WORK IN PROGRESS

# Let's Be Realistic: AI-Recommender Use in a Complex Management Setting

**JANA GONNERMANN-MÜLLER**, University of Potsdam, Potsdam, Brandenburg, Germany

**KRISTINA SAHLING**, Humboldt University of Berlin, Berlin, Germany

**JENNIFER HAASE**, Humboldt University of Berlin, Berlin, Germany

**Open Access Support** provided by:

**University of Potsdam**

**Humboldt University of Berlin**



**Published:** 26 April 2025

**Citation in BibTeX format**

CHI EA '25: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems
*April 26 - May 1, 2025*
*Yokohama, Japan*

**Conference Sponsors:**
SIGCHI

# Let's Be Realistic: AI-Recommender Use in a Complex Management Setting

Jana Gonnermann-Müller
University of Potsdam
Potsdam, Germany
Weizenbaum Institute
Berlin, Germany
jana.gonnermann@wi.uni-potsdam.de

Kristina Sahling
Humboldt University
Berlin, Germany
Weizenbaum Institute
Berlin, Germany
kristina.sahling@hu-berlin.de

Jennifer Haase
Humboldt University
Berlin, Germany
Weizenbaum Institute
Berlin, Germany
jennifer.haase@hu-berlin.de

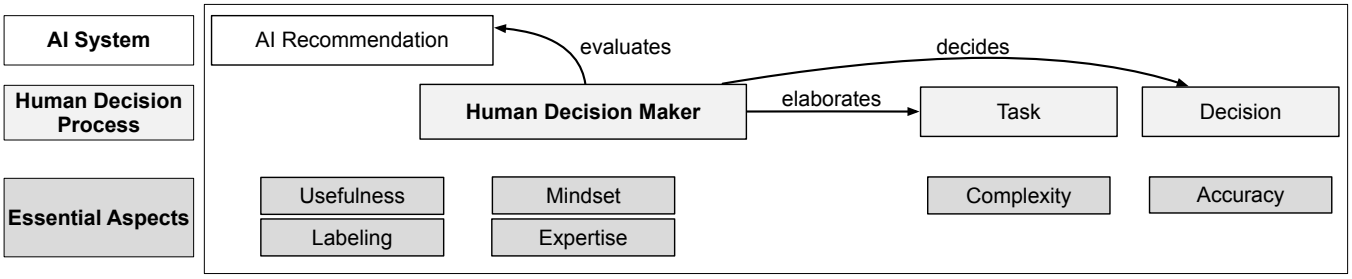


**Figure 1: This Experimental Study Investigates the Influence of Key Elements (Human Decision-Maker Characteristics, AI Recommendation Labeling, Task Attributes) on Decision Accuracy**

**The figure depicts the experimental setup, where an AI system provides recommendations assessed by a human decision-maker whose decision-making process is influenced by factors such as usefulness, labeling, mindset, expertise**

## Abstract

Labels like "AI-powered" or "Human-Expert" activate mental models and shape user decisions. Yet, the transferability of these labels on performance in complex, realistic tasks needs investigation. This study examines how recommender labeling and human factors (mindset, expertise) impact performance in a complex business management scenario. We conducted an online experiment employing a management dashboard, where participants (N = 395) received recommendations labeled as either Artificial Intelligence (AI) or Human-Expert-generated. Unlike previous research, labeling did not significantly influence task performance. Instead, graph literacy and cognitive load were key predictors of performance. Participants with positive attitudes toward AI found recommendations helpful, but their performance did not improve with their use. Expertise seems to be dominant in AI labeling in this context. These findings highlight the interaction between expertise, mindset, and labeling, advocating for further research investigating in which contexts labeling and human factors critically influence performance when using AI recommendations.


*CHI EA '25, Yokohama, Japan*

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720131

## CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

## Keywords

Artificial Intelligence, Decision-Making, Recommender Systems, Management Dashboard, Labeling, Human Factors

**ACM Reference Format:**
Jana Gonnermