 2023)	Goddard et al. (2014)	"people"	"propensity to over rely on"	"automated advice"	threshold of appropriate reliance (implicit)	imperfect system (implicit)	decision-making (implicit)
"[…] tendency to over-trust HIT [Healthcare Information Technology] leading a physician to make an incorrect decision in order to follow the advice provided by a CDSS [Clinical Decision Support System]" (Cabitza et al. 2023)	Bouaud et al. (2015)	"a physician"	"tendency to over-trust"	"Healthcare Information Technology", "advice provided by a Clinical Decision Support System"	correct decision (implied by "to make an incorrect decision")	erroneous system (implicit)	physician tasks
"[…] the human tendency . . . which occurs when a human decision maker disregards or does not search for contradictory information in light of a computer-generated solution which is accepted as correct […]"(Cabitza et al. 2023)	Cummings (2004)	"human decision maker"	"disregards or does not search for"	contradictory information	"a computer-generated solution "	"computer generated solution which is accepted as correct [but is not]"	decision-making (implicit)
[…] the tendency "by which users tend to over-accept computer output 'as a heuristic replacement of vigilant information seeking and processing"' (Cabitza et al. 2023)	Mosier and Skitka (1996)	"users"	"tend to over-accept", "as a heuristic replacement"	"computer output"	"of vigilant information seeking and processing"	imperfect system (implicit)	human-computer-interaction
"users [...] rely on the system more than they should […]" (Nourani et al. 2021)	-	"users"	"rely on"	"the system"	"more than they should"	imperfect system (implicit)	system use (implicit)
"[…] the over-trusting of technology […]" (Schaffer et al. 2019)	Cummings (2004)	-	"the over-trusting of"	"technology"	threshold of appropriate reliance (implicit)	imperfect system (implicit)	technology use (implicit)

Appendix D – Data Tables

Table D. 1 presents all empirical articles included in the sample, organized by publication date and author. For each study, it indicates the phenomena under investigation and categorizes the applied measures as either subjective or objective.

Table D. 1: Measurement types of Empirical Articles

Author(s) & Phenomenon	Subjective Measurements	Objective Measurements
Cabitza et al. (2025), Automation Bias	confidence; perceived utility; perceived complexity;	diagnostic accuracy (before and after support);
Chacon et al. (2025), Algorithm Aversion	perceived learning capabilities	adjustment (WOA); binary choice; accuracy (over time)
Jain et al. (2025), Algorithm Aversion	variant of xenophobic scale (e.g., with increased reliance on the algorithm I fear that organizational life would change for the worse"); trust; neuroticism;	-
Li et al. (2025), Algorithm Aversion	purchase intention; green-identity; perceived objectivity; perceived scarcity;	-
Rix et al. (2025), Algorithm Aversion	general attitudes towards digital news (interviews), beliefs about love, perceived effort, perceived quality, curiosity, AI knowledge	binary choice (CBC experiment);
Silber et al. (2025), Algorithm Aversion	high standards (9 items); alternative search (12 items); maximizing mindset (5 items); likelihood of using AI advisor; perceived effectiveness (3 items); algorithm aversion (7 items); AI usage experience (5 open questions)	-
SimanTov-Nachlieli (2025), Algorithm Aversion	Preimplementation AI attitude (2 items); Behavioral construct, would integrate AI advice (y/n/indifferent); would recommend AI integration; perceived potential loss of standing; confidence in own vs AI ability (difference)	performance (ranking)
Talebi et al. (2025), Algorithm Appreciation	preference; perceived embarrassment (3 items); perceived exploitation; perceived itentity threat;	binary choice; donation behavior;
Baek et al. (2024), Algorithm Aversion	perceived ad credibility (scale); attitude towards add (negative/positive, scale); AI-perception;	donation amount
Brüns and Meißner (2024), Algorithm Aversion	brand authenticity (3 items); (social media) post credibility (3 items); EWOM intentions (3 items); brand loyalty (1 item); perceived passion (2 items);	-
Castelo (2024), Algorithm Aversion	preference for human or algorithm for different tasks; perceived corruption (5 items); perceived algorithm transparency; perceived productivity boost; honesty, competence of humans decision-makers; worry about unfair algorithms	-
Chávez et al. (2024), Algorithm Aversion	Algorithm Aversion (3 items); Perceived hedonism (3 items); Behavioral intention to use (3 items)	-
Commerford et al. (2024), Algorithm Aversion	Locus if Control (external & internal); expected adjustment;	-
Downen et al. (2024), Algorithm Aversion	emotional response as function of [interested, surprised, sad, angry, challenged, ashamed, proud, frustrated, fear, disgusted, guilty, bored, happy, contempt, hope]; reliability; competence; trustworthiness	Adjustment; (investment) performance

Bankuoru Egala and Liang (2024), Algorithm Aversion	Algorithm Aversion (6 items CDC Analysis)	-
Gill et al. (2024), Algorithm Aversion	algorithm familiarity; domain experience; competitiveness (big five); cognitive reflection (3 items)	advice rejection (binary); performance (score)
Jenkin et al. (2024), Algorithm Aversion	experience with rental platforms	advice adherence (binary); adjustment (WoA); explanation seeking (binary)
Keppeler (2024), Algorithm Aversion	-	response to message (binary); interest in job (binary); click link (binary); person job fit (percentage)
Koo (2024), Algorithm Aversion	-	binary choice (human vs. AI)
Sachin and Schecter (2024), Algorithm Appreciation	preference (perceived usefulness; satisfaction; trust); confidence in advisor;	adjustment (WOA)
Tse et al. (2024), Algorithm Aversion	-	adjustment (MSHIFT)
Wang et al. (2024b), Algorithm Aversion	-	safety stock adjustment (deviation from algorithmic recommendations); performance (sales volume); availability (days of positive inventory level); waste (avg daily expired inventory); heteroskedasticity (Engle's ARCH test)
Wu et al. (2024), Algorithm Aversion	perceived benefit (5 items); perceived risk (4 items); perceived equity (3 items); social trust (3 items); intention to click (3 items);continuance intention (3 items); algorithm aversion (Melick, 2020)	-
Xu and Wang (2024), Algorithm Aversion	perceived explainability (4 items); trust; self-confidence; decision satisfaction; decision regret	behavioral trust (adherence after error); behavioral trust (MAD, PtC bias)
Cabitza et al. (2023), Automation Bias	average perceived complexity	adjustment; case complexity
Cheng and Chouldechova (2023), Algorithm Aversion	perceived average error; confidence; perceived transparency; perceived fairness; perceived representation	adjustment (AAD); prediction performance (AAE)
Dennis et al. (2023), Algorithm Aversion	trustworthiness (3 items); perceived ability (2 items); perceived integrity (2 items); perceived benevolence (3 items); willingness to wirk with (4 items); process satisfaction (3 items); perceived conflict (9 items)	-
Germann and Merkle (2023), Algorithm Aversion	self reported trust, risk-aversion, expertise; familiarity	binary choice; investment performance
Horowitz and Kahn (2023), Automation Bias	trust in ai (quiz); confidence in decision aid; self-confidence; overall ai knowledge (quiz); overall ai familiarity (quiz); overall ai experience (quiz); ai background index (3 items)	binary choice (rate of switching)
Liu et al. (2023), Algorithm Aversion	Algorithm Aversion (difference in trust between human & algorithm advisor); uniqueness neglect; familiarity; objectivity; consequentiality; individualistm-collectivism (6items, Yoo et al., 2011)	-
Reich et al. (2023), Algorithm Aversion, Algorithm Appreciation	trust; perceived learning from mistakes (2 items); perceived impressiveness of art; preference (human vs. algorithm)	binary choice (human vs. algorithm); binary choice (self vs. algorithm)
Turel and Kalhan (2023), Algorithm Aversion	attitude toward AI; trust in AI; familiarity with AI; self-efficacy in estimating weight	adjustment (WOA);
Commerford et al. (2022), Algorithm Aversion	concern about perceived expertise of source; perceived likelihood of management convincing; perception of managements willingness to adjust;	adjustment
Ganbold et al. (2022), Algorithm Aversion	perceived algorithm complexity; perceived avatar competence; estimated likelihood to follow advice;	-
Xie et al. (2022), Algorithm Aversion	purchase likelihood;	binary choice (accept or reject recommendation); ERP (cognitive conflict)

You et al. (2022), Algorithm Appreciation	trust in advice (3 items); need for cognition (3 items);	adjustment (WOA); cognitive load (adjusted response time)
Hou and Jung (2021), Algorithm Aversion, Algorithm Appreciation	social power (expert power, referent power, information power, legitimate power-position)	adjustment (influence factor)
Nourani et al. (2021), Automation Bias	confidence; perceived utility; perceived complexity; estimated detection accuracy; helpfulness;	task time; task error;
Shariff et al. (2021), Algorithm Aversion	self-perceived driving safety relative to others;	minimum safety threshold;
Dietvorst and Bharti (2020), Algorithm Aversion	bid on superior model; model rate of error; confidence in model; own rate of error; participant confidence; distribution of model forecasts; distribution of own forecasts;	-
Feng and Gao (2020), Algorithm Aversion	-	Humans deviation from optimum (PtC bias); Regret aversion (adjustment after feedback); PtC asymmetry (high vs low profit); Delta profit from optimum;
Castelo et al. (2019), Algorithm Aversion	perceived objectivity; perceived consequentialness; familiarity; trust; choice preference; cognitive trust (3 items); affective (dis)trust (3 items); perceived effectiveness; perceived discomfort; perceived human-likeness	click through rate; adjustment (WOA) as theoretical estimate;
Logg et al. (2019), Algorithm Appreciation	-	adjustment (WOA); binary choice
Schaffer et al. (2019), Automation Bias	familiarity; expertise; trust;	adherence to advice (proportion of rounds where algorithmic advice was followed); situation awareness (inverted sum of estimation errors); performance
Dietvorst et al. (2018), Algorithm Averison	satisfaction; confidence;	binary choice (agent); adjustment (limited); forecasting accuracy
Prahl and Van Swol (2017), Algorithm Aversion	confidence; emotional reactions (10 items); perceived similarity	adjustment (SHIFT)
Dietvorst et al. (2015), Algorithm Aversion	confidence in judgment, perceived accuracy, likelihood of perfect prediction, likelihood of bad estimate	binary choice (self vs. algorithm); binary choice (human vs. algorithm); average absolute error (AAE)
Skitka et al. (1999), Automation Bias	confidence; perceived difficulty; perceived reliability; perceived judgment superiority;	omission errors; commission errors; verification instances (secondary display use); response time; tracking performance (weighted average);

Appendix E – Complete Sample List

Table E. 1 provides the full list of sampled studies, sorted by year (newest first) and then alphabetically by author. It reports the title, the outlet of publication (journal or conference), and the disciplinary focus of that outlet.

Table E. 1: Complete Sample List

Author(s)	Title	Outlet	Subject
(Cabitza et al. 2025)	From Oracular to Judicial: Enhancing Clinical Decision Making through Contrasting Explanations and a Novel Interaction Protocol	Proceedings of the 30th International Conference on Intelligent User Interfaces	Human-Computer Interaction
(Chacon et al. 2025)	Preventing algorithm aversion: People are willing to use algorithms with a learning label.	Journal of Business Research	Marketing
(Jain et al. 2025)	Adaption and validation of the algorithm aversion scale and its relationship with neuroticism and trust.	Journal of Organizational Change Management	General & Strategy
(Li et al. 2025)	The interactive effect of recommendation subjects and message types on consumers' suboptimal food purchase intentions.	Journal of Retailing & Consumer Services	Marketing
(Rix et al. 2025)	The Algorithm Discount: Explaining Consumers' Valuation of Human- versus Algorithm-Created Digital Products.	Journal of Management Information Systems	Management Information Systems, Knowledge Management
(Silber et al. 2025)	Embracing AI advisors for making (complex) financial decisions: an experimental investigation of the role of a maximizing decision-making style.	International Journal of Bank Marketing	Marketing
(SimanTov-Nachlieli 2025)	More to Lose: The Adverse Effect of High Performance Ranking on Employees' Preimplementation Attitudes Toward the Integration of Powerful AI Aids.	Organization Science	Organization Behavior/Studies, Human Resource Management, Industrial Relations
(Talebi et al. 2025)	Unveiling coping mechanisms in marketplace discrimination: The allure of artificial intelligence recommendations.	Journal of Product Innovation Management	Innovation
(Baek et al. 2024)	Effect of disclosing AI-generated content on prosocial advertising evaluation.	International Journal of Advertising	Marketing
(Brüns and Meißner 2024)	Do you create your content yourself? Using generative artificial intelligence for social media content creation diminishes perceived brand authenticity.	Journal of Retailing & Consumer Services	Marketing
(Castelo 2024)	Perceived corruption reduces algorithm aversion.	Journal of Consumer Psychology	Marketing
(Chávez et al. 2024)	Opening the moral machine's cover: How algorithmic aversion shapes autonomous vehicle adoption.	Transportation Research Part A: Policy & Practice	Operations Research, Management Science, Production & Operations Management
(Commerford et al. 2024)	Control issues: How providing input affects auditors' reliance on artificial intelligence.	Contemporary Accounting Research	Finance & Accounting
(Downen et al. 2024)	Algorithm aversion, emotions, and investor reaction: Does disclosing the use of AI influence investment decisions?	International Journal of Accounting Information Systems	Finance & Accounting
(Bankuoru Egala and Liang 2024)	Algorithm aversion to mobile clinical decision support among clinicians: a choice-based conjoint analysis.	European Journal of Information Systems	Management Information Systems, Knowledge Management
(Gill et al. 2024)	Dynamics of Reliance on Algorithmic Advice.	Journal of Behavioral Decision Making	Psychology

(Jenkin et al. 2024)	Explanation seeking and anomalous recommendation adherence in human-to-human versus human-to-artificial intelligence interactions.	Decision Sciences	Operations Research, Management Science, Production & Operations Management
(Jussupow et al. 2024)	An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making	Management Information Systems Quarterly	Management Information Systems, Knowledge Management
(Keppeler 2024)	No Thanks, Dear AI! Understanding the Effects of Disclosure and Deployment of Artificial Intelligence in Public Sector Recruitment.	Journal of Public Administration Research & Theory	Public Sector Management
(Koo 2024)	AI is not careful: approach to the stock market and preference for AI advisor.	International Journal of Bank Marketing	Marketing
(Rebholz et al. 2024)	Mixed-effects regression weights for advice taking and related phenomena of information sampling and utilization.	Journal of Behavioral Decision Making	Psychology
(Sachin and Schecter 2024)	Advice Utilization in Combined Human-Algorithm Decision-Making: An Analysis of Preferences and Behaviors.	Journal of the Association for Information Systems	Management Information Systems, Knowledge Management
(Tse et al. 2024)	Beware the performance of an algorithm before relying on it: Evidence from a stock price forecasting experiment.	Journal of Economic Psychology	Economics
(Wang et al. 2024a)	A dimensional exploration and scale development study of algorithm aversion.	Journal of the Operational Research Society	Operations Research, Management Science, Production & Operations Management
(Wang et al. 2024b)	Algorithm aversion during disruptions: The case of safety stock.	International Journal of Production Economics	Operations Research, Management Science, Production & Operations Management
(Wu et al. 2024)	Social trust and algorithmic equity: The societal perspectives of users' intention to interact with algorithm recommendation systems.	Decision Support Systems	Management Information Systems, Knowledge Management
(Xu and Wang 2024)	Explainability increases trust resilience in intelligent agents.	British Journal of Psychology	Psychology
(Cabitza et al. 2023)	AI Shall Have No Dominion: on How to Measure Technology Dominance in AI-supported Human decision-making	Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems	Human-Computer Interaction
(Cheng and Chouldechova 2023)	Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control	Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems	Human-Computer Interaction
(Dennis et al. 2023)	AI Agents as Team Members: Effects on Satisfaction, Conflict, Trustworthiness, and Willingness to Work With.	Journal of Management Information Systems	Management Information Systems, Knowledge Management
(Germann and Merkle 2023)	Algorithm aversion in delegated investing	Journal of Business Economics	General & Strategy
(Horowitz and Kahn 2023)	Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts.	International Studies Quarterly	Economics
(Liu et al. 2023)	Is algorithm aversion WEIRD? A cross-country comparison of individual-differences and algorithm aversion.	Journal of Retailing & Consumer Services	Marketing
(Reich et al. 2023)	How to overcome algorithm aversion: Learning from mistakes.	Journal of Consumer Psychology	Marketing
(Turel and Kalhan 2023)	Prejudiced against the Machine? Implicit Associations and the Transience of Algorithm Aversion	Management Information Systems Quarterly	Management Information Systems, Knowledge Management

