U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

# EVALUATING INNOVATION WITH AI: A BEHAVIORAL ANALYSIS OF DECISION-MAKING IN STARTUP ASSESSMENTS

**Ana Sofia Costa**

Dissertation

Master's in Innovation and Technological Entrepreneurship

Supervised by

**Aurora Teixeira**

**Jacqueline Lane**

2024

## Acknowledgements

I would like to express my deepest gratitude to everyone who supported me throughout the development of this dissertation.

To my academic advisor at FEUP, Professor Aurora Teixeira, for her trust, encouragement, and guidance throughout this journey. Her support helped me believe in the value of this work and pushed me to pursue the most ambitious version of it.

To Professor Jacqueline Lane and the Laboratory for Innovation Science at Harvard (LISH), for welcoming me as a visiting researcher and for providing the research environment that made this dissertation possible, as well as the idea for a topic I truly loved working on. A special thanks to Professor Léonard Boussioux, for providing access to the data, for offering invaluable feedback along the way, and for encouraging me to aim for both rigor and originality.

I also thank the University of Harvard, and especially Harvard Business School, where the research group was based and where I was warmly welcomed during my stay.

My appreciation extends to the LISH community and my research peers, whose inspiring discussions and thoughtful questions enriched this work, and whose participation in the evaluation of the framework was essential to its final validation stage.

I gratefully acknowledge the Fulbright Commission and the Luso-American Development Foundation (FLAD) for supporting my research stay in the United States and making this international experience possible.

Finally, to my family and friends, thank you for your unconditional support and belief in me, especially during the most demanding phases of this journey.

## Abstract

This dissertation investigates how artificial intelligence (AI) influences human decision-making behavior during startup idea evaluations, with a focus on developing a methodology to detect one of the risks of this collaboration: the potential for overreliance on AI. The study is based on a dataset provided through a research affiliation with the Laboratory for Innovation Science at Harvard (LISH), in which students from the University of Washington evaluated startup applications submitted to the MIT Solve Global Health Equity Challenge, where three conditions were randomly assigned: No AI, Black Box AI (binary recommendation), and Narrative AI (recommendation with explanation). Two types of data were extracted from screen recordings of the evaluation sessions: session level behavioral data (clicks, scrolls, tab switches, hovers) and decision data (final answer, confidence). These inputs were used not only to examine behavioral shifts across AI conditions, but also to support the creation of an innovative framework to detect overreliance. As startup evaluation is a task with subjective correctness, traditional operationalizations of overreliance based on disagreement with ground truth are inadequate. This work instead proposes a novel behavioral operationalization in which overreliance is identified when evaluators follow AI recommendations while displaying low cognitive effort. The framework computes a score for each AI-assisted evaluation based on its projection along a fixed direction manually defined to represent low effort behavior, using predefined weights for each interaction feature. The development process is thoroughly documented across five refinement iterations. Results of this investigation indicate that AI assistance was associated with evaluations that were shorter in duration, with reduced maximum scroll depth, scroll frequency, and number of tabs opened. This signals lower cognitive effort and engagement. Additionally, providing explanations did not prove to significantly alter evaluators' reliance on AI compared to simply presenting the recommendation. The final version of the framework demonstrates robustness and good alignment with human judgments of evaluator overreliance. Ultimately, this work contributes a generalizable approach to understanding responsible human AI collaboration in contexts of high subjectivity.

**Keywords:** overreliance detection, human-AI collaboration, startup idea evaluation, behavioral interaction data, AI-assisted decision making

## Resumo

Esta dissertação investiga de que forma a inteligência artificial (IA) influencia o comportamento humano na tomada de decisão durante a avaliação de ideias de startups, com especial foco no desenvolvimento de uma metodologia para detetar um dos riscos associados a esta colaboração: o potencial de dependência excessiva da IA. O estudo baseia-se num conjunto de dados obtido através de uma afiliação de investigação com o Laboratory for Innovation Science at Harvard (LISH), no qual estudantes da University of Washington avaliaram candidaturas submetidas ao MIT Solve Global Health Equity Challenge, sob três condições atribuídas aleatoriamente: No AI, Black Box AI (recomendação binária) e Narrative AI (recomendação com explicação). A partir das gravações de ecrã das sessões de avaliação, foram extraídos dois tipos de dados: dados comportamentais ao nível da sessão (cliques, scrolls, mudanças de separador, hovers) e dados de decisão (resposta final, confiança). Estes dados permitiram não só analisar alterações comportamentais entre as diferentes condições de IA, como também suportar a criação de uma framework inovadora para detetar dependência excessiva. Dado que a avaliação de startups é uma tarefa com correção subjetiva, abordagens tradicionais baseadas em discordância com a resposta correta revelam-se inadequadas. Este trabalho propõe, assim, uma operacionalização comportamental alternativa, em que a dependência excessiva é identificada quando o avaliador segue a recomendação da IA demonstrando simultaneamente baixo esforço cognitivo. A framework calcula um score para cada avaliação assistida por IA com base na projeção dessa avaliação ao longo de uma direção fixa, manualmente definida para representar comportamentos de baixo esforço, utilizando pesos predefinidos para cada variável de interação. O processo de desenvolvimento é documentado em cinco iterações de refinamento. Os resultados desta investigação indicam que a assistência por IA foi associada a avaliações de menor duração, com uma redução da profundidade máxima de scroll, da frequência de scroll e o número de separadores abertos, sinalizando um menor esforço cognitivo e envolvimento. Além disso, a disponibilização de explicações não alterou significativamente a dependência dos avaliadores em relação à IA quando comparada com a simples apresentação da recomendação. A versão final da framework revela-se robusta e bem alinhada com os julgamentos humanos sobre casos de dependência excessiva. Em última análise, este trabalho contribui com uma abordagem generalizável para compreender uma colaboração responsável entre humanos e IA em contextos de elevada subjetividade.

**Palavras-chave:** deteção de dependência excessiva, colaboração humano-IA, avaliação de ideias de startups, dados de interação comportamental, tomada de decisão assistida por IA.

# Index

## Index of Tables

## Index of Figures

# 1. Introduction

## 1.1. Problem Statement

The concept of idea evaluation entails a systematic assessment of potential business opportunities with the goal of ascertaining their feasibility and likelihood of success (Scheaf et al., 2019). The traditional process of evaluating startup ideas is heavily reliant on human judgment, making it inherently inconsistent and subjective (Huang & Pearce, 2015). This inconsistency arises because each evaluator's decision-making processes differ, assigning varying levels of importance to different evaluation factors. Additionally, the decision-making process can be influenced by biases unrelated to the idea's intrinsic merit, such as the evaluator's experience or their subjective familiarity with the idea (Li, 2017; Greul et al., 2023). These biases can undermine the goal of accurately assessing an idea's potential, potentially leading to the undervaluation or exclusion of innovative concepts. Consequently, they not only affect the fairness of evaluations but also restrict the diversity of ideas receiving support (Greul et al., 2023), ultimately limiting the innovative capacity of entrepreneurial ecosystems.

Considering these challenges, integrating artificial intelligence (AI) into the evaluation process emerges as a promising avenue to enhance objectivity and consistency. AI systems possess the capability to analyze vast amounts of data and assess ideas based on predefined criteria in a systematic and replicable manner (Dwivedi et al., 2021). By providing data-driven insights and evaluations grounded in objective metrics, AI can support decision-makers in identifying high-potential opportunities that might otherwise be overlooked due to subjective judgments.

Despite these advantages, reliance on AI introduces distinct risks. Blind acceptance of AI recommendations can diminish human critical thinking and autonomy in decision-making. While there is substantial literature on using AI in various decision-making contexts, few studies investigate behavioral shifts when AI delivers direct pass or fail recommendations in startup screening. This gap is further magnified by the fact that existing research tends to focus on model accuracy or predictive performance, leaving open the question of how evaluators' decision-making patterns shift in the presence of AI. Understanding whether evaluators' critical thinking diminishes or whether they engage in low-effort acceptance of AI outputs remains an underexplored dimension.

### 1.2. Motivation and Significance

To address the problem outlined above, this research introduces a structured methodology for analyzing evaluator behavior through interaction data captured from session recordings. The framework is designed to identify cases where evaluators align with AI recommendations while displaying behavioral patterns associated with low cognitive engagement. This provides a practical lens for interpreting how evaluators interact with AI during startup assessments and detecting potential instances of overreliance.

Through this methodological contribution and practical tool, the study seeks to address both conceptual and methodological gaps in the literature, providing new insights into how AI influences human decision-making behavior in entrepreneurial contexts. Ultimately, this research aims to support the responsible integration of AI in idea evaluation systems, helping to ensure that AI complements, rather than replaces, human judgment.

### 1.3. Scope of the Work

This research introduces a methodological framework to analyze evaluator behavior through session-level interaction logs as they assess startup applications with varying types of AI assistance. The artifact is designed to detect potential overreliance on AI when evaluators are exposed to AI outputs. This offers a comprehensive approach to studying human-AI interaction in evaluative contexts.

The study is guided by the primary research question: How does AI assistance influence the behaviors of evaluators in idea evaluation? This question seeks to understand if and how AI recommendations impact evaluators' decision-making patterns. By using interaction analysis based on session recordings, the study will capture behavioral shifts that may occur when AI is introduced into the evaluation process.

To deepen this analysis, two sub-questions are posed:

1. How does the type of AI input (simple recommendation versus recommendation with explanation) affect evaluators' reliance on AI? This subquestion explores whether different levels of AI input influenced how much evaluators relied on AI, and whether the presence of explanations amplified that effect.
2. How does the presence of AI support contribute to the potential for overreliance in this context? While agreement with AI does not necessarily imply low-quality decision-making, this subquestion examines whether evaluators are more likely to

accept AI recommendations without strong behavioral evidence of independent analysis.

Through these questions, the methodology will provide practical insights into the behavioral impacts of AI assistance, helping to ensure that AI supports rather than overshadows human judgment in startup evaluations.

### 1.4. Overview of Methods & Anticipated Findings

This dissertation adopted an interaction-based approach to examine how AI assistance influences evaluator behavior during idea evaluation tasks. By analyzing session recordings, the study introduced a structured framework to identify potential overreliance on AI, defined as the combination of behavioral passivity and agreement with AI recommendations. The approach is grounded in behavioral theory and was iteratively refined to ensure that the patterns it detects are both meaningful and interpretable.

Findings from the application of this approach to a dataset of 555 evaluations suggest that, in this context, AI assistance was associated with reduced behavioral indicators of cognitive effort. However, reliance did not significantly differ between AI recommendation formats, indicating that the addition of justifications did not substantially alter evaluators' reliance on the system.

The behavioral framework successfully distinguished between interaction profiles across experimental groups and identified sessions with high-risk patterns of low-effort agreement. Its output showed strong alignment with human judgments of passivity and revealed considerable variability in overreliance scores among AI-Followed sessions. These results suggest that while AI support can increase the likelihood of passive agreement, overreliance is not a uniform outcome and depends on how evaluators engage with the system.

Theoretically, the research contributes to the understanding of how AI shapes human decision-making behavior in high-subjectivity contexts. It builds on existing literature on automation bias by demonstrating how the presence and type of AI input, such as recommendations with or without explanations, affect evaluators' behavioral engagement with the task. These findings extend current models of human-AI interaction by highlighting not just whether evaluators follow AI, but how they do so and with what degree of engagement. Rather than framing AI agreement as inherently problematic, this work proposes a new operationalization of overreliance as the co-occurrence of low-effort

behavior and agreement with AI recommendations. This operationalization offers a novel lens through which to study overreliance during AI-assisted evaluation, even in the absence of ground-truth correctness.

Practically, the work offers a generalizable method for organizations and researchers seeking to monitor reliance behaviors in AI-supported environments. The proposed framework can help detect when evaluators may be depending on AI recommendations with insufficient interaction or critical review, especially in contexts where thoughtful decision-making is expected. Additionally, the findings will provide actionable insights into how different types of AI input affect user behavior, offering guidance for designing future decision-support systems that promote thoughtful human-AI collaboration.

### 1.5. Structure

This dissertation is organized into an Introduction, three chapters, and a Conclusion. Each chapter and conclusion are built on previous chapters to provide a clear exploration of the research topic. The Introduction outlines the problem, motivation, objectives, and scope of the study. It introduces the research questions and situates the work within the broader context of AI in startup idea evaluation, emphasizing the importance of understanding behavioral impacts such as overreliance on AI. The literature review (Chapter 2) examines existing research relevant to the study. It is divided into two sections: the first (Section 2.1.) explores the role of AI in idea evaluation and the broader concept of overreliance on AI, while the second (Section 2.2.) focuses on methodologies for analyzing behavioral changes in human-computer interactions, particularly using video analysis. This section identifies gaps in the literature and establishes the foundation for the research. Chapter 3, Research Design, describes the research design and the development of the artifact, a methodology based on session-level interaction data captured through screen recordings, used to study evaluator behaviors during AI-assisted idea evaluation. It details how the artifact will be constructed, tested, and iterated upon. Chapter 4, Discussion, interprets results from pilot or preliminary analyses, relating findings to the literature on overreliance and highlighting potential implications for startup evaluators. In the Conclusion, the key findings are summarized, reflected on their implications for both theory and practice, and point to future research directions for AI-assisted startup evaluation.

# 2. Literature Review

## 2.1. Initial considerations

This chapter presents the literature review, outlining the searches conducted, the methodology for selecting relevant papers, and a synthesis of key themes. By examining existing theories, empirical research, and practical findings, this review aims to provide a comprehensive theoretical framework for the research topic addressed in this dissertation.

The review is structured in two main sections. The first section (Section 2.2.) explores the research context, focusing on the role of artificial intelligence (AI) in the idea evaluation process within entrepreneurship and the broader concept of overreliance on AI. It examines how AI tools are increasingly integrated into the evaluation of startup ideas, while also addressing the potential risks posed by overreliance on AI, a phenomenon studied more broadly in human-AI interaction research. While overreliance has not been extensively explored within the specific domain of idea evaluation, understanding its mechanisms is critical for this study. By synthesizing insights from related fields, this section establishes a theoretical basis for investigating how AI's influence might inadvertently diminish evaluators' critical thinking and autonomy when assessing the viability of startup ideas. The second section (Section 2.3) shifts to the methodological core of the study, reviewing techniques for analyzing behavioral changes through screen-recorded interaction data. This is essential for developing the artifact that will guide this research: a methodology based on interaction data from session recordings, used to detect evaluator behaviors such as overreliance and reductions in critical engagement during AI-assisted decision-making. By synthesizing existing approaches and highlighting gaps in the literature, this section provides a foundation for designing the artifact and aligning it with the study's objectives.

This dual approach ensures that the review not only contextualizes the relevance of AI in idea evaluation but also narrows the focus to the technical and methodological tools required to analyze behavioral changes. In doing so, this chapter sets the stage for the research methodology and contributes to a deeper understanding of how evaluators interact with AI systems in entrepreneurial settings.

## 2.2. Contextual Themes

The contextual themes of this research explore the broader theoretical background necessary to understand the role of artificial intelligence (AI) in the startup idea evaluation process

(Section 2.2.1.) and the associated behavioral phenomena such as overreliance on AI. (Section 2.2.2.).

### 2.2.1 AI in Idea Evaluation

To understand the previously studied applications of Artificial Intelligence (AI) in the process of evaluating entrepreneurial ideas, a structured bibliometric search and analysis were conducted.

At the beginning of the search, a broad approach was undertaken by using general terms such as "artificial intelligence" and "entrepreneur*" as keywords on the search strings. This phase aimed to provide an overview of AI's application in entrepreneurship and to understand if the evaluation of startup ideas was a prevalent topic. While it provided valuable insights, it did not yield substantial findings directly addressing startup idea evaluation.

To overcome such a problem, the search strategy was refined in a second phase. This time, the focus was on the intersection of idea evaluation and artificial intelligence by incorporating more specific terms. The research query included multiple keywords that represented the concept of idea evaluation (such as idea assessment, startup evaluation, innovation evaluation, concept evaluation, and new product evaluation) alongside different facets of AI (including machine learning, neural networks, and large language models). Table A1 from appendix summarizes the search strings, the academic databases used, the number of results recorded, and the exact search dates, providing a clear overview of the search strategies employed.

A total of 686 records were retrieved from all the searches. After eliminating duplicate entries (86) and excluding articles with missing abstracts (18), 582 records were left. These records were then screened by reviewing their titles and abstracts to ensure they were relevant. Articles that did not specifically focus on the application of AI in evaluating entrepreneurial ideas were removed from consideration. This thorough screening process resulted in 32 articles that were found to be directly relevant to this research. The selection process is illustrated in Figure 1, in a modified version of the PRISMA 2020 flow diagram (Page et al., 2021).

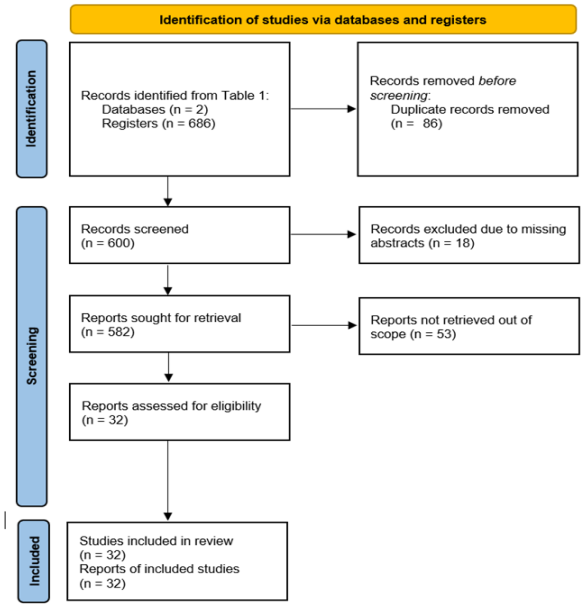


**Figure 1: PRISMA 2020 Flow Diagram for AI and Idea Evaluation Literature**
*Source:* Adapted from Page et al. (2021).

These selected studies span a publication period from 2013 to 2024, showcasing the evolution of research focus and methodologies in AI-driven entrepreneurial idea evaluation. Several key themes emerged, of which the most relevant to this study are detailed below.

### Predictive and Feasibility Models

Early research efforts were focused on creating predictive frameworks aimed at reducing uncertainty and improving the evaluation of entrepreneurial concepts in their early stages. Bayesian networks enhanced the accuracy of crowd-based evaluations (Burnap et al., 2013), while bio-artificial neural network simulators aided in predicting growth potential (Fedriani & Chaves-Maza, 2014). Later studies adopted more advanced machine learning and fuzzy logic techniques, allowing evaluators to systematically convert intangible investor criteria into decision rules (Arsenyan, 2017) or forecast startup success probabilities prior to market entry (Tomy & Pardede, 2017). As time progressed, the range of AI techniques, including neural networks and large language models, grew to provide increasingly context-sensitive feasibility assessments (Zhang et al., 2023).

### Scalability and Large-Scale Idea Evaluation

As the number of entrepreneurial ideas and related data has grown, research has addressed the challenge of evaluating a large volume of concepts. Neuro-fuzzy networks have been

effective in managing extensive collections of crowd-sourced ideas (Chang et al., 2016), and machine learning-based evaluation methods have in some cases outperformed time-limited human judgment in selecting high-quality concepts (Camburn et al., 2020). Researchers have combined various datasets, from knowledge bases to social media sentiment, to enhance the evaluation process (Akbari et al., 2020; Ferrati & Muffatto, 2020). Furthermore, frameworks that automate aspects of the venture capital workflow (Tewari et al., 2020) demonstrate how AI's scalability facilitates the identification of promising ideas within large repositories.

**Human-AI Collaboration and Hybrid Models**

A key theme emerging in recent research is the changing relationship between human evaluators and AI systems, with studies exploring how AI can support rather than replace human judgment. Mesbah et al. (2023) demonstrated that combining machine learning with crowd assessments improves consistency by modeling worker bias and reliability, reducing the need for extensive expert input. However, while this confirms AI's potential to enhance outcomes, it does not clarify how its presence affects evaluators' decision-making strategies or critical thinking. Similarly, Shaer et al. (2024) showed that integrating large language models into group ideation supports both idea generation and ranking, suggesting that human–AI collaboration may improve creative outcomes. Yet, across these studies, changes in evaluators' behavior between AI-supported and non-AI conditions are rarely examined.

### 2.2.2 Overreliance on AI in Decision-Making

This section addresses the important issue of excessive dependence on AI in decision-making. To investigate this phenomenon, a bibliometric search was conducted using Scopus and Web of Science. The search strategy combined keywords related to trust and reliance (e.g., "overreliance," "automation bias," "overtrust") with AI-related terms (e.g., "artificial intelligence," "large language models") and decision-making concepts (e.g., "judgment," "critical thinking"). Full search details are available in Appendix Table A2.

An initial dataset of 166 articles was compiled, which was subsequently narrowed down by removing duplicates (30) and articles lacking abstracts (5). Additionally, 16 studies were excluded because, although they explored the use of AI or the concept of bias, the focus of the studies were exclusively on algorithmic performance metrics or technical optimizations without addressing human behavioral aspects of overreliance or trust in AI-driven decision-

making. This left a total of 115 articles for further analysis. The selection process is depicted in Figure 2, in a modified version of the PRISMA 2020 flow diagram (Page et al., 2021).

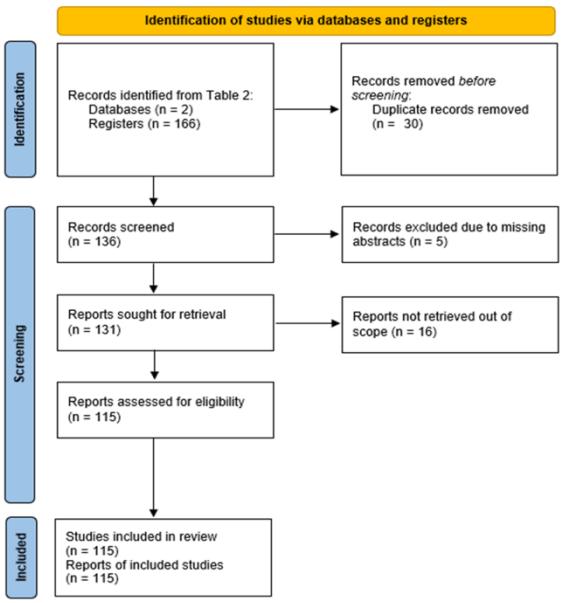


**Figure 2: PRISMA 2020 Flow Diagram for Overreliance on AI Literature**
*Source:* Adapted from Page et al. (2021).

These works collectively explore the phenomenon of overreliance on artificial intelligence (AI) in various sectors, including healthcare (Evans & Snead, 2024; Lee et al., 2024; Meyer et al., 2022; Gu et al., 2024; Frazer et al., 2024; Bouaud et al., 2015; Shafique et al., 2023), innovation-oriented decision-making (Keding & Meissner, 2021), security operations (Tilbury & Flowerday, 2024), financial or sales forecasting (Fahse & Schmitt, 2023), and human-computer interaction (He et al., 2023; Chiang & Yin, 2022). Despite the differing contexts, a common thread runs through them: as AI systems become more advanced and widespread, users frequently find it challenging to strike the right balance between trusting automated suggestions and exercising necessary critical judgment. The next paragraphs will address the key topics found pertinent to this dissertation.

## Defining Overreliance and Related Concepts

Some studies have implicitly treated trust and reliance as distinct constructs, where trust refers to the user's attitudinal stance toward the system, often assessed through self-reported measures, and reliance refers to the behavioral act of following AI recommendations (Bussone et al., 2015). Lu and Yin (2021) further highlight that trust and reliance do not always align, showing that users may rely on AI recommendations even when they report low trust.

From this distinction, the concept of overreliance emerges. It refers to the behavioral tendency to follow AI recommendations excessively or inappropriately, particularly when the advice is inaccurate or contextually unfit (Keding & Meissner, 2021; Klingbeil et al., 2024). Closely related concepts include automation bias, which encompasses both commission (acting on incorrect advice) and omission errors (failing to act when no alert is given) due to overdependence on AI (Parasuraman & Manzey, 2010; Vered et al., 2023; Goddard et al., 2014); complacency, understood as reduced attentiveness or critical monitoring when automation is perceived as reliable (Parasuraman & Manzey, 2010; Rodriguez et al., 2019); and deep automation bias, which captures a persistent, context-independent tendency to accept AI input without scrutiny (Strauß, 2021).

From the concept of overreliance and its overlapping constructs, the issue of low effort or reduced engagement emerges as a recurring concern. Keding and Meissner (2021) characterize overreliance as a passive behavioral tendency, often involving a lack of scrutiny toward AI recommendations. Klingbeil et al. (2024) similarly highlight that such behavior frequently arises under cognitive load or time pressure, conditions that reduce users' capacity for deliberate analysis. Vered et al. (2023) reinforce this view by linking automation bias to interface designs that promote shallow interaction and minimal user input. These patterns suggest that overreliance is not only a function of trust miscalibration but may also reflect a broader reduction in cognitive effort during AI-supported decision-making.

**Experimental Approaches and Behavioral Drivers of Overreliance**

Most existing operationalizations of overreliance are drawn from domains where task correctness can be explicitly defined. In such settings, researchers have measured the proportion of decisions aligned with incorrect AI advice (Keding & Meissner, 2021; Vered et al., 2023), or the degree of behavioral adjustment toward AI suggestions under varying levels of reliability (Klingbeil et al., 2024).

A growing body of work has explored how various factors shape reliance behavior. Explanations, in particular, have received considerable attention, with multiple types of AI-generated justifications being tested across different experimental settings. Vered et al. (2023) examined how local and global explanations affect users in two visual decision-making tasks. They found that neither type of explanation reduced participants' tendency to follow incorrect AI recommendations, and in some conditions, explanations increased this tendency, amplifying overreliance. In contrast, Lee and Chew (2023) investigated whether

counterfactual explanations, which describe how the AI's decision would change under different input conditions, could support more calibrated use of AI advice during rehabilitation assessment based on patient motion data. Their results showed that counterfactuals were more effective than salient feature explanations at helping users reject incorrect AI recommendations and better calibrate their trust. From another perspective, Cabitza et al. (2023) observed that explainable AI features, such as pixel attribution maps and textual justifications, could sometimes increase overreliance by inducing overtrust. These contrasting findings suggest that there is still no clear consensus on the impact of explanations, with some appearing to promote discernment while others may undermine the quality of human-AI collaboration.

Other system- and user-level factors have been studied such as uncertainty visualizations (Zhao et al., 2024), interface design (Raikwar et al., 2024), AI accuracy disclosure (He et al. 2023), and psychological traits (Küper & Krämer, 2024) to understand their impact on reliance behaviors. However, these individual characteristics are not modeled in the present work. These approaches nonetheless reflect the diversity of ongoing research and highlight the many variables that can influence reliance behavior.

**Proposed Strategies to Reduce Overreliance**

A variety of strategies have been proposed to mitigate overreliance, targeting both system level and user level interventions. On the system side, Fukuchi and Yamada (2023) proposed selectively presenting reliance calibration cues (RCCs) based on a cognitive model predicting user reliance behavior. Lee and Chew (2023) promote the use of counterfactual explanations to support more discerning acceptance or rejection of AI advice. Other approaches propose structural modifications to the decision process, including majority voting schemes that aggregate decisions from multiple humans assisted by AI to reduce overdependence on a single source (Gu et al. 2024) and frictional AI designs that introduce conflicting perspectives to prompt deeper engagement (Fregosi and Cabitza 2024). Lu et al. (2024) propose incorporating second opinions from either alternative models or human peers to foster reflective decision-making.

On the user side, interventions that aim to improve users' understanding of AI have also gained traction, with Chiang and Yin (2022) advocating for tutorials that build machine learning literacy and help users form more accurate mental models of AI systems.

In summary, these studies highlight that overreliance on AI in decision-making is a complex human-AI interaction issue shaped by multiple behavioral and contextual factors. Due to the diversity of domains in which this concern arises, it has been explored through a range of experimental and qualitative methods. As humans remain indispensable in decision-making loops, ensuring appropriate reliance on AI has become a central objective. As a result, several works propose strategies to reduce overreliance and support a more thoughtful and responsible use of AI.

### 2.3. A Review of the Methodologies for Behavioral Analysis

This section outlines the systematic process of selecting and reviewing literature to identify methodologies for analyzing behavioral changes through interaction data in human-computer and human-AI interaction contexts. These methodologies are foundational for designing the artifact that will guide this study's analysis.

#### 2.3.1 Research Paper Selection Process

The selection of relevant studies for analyzing behavioral changes in AI-assisted human-computer interaction was conducted through a structured and multi-stage approach to ensure comprehensive coverage and methodological relevance. The process began with the formulation of four distinct topic queries, each targeting a different methodological component necessary for analyzing user behavior and overreliance in decision-making settings.

The first topic focused on identifying studies that examined interaction features as proxies for cognitive effort, aiming to uncover how user actions such as clicks, scrolls, time on task, or mouse movement correlate with engagement and mental effort. The second topic explored the use of session log data to infer user states, concentrating on methods that process screen recordings or event streams to model attention, engagement, or fatigue. The third topic centered on scoring and quantification approaches, retrieving studies that develop composite indices or scoring systems to interpret behavioral signals in a structured way. Finally, the fourth topic targeted behavioral deviation and projection-based detection methodologies, with a particular focus on work that applies statistical distances, PCA, or vector projections to identify anomalies or behavioral shifts.

All searches were conducted using the Scopus database, yielding a total of 366 records across the four topics. Table A3 in the appendix provides a detailed overview of the queries used for each topic and their respective result counts.

Following the initial retrieval of articles, a rigorous screening process was undertaken to enhance the reliability and relevance of the selected studies. First, duplicates and entries with missing information were identified and removed, reducing the dataset to 363 records. Subsequently, these records were subjected to an abstract screening phase, where titles and abstracts were reviewed against predefined inclusion and exclusion criteria. Articles that did not specifically address methodologies for modeling user behavior from interaction data, particularly in the context of human-AI interaction, effort modeling, and decision-making support were excluded at this stage. This thorough screening process resulted in a total of 12 articles. The selection process is illustrated in Figure 3, an adaptation of the PRISMA 2020 flow diagram (Page et al., 2021).

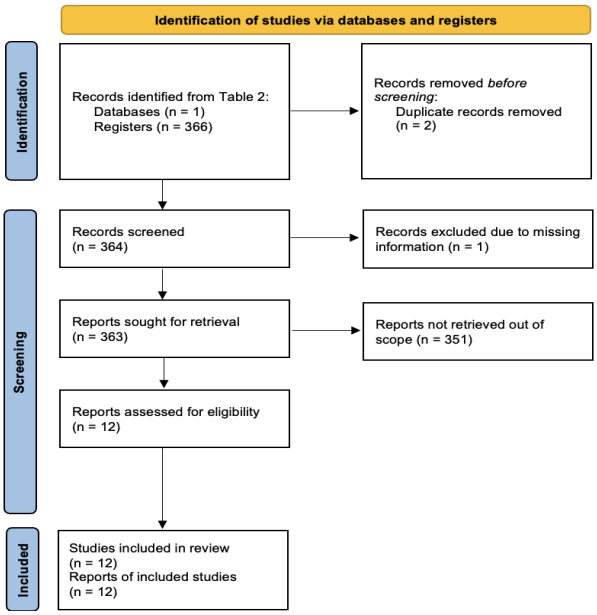


**Figure 3: PRISMA 2020 Flow Diagram for Methodologies for Behavioral Analysis**
*Source:* Adapted from Page et al. (2021).

These studies will be reviewed in detail in the following sections to identify methodologies that can effectively capture and interpret behavioral changes. This analysis will inform the methodological framework for examining evaluator behaviors in startup idea assessments.

### 2.3.2. Methodological Insights

Following a thorough review, this section synthesizes methodologies from a subset of the selected studies. These methodologies were identified as particularly relevant for developing the proposed artifact. Each study offers conceptual or technical guidance for interpreting user behavior in ways that align with the design of the behavioral scoring system introduced in this thesis. Table 1 summarizes these methodologies and outlines how their approaches inform various aspects of the artifact's behavioral modeling, including feature engineering, deviation scoring, and proxy selection for effort and engagement.

**Table 1: Summary of Relevant Behavioral Analysis Methodologies**

| Study | Methodology Summary | Contribution to This Study |
|---|---|---|
| Wei et al. (2022), "User Behavior Profile: A Key to Database Anomaly Access Detection" | Constructed user behavior profiles using features such as query type, access frequency, and time of access, derived from TPC-C and TPC-E datasets. Used K-means clustering (TPC-C) to group users by behavioral similarity and PCA (TPC-E) for dimensionality reduction. A Random Tree classifier was then trained to detect anomalies as deviations from normal profile patterns. | Demonstrates a pipeline for modeling user behavior from interaction logs and classifying anomalies based on deviation from clustered norms. Offers methodological precedent for using session-derived features to detect unusual or low-consistency behavioral patterns. |
| Carlton et al. (2019), "Inferring User Engagement from Interaction Data" | Collected self-reported engagement scores using User Engagement Scale and interaction data from users completing an interactive video-based origami task. Extracted 43 application-level features (e.g., button clicks, time spent, time between actions) and 87 low-level features (e.g., cursor speed, distance, pauses). Temporal pauses were categorized into short, medium, long, very long. Correlation and ANOVA were used to relate features to engagement levels. | Shows how session-level interaction data can differentiate levels of user engagement. Validates the use of input frequency metrics as behavioral indicators, which can inform models of user engagement intensity. |
| Carlton et al. (2021), "Using Interaction Data to Predict Engagement with Interactive Media" | Developed a methodology to predict user engagement by analyzing temporal features from interaction logs. Features included time spent, decision timing, and intervals between actions. Found that longer time durations were positively associated with higher engagement. Used Random Forest classifiers to distinguish between high, medium, and low engagement levels, and applied SHAP for feature importance analysis. | Provides strong evidence for using temporal features to quantify engagement. Supports the use of behavioral logging and predictive modeling to distinguish between higher- and lower-effort interactions. |
| Li & Zhan (2020), "Integrated infrared imaging techniques and | Introduced a multimodal interaction analysis framework for learning engagement classification, combining infrared facial expressions, mouse interaction data (including scroll activity, page navigation, reading speed, and dwell time), and LMS | Demonstrates how scroll activity and page navigation, as components of mouse interaction data, can serve as behavioral signals for engagement, supporting their inclusion in scoring-based evaluation frameworks. |

| Study | Methodology Summary | Contribution to This Study |
|---|---|---|
| multi-model information via convolution neural network for learning engagement evaluation" | logs (e.g., logins, content views, and participation). Each data stream was processed through convolutional neural networks, and their outputs were fused to predict engagement levels, which showed strong correlation with self-reported engagement scores. | |

Overall, the reviewed methodologies offer valuable strategies for capturing user engagement and effort through interaction data. They inform key design choices for this study's artifact, particularly in terms of feature selection, behavioral scoring, and classification techniques.