(Cabiddu et al. 2022)	Why do users trust algorithms? A review and conceptualization of initial trust and trust over time.	European Management Journal	Management
(Commerford et al. 2022)	Man Versus Machine: Complex Estimates and Auditor Reliance on Artificial Intelligence.	Journal of Accounting Research	Finance & Accounting
(Ganbold et al. 2022)	Increasing Reliance on Financial Advice with Avatars: The Effects of Competence and Complexity on Algorithm Aversion.	Journal of Information Systems	Finance & Accounting
(Mahmud et al. 2022)	What influences algorithmic decision-making? A systematic literature review on algorithm aversion	Technological Forecasting and Social Change	Innovation
(Xie et al. 2022)	The searching artificial intelligence: Consumers show less aversion to algorithm-recommended search product.	Psychology & Marketing	Marketing
(You et al. 2022)	Algorithmic versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation?	Journal of Management Information Systems	Management Information Systems, Knowledge Management
(Hou and Jung 2021)	Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making	Proc. ACM Hum.-Comput. Interact.	Management Information Systems, Knowledge Management
(Nourani et al. 2021)	Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems	Proceedings of the 26th International Conference on Intelligent User Interfaces	Human-Computer Interaction
(Shariff et al. 2021)	How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars.	Transportation Research Part C: Emerging Technologies	Operations Research, Management Science, Production & Operations Management
(Burton et al. 2020)	A systematic review of algorithm aversion in augmented decision making.	Journal of Behavioral Decision Making	Psychology
(Dietvorst and Bharti 2020)	People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Erro	Psychological Science	Psychology
(Feng and Gao 2020)	Is optimal recommendation the best? A laboratory investigation under the newsvendor problem.	Decision Support Systems	Management Information Systems, Knowledge Management
(Castelo et al. 2019)	Task-Dependent Algorithm Aversion.	Journal of Marketing Research (JMR)	Marketing
(Logg et al. 2019)	Algorithm appreciation: People prefer algorithmic to human judgment.	Organizational Behavior and Human Decision Processes	Organization Behavior/Studies, Human Resource Management, Industrial Relations
(Schaffer et al. 2019)	I can do better than your AI: expertise and explanations	Proceedings of the 24th International Conference on Intelligent User Interfaces	Human-Computer Interaction
(Schaffer et al. 2019)	Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them.	Management Science	Operations Research, Management Science, Production & Operations Management
(Prahl and Van Swol 2017)	Understanding algorithm aversion: When is advice from automation discounted?	Journal of Forecasting	General & Strategy
(Dietvorst et al. 2015)	Algorithm aversion: People erroneously avoid algorithms after seeing them err.	Journal of Experimental Psychology: General	Psychology
(Skitka et al. 1999)	Automation Use and Automation Bias	International Journal of Human Computer Studies	Operations Research, Management Science, Production & Operations Management

References

Akmajian, A., Farmer, A. K., Bickmore, L., Demers, R. A., and Harnish, R. M. 2017. *Linguistics: An Introduction to Language and Communication*. MIT press.

Baek, T. H., Kim, J., and Kim, J. H. 2024. "Effect of Disclosing Ai-Generated Content on Prosocial Advertising Evaluation," *International Journal of Advertising*), pp. 1-22.

Bankuoru Egala, S., and Liang, D. 2024. "Algorithm Aversion to Mobile Clinical Decision Support among Clinicians: A Choice-Based Conjoint Analysis," *European Journal of Information Systems* (33:6), pp. 1016-1032.

Bouaud, J., Spano, J.-P., Lefranc, J.-P., Cojean-Zelek, I., Blaszka-Jaulerry, B., Zelek, L., Durieux, A., Tournigand, C., Rousseau, A., and Vandenbussche, P.-Y. 2015. "Physicians' Attitudes Towards the Advice of a Guideline-Based Decision Support System: A Case Study with Oncodoc2 in the Management of Breast Cancer Patients," in *Medinfo 2015: Ehealth-Enabled Health*. IOS Press, pp. 264-269.

Brüns, J. D., and Meißner, M. 2024. "Do You Create Your Content Yourself? Using Generative Artificial Intelligence for Social Media Content Creation Diminishes Perceived Brand Authenticity," *Journal of Retailing and Consumer Services* (79), p. 103790.

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).

Cabiddu, F., Moi, L., Patriotta, G., and Allen, D. G. 2022. "Why Do Users Trust Algorithms? A Review and Conceptualization of Initial Trust and Trust over Time," *European management journal* (40:5), pp. 685-706.

Cabitza, F., Campagner, A., Angius, R., Natali, C., and Reverberi, C. 2023. "Ai Shall Have No Dominion: On How to Measure Technology Dominance in Ai-Supported Human Decision-Making," *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1-20.

Cabitza, F., Famiglini, L., Fregosi, C., Pe, S., Parimbelli, E., La Maida, G. A., and Gallazzi, E. 2025. "From Oracular to Judicial: Enhancing Clinical Decision Making through Contrasting Explanations and a Novel Interaction Protocol," *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 745-754.

Castelo, N. 2024. "Perceived Corruption Reduces Algorithm Aversion," *Journal of Consumer Psychology* (34:2), pp. 326-333.

Castelo, N., Bos, M. W., and Lehmann, D. R. 2019. "Task-Dependent Algorithm Aversion," *Journal of marketing research* (56:5), pp. 809-825.

Chacon, A., Kausel, E. E., Reyes, T., and Trautmann, S. 2025. "Preventing Algorithm Aversion: People Are Willing to Use Algorithms with a Learning Label," *Journal of Business Research* (187), p. 115032.

Chávez, D. G., Cloarec, J., and Meyer-Waarden, L. 2024. "Opening the Moral Machine's Cover: How Algorithmic Aversion Shapes Autonomous Vehicle Adoption," *Transportation Research Part A: Policy and Practice* (187), p. 104193.