### 2.4. Research Gaps and Final Remarks

This chapter has provided a thorough review of the literature, synthesizing both contextual and methodological insights essential for addressing the research questions of this study. While the existing body of work offers a strong foundation for understanding overreliance on AI and behavioral modeling through interaction data, some important research gaps remain.

A central limitation of existing research lies in its focus on domains where task correctness can be objectively defined, such as healthcare or classification tasks. In these settings, overreliance is typically operationalized as agreement with incorrect AI outputs. However, this approach is not transferable to judgment-based tasks where correctness is subjective or undefined, such as startup idea evaluation. Consequently, there is a need to reconsider how overreliance is conceptualized and detected in such contexts. Rather than relying on outcome accuracy, alternative approaches could focus on behavioral indicators, capturing how users interact with AI and whether their decision strategies reflect reduced critical engagement.

In addition, although several studies propose proxies for user effort and engagement based on interaction data, few outline structured methodologies for scoring behavioral deviation in a continuous and interpretable manner. Techniques such as projection or distance-based modeling remain relatively underexplored in this context, particularly when applied to detecting subtle shifts in behavior induced by AI presence.

In response to these limitations, this study proposes a novel behavioral analytics artifact designed to detect overreliance in environments where correctness is undefined. It draws

upon the methodologies reviewed in this chapter, spanning feature extraction, behavioral scoring, and deviation modeling, to construct a framework grounded in empirical interaction data. By integrating insights from both contextual and methodological perspectives, this chapter lays the foundation for the development and application of the artifact, ensuring a novel contribution to understanding human–AI interaction in entrepreneurial evaluation settings.

# 3. Research Design

This chapter outlines the methodological approach employed to address the research questions presented in Chapter 1. The chapter is organized into three sections. Section 3.1. describes the methodological approach and its theoretical foundation. Section 3.2. describes the developed artifact and its components. Section 3.3. outlines the evaluation strategy used to assess the artifact's performance and contribution.

## 3.1. Research Method

This dissertation adopts the Design Science Research (DSR) methodology outlined by Hevner et al. (2004) and tailors it to the context of entrepreneurial idea evaluation augmented by artificial intelligence (AI). The approach is iterative by nature and emphasizes the creation of artifacts that address practical problems while contributing to theoretical knowledge, which makes it well-suited to the study's dual objective of developing a behavioral analytics framework for detecting potential behavioral overreliance on AI in subjective evaluation settings and using it as a lens to investigate how AI assistance influences human decision-making behavior.

The methodology is structured around three core components: the relevance environment, the rigor of the knowledge base, and the central design cycle of building and evaluating artifacts. Figure 4 illustrates the methodology specification for this dissertation.

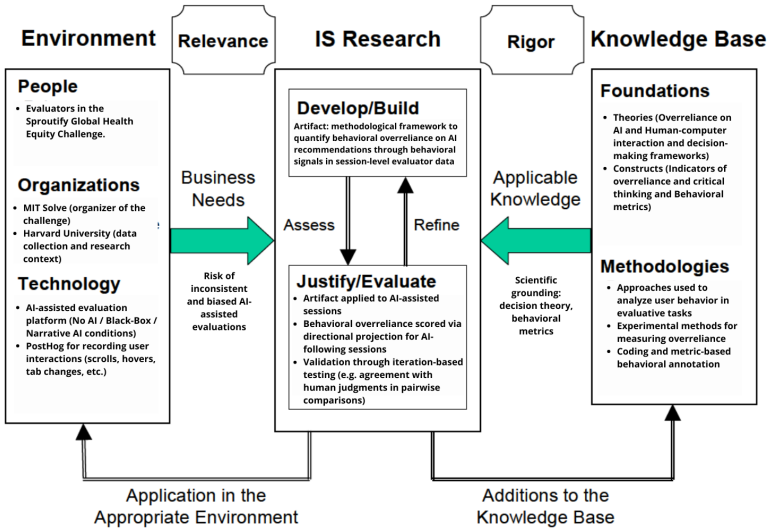


**Figure 4: Design Science Research Process for Developing and Evaluating the Artifact**
*Source:* Adapted from Hevner et al. (2004).

The Environment represents the real-world context in which the problem is situated. In this study, students served as evaluators in the MIT Solve Global Health Equity Challenge, where they assessed startup applications under different AI conditions. The evaluations were conducted through an AI-assisted platform, with three treatment groups: No AI, Black Box AI, and Narrative AI. User interactions were recorded using PostHog, providing the session-level behavioral data needed for artifact development.

The Knowledge Base includes theoretical and methodological foundations that informed both the construction and interpretation of the artifact. From a theoretical perspective, the study draws on prior work in automation bias, overreliance on AI, and decision-making in human-computer interaction contexts. Methodologically, the research is informed by previous work on analyzing behavior in evaluative tasks, coding screen interactions, and measuring trust and critical thinking through structured behavioral indicators.

At the core of the framework is the IS Research cycle, composed of iterative phases of building and evaluating the artifact. In the Build phase, a behavioral analytics framework was developed to detect potential behavioral overreliance on AI. This artifact processes session-level interaction logs and models behavioral deviation using directional projections aligned with low-effort behavior, benchmarked against a No-AI baseline. In the Justify/Evaluate phase, the framework was iteratively refined and validated through empirical testing of thresholding, baseline robustness, and behavior flagging consistency.

Together, these components ensure that the artifact is both theoretically grounded and empirically tested. The process enables practical application in the evaluation setting while contributing new methodological knowledge to the study of AI-human decision dynamics.

### 3.2. Artifact Description

This section presents the components and structure of the developed artifact. It includes a description of the data and inputs used, the process of constructing behavioral features, the internal logic that governs its operation, and the iterative refinements applied throughout its development.

#### Overview of the Artifact

The artifact developed in this study is a behavioral analytics framework that processes session-level interaction data captured from evaluators navigating a platform where AI

recommendations were either absent, binary (Black Box AI), or accompanied by explanations (Narrative AI). Its purpose is to analyze evaluator behavior and detect potential instances of overreliance and reductions in critical thinking when evaluators are exposed to AI suggestions.

The framework is grounded in an operational logic that defines overreliance as occurring when an evaluator agrees with the AI's recommendation while exhibiting behavioral signs of low cognitive effort. The artifact is designed to detect this pattern by combining decision alignment with the AI and indicators of minimal interaction.

To support this logic, both decision-context and behavioral features are extracted from the session logs. Behavioral features are standardized and compared against a reference baseline of unaided evaluations, and used to compute a behavioral overreliance score, derived from a directional projection aligned with behavioral patterns indicative of low cognitive effort.

The system operates through a modular pipeline that includes data preprocessing, feature construction, deviation scoring, and rule-based classification logic. While grounded in statistical and theoretical principles, the artifact is designed to be interpretable and adaptable across different types of AI-assisted evaluation contexts.

**Data Context and Collection Process**

The dataset analysed in this study was provided through a research affiliation with the Laboratory for Innovation Science at Harvard (LISH). It was collected during a classroom experiment conducted as part of the Generative AI for Business Applications course, taught by Professor Léonard Boussioux at the University of Washington. As part of the course, 95 students from the Master of Science in Information Systems (MSIS) participated in a decision-making activity that simulated the triage process of the MIT Solve Global Health Equity Challenge, a real competition in which startups submit proposals for innovative solutions to pressing global health issues.

The evaluations were conducted through a custom-built web application developed by MSIS alumni Ian Chen, Amy Wang, and Camila Lin. The platform replicated the official screening process used by MIT Solve and presented five core evaluation criteria:

1. Completeness and intelligibility of the application
2. Prototype stage verification

3. Alignment with the challenge question
4. Use of technology
5. Quality threshold for further review

Each student was expected to assess ten startup solutions, following two initial practice examples. The experiment followed a design in which each evaluation was randomly assigned to one of three experimental conditions:

- No AI: Evaluators made decisions independently, without any AI assistance.
- Black Box AI: Evaluators received a binary Pass/Fail recommendation from an AI system for the overall solution, as well as binary judgments for each of the five criteria.
- Narrative AI: Evaluators received the same overall and per-criterion Pass/Fail judgments, but in this case, each criterion-level decision was accompanied by a brief textual explanation generated by the AI.

The platform interface is illustrated in Figure 5. It is divided into three main areas: the Screening Criteria, the Human Screener Decision and the Solution Application.

The Screening Criteria area presents the evaluation dimensions previously described. As shown in Figure 5, these criteria may appear alone, alongside binary AI recommendations, or with AI-generated explanations, depending on the condition assigned.

The Solution Application area allows users to review information from the startup submission through multiple navigable tabs. The tabs labeled "All", "Criterion 1–5", and "Summary" contained startup-provided content. The "All" tab displayed the full application, while the individual Criterion tabs filtered content relevant to each criterion. A "Challenge Question" tab enabled evaluators to revisit the specific challenge prompt at any point.

The Human Screener Decision area is where users submitted their final decision: "Pass" if all five criteria were met, or "Fail" if any criterion was not satisfied. If "Fail" was selected, a dropdown menu prompted the evaluator to indicate which criterion had failed. In this same section, evaluators also rated their confidence in the decision on a 1–5 scale. Once completed, the evaluator advanced to the next case by clicking the Next button.

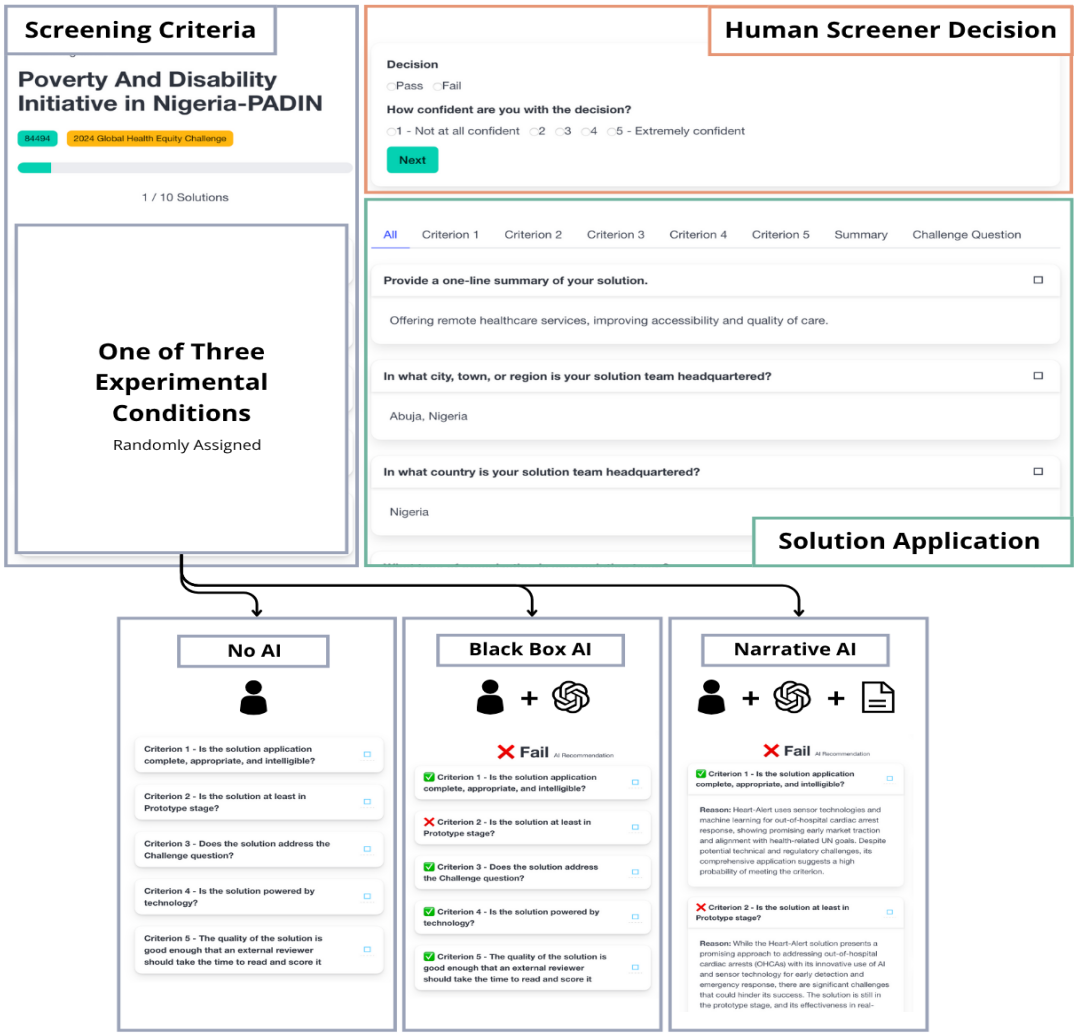


**Figure 5: Platform interface and overview of the three experimental conditions**

During the evaluation, user interactions were tracked using PostHog, a behavioral analytics tool embedded in the platform. PostHog recorded low-level event logs, including clicks, scrolls, hovers, mouse movements, tab switches, dropdown interactions, confidence selections, and precise timestamps.

Two types of access were available for analyzing evaluator behavior:

1. Video replays of each session on the PostHog interface;
2. Structured JSON logs, containing raw event data with consistent formatting for computational processing.

This dataset was made available through a research collaboration with Harvard University, under appropriate consent and ethical approval procedures, and served as the starting point for artifact development.

**Interaction Data Extraction**

The process of interaction data extraction was initiated with a manual, exploratory phase in which a subset of PostHog session videos was reviewed. This initial step allowed for direct observation of how evaluators interacted with the platform under different AI conditions, leading to the identification of potential behavioral signals such as time spent on the evaluation, number of tabs opened, response changes, and the timing of decision. These early insights provided the foundation for defining behavioral patterns and inspired the design of the automated extraction logic.

To scale the analysis, a fully automated extraction pipeline was developed using the structured JSON logs generated by PostHog. These logs captured low-level user interaction data with millisecond precision, including click events, scroll activity, hovers, input selections, and tab switches. At the beginning of each solution evaluation, the logs included a specific event containing the full HTML structure of the page along with the IDs of all visible interface components. This enabled the development of a dynamic parsing script to extract the identifiers of relevant interface elements such as the Pass/Fail buttons, confidence scale, Fail Reason dropdown, Next button, and criterion tabs to later associate clicks and hovers to these events. This allowed for robust element mapping across sessions, regardless of variation in ID structures.

The complete pipeline was modular and implemented entirely in Python, using libraries such as "pandas" for data manipulation, "json" for structured parsing, "re" for pattern matching, and "datetime" for precise timestamp handling. The code was distributed across multiple custom scripts that handled distinct components of the process, including scroll and hover tracking, click event classification, dynamic HTML parsing, tab and trust detection, solution segmentation, feature synthesis, and consistency validation. The system was iteratively refined over several development cycles to ensure accuracy and generalizability across sessions. Special attention was given to handling edge cases, including truncated sessions, missing events, or sessions involving multiple consecutive solutions, to ensure reliable and reproducible extraction.

Importantly, the automated pipeline enabled the discovery of behavioral dimensions that would have been infeasible to capture through manual video review alone. For example, exact hover durations over specific elements, precise scroll depth, and timestamp-based latency metrics were only identifiable through the log structure. This expanded the behavioral data available and enabled more granular modeling of user engagement and interaction effort during the evaluation process.

Although 95 evaluators participated and were each expected to assess ten solutions, the final dataset consisted of 555 valid solution-level evaluations. Several cases were excluded due to video truncation, incomplete or missing answers, or absence of a final "Next" action that marked session completion. Additionally, evaluations in which the participant navigated away from the evaluation window for more than five consecutive seconds were removed, based on the assumption that meaningful engagement with the solution was no longer occurring during that time.

The cleaned and structured interaction logs derived from this process served as the foundation for constructing session-level behavioral features, described in the following subsection.

### Feature Engineering

Based on the structured interaction logs obtained through the extraction process, a comprehensive set of features was engineered to represent evaluators' behavior and decision context during solution assessments. These features were grouped into two categories: behavioral features and decision-context features.

Behavioral features reflect how evaluators interacted with the interface. Raw behavioral metrics collected included session duration, the timestamp of the first clicked decision, the time between final answer and leaving the evaluation, hovering time on the sidebar (which included AI content), the opening of tabs with respective timestamps, click and scroll logs, time spent at the top of the page, and the maximum vertical scroll reached. To enrich this dataset, several derived variables were constructed, such as the number of returns to the top of the page, the attention ratio to the top of the page, the total number of clicks, the number of scrolls, and the number of tabs opened.

Decision-context features capture the evaluator's final choices and their relationship with AI input. These include the final answer, the answer recommended by AI, the criteria flagged as passed or failed by the AI, the chosen failure reason (if applicable), confidence scores, as well as the sequence of answers and confidence changes with their corresponding timestamps. From these variables, derived indicators were created, including the number of answer changes, the number of confidence changes, whether the selected fail reason matched one of the AI-failed criteria, and whether the final decision aligned with the AI's recommendation.

While all these features were extracted to support a detailed understanding of evaluator behavior, only a subset was incorporated into the artifact. The selection of these features was guided by the framework's operational logic of detecting potential overreliance by identifying cases where the evaluator follows the AI's recommendation while also exhibiting behavioral patterns indicative of low engagement. To ensure theoretical coherence, the behavioral variables included were those with consistent support in prior literature as proxies for user effort and engagement, as discussed in Section 2.3..

Accordingly, the final set of behavioral features integrated into the artifact includes: session duration, number of scrolls, number of clicks, number of tabs opened, attention ratio to the top of the page, and scroll depth. The only decision-context variable directly included in the deviation model was the binary "AI Followed" indicator, used to subset the data for score computation. The remaining features, while excluded from the artifact, are retained for further analysis in Chapter 4. A summary of the selected variables is presented in Table 2.

**Table 2: List of features**

| Feature Name | Description |
|---|---|
| AI Followed | Binary indicator equal to 1 if the evaluator's final decision matched the AI's overall recommendation, and, in the case of a Fail decision, the selected failure reason matched at least one of the AI-flagged criteria. |
| Duration (s) | Total time spent on the evaluation from load to submission |
| Number of Tabs Opened | Number of distinct tabs opened during the evaluation |
| Number of Scrolls | Number of scrolls during the evaluation |
| Number of Clicks | Number of clicks during the evaluation |
| Max Y Reached | Maximum vertical scroll position reached during the session |
| Attention Top Ratio | Proportion of time spent with the viewport focused on the top of the screen |

### Method Selection and Rationale

With the set of behavioral variables defined, the next step involved selecting an appropriate methodological approach to translate these features into a reliable and interpretable measure of overreliance. With the goal of operationalizing overreliance in subjective evaluation contexts, the central operationalization adopted in this framework was the dual condition of alignment with AI and low-effort interaction. To capture the low-effort component, three major methodological families were considered: supervised classification models, unsupervised clustering, and scoring-based approaches.

Supervised classification models (Carlton et al., 2021) and unsupervised clustering techniques (Wei et al., 2022) had been applied in prior work. Nevertheless, both were deemed unsuitable for the purpose and context of this study. Supervised models require labeled data to learn patterns of overreliance, which was not available in this dataset and, due to time constraints, was not collected. Clustering approaches, in turn, group sessions based on overall similarity across behavioral features. While they can reveal patterns or subtypes of interaction, these patterns are not necessarily aligned with the specific notion of low-effort behavior targeted in this study. As such, they offer limited explanatory value for this construct and lack the precision needed for targeted detection.

The scoring-based approach, although not proposed in the reviewed literature, was considered a promising response to the specific research objective. It provided a transparent, interpretable, and flexible way to translate continuous behavioral interaction data into a structured representation of overreliance risk, allowing the detection of the phenomenon directly through behavior. Within this family, both distance-based and directional approaches

were considered. Directional projection was adopted because it provided a conceptually precise way to quantify how closely a session aligned with the low-effort behavior pattern at the core of the framework. In contrast, general distance-based methods lacked this specificity, potentially flagging any form of deviation rather than behavior aligned with reduced engagement, which was the central target of this framework.

Ultimately, the scoring-based approach using directional projection was chosen. Unlike the other methods, it allowed the detection of potential overreliance based on the level of low-effort behavior, enabled more nuanced comparisons across sessions by producing an individual score for each case, and made the reasoning behind each evaluation interpretable through behavioral traceability.

**Framework Architecture**

The artifact operates as a modular behavioral analytics pipeline designed to detect potential overreliance on AI recommendations. Its structure integrates session-level behavioral data with decision alignment information and a directional deviation model grounded in cognitive effort theory. This subsection outlines the architecture of the system, from input to detection.

The framework takes as input a structured dataset where each row represents a single evaluation session and contains the precomputed values of a selected subset of features described in the previous section. These values are the direct input to the artifact and are used without further modification.

In the first step of the pipeline, sessions are filtered to retain only those in which the evaluator's final decision matches the AI recommendation (captured by the binary AI Followed feature). For these sessions, the continuous behavioral features are standardized using z-score normalization to ensure comparability across users. These standardized values are then passed into the deviation modeling component, which computes a projection along a direction vector that quantifies behavioral alignment with low cognitive effort.

To accommodate variations in how low-effort behavior may manifest across different decision outcomes, the framework applies outcome-specific direction vectors. These vectors were precomputed during the development phase of the artifact and were based on the dominant behavioral patterns observed within each outcome group. This vector approach was selected among other options due to its better alignment with human judgments of

evaluator overreliance. The resulting scalar value reflects the degree to which an evaluator's behavior corresponds to low-effort agreement with AI recommendations.

The framework outputs this projected behavioral deviation as a behavioral overreliance score, which reflects the evaluator's degree of behavioral alignment with low-effort agreement when following AI recommendations. Rather than applying a predefined threshold to flag sessions, the framework leaves the interpretation of the behavioral overreliance score to the user. However, the output can be flexibly operationalized through custom thresholds to identify sessions of interest, enabling real-world applications and providing flexibility for each use case.

This architecture enables consistent and interpretable quantification of behavioral overreliance across a large number of sessions, while remaining grounded in theoretical constructs, validated projection logic, and empirical behavioral data. The complete structure of the artifact, including the sequence of operations from input to output, is illustrated in Figure 6.

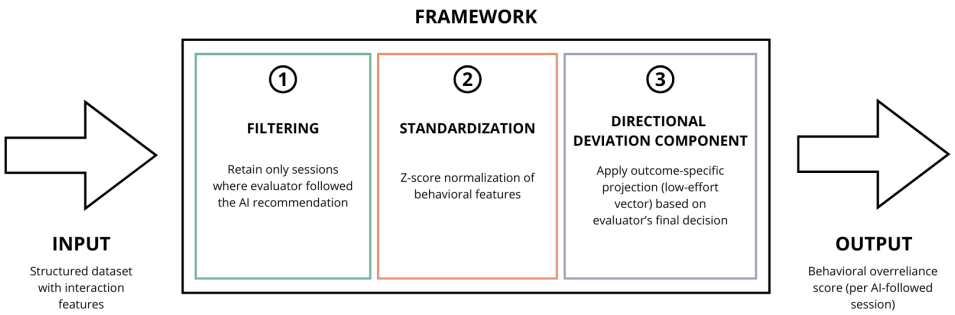


**Figure 6: Framework architecture for detecting overreliance on AI**
*Source*: Own elaboration.

## 3.3. Approach Assessment

### 3.3.1. Overview of the Evaluation Strategy

The evaluation of this artifact follows the principles of Design Science Research (DSR), which emphasizes the assessment of artifacts based on their usefulness, correctness, and improvement. In this context, the goal of the artifact is to quantify behavioral patterns suggestive of overreliance on AI in a way that is theoretically grounded and practically interpretable.

Since the dataset did not provide usable objective labels indicating which evaluations were correct or high-quality, the artifact is evaluated through an internally validated, iterative process. Each iteration is designed to refine the logic and robustness of the scoring

mechanism by adjusting the structure of the directional projection and improving the representation of behavioral passivity.

Usefulness is addressed by demonstrating that the artifact generates interpretable behavioral insights from real-world evaluation data, offering a systematic way to surface potentially problematic user-AI dynamics in startup assessment. Correctness is supported by internal coherence checks, including feature selection justification, correlation and variance diagnostics, and logical alignment between inputs and scoring logic. Improvement is shown through a series of artifact iterations that enhance sensitivity, contextual specificity, and robustness of the framework, culminating in a refined directional model aligned with low-effort behavior.

To externally assess the validity of the behavioral overreliance score, a manual annotation task was conducted with researchers, who reviewed pairs of AI-following evaluation sessions and selected the one that appeared to reflect greater overreliance. Agreement between human judgments and the ordering produced by the framework's scores provides an additional layer of validation, confirming that the scoring model aligns with human perceptions of low engagement and deference to AI.

The remainder of this section details the evolution of the framework across refinement iterations. Each step presents the rationale for the change, the technical adjustments made, and a brief analysis of the impact on detection patterns. Together, these assessments provide evidence that the artifact is both conceptually valid and operationally effective in capturing low-effort alignment with AI recommendations.

### 3.3.2. Iterative Refinement Process

The following subsections describe the five iterations that shaped the artifact, outlining the rationale, changes implemented, and effects on detection behavior.

#### 3.3.2.1. Iteration 1: Feature Validation

To support the robustness of the behavioral overreliance score, the combination of behavioral features used to define the projection direction was tested in terms of correlation, multicollinearity, and contribution to variance. The goal was to ensure that each feature captured distinct aspects of user interaction and contributed meaningfully to the variance in effortful behavior.

Given the framework's logic, the behavioral deviation model is only applied to sessions where evaluators followed the AI recommendation. Accordingly, from the full dataset of 555 real-world evaluations collected for this study, only the 253 evaluations in which the evaluator aligned with the AI recommendation were retained for feature validation. This ensures conceptual consistency between the space used to construct the projection vector and the behavioral population the model is designed to characterize. Including sessions where evaluators disagreed with the AI or received no AI input would introduce behavioral noise unrelated to overreliance dynamics, weakening the test's relevance. Since the model targets low-effort agreement with AI, this filtered subset best reflects the behavioral conditions the framework is meant to detect.

To assess the independence and complementarity of the six behavioral features used in the framework, a correlation matrix was first computed within the AI Followed subset. As shown in Table 3, most pairwise relationships were statistically significant at the $p < 0.05$ level, indicated by an asterisk. This significance suggests that the features capture consistent patterns of interaction across users, rather than random variation, reinforcing their relevance for modeling effortful behavior. The highest observed correlation was 0.72 between Duration and Number of Scrolls, followed by 0.70 between Tabs Opened and Clicks. These values reflect moderate associations but remain below the standard multicollinearity threshold of 0.9, indicating the features capture overlapping yet distinct dimensions of interaction effort. Additionally, low and non-significant correlations such as 0.03 between Max Y Reached and Number of Clicks support the interpretation that the variables are not redundant but contribute unique behavioral signals.

**Table 3: Correlation Matrix of Behavioral Variables (AI Followed Group)**

| | Number of Tabs Opened | Max Y Reached | Attention Top Ratio | Duration (s) | Number of Scrolls | Number of Clicks |
|---|---|---|---|---|---|---|
| Number of Tabs Opened | 1.00* | -0.01 | -0.16* | 0.36* | 0.39* | 0.70* |
| Max Y Reached | -0.01 | 1.00* | -0.51* | 0.41* | 0.54* | 0.03 |
| Attention Top Ratio | -0.16* | -0.51* | 1.00* | -0.49* | -0.66* | -0.12 |
| Duration (s) | 0.36* | 0.41* | -0.49* | 1.00* | 0.72* | 0.38* |
| Number of Scrolls | 0.39* | 0.54* | -0.66* | 0.72* | 1.00* | 0.36* |
| Number of Clicks | 0.70* | 0.03 | -0.12 | 0.38* | 0.36* | 1.00* |

To complement this analysis, a Variance Inflation Factor (VIF) test was conducted to measure multicollinearity more rigorously. All VIF values remained below the critical

threshold of 5, with the highest being 3.28 for Number of Scrolls (Table A4). These results confirm that no individual feature can be linearly predicted with high accuracy from the others, supporting the inclusion of all six variables in the projection space without risk of collinearity inflating behavioral deviation measurements.

To evaluate each feature's unique contribution to behavioral variability, a Principal Component Analysis (PCA) was also performed on the standardized values of the same AI Followed sessions. The first two principal components together explained 75% of the total variance, and the first four components accounted for over 90% (Figure 7). Component loadings confirmed that the features were well distributed across components: Max Y Reached dominated PC3 (0.78), Duration was most prominent in PC4 (0.75), Scrolls in PC6 (0.81), and Clicks in PC5 (0.68), while Tabs Opened contributed heavily to PC2 (0.58) (Table 4). These results reinforce that no single component dominates the behavioral space, and each variable contributes to a different interaction dimension.

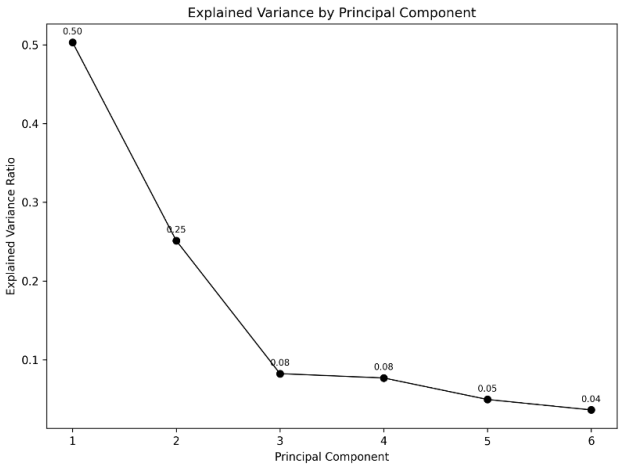


**Figure 7: Explained Variance by Principal Component (PCA on AI Followed Group)**

**Table 4: Component Loadings of Behavioral Variables (AI Followed Group)**

| | Number of Tabs Opened | Max Y Reached | Attention Top Ratio | Duration (s) | Number of Scrolls | Number of Clicks |
|---|---|---|---|---|---|---|
| PC1 | 0.33 | 0.35 | -0.42 | 0.48 | 0.52 | 0.32 |
| PC2 | 0.58 | -0.46 | 0.33 | -0.03 | -0.12 | 0.57 |
| PC3 | 0.01 | 0.78 | 0.48 | -0.21 | -0.13 | 0.3 |
| PC4 | -0.27 | -0.1 | 0.59 | 0.75 | 0.09 | -0.09 |
| PC5 | -0.66 | -0.11 | -0.25 | 0.04 | -0.17 | 0.68 |
| PC6 | -0.23 | -0.17 | 0.28 | -0.41 | 0.81 | 0.08 |

Together, these diagnostics support the inclusion of all six features in the projection vector. They provide a rich and complementary basis for capturing interaction behavior in the AI Followed condition, the precise behavioral space the framework seeks to analyze.

#### 3.3.2.2. Iteration 2: Alternative Projection Weights and Validation Against Human Judgments

Following the validation of the feature set in Iteration 1, this second iteration investigates the optimal construction of the projection vector used to compute behavioral overreliance scores. While the initial version relied on a theoretically defined direction based on expected signs for each feature, this step explores whether alternative formulations may better align with human perceptions of overreliance. Three projection strategies were considered:

1. Theoretical Vector (Theoretical): Constructed using fixed weights of ±1 based on prior expectations from the literature about how each behavioral feature relates to cognitive effort. Features such as fewer tabs opened, shorter durations, fewer clicks, fewer scrolls, and less scroll depth are associated with lower engagement and are thus assigned a weight of -1, whereas features like a higher Attention Top Ratio (indicating more time spent at the top of the page) are given a weight of +1. This approach is simple, highly interpretable, and grounded in theory; however, it does not account for correlations between variables or differences in their scale and contribution.
2. PCA-Based Vector (PCA_Component1): Derived from the first principal component obtained through Principal Component Analysis of the AI Followed group. This component captures the direction of maximum variance within the behavioral space, offering an empirically driven projection direction. While it reflects the dominant structure in the data, its limitation lies in its potentially weak semantic alignment with effort or overreliance, as it may capture variance unrelated to cognitive engagement.
3. Empirical Difference Vector (MeanDiff_NoAI_AIFollowed): Constructs the projection vector by computing the difference in the mean values of each behavioral feature between the No-AI and AI Followed groups. It captures the empirical shift in behavior when AI recommendations are followed, assuming that unaided evaluations serve as a reasonable benchmark for higher cognitive effort. While this assumption may not universally hold, the vector directly reflects real-world behavioral changes that the framework aims to detect.