Cheng, L., and Chouldechova, A. 2023. "Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control," *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1-27.

Commerford, B. P., Dennis, S. A., Joe, J. R., and Ulla, J. W. 2022. "Man Versus Machine: Complex Estimates and Auditor Reliance on Artificial Intelligence," *Journal of Accounting Research* (60:1), pp. 171-201.

Commerford, B. P., Eilifsen, A., Hatfield, R. C., Holmstrom, K. M., and Kinserdal, F. 2024. "Control Issues: How Providing Input Affects Auditors' Reliance on Artificial Intelligence," *Contemporary Accounting Research* (41:4), pp. 2134-2162.

CORE. 2023. "International Computing Research & Eductaion Conference Rankings." from https://portal.core.edu.au/conf-ranks/

Cummings, M. 2004. "Automation Bias in Intelligent Time Critical Decision Support Systems," *AIAA 1st Intelligent Systems Technical Conference*: American Institute of Aeronautics and Astronautics.

Dennis, A. R., Lakhiwal, A., and Sachdeva, A. 2023. "Ai Agents as Team Members: Effects on Satisfaction, Conflict, Trustworthiness, and Willingness to Work With," *Journal of Management Information Systems* (40:2), pp. 307-337.

Dietvorst, B. J., and Bharti, S. 2020. "People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error," *Psychological science* (31:10), pp. 1302-1314.

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), p. 114.

Dietvorst, B. J., Simmons, J. P., and Massey, C. 2018. "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them," *Management science* (64:3), pp. 1155-1170.

Dijkstra, J. J. 1999. "User Agreement with Incorrect Expert System Advice," *Behaviour & Information Technology* (18:6), pp. 399-411.

Dijkstra, J. J., Liebrand, W. B., and Timminga, E. 1998. "Persuasiveness of Expert Systems," *Behaviour & Information Technology* (17:3), pp. 155-163.

Downen, T., Kim, S., and Lee, L. 2024. "Algorithm Aversion, Emotions, and Investor Reaction: Does Disclosing the Use of Ai Influence Investment Decisions?," *International Journal of Accounting Information Systems* (52), p. 100664.

Eastwood, J., Snook, B., and Luther, K. 2012. "What People Want from Their Professionals: Attitudes toward Decision-Making Strategies," *Journal of Behavioral Decision Making* (25:5), pp. 458-468.

Feng, X., and Gao, J. 2020. "Is Optimal Recommendation the Best? A Laboratory Investigation under the Newsvendor Problem," *Decision Support Systems* (131), p. 113251.

Ganbold, O., Rose, A. M., Rose, J. M., and Rotaru, K. 2022. "Increasing Reliance on Financial Advice with Avatars: The Effects of Competence and Complexity on Algorithm Aversion," *Journal of Information Systems* (36:1), pp. 7-17.

Germann, M., and Merkle, C. 2023. "Algorithm Aversion in Delegated Investing," *Journal of Business Economics* (93:9), pp. 1691-1727.

Gill, A., Gillenkirch, R. M., Ortner, J., and Velthuis, L. 2024. "Dynamics of Reliance on Algorithmic Advice," *Journal of Behavioral Decision Making* (37:4), p. e2414.

Goddard, K., Roudsari, A., and Wyatt, J. C. 2012. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," *Journal of the American Medical Informatics Association* (19:1), pp. 121-127.

Goddard, K., Roudsari, A., and Wyatt, J. C. 2014. "Automation Bias: Empirical Results Assessing Influencing Factors," *International journal of medical informatics* (83:5), pp. 368-375.

Harzing, A. W. 2024. "Journal Quality List (Jql)." *71st Edition*, from https://harzing.com/resources/journal-quality-list

Horowitz, M., and Kahn, L. 2023. "Bending the Automation Bias Curve: A Study of Human and Ai-Based Decision Making in National Security Contexts. Arxiv. Org."

Hou, Y. T.-Y., and 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), pp. 1-25.

Hund, A., Wagner, H.-T., Beimborn, D., and Weitzel, T. 2021. "Digital Innovation: Review and Novel Perspective," *The Journal of Strategic Information Systems* (30:4), p. 101695.

Jain, R. M., Garg, N., and Khera, S. N. 2025. "Adaption and Validation of the Algorithm Aversion Scale and Its Relationship with Neuroticism and Trust," *Journal of Organizational Change Management* (38:2), pp. 458-470.