To evaluate the projection vectors, a human annotation task was conducted using 50 carefully selected comparisons drawn from the 253 AI-Followed sessions. The comparisons were programmatically generated to ensure diversity in score contrasts and outcomes, including pairs with very similar behavioral deviation scores, large differences, and same or differing final decisions.

Each comparison consisted of two session videos (A and B), which were presented to eleven independent raters affiliated with the Laboratory for Innovation Science at Harvard (LISH). Before beginning the task, raters were introduced to the evaluation platform, the user interface, and the decision-making process shown in the videos. They were also provided with the definition of overreliance adopted in this study, as presented in the literature review (Section 2.2.2). Raters were asked to select the session they perceived as more overreliant in each pair. They were not involved in the development of the scoring model and were blinded to the behavioral deviation scores. For each comparison, the session receiving the majority of votes was considered the ground truth label.

All vectors were then used to compute projection scores for all sessions. For each annotated pair, the session with the higher score was taken as the model's prediction of greater overreliance. Accuracy was defined as the proportion of times the model's prediction matched the human judgment. This evaluation method allowed for a direct comparison between projection strategies based on their alignment with perceived behavioral passivity.

### Results and Interpretation

Table 5 presents the projection vectors used to score the 253 AI-Followed sessions.

**Table 5: Feature Weights for Projection Vectors**

| Feature | Theoretical | PCA-Based (PC1) | Empirical Difference |
|---|---|---|---|
| Duration (s) | -1 | 0.475 | 0.283 |
| Number of Tabs Opened | -1 | 0.329 | 0.381 |
| Number of Scrolls | -1 | 0.519 | 0.445 |
| Number of Clicks | -1 | 0.323 | 0.271 |
| Max Y Reached | -1 | 0.345 | 0.528 |
| Attention Top Ratio | +1 | -0.417 | -0.505 |

The PCA-Based vector, extracted from the first principal component of the AI-Followed group, reflects the dominant axis of variation in user behavior. Interestingly, the weights it assigns closely mirror the expectations embedded in the theoretical formulation. Higher values are given to interaction-related features like scrolls, clicks, and duration, while Attention Top Ratio carries an opposite weight. This indicates that the principal behavioral variance within AI-Followed sessions naturally captures a gradient of cognitive engagement.

The Empirical Difference vector follows a strikingly similar pattern. It encodes the direction of behavioral change between No-AI and AI-Followed conditions, based on the assumption that unaided evaluations demand more deliberate engagement. Here again, the weights reflect the same polarity: interaction features are higher in the No-AI group, while Attention Top Ratio tends to be lower.

The most notable finding is that, despite originating from distinct methods (one unsupervised and the other based on cross-condition contrasts), both empirical vectors end up aligned in sign with the theoretical model. This convergence suggests that the underlying behavioral structure of the data naturally supports the theoretical framing of overreliance as a passive, surface-level pattern. To maintain consistency in interpretation and comparability, both vectors were inverted (multiplied by –1) so that higher scores consistently indicate stronger alignment with low-effort behavior. This standardization allows their outputs to be directly assessed against human-labeled judgments of overreliance.

Table 6 summarizes the accuracy of each vector in predicting which session in each pair was judged by humans to reflect stronger signs of overreliance.

**Table 6: Accuracy of Projection Vectors Against Human Annotations**

| Projection Vector | Accuracy (%) |
|---|---|
| Theoretical | 86 |
| PCA_Component1_Inverted | 86 |
| DiffMean_LowEffortProjection_Inverted | 82 |

Both the PCA_Component1_Inverted and Theoretical vectors achieved an accuracy of 86 percent, indicating strong alignment with human judgments of low-effort behavior. The Empirical Difference vector, while still performing well, showed slightly lower agreement at 82 percent. This suggests that, although grounded in actual shifts between AI and No-AI conditions, the difference-based projection may capture a wider range of behavioral change that is not always interpreted as passivity by human observers.

Given these results, the PCA_Component1_Inverted vector was retained for subsequent iterations not only due to its strong alignment with human perception but also because it offers an empirically grounded yet unsupervised alternative that validates and reinforces the theoretical assumptions. Its performance confirms that the dominant behavioral patterns in the data naturally reflect the structure the framework seeks to capture, making it a compelling and theoretically coherent projection direction for detecting potential overreliance.

#### 3.3.2.3. Iteration 3: Outcome-Specific Projection Vectors

This third iteration extends the framework by testing whether behavioral expressions of overreliance vary depending on the evaluator's final decision outcome. While the previous iteration constructed a single projection vector to quantify low-effort alignment with AI recommendations, this iteration explores whether applying one shared vector across all outcomes may obscure important behavioral differences. To address this, the current step introduces the logic of outcome-specific projection vectors, where each possible decision outcome is associated with its own behavioral deviation model.

The motivation stems from both contextual and methodological considerations. In the present dataset, evaluators produce one of two final outcomes: Pass or Fail. However, the structural logic of these outcomes is asymmetric. A Pass requires that all five criteria be satisfied, whereas a Fail is triggered when any single criterion is not met. This asymmetry can influence how evaluators interact with the platform. For instance, reaching a Fail may involve focusing on a narrow portion of the information, while a Pass may reflect either comprehensive engagement or shallow acceptance across all criteria. As such, what constitutes low-effort behavior may differ significantly between these two decision outcomes.

To account for this, the framework was modified to include a distinct projection vector for each outcome category. Although this dataset only has two outcomes, the design principle is intended to be generalizable. In any evaluation setting with multiple categorical outcomes, each decision class can be modeled independently using outcome-specific vectors. This enables the framework to adapt to a wider range of evaluative environments while maintaining theoretical consistency.

## Methodology

Building on the previous iteration, which identified the PCA_Component1_Inverted vector as the most appropriate projection direction, this step applied the same methodology separately to two subsets of the data: evaluations that ended in a Pass and those that ended in a Fail. The goal was to construct outcome-specific projection vectors (Vector_Pass and Vector_Fail) each capturing the dominant behavioral variance within its respective outcome group.

For each subset, a Principal Component Analysis was conducted using the same six standardized behavioral features. The first principal component from each group was extracted and used to define the projection vector corresponding to that decision type. As with the earlier PCA-Based vector, the resulting components were multiplied by –1 to align their direction with the theoretical notion of behavioral passivity. This inversion ensures that higher scores continue to reflect lower engagement, regardless of outcome type.

By applying PCA separately to Pass and Fail decisions, this iteration aims to capture outcome-specific behavioral signals that may be masked when using a unified projection across all evaluations. The resulting vectors are presented in Table 7, allowing for direct comparison of how different features contribute to perceived low effort depending on the nature of the evaluator's final decision.

**Table 7: Feature Weights in Vector_Pass vs Vector_Fail**

| Feature | Vector_Pass | Vector_Fail |
|---|---|---|
| Duration (s) | 0.442 | 0.501 |
| Number of Tabs Opened | 0.338 | 0.329 |
| Number of Scrolls | 0.504 | 0.527 |
| Number of Clicks | 0.315 | 0.324 |
| Max Y Reached | 0.383 | 0.315 |
| Attention Top Ratio | -0.436 | -0.399 |

### Evaluation Procedure and Results

To assess whether this outcome-specific approach improved alignment with human perception, each session in the 50 annotated comparison pairs was scored using the corresponding vector for its decision outcome (Vector_Pass_Inverted or Vector_Fail_Inverted). These predictions were then again compared to the human annotations.

The outcome-specific modeling yielded an overall accuracy of 88 percent, with 44 out of 50 comparisons matching the majority human judgment. This marks a modest improvement over the single-vector PCA approach (86 percent), suggesting that low-effort behavior manifests differently depending on the evaluator's final decision. A comparison of the vector structures (see Table 7) supports this, as features such as Max Y Reached and Attention Top Ratio shift in relative importance across the two outcomes.

These findings highlight that even within a shared evaluation interface, the behavioral cues associated with overreliance are not uniform. Tailoring the projection logic to the decision outcome allows the framework to capture more nuanced expressions of low cognitive effort, reinforcing the value of an adaptive, modular design.

#### 3.3.2.4. Iteration 4

This final iteration focuses on validating the expressiveness and practical utility of the behavioral projection score developed in this study. While earlier iterations optimized the projection vector for its alignment with human perception, this iteration assesses whether the score can meaningfully distinguish evaluator behavior at scale across experimental conditions.

The central hypothesis is that if the projection score effectively captures behavioral patterns associated with cognitive effort and engagement, then it should vary systematically across the three evaluator groups:

- **No AI**: sessions performed without any AI support
- **AI Followed**: sessions where the evaluator aligned with the AI recommendation
- **AI Not Followed**: sessions where the evaluator disagreed with the AI

These groups are mutually exclusive and cover the full dataset of 555 evaluations. The score is expected to be highest in the AI Followed group, where the risk of overreliance is presumed to be greater, and lowest in the No AI group, where evaluators rely exclusively on their own judgment. This step serves as a broader validation of the framework by testing whether the behavioral deviation scores vary systematically across experimental conditions.

To test whether the behavioral projection score captured meaningful distinctions across the three evaluator groups, the behavioral deviation scores established in the previous step were used for pairwise comparisons. Given the non-normal distribution of scores in all groups

(Shapiro–Wilk test, $p < 0.05$; see Appendix Table A9), non-parametric tests were used. Table 8 presents the results of Mann–Whitney U tests.

**Table 8: Pairwise Comparisons of Behavioral Projection Scores**

| Group A | Group B | U-Statistic | p-value | Direction of Difference |
|---|---|---|---|---|
| AI Followed | No AI | 25084.0 | 0.0000 | AI Followed > No AI |
| AI Followed | AI Not Followed | 25120.0 | 0.0002 | AI Followed > AI Not Followed |
| AI Not Followed | No AI | 13508.0 | 0.0040 | AI Not Followed > No AI |

These results are visualized in Figure 8, which displays the distribution of projection scores for each group using outcome-specific PCA vectors. As expected, AI Followed sessions showed the highest scores while No AI sessions had the lowest. The AI Not Followed group fell in between, reflecting the fact that these evaluators had AI available but chose to diverge from its recommendation, likely requiring more independent engagement.

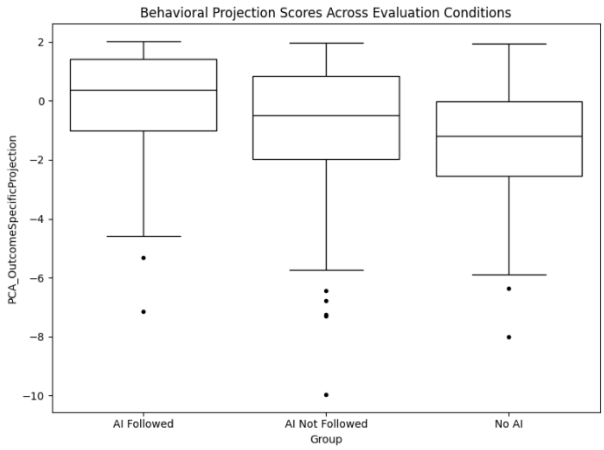


**Figure 8: Distribution of Behavioral Deviation Scores Across Experimental Groups**

Together, these results support the expressiveness and practical utility of the projection score: it distinguishes evaluator behavior across real-world conditions in line with the expected gradient of cognitive effort. The score not only reflects theoretically grounded engagement patterns but also scales effectively across a diverse set of evaluation scenarios.

# 4. Empirical results and discussion

This chapter presents a comprehensive interpretation of the study's findings in relation to the research questions and objectives outlined earlier. It is structured in four parts. Section 4.1. focuses on a descriptive analysis of evaluator behavior across the different AI assistance conditions, addressing the primary and first sub-research questions. Section 4.2. builds on this by applying the behavioral analytics framework developed in this dissertation to examine the potential for overreliance on AI, answering the second sub-research question.

The discussion is grounded in a dataset comprising a total of 555 solution evaluations, distributed across three experimental conditions: 216 under Narrative AI, 200 under Black Box AI, and 139 with No AI. By leveraging multiple behavioral variables, it aims to characterize both the extent and nature of AI influence on human evaluators.

Section 4.3. reflects on the main limitations of the study and identifies opportunities for future work, while Section 4.4. outlines the research contributions and broader implications of the findings. Throughout the discussion, results are interpreted considering existing literature, and both theoretical and practical significance are considered.

## 4.1. Descriptive Analysis of Evaluator Behavior

### 4.1.1 Behavioral Patterns Across AI Types

To investigate how AI assistance influences evaluator behavior, ten behavioral variables were analyzed across the three conditions: No AI, Black Box AI, and Narrative AI. These variables reflect dimensions of time investment, navigation effort, interface engagement, and visual attention.

The analysis accounted for the fact that the AI condition was randomly assigned prior to each evaluation and was therefore treated as the independent variable. Accordingly, behavioral variables were examined as outcomes influenced by the presence and format of AI assistance. Given this causal structure, predictive modeling techniques such as logistic regression were not applied, as they would invert the direction of inference and misrepresent the experimental design. Instead, an observational approach based on pairwise comparisons and summary statistics was implemented.

Since normality assumptions were violated (see Appendix Table A7), Mann-Whitney U tests were used to compare groups. Full results of the pairwise comparisons are provided in

Appendix Table A6. Summary statistics are presented in Appendix Table A5, and distributions are illustrated through boxplots from Figure 9 to Figure 18.

**Time Investment**

Evaluation duration was markedly longer in the No-AI condition (mean = 85.63 s, median = 64.25 s, SD = 93.16) than in either Black Box AI (70.34 s, 49.32 s, 76.50) or Narrative AI (59.32 s, 38.74 s, 61.44), with both contrasts reaching significance ($p = 0.0054$ and $p < 0.0001$, respectively; see Figure 9). The two AI formats themselves did not differ ($p = 0.171$). The same ordering holds for the time to the first answer selection. Participants without AI took the longest (mean = 76.14 s, median = 58.19 s) versus Black-Box (60.16 s, 38.89 s) and Narrative AI (46.39 s, 24.13 s). No-AI versus AI differences were significant ($p = 0.0013$ and $p < 0.001$; Figure 10), while the two AI conditions again showed no reliable difference ($p = 0.166$). Finally, the interval between the final answer and pressing "Next" was shortest without AI (median = 3.56 s, SD = 13.40) and longer in Black-Box (4.47 s) and Narrative AI (4.77 s) sessions. While both contrasts with No AI reached significance, with Black Box AI ($p = 0.045$) and Narrative AI ($p = 0.006$), the result for Black Box AI should be interpreted with caution, as it lies close to the conventional significance threshold and is one of many comparisons examined without correction for multiple testing. The two AI formats did not differ ($p = 0.339$; Figure 11). Taken together, these results suggest that the presence of AI support tends to shorten deliberation up to the point of choosing an answer. After the final answer is selected, there is some evidence, particularly in the Narrative AI condition, that evaluators pause slightly longer before moving on.

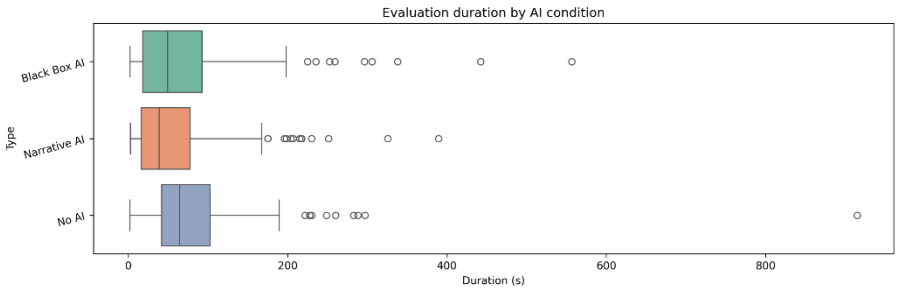


**Figure 9: Evaluation duration by AI condition**

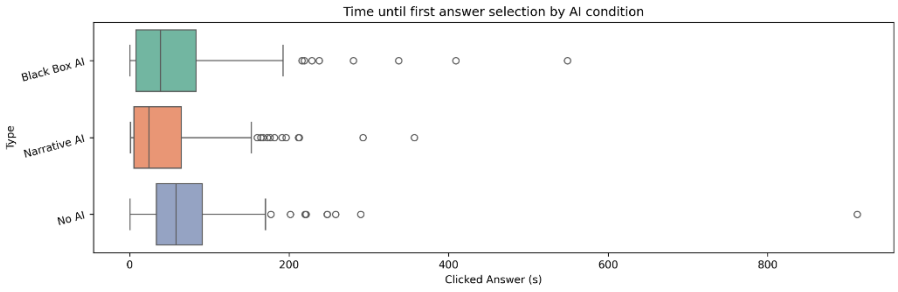


**Figure 10: Time until first answer selection by AI condition**

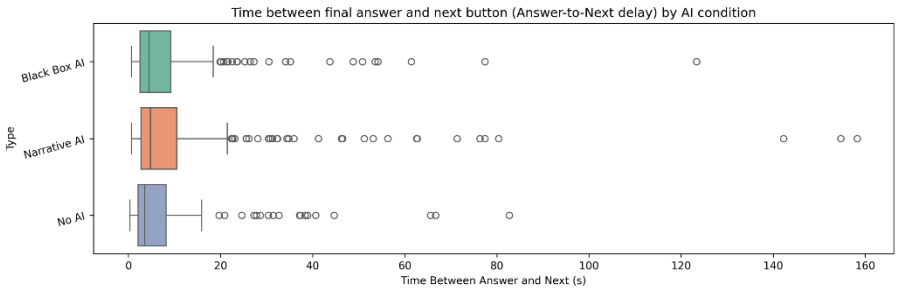


**Figure 11: Time between final answer and next button by AI condition**

## Exploratory and Navigational Effort

The amount of vertical exploration, measured by maximum scroll depth, was highest without AI, averaging 4 574 px (median = 3 258 px, SD = 4 499). Depth fell to 2 452 px (1 333 px, 3 369) with Black Box AI and to 2 245 px (1 321 px, 3 121) with Narrative AI. Both contrasts with No-AI were highly significant ($p < 0.001$; Figure 12), whereas the two AI formats did not differ ($p = 0.989$). Scroll frequency showed the same pattern. No-AI sessions averaged 13.52 scrolls (median = 12), versus 10.76 (median = 7) with Black Box AI and 9.20 (median = 6) with Narrative AI. The No-AI versus AI gaps were significant ($p = 0.001$ and $p < 0.001$; Figure 13), and again the two AI formats did not differ ($p = 0.719$). Interestingly, tab exploration breaks the earlier pattern. The decrease becomes significant only when the AI provides an explanation: openings fall from 2.56 in No-AI to 1.83 in Narrative AI ($p = 0.0014$, see Figure 14), whereas Black Box AI shows no reliable change (mean = 2.19, $p = 0.094$). Taken together, in this dataset, all forms of AI support were associated with reduced deep scrolling and scroll counts, while explanatory text appeared to further limit cross-tab searching. This suggests a shift toward quicker and potentially less thorough information gathering in this context.

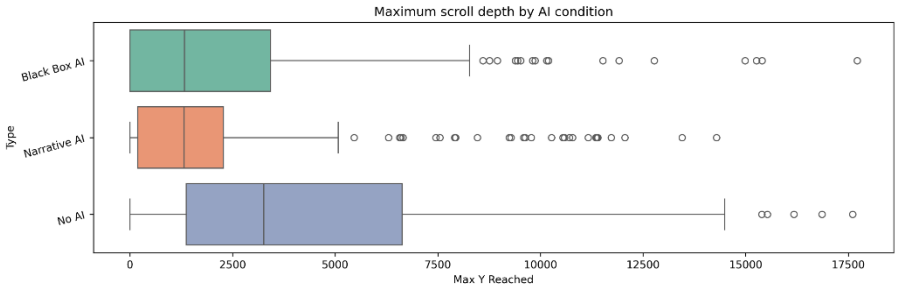


**Figure 12: Maximum scroll depth by AI condition**

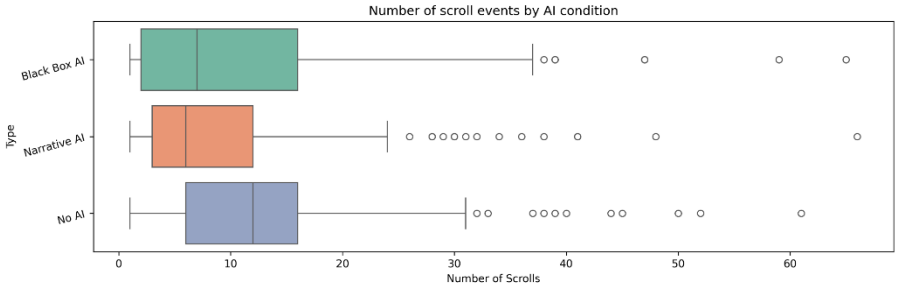


**Figure 13: Number of scroll events by AI condition**

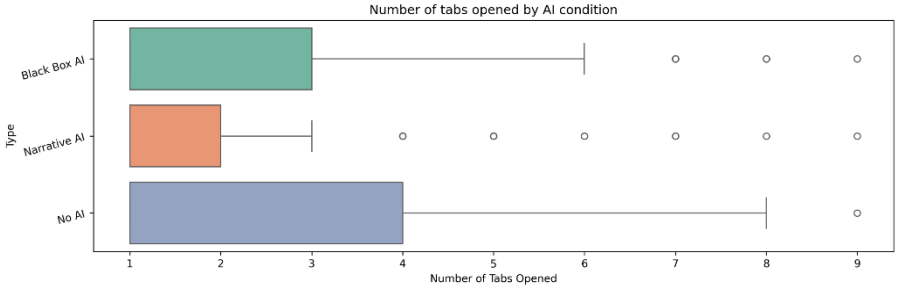


**Figure 14: Number of tabs opened by AI condition**

## Interface Engagement

Click-related engagement does not decline uniformly across AI formats. Click activity peaks in the No-AI condition (mean = 8.55, median = 7) and falls to its lowest level with Narrative AI (mean = 6.89, median = 6, p = 0.0007, see Figure 15). In contrast, Black Box AI mirrors the No-AI baseline (mean = 8.64, median = 6) and shows no significant change (p = 0.072). A different pattern emerges for rereading behavior. Returns to the top of the page average 3.10 in No-AI sessions, drop to 2.74 with Black Box AI, and sit at 2.69 with Narrative AI. Only the No-AI versus Black-Box comparison is significant (p = 0.031, Figure 16). Neither No-AI versus Narrative (p = 0.083) nor Black-Box versus Narrative (p = 0.472) shows a reliable difference. Together, these findings suggest that, in this dataset, explanatory AI may

dampen overall clicking, while the simpler pass–fail cue appears to primarily reduce the need to return to the top of the page.

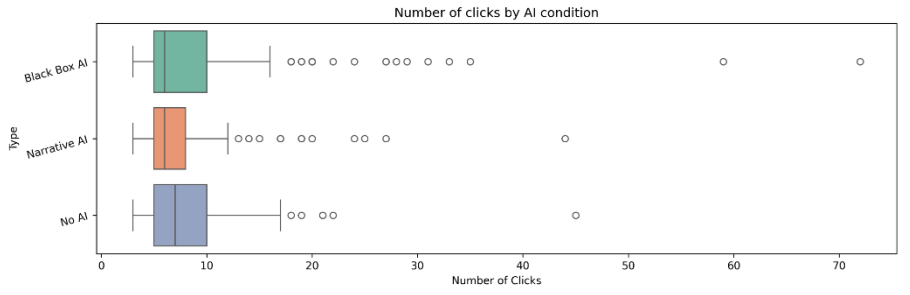


**Figure 15: Number of clicks by AI condition**

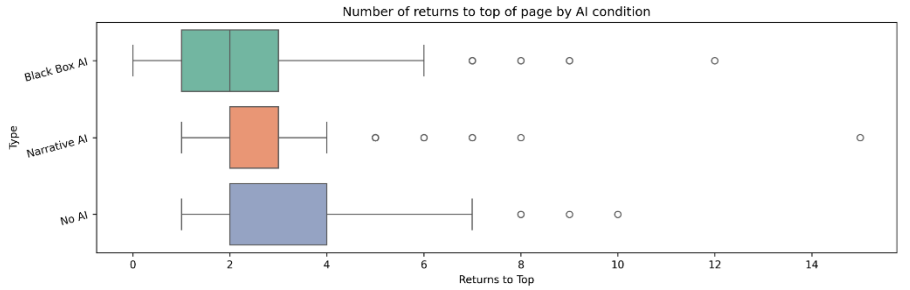


**Figure 16: Number of returns to top of page by AI condition**

### Attention Distribution

The Attention Top Ratio, which reflects the proportion of time spent viewing the top portion of the screen, was significantly lower in No AI (mean = 0.52; median=0.44) compared to both Black Box and Narrative AI (0.64 in both conditions; 0.70 in Black-Box and 0.64 in Narrative), with both differences being statistically significant ($p = 0.0006$ and $p = 0.0003$). The two AI formats do not differ ($p = 0.743$). Hover duration over the sidebar, which either contained static criteria or AI content depending on the condition, was highest in Narrative AI (mean = 6.52s, median = 1.13s), followed by Black Box AI (3.98 s, 0 s), and lowest in No AI (3.11 s, 0 s). The difference between Narrative AI and No AI was statistically significant ($p < 0.001$), whereas No AI and Black-Box do not differ ($p = 0.860$). Figure 17 and Figure 18 visualise these shifts. Taken together, in this dataset, all AI cues appeared to shift evaluators' gaze toward the top of the screen, while only explanatory AI was associated with sustained attention on the sidebar. This pattern aligns with broader indications of reduced deep reading and greater focus on areas where the model's guidance was displayed.

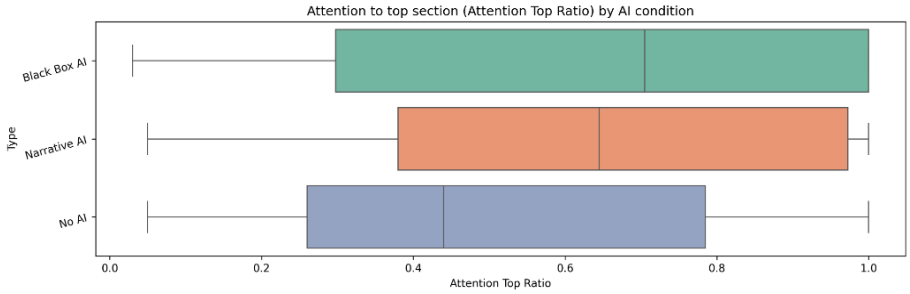


**Figure 17: Attention to top section by AI condition**

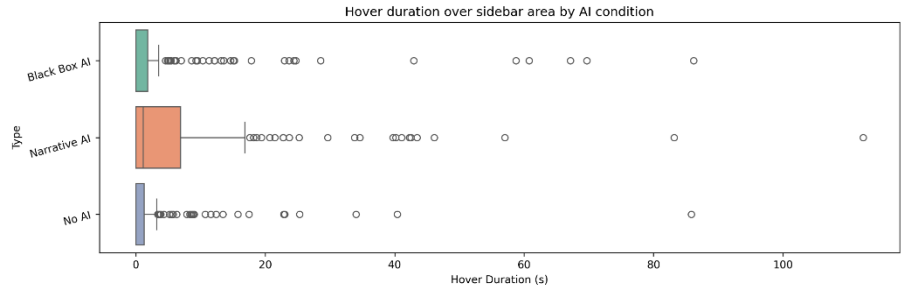


**Figure 18: Hover duration over sidebar area by AI condition**

## Summary

In summary, the results provide evidence that, in this sample, the presence of AI assistance is associated with consistent behavioral changes in evaluator behavior and may influence how deeply and critically evaluators engage with the task. Across all ten behavioral measures, both AI formats reduced decision time, scrolling depth, and most navigation metrics compared with the No-AI baseline, while directing visual attention toward the upper portion of the screen. Explanations intensified these effects by further lowering click counts and tab openings and by anchoring gaze on the sidebar, whereas the Black Box cue showed a milder pattern, with modest changes in temporal and scrolling behavior. In this sense, behavioral changes under AI assistance may reflect a shift from deliberate, exploratory appraisal to faster, surface-level interaction focused on the model's output. This raises important implications about overreliance and the potential trade-offs between efficiency and thoroughness in decision-making.

### 4.1.2. Influence of AI Format on Evaluator Reliance

This section explores how the type of AI input, namely a simple recommendation or a recommendation accompanied by an explanation, shapes evaluator reliance on AI. As discussed in the literature review, reliance is defined behaviorally as the act of following the AI's recommendation, regardless of internal trust or confidence.

With 216 observations under Narrative AI and 200 under Black Box AI, the dataset provided a solid basis for comparing evaluator behavior across explanation formats. Evaluator reliance was operationalized using the AI Follow Rate, defined as the proportion of times evaluators aligned their final answer with the AI's suggestion. As presented in Table 9, evaluators under Narrative AI followed the AI recommendation in 62.5% of the cases, compared to 59% under Black Box AI. Although this reflects a slight increase when explanations were provided, a chi-squared test of independence revealed no statistically significant difference between the two groups ($\chi^2 = 0.40$, $p = 0.53$). The effect size was small (Cohen's $h = 0.07$), and the post hoc power analysis indicated low statistical power (31%), suggesting that the study may not have been sufficiently powered to detect such a subtle difference. The test result for this overall comparison is reported in Appendix Table A8.

**Table 9: AI Follow Rate by Recommendation Type**

| AI Format | N (Followed) | N (Total) | Follow Rate (%) |
|---|---|---|---|
| Black Box AI | 118 | 200 | 59 |
| Narrative AI | 135 | 216 | 62.5 |

A more detailed analysis, presented in Table 10, examines follow rates by the type of AI recommendation (Pass vs. Fail). For Pass recommendations, reliance was extremely high across both conditions: 98.1% under Narrative AI and 89.5% under Black Box AI, resulting in a notable gap of nearly 9 percentage points. A chi-squared test did not confirm this difference as statistically significant ($\chi^2 = 2.22$, $p = 0.14$). However, the observed effect size was moderate (Cohen's $h = 0.39$), suggesting a potentially meaningful difference that the present sample may not have been large enough to confirm with confidence. As such, while explanations may appear to increase reliance when recommending "Pass," this trend was not statistically robust in this dataset.

**Table 10: AI Follow Rate by Format and Recommendation Type**

| AI Format | AI Recommendation | N (Followed) | N (Total) | Follow Rate (%) |
|---|---|---|---|---|
| Black Box AI | Pass | 51 | 57 | 89.5 |
| | Fail | 67 | 143 | 46.9 |
| Narrative AI | Pass | 53 | 54 | 98.1 |
| | Fail | 82 | 162 | 50.6 |

For Fail recommendations, follow rates were lower overall, at 50.6% in the Narrative AI condition and 46.9% in the Black Box condition. In this case, the difference was minimal and not statistically significant ($\chi^2 = 0.29$, $p = 0.59$). The effect size was small (Cohen's $h = 0.08$), offering little indication of a meaningful difference between formats in this case. These test results for the conditional analyses are reported in Appendix Table A8.

These results suggest that, while narrative explanations may visually appear to increase follow rates, these effects were not statistically robust in this dataset. In that sense, based on the present results, the type of AI input does not appear to significantly affect evaluators' reliance on AI.

Additionally, due to the observed differences in follow rates between types of AI recommendation, future analyses could test whether the nature of the AI suggestion itself (Pass or Fail) meaningfully shapes evaluator reliance, potentially in interaction with the explanation format.

### 4.2. Framework-Based Analysis

This section applies the behavioral deviation framework developed in Chapter 3 to the experimental data.

#### 4.2.1. Distribution of Framework Scores

The behavioral analytics framework was applied to the full dataset of 555 solution evaluations. As designed, the framework filters the dataset internally and delivers behavioral deviation scores for the subset of 253 evaluations in which the AI recommendation was followed. These scores reflect the degree of alignment between each evaluator's interaction behavior and the low-effort patterns captured by outcome-specific PCA vectors.

The resulting distribution is presented in Figure 19. Scores ranged from -7.16 to 2.01, with a mean near zero, a median of 0.38, and a standard deviation of 1.74. Most evaluations fell between –1.01 and 1.43, though a subset exhibited particularly high values, suggesting behavioral passivity. On the lower end, a smaller number of sessions reflect more deliberate interaction patterns despite alignment with AI, highlighting the heterogeneity in engagement levels even within AI-Followed cases.

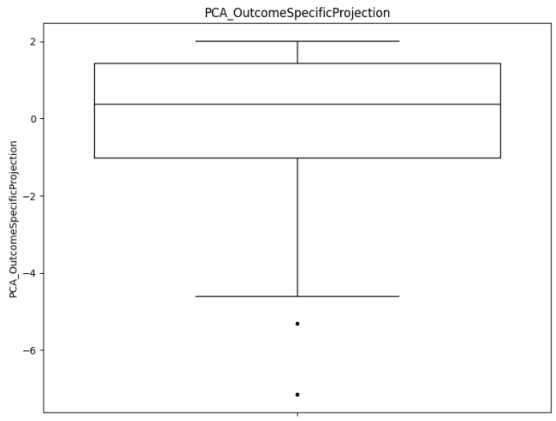


**Figure 19: Distribution of Behavioral Deviation Scores**

To assess whether the type of AI input influenced behavioral deviation, scores were compared between Black Box AI (n = 118) and Narrative AI (n = 135) sessions. As shown in Figure 20, both formats produced similar distributions. Narrative AI displayed more extreme low-end outliers, but no substantial differences were observed in central tendency or overall spread.