Jenkin, T., Kelley, S., Ovchinnikov, A., and Ying, C. 2024. "Explanation Seeking and Anomalous Recommendation Adherence in Human-to-Human Versus Human-to-Artificial Intelligence Interactions," *Decision Sciences* (55:6), pp. 653-668.

Jussupow, E., Benbasat, I., and Heinzl, A. 2020. "Why Are We Averse Towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion,").

Jussupow, E., Benbasat, I., and Heinzl, A. 2024. "An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making," *MIS Quarterly* (48:4).

Keppeler, F. 2024. "No Thanks, Dear Ai! Understanding the Effects of Disclosure and Deployment of Artificial Intelligence in Public Sector Recruitment," *Journal of Public Administration Research and Theory* (34:1), pp. 39-52.

Koo, J. 2024. "Ai Is Not Careful: Approach to the Stock Market and Preference for Ai Advisor," *International Journal of Bank Marketing* (42:7), pp. 2117-2142.

Leidner, D. E. 2018. "Review and Theory Symbiosis: An Introspective Retrospective," *Journal of the Association for Information Systems* (19:6), p. 1.

Li, S., Huang, X., Sheng, Y., and Chen, K. 2025. "The Interactive Effect of Recommendation Subjects and Message Types on Consumers' Suboptimal Food Purchase Intentions," *Journal of Retailing and Consumer Services* (84), p. 104200.

Liu, N. T. Y., Kirshner, S. N., and Lim, E. T. 2023. "Is Algorithm Aversion Weird? A Cross-Country Comparison of Individual-Differences and Algorithm Aversion," *Journal of Retailing and Consumer Services* (72), p. 103259.

Logg, J. M., Minson, J. A., and Moore, D. A. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes* (151).

Longoni, C., Bonezzi, A., and Morewedge, C. K. 2020. "Resistance to Medical Artificial Intelligence Is an Attribute in a Compensatory Decision Process: Response to Pezzo and Beckstead (2020)," *Judgment and Decision Making* (15:3), pp. 446-448.

Longoni, C., Cian, L., and Kyung, E. 2022. "Artificial Intelligence in the Government: Responses to Failures and Social Impact," in: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. Oxford, United Kingdom: Association for Computing Machinery, p. 446.

Lyell, D., and Coiera, E. 2017. "Automation Bias and Verification Complexity: A Systematic Review," *Journal of the American Medical Informatics Association* (24:2), pp. 423-431.

Magni, F., Park, J., and Chao, M. M. 2024. "Humans as Creativity Gatekeepers: Are We Biased against Ai Creativity?," *Journal of Business and Psychology* (39:3), pp. 643-656.

Mahmud, H., Islam, A. N., 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), p. 121390.

Meehl, P. E. 1954. "Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence,").

Morewedge, C. K. 2022. "Preference for Human, Not Algorithm Aversion," *Trends in Cognitive Sciences* (26:10), pp. 824-826.

Mosier, K. L., and Skitka, L. J. 1996. "Human Decision Makers and Automated Decision Aids: Made for Each Other?,").

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., and Gogate, V. 2021. "Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable Ai Systems," *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 340-350.

Petropoulos, F., Fildes, R., and Goodwin, P. 2016. "Do 'Big Losses' in Judgmental Adjustments to Statistical Forecasts Affect Experts' Behaviour?," *European Journal of Operational Research* (249:3), pp. 842-852.

Prahl, A., and Van Swol, L. 2017. "Understanding Algorithm Aversion: When Is Advice from Automation Discounted?," *Journal of Forecasting* (36:6), pp. 691-702.

Rebholz, T. R., Biella, M., and Hütter, M. 2024. "Mixed-Effects Regression Weights for Advice Taking and Related Phenomena of Information Sampling and Utilization," *Journal of Behavioral Decision Making* (37:2), p. e2369.

Reich, T., Kaju, A., and Maglio, S. J. 2023. "How to Overcome Algorithm Aversion: Learning from Mistakes," *Journal of Consumer Psychology* (33:2), pp. 285-302.

Rix, J., Berger, B., Hess, T., and Rzepka, C. 2025. "The Algorithm Discount: Explaining Consumers' Valuation of Human-Versus Algorithm-Created Digital Products," *Journal of Management Information Systems* (42:2), pp. 633-668.

Sachin, P. K., and Schecter, A. 2024. "Advice Utilization in Combined Human-Algorithm Decision-Making: An Analysis of Preferences and Behaviors," *Journal of the Association for Information Systems* (25:6), pp. 1439-1465.

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., and Höllerer, T. 2019. "I Can Do Better Than Your Ai: Expertise and Explanations," *Proceedings of the 24th international conference on intelligent user interfaces*, pp. 240-251.