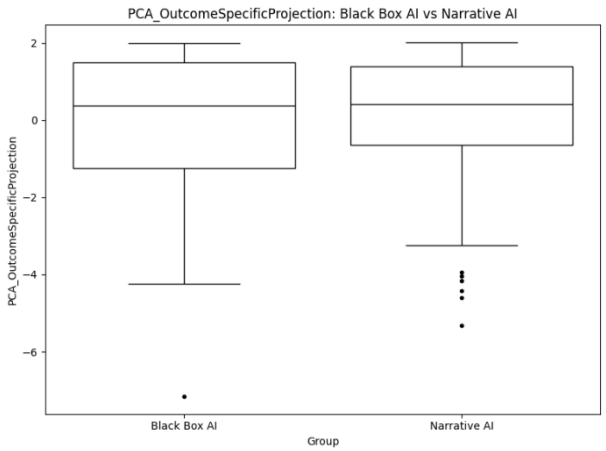


**Figure 20: Behavioral Deviation Scores by AI Format**

Shapiro–Wilk tests confirmed non-normality in both groups ($W = 0.887$ for Black Box AI and $W = 0.891$ for Narrative AI, both $p < 0.001$), justifying the use of non-parametric testing.

A Mann–Whitney U test revealed no statistically significant difference between conditions ($U = 8018.0$, $p = 0.9280$), confirming that the level of behavioral deviation does not depend meaningfully on whether AI input was presented with or without explanation. These results underscore the framework's capacity to differentiate levels of potential overreliance among AI-Followed sessions and suggest that such risk is not solely driven by format, but rather by the evaluator's behavioral response.

Finally, this framework makes it possible to capture and analyze the variability of evaluator behavior in the presence of AI recommendations in a systematic and interpretable manner. It offers a behavioral lens through which overreliance can be understood not as a binary outcome, but as a spectrum of engagement. As concluded in Section 4.1, in this sample, the presence of AI assistance was associated with consistently lower behavioral effort across evaluators. However, the application of the framework revealed that not all instances of agreement with AI reflect overreliance. Instead, behavioral deviation scores varied widely among AI-Followed sessions, indicating that overreliance emerges under specific interaction patterns rather than being uniformly caused by AI support. These findings provide a nuanced answer to the second sub-research question, showing that while AI assistance increases the potential for low-effort agreement, it does not deterministically lead to overreliance. The next section builds on this by applying thresholding strategies to identify sessions that exhibit especially pronounced signs of low-effort agreement.

### 4.2.2. Thresholding and Identification of High-Risk Cases

While the framework does not impose fixed cutoffs or prescribe normative classifications of overreliance, real-world implementations may require the definition of operational thresholds on the behavioral deviation score. These thresholds can serve various purposes, such as monitoring evaluator performance, triggering feedback mechanisms, or flagging decision instances for further audit and review.

To explore this practical dimension, the 90th and 95th percentiles of the behavioral deviation score distribution were applied to the 253 AI-Followed sessions. These thresholds are not an integral component of the framework itself but serve to demonstrate how the score can be adapted to different institutional goals, risk tolerances, or operational constraints. Their application here is purely illustrative and intended to show how decision-makers might flexibly interpret and act upon the behavioral signal. Depending on context, institutions may

wish to adopt more lenient or stricter thresholds or combine the behavioral score with other indicators such as decision quality, evaluator experience, or contextual risk levels.

The 90th percentile corresponded to a score of 1.743 and flagged 26 sessions. The 95th percentile, at 1.881, isolated the top 13 sessions. These high-scoring evaluations were characterized by extremely low interaction depth: nearly all had only one tab opened, minimal scrolling, and consistently high attention to the top portion of the interface. Among those above the 95th percentile, the median number of scrolls was just one, the average number of clicks was below five, and Max Y Reached was zero in every case, indicating no vertical navigation beyond the visible top of the page. A more detailed summary of interaction metrics for these high-scoring sessions is provided in Appendix Table A10.

This behavioral profile aligns with a pattern of passive acceptance, in which evaluators may have followed the AI recommendation without engaging critically with the information provided. Although these sessions cannot be definitively classified as overreliant, they reflect meaningful outliers that the framework is uniquely equipped to detect. In this way, the thresholding analysis highlights how the framework can support more attentive, context-sensitive oversight, flagging cases where decision processes may warrant reflection, review, or intervention.

### 4.3. Limitations and Future Work

This study presents a structured framework to detect potential overreliance on AI by analyzing evaluators' screen-based interaction patterns. While the results demonstrate the framework's capacity to capture relevant behavioral variation, several limitations and directions for future research should be noted.

First, the behavioral analysis relies exclusively on-screen recordings, which restricts observable features to clicks, scrolls, tab switches, and timing. This approach excludes richer behavioral signals such as eye movements, hand gestures, facial expressions, or verbal cues, which could improve the precision and depth of engagement detection. Future work could explore integrating multimodal data sources, especially in settings where more advanced instrumentation is available.

Second, although the framework was designed around the concept of low-effort agreement with AI, its core logic of capturing behavioral passivity could be extended beyond reliance scenarios. Future applications may adapt the same projection logic to flag evaluations that

exhibit minimal exploration or surface-level interaction, even in the absence of AI influence. In the context of startup assessment, such shallow engagement could indicate insufficient diligence and merit further scrutiny, regardless of AI involvement.

Third, the human-labeled comparison set used to validate projection vectors was relatively small and based on subjective judgments. While these annotations proved valuable, future studies could benefit from expanding the number and diversity of annotators and integrating more structured performance assessments. Additionally, with a larger dataset, it would be feasible to explore alternative methods for constructing projection vectors, including supervised learning approaches that optimize alignment with known outcomes.

Fourth, although this study focused on the general influence of AI presence and explanation format, it did not distinguish between the type of AI recommendation received. Given the observed differences in follow rates between sessions with positive (Pass) and negative (Fail) AI suggestions, future research could examine whether the nature of the AI's advice interacts with format to shape reliance patterns. This may uncover more nuanced dynamics in how evaluators respond to AI input.

Fifth, this study did not find evidence in the dataset that the presence of AI assistance uniformly leads to overreliance, even though all participants were students engaged in the same academic setting and task. This behavioral variability raises important questions about how individual-level cognitive differences may shape how people respond to AI recommendations. Future research could explore factors such as cognitive reflection ability, prior experience with AI systems, or personality traits by incorporating psychometric assessments or pre-task surveys, in order to investigate whether certain cognitive profiles are more prone to overreliance and to identify user-level moderators that support more effective human AI collaborations.

Finally, the analysis did not control for potential position effects in the evaluation sequence. It remains possible that the order in which startup profiles appeared, or cumulative cognitive fatigue over multiple evaluations, influenced the observed behavioral patterns. Accounting for these temporal dynamics would strengthen the interpretability of the results and clarify whether low-effort behaviors reflect inherent passivity or situational decline in engagement.

### 4.4. Research Contributions and Implications

This study makes both theoretical and practical contributions by introducing a behavioral analytics framework capable of detecting potential overreliance on AI during idea evaluation tasks. The framework, built on interaction data and iteratively refined, defines overreliance as the combination of behavioral passivity and agreement with AI recommendations, offering a structured way to detect such patterns in the dynamics of human-AI collaboration in subjective decision-making contexts.

From a theoretical perspective, the study advances the conceptualization of overreliance by moving beyond traditional error-based definitions, which depend on the availability of ground-truth correctness. Instead, it operationalizes overreliance as a spectrum of low-effort behavior that can be observed and quantified even in domains where outcomes are subjective. The framework also contributes methodologically by adapting dimensionality-reduction techniques, such as outcome-specific PCA vectors, to construct interpretable and empirically grounded engagement scores, validated against human annotations and experimental group differences. These methodological choices expand the toolkit for researchers studying behavioral engagement and AI reliance in complex, non-binary environments.

Practically, the framework provides organizations with a lightweight and generalizable tool for monitoring interaction quality in AI-assisted evaluation settings. By capturing granular behavioral patterns, the framework identifies sessions where users may be defaulting to AI suggestions without sufficient independent scrutiny. This supports applications in evaluator training, interface design, and decision auditing, where transparency and judgment quality are critical. Importantly, the score remains interpretable and adaptable, allowing institutions to define operational thresholds based on contextual priorities or risk profiles.

Finally, the work offers broader implications for the responsible deployment of AI in evaluative workflows. It underscores that overreliance may arise from subtle behavioral shifts that can become visible through interaction-level analysis. By illuminating these patterns, the study supports the design of decision-support systems that foster thoughtful engagement, rather than passive acceptance, and contributes to a more nuanced understanding of reliance and cognitive effort in human-AI collaboration.

## 5. Conclusions

This dissertation set out to examine how AI assistance shapes evaluator behavior during startup idea assessments and to develop a method for identifying potential overreliance on AI in this context. The results revealed that, in the dataset of this study, the presence of AI was associated with reductions in behavioral indicators of cognitive effort, including shorter evaluation times, less scrolling, and fewer interactions with alternative information tabs. However, evaluators' reliance on the AI system, as operationalized through the follow rate, did not significantly differ between the simple and explanatory formats.

Given the lack of objective correctness in subjective evaluation tasks, this study introduced a behavioral framework that identifies potential overreliance as the co-occurrence of low-effort interaction patterns and agreement with AI recommendations. The framework successfully distinguished behavioral profiles across experimental groups, flagged high-risk sessions through interpretable deviation scores, and showed strong alignment with human judgments of evaluator passivity. Its application to the dataset revealed considerable variability in behavioral deviation scores among AI-Followed sessions, demonstrating that while AI support can influence engagement, it does not uniformly lead to overreliance.

This study offers both theoretical and practical contributions. It reconceptualizes overreliance as a behavioral phenomenon observable through interaction patterns, even in the absence of ground-truth correctness. Methodologically, it introduces a framework built on interaction data and dimensionality-reduction techniques to construct interpretable deviation scores. Practically, the framework provides a lightweight, adaptable tool for monitoring engagement in AI-assisted evaluations, with potential applications in evaluator training, interface design, and decision auditing.

The findings underscore the importance of detecting subtle behavioral signals that may indicate passive acceptance of AI recommendations, supporting the design of decision-support systems that foster critical engagement. In doing so, this study contributes to more transparent and responsible forms of AI-human collaboration in evaluative contexts.

Nonetheless, the framework has several limitations. It relied exclusively on screen-based data, excluding richer behavioral signals such as eye movements or facial expressions. The projection logic was validated using a relatively small set of human-labeled comparisons. Additionally, while the analysis considered the presence and format of AI assistance, it did

not isolate the influence of recommendation polarity (e.g., Pass vs. Fail suggestions). Future research could expand the annotated dataset, incorporate multimodal data sources, explore applications beyond AI reliance, and examine individual cognitive differences or position effects to better understand how engagement and reliance evolve throughout evaluation tasks.

## Contribution to the United Nations Sustainable Development Goals

This dissertation supports the United Nations Sustainable Development Goal 9: Industry, Innovation and Infrastructure, which promotes inclusive and sustainable industrial development, fosters innovation, and emphasizes the importance of building resilient infrastructure not only physical but also institutional and digital.

In the context of this goal, ensuring that innovative solutions are fairly and effectively evaluated becomes essential. Many high-potential ideas depend on early selection processes such as grant competitions, accelerators, or innovation challenges to gain visibility and support. However, these evaluation mechanisms are often inconsistent and subjective, creating barriers to inclusive and sustainable innovation. SDG 9 calls for building not only the innovations themselves but also the systems that identify, select, and promote them responsibly.

This dissertation contributes to that objective by studying the effects of human–AI collaboration in innovation evaluation. As artificial intelligence is increasingly used to bring consistency and scale to startup assessments, this work investigates how such collaboration affects evaluator behavior. Crucially, it identifies a risk: that AI assistance, while meant to support human judgment, can lead to overreliance, where evaluators agree with AI suggestions without sufficient critical engagement. This behavioral tendency undermines the very purpose of the collaboration and weakens the integrity of innovation infrastructures.

To address this challenge, the dissertation introduces a behavioral analytics framework capable of detecting signs of overreliance in AI-assisted evaluations. Rather than rejecting AI integration, the work promotes responsible use by offering a practical tool to monitor interaction quality. In doing so, it contributes to building stronger, fairer, and more transparent decision systems, which are a key component of SDG 9's vision for resilient innovation ecosystems. The alignment between the dissertation and SDG 9 is summarized in Table 11.

**Table 11: Overview of Bibliometric Search Results for AI and Idea Evaluation Literature**

| SDG | Goal | Contribution | Performance indicators and metrics |
|---|---|---|---|
| 9 | Industry, Innovation and Infrastructure | Strengthens the evaluation infrastructure behind innovation systems by studying human–AI collaboration and introducing a framework that helps safeguard thoughtful human judgment in AI-supported assessments | Framework adoption, reduction in passive AI-following behaviors, improved behavioral engagement during innovation evaluations |

## References


Akbari, F., Saberi, M., & Hussain, O. (2020). Social network structure-based framework for innovation evaluation and propagation for new product development. *Service Oriented Computing and Applications, 14*, pp. 189–201.

Arsenyan, J. (2017, September). Fuzzy Rule-Based Decision Support System for Technology Start-Up Selection Problem. *European Conference on Innovation and Entrepreneurship*.

Bouaud, J., Spano, J.-P., Lefranc, J.-P., Cojean-Zelek, I., Blaszka-Jaulerry, B., Zelek, L., . . . Séroussi, B. (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. *Studies in Health Technology and Informatics, 216*, pp. 264-269.

Burnap, A., Ren, Y., Papalambros, P., Gonzalez, R., & Gerth, R. (2013, August). A Simulation Based Estimation of Crowd Ability and its Influence on Crowdsourced Evaluation of Design Concepts. *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, 3B: 39th Design Automation Conference*.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. *2015 International Conference on Healthcare Informatics*, pp. 160-169.

Cabitza, F., Campagner, A., Angius, R., Natali, C., & 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 (CHI '23)*, pp. 1-20.

Camburn, B., He, Y., Raviselvam, S., Luo, J., & Wood, K. (2020, Mach). Machine Learning-Based Design Concept Evaluation. *Journal of Mechanical Design, 142*(3), p. 031113.

Carlton, J., Brown, A., Jay, C., & Keane, J. (2019). Inferring User Engagement from Interaction Data. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*, pp. 1-6.

Carlton, J., Brown, A., Jay, C., & Keane, J. (2021). Using Interaction Data to Predict Engagement with Interactive Media. *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)*, pp. 1258–1266.

Chang, D., Chen, C.-H., & Kuo, J.-Y. (2016, January). A product conceptualization strategy based on crowd-innovation. *Transdisciplinary Engineering: Crossing Boundaries*, pp. 1038 - 1047.

Chiang, C.-W., & Yin, M. (2022). Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22)*, pp. 148–161.

Dwivedi, Y. K., Hughes, D. L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice, and policy. *International Journal of Information Management, 57*, p. 101994.

Evans, H., & Snead, D. (2024). Why do errors arise in artificial intelligence diagnostic tools in histopathology and how can we minimize them? *Histopathology, 84*(2), pp. 279–287.

Fahse, T., & Schmitt, A. (2023, August). Exploring the Synergies in Human-AI Hybrids: A Longitudinal Analysis in Sales Forecasting. *Americas Conference on Information Systems*.

Fedriani, E. M., & Chaves-Maza, M. (2014). Entrepreneurship Support Based on Mixed Bio-Artificial Neural Network Simulator (ESBBANN). *LNCS, 8681*, pp. 845-846.

Ferrati, F., & Muffatto, M. (2020). Using Crunchbase for research in Entrepreneurship: data content and structure. *19th European Conference on Research Methodology for Business and Management Studies*, pp. 342-351.

Frazer, H. M., Peña-Solorzano, C. A., Kwok, C. F., Elliott, M. S., Chen, Y., Wang, C., . . . Hopp. (2024). Comparison of AI-integrated pathways with human-AI interaction in population mammographic screening for breast cancer. *Nat Commun., 15*(1), p. 7525.

Fregosi, C., & Cabitza, F. (2024). A Frictional Design Approach: Towards Judicial AI and its Possible Applications. *HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI)*.

Fukuchi, Y., & Yamada, S. (2023). Dynamic Selection of Reliance Calibration Cues With AI Reliance Model. *IEEE Access, 11*, pp. 138870-138881.

Goddard, K., Roudsari, A., & Wyatt, J. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics, 83*(5), pp. 368-375.

Greul, A., Schweisfurth, T. G., & Raasch, C. (2023). Does familiarity with an idea bias its evaluation? *PLOS ONE, 18*(7).

Gu, H., Yang, C., Magaki, S., Zarrin-Khameh, N., Lakis, N. S., Cobos, I., . . . Stevens, T. M. (2024). Majority voting of doctors improves appropriateness of AI reliance in pathology. *International Journal of Human-Computer Studies, 190*, p. 103315.

He, G., Buijsman, S., & Gadiraju, U. (2023, October). How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW2), pp. 1 - 29.

Hevner, A., March, S., Park, J., & Ram, S. (2004, March). Design science in information systems research. *MIS Quarterly, 28*(1), pp. 75 - 105.

Huang, L., & Pearce, J. (2015). Managing the unknowable: The effectiveness of early-stage investor gut feel in entrepreneurial investment decisions. *Administrative Science Quarterly, 60*(4), pp. 634–670.

Küper, A., & Krämer, N. (2024). Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI. *International Journal of Human–Computer Interaction, 41*(7), pp. 4115–4131.

Keding, C., & Meissner, P. (2021). Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. *Technological Forecasting and Social Change, 171*, p. 120970.

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior, 160*, p. 108352.

Lee, B., Kim, J., Chang, Y., Choi, S., Park, J., Byun, H., . . . Chang, J. (2024). Experience of Implementing Deep Learning-Based Automatic Contouring in Breast Radiation Therapy Planning: Insights From Over 2000 Cases. *International Journal of Radiation Oncology*Biology*Physics, 119*(5), pp. 1579-1589.

Lee, M. H., & Chew, C. J. (2023). Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW2), pp. 1 - 22 .

Li, D. (2017, April). Expertise versus Bias in Evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics, 9*(2), pp. 60–92.

Li, Z., & Zhan, Z. (2020). Integrated infrared imaging techniques and multi-model information via convolution neural network for learning engagement evaluation. *Infrared Physics & Technology, 109*, p. 103430.

Lu, Z., & Yin, M. (2021). Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, pp. 1–16.

Lu, Z., Wang, D., & Yin, M. (2024). Does More Advice Help? The Effects of Second Opinions in AI-Assisted Decision Making.

Mesbah, S., Arous, I., Yang, J., & Bozzon, A. (2023, April 30). HybridEval: A Human-AI Collaborative Approach for Evaluating Design Ideas at Scale. *WWW '23: Proceedings of the ACM Web Conference 2023* , pp. 3837 - 3848.

Meyer, J., Khademi, A., Têtu, B., Han, W., Nippak, P., & Remisch, D. (2022). Impact of artificial intelligence on pathologists' decisions: an experiment. *Journal of the American Medical Informatics Association : JAMIA, 29*(10), pp. 1688–1695.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . Mayo-Wilso. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*(71).

Parasuraman, R., & Manzey, D. (2010). Complacency and bias in human use of automation: an attentional integration. *Human factors, 52*(3), pp. 381–410.

Raikwar, A., Mifsud, D., Wickens, C., Batmaz, A., Warden, A., Kelley, B., . . . Ortega, F. (2024, May). Beyond the Wizard of Oz: Negative Effects of Imperfect Machine Learning to Examine the Impact of Reliability of Augmented Reality Cues on Visual Search Performance. *IEEE Transactions on Visualization and Computer Graphics, 30*(5), pp. 2662-2670.

Rodriguez, S., O'Donovan, J., Schaffer, J., & Schaffer, T. (2019). Knowledge Complacency and Decision Support Systems. *2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pp. 43-51.

Scheaf, B., Townsend, D., & Klein, P. (2019). Measuring opportunity evaluation: Conceptual synthesis and scale development. *Journal of Business Venturing, 35*(2), p. 105930.

Shaer, O., Cooper, A., Mokryn, O., Kun, A., & Shoshan, H. (2024, May). AI-Augmented Brainwriting: Investigating the use of LLMs in group ideation. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1-17.

Shafique, U., Chaudhry, U., & Towbin, A. (2023, August). Are the Pilots Onboard? Equipping Radiologists for Clinical Implementation of AI. *J Digit Imaging, 36*, pp. 2329–2334.

Strauß, S. (2021, April 20). Deep Automation Bias: How to Tackle a Wicked Problem of AI? *Big Data Cogn. Comput., 5*(2), p. 18.

Tewari, A., Gabarro, J., Sole, J., Lapouble, B., & Montull, L. (2020). Artificial Intelligence Based Decision Making for Venture Capital Platform. *International Conference on Decision Support System Technology*, pp. 136-149.

Tilbury, J., & Flowerday, S. (2024, July). Automation Bias and Complacency in Security Operation Centers. *Computers, 13*(7), p. 165.

Tomy, S., & Pardede, E. (2017). Uncertainty analysis and success prediction for start-ups. *The 5th International Conference on Innovation and Entrepreneurship (ICIE).*

Vered, M., Livni, T., Howe, P., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. *Artificial Intelligence, 322*, p. 103952.

Wei, X., He, S., Han, Y., Feng, A., Yang, P., & Wang, X. (2022). User Behavior Profile: A key to Database Anomaly Access Detection. *2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC)*, pp. 328-334.

Zhang, Q., Zeng, X., Lo, W., & Fan, B. (2023). Enterprise innovation evaluation method based on swarm optimization algorithm and artificial neural network. *Neural Computing and Applications, 35*, pp. 25143–25156.

Zhao, J., Wang, Y., Mancenido, M. V., Chiou, E. K., & Maciejewski, R. (2024, July). Evaluating the Impact of Uncertainty Visualization on Model Reliance. *IEEE Transactions on Visualization and Computer Graphics, 30*(7), pp. 4093-4107.

# Appendix

**Table A1: Overview of Bibliometric Search Results for AI and Idea Evaluation Literature**

| Search String | Number of Results Found | Bibliographic database | Date |
|---|---|---|---|
| KEY ( artificial AND intelligence AND entrepreneur* ) | 405 | Scopus | 21-10-2024 |
| AK=(artificial) AND AK=(intelligence) AND AK=(entrepreneur*) | 131 | WoS | 21-10-2024 |
| KEY ( artificial AND intelligence AND startup ) | 80 | Scopus | 21-10-2024 |
| AK=(artificial) AND AK=(intelligence) AND AK=(startup) | 24 | WoS | 21-10-2024 |
| KEY ( ( "artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "large language model" OR "neural networks" ) AND ( "idea evaluation" OR "idea assessment" OR "startup evaluation" OR "innovation evaluation" OR "concept evaluation" OR "new product evaluation" )) | 46 | Scopus | 25-11-2024 |

**Table A2: Overview of Bibliometric Search Results for Overreliance on AI Literature**

| Search String | Number of Results Found | Bibliographic database | Date |
|---|---|---|---|
| AK = ( ( overreliance OR "over reliance" OR reliance OR overtrust OR overdependence OR "undue trust" OR "automation bias" OR complacency ) AND ( "artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "large language model" OR "neural networks") AND ( "decision-making" OR "human" OR "judgment" OR "critical thinking" OR "startup evaluation" )) | 33 | WoS | 14-12-2024 |
| KEY ( ( overreliance OR "over reliance" OR reliance OR overtrust OR overdependence OR "undue trust" OR "automation bias" OR complacency ) AND ( "artificial intelligence" OR "AI" OR "machine learning" OR "deep learning" OR "large language model" OR "neural networks" ) AND ( "decision-making" OR "human" OR "judgment" OR "critical thinking" OR "startup evaluation" ) ) | 133 | Scopus | 14-12-2024 |

**Table A3: Search Details of Methodologies for Behavioral Analysis**

| Topic | Search String | Number of Results | Bibliographic database | Date |
|---|---|---|---|---|
| Topic 1 – Interaction Features as Proxies for Cognitive Effort | KEY ( ( engagement OR "cognitive effort" OR "mental effort" ) AND ( clicks OR scroll* OR dwell OR "time on task" OR "mouse movement" OR "hover" OR "tab switch" OR "interaction logs" ) ) | 104 | Scopus | 11/03/2024 |
| Topic 2 – Session Log-Based Inference of Engagement | KEY ( ( "user modeling" OR "user state inference" OR "behavioral analytics" ) AND ( "interaction data" OR "session logs" OR "event logs" OR "screen recordings" ) AND ( attention OR engagement OR fatigue OR trust ) ) | 1 | Scopus | 11/03/2024 |
| Topic 3 – Scoring and Quantification of Behavioral Effort | KEY ( ( engagement OR "cognitive effort" ) AND ( scor* OR "composite index" OR "engagement metric" OR "engagement score" ) AND ( "interaction data" OR logs OR scroll* OR clicks OR time )) | 44 | Scopus | 11/03/2024 |
| Topic 4 – Behavioral Deviation and Projection-Based Detection | KEY (("behavioral deviation" OR "anomaly detection" OR "projection score" OR "Mahalanobis distance" ) AND ( "interaction data" OR "user behavior" OR "behavioral analytics" )) | 217 | Scopus | 11/03/2024 |

**Table A4: Variance Inflation Factor of Behavioral Variables**

| Feature | VIF |
|---|---|
| Duration (s) | 2.15 |
| Number of Tabs Opened | 2.12 |
| Number of Scrolls | 3.28 |
| Number of Clicks | 2.03 |
| Max Y Reached | 1.62 |
| Attention Top Ratio | 1.91 |

**Table A5: Descriptive Statistics of Behavioral Variables by Condition**

| Variable | Group | Mean | Median | Std Dev |
|---|---|---|---|---|
| Evaluation Duration (s) | Black Box AI | 70.34 | 49.32 | 76.50 |
| | Narrative AI | 59.32 | 38.74 | 61.44 |
| | No AI | 85.63 | 64.25 | 93.16 |
| Timestamp First Answer (s) | Black Box AI | 60.16 | 38.89 | 73.13 |
| No AI | Narrative AI | 46.39 | 24.13 | 56.07 |
| | No AI | 76.14 | 58.19 | 91.11 |
| Time Between Last Answer and Next (s) | Black Box AI | 9.10 | 4.47 | 13.85 |
| | Narrative AI | 12.39 | 4.77 | 22.09 |
| | No AI | 8.71 | 3.56 | 13.40 |
| Hover Duration (s) | Black Box AI | 3.98 | 0.00 | 11.94 |
| | Narrative AI | 6.52 | 1.13 | 13.49 |
| | No AI | 3.11 | 0.00 | 9.46 |
| Number of Tabs Opened | Black Box AI | 2.19 | 1.00 | 1.92 |
| | Narrative AI | 1.83 | 1.00 | 1.50 |
| | No AI | 2.56 | 1.00 | 2.08 |
| Number of Clicks | Black Box AI | 8.64 | 6.00 | 8.27 |
| | Narrative AI | 6.89 | 6.00 | 4.40 |
| | No AI | 8.55 | 7.00 | 5.37 |
| Number of Scrolls | Black Box AI | 10.76 | 7.00 | 11.37 |
| | Narrative AI | 9.20 | 6.00 | 9.38 |
| | No AI | 13.52 | 12.00 | 11.16 |
| Max Y Reached | Black Box AI | 2452.22 | 1333.50 | 3369.28 |
| | Narrative AI | 2244.57 | 1321.00 | 3121.49 |
| | No AI | 4574.33 | 3258.00 | 4498.67 |
| Returns to Top | Black Box AI | 2.74 | 2.00 | 1.88 |
| | Narrative AI | 2.69 | 2.00 | 1.59 |
| | No AI | 3.10 | 2.00 | 1.87 |
| Attention Top Ratio | Black Box AI | 0.64 | 0.70 | 0.33 |
| | Narrative AI | 0.64 | 0.64 | 0.31 |
| | No AI | 0.52 | 0.44 | 0.30 |

**Table A6: Mann-Whitney U Test Results for Behavioral Variables**

| Variable | Group 1 | Group 2 | U Statistic | p-value |
|---|---|---|---|---|
| Evaluation Duration (s) | No AI | Black Box AI | 16368.0 | 0.0054 |
| | No AI | Narrative AI | 19228.5 | 0.0000 |
| | Black Box AI | Narrative AI | 23280.0 | 0.1705 |
| Timestamp First Answer (s) | No AI | Black Box AI | 16751.5 | 0.0013 |
| | No AI | Narrative AI | 20068.5 | 0.0000 |
| | Black Box AI | Narrative AI | 23863.0 | 0.0648 |
| Time Between Last Answer and Next (s) | No AI | Black Box AI | 12124.0 | 0.0454 |
| | No AI | Narrative AI | 12395.5 | 0.0056 |
| | Black Box AI | Narrative AI | 20428.0 | 0.3390 |
| Hover Duration (s) | No AI | Black Box AI | 13762.0 | 0.8601 |
| | No AI | Narrative AI | 11301.5 | 0.0000 |
| | Black Box AI | Narrative AI | 16523.0 | 0.0000 |
| Number of Tabs Opened | No AI | Black Box AI | 15228.5 | 0.0940 |
| | No AI | Narrative AI | 17642.0 | 0.0014 |
| | Black Box AI | Narrative AI | 23209.0 | 0.1253 |
| Number of Clicks | No AI | Black Box AI | 15490.0 | 0.0717 |
| | No AI | Narrative AI | 18173.0 | 0.0007 |
| | Black Box AI | Narrative AI | 23062.5 | 0.2283 |
| Number of Scrolls | No AI | Black Box AI | 16859.0 | 0.0008 |
| | No AI | Narrative AI | 19345.0 | 0.0000 |
| | Black Box AI | Narrative AI | 22040.0 | 0.7188 |
| Max Y Reached | No AI | Black Box AI | 18532.0 | 0.0000 |
| | No AI | Narrative AI | 20639.5 | 0.0000 |
| | Black Box AI | Narrative AI | 21617.5 | 0.9889 |
| Returns to Top | No AI | Black Box AI | 15762.5 | 0.0312 |
| | No AI | Narrative AI | 16586.0 | 0.0830 |
| | Black Box AI | Narrative AI | 20745.5 | 0.4717 |
| Attention Top Ratio | No AI | Black Box AI | 10867.0 | 0.0006 |
| | No AI | Narrative AI | 11566.5 | 0.0003 |
| | Black Box AI | Narrative AI | 21999.0 | 0.7427 |

**Table A7: Normality Test Results (Shapiro-Wilk Test) for Behavioral Variables by Condition**

| Variable | No AI (p-value) | Black Box AI (p-value) | Narrative AI (p-value) |
|---|---|---|---|
| Evaluation Duration (s) | < 0.0001 | < 0.0001 | < 0.0001 |
| Timestamp First Answer (s) | < 0.0001 | < 0.0001 | < 0.0001 |
| Time Between Last Answer and Next (s) | < 0.0001 | < 0.0001 | < 0.0001 |
| Hover Duration (s) | < 0.0001 | < 0.0001 | < 0.0001 |
| Number of Tabs Opened | < 0.0001 | < 0.0001 | < 0.0001 |
| Number of Clicks | < 0.0001 | < 0.0001 | < 0.0001 |
| Number of Scrolls | < 0.0001 | < 0.0001 | < 0.0001 |
| Max Y Reached | < 0.0001 | < 0.0001 | < 0.0001 |
| Returns to Top | < 0.0001 | < 0.0001 | < 0.0001 |
| Attention Top Ratio | < 0.0001 | < 0.0001 | < 0.0001 |

**Table A8: Statistical Tests Comparing AI Formats**

| Comparison | $\chi^2$ | p-value | Cohen's h | Power |
|---|---|---|---|---|
| Overall Follow Rate | 0.40 | 0.53 | 0.07 | 0.31 |
| Follow Rate (Pass Only) | 2.22 | 0.14 | 0.39 | 0.98 |
| Follow Rate (Fail Only) | 0.29 | 0.59 | 0.08 | 0.26 |

**Table A9: Shapiro–Wilk Normality Test for Outcome-Specific Projection Scores by Group**

| Group | W-statistic | p-value | Interpretation |
|---|---|---|---|
| Overall Follow Rate | 0.892 | 0.0000 | Not normally distributed |
| Follow Rate (Pass Only) | 0.915 | 0.0000 | Not normally distributed |
| Follow Rate (Fail Only) | 0.971 | 0.0048 | Not normally distributed |

**Table A10: Descriptive Statistics for High-Scoring Sessions (≥ 90th and ≥ 95th Percentiles)**

| Feature | Metric Type | ≥ 90th Percentile (n = 26) | ≥ 95th Percentile (n = 13) |
|---|---|---|---|
| Score Threshold | _ | 1.743 | 1.881 |
| Duration (s) | Mean | 9.688 | 8.412 |
| | Median | 8.410 | 8.280 |
| | Std | 4.623 | 3.013 |
| Number of Tabs Opened | Mean | 1.0 | 1.0 |
| | Median | 1.0 | 1.0 |
| | Std | 0.0 | 0.0 |
| Number of Scrolls | Mean | 1.346 | 1.077 |
| | Median | 1.000 | 1.000 |
| | Std | 0.562 | 0.277 |
| Number of Clicks | Mean | 5.000 | 4.923 |
| | Median | 5.000 | 5.000 |
| | Std | 0.980 | 0.641 |
| Max Y Reached | Mean | 5.154 | 0.0 |
| | Median | 0.000 | 0.0 |
| | Std | 17.861 | 0.0 |
| Attention Top Ratio | Mean | 1.0 | 1.0 |
| | Median | 1.0 | 1.0 |
| | Std | 0.0 | 0.0 |