Shariff, A., Bonnefon, J.-F., and Rahwan, I. 2021. "How Safe Is Safe Enough? Psychological Mechanisms Underlying Extreme Safety Demands for Self-Driving Cars," *Transportation research part C: emerging technologies* (126), p. 103069.

Silber, D., Hoffmann, A., and Belli, A. 2025. "Embracing Ai Advisors for Making (Complex) Financial Decisions: An Experimental Investigation of the Role of a Maximizing Decision-Making Style," *International Journal of Bank Marketing* (43:6), pp. 1325-1346.

SimanTov-Nachlieli, I. 2025. "More to Lose: The Adverse Effect of High Performance Ranking on Employees' Preimplementation Attitudes toward the Integration of Powerful Ai Aids," *Organization Science* (36:1), pp. 1-20.

Skitka, L. J., Mosier, K. L., and Burdick, M. 1999. "Does Automation Bias Decision-Making?," *International Journal of Human-Computer Studies* (51:5), pp. 991-1006.

Suddaby, R. 2010. "Editor's Comments: Construct Clarity in Theories of Management and Organization." Academy of Management Briarcliff Manor, NY, pp. 346-357.

Talebi, A., Mukherjee, S., Gera, N., Kaur, K., and Das, G. 2025. "Unveiling Coping Mechanisms in Marketplace Discrimination: The Allure of Artificial Intelligence Recommendations," *Journal of Product Innovation Management*).

Templier, M., and Pare, G. 2018. "Transparency in Literature Reviews: An Assessment of Reporting Practices across Review Types and Genres in Top Is Journals," *European Journal of Information Systems* (27:5), pp. 503-550.

Templier, M., and Paré, G. 2015. "A Framework for Guiding and Evaluating Literature Reviews," *Communications of the Association for Information Systems* (37:1), p. 6.

Tse, T. T. K., Hanaki, N., and Mao, B. 2024. "Beware the Performance of an Algorithm before Relying on It: Evidence from a Stock Price Forecasting Experiment," *Journal of Economic Psychology* (102), p. 102727.

Turel, O., and Kalhan, S. 2023. "Prejudiced against the Machine? Implicit Associations and the Transience of Algorithm Aversion," *Mis Quarterly* (47:4).

Vial, G. 2019. "Understanding Digital Transformation:: A Review and a Research Agenda,").

Wang, L., Li, X., Zhu, H., and Zhao, Y. 2024a. "A Dimensional Exploration and Scale Development Study of Algorithm Aversion," *Journal of the Operational Research Society*), pp. 1-22.

Wang, X., Rodrigues, V. S., Demir, E., and Sarkis, J. 2024b. "Algorithm Aversion During Disruptions: The Case of Safety Stock," *International Journal of Production Economics* (278), p. 109442.

Webster, J., and Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review," *MIS quarterly*), pp. xiii-xxiii.

Wolfswinkel, J. F., Furtmueller, E., and Wilderom, C. P. 2013. "Using Grounded Theory as a Method for Rigorously Reviewing Literature," *European journal of information systems* (22:1), pp. 45-55.

Wu, W., Huang, Y., and Qian, L. 2024. "Social Trust and Algorithmic Equity: The Societal Perspectives of Users' Intention to Interact with Algorithm Recommendation Systems," *Decision Support Systems* (178), p. 114115.

Xie, Z., Yu, Y., Zhang, J., and Chen, M. 2022. "The Searching Artificial Intelligence: Consumers Show Less Aversion to Algorithm-Recommended Search Product," *Psychology & Marketing* (39:10), pp. 1902-1919.

Xu, M., and Wang, Y. 2024. "Explainability Increases Trust Resilience in Intelligent Agents," *British Journal of Psychology*).

You, S., Yang, C. L., and Li, X. 2022. "Algorithmic Versus Human Advice: Does Presenting Prediction Performance Matter for Algorithm Appreciation?," *Journal of Management Information Systems* (39:2), pp. 336-365.

Eidesstattliche Erklärung

Ich versichere, dass ich die vorliegende Abschlussarbeit selbständig angefertigt, nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe und die Überprüfung mittels Anti-Plagiatssoftware dulde.

Ich versichere außerdem, dass ich beim Einsatz von IT-/KI-gestützten Schreibwerkzeugen (falls zulässig) diese Werkzeuge in einem Textabschnitt namens "Übersicht verwendeter Hilfsmittel" mit ihrem Produktnamen, meiner Bezugsquelle und einer Übersicht des im Rahmen dieser Abschlussarbeit genutzten Funktionsumfangs vollständig aufgeführt habe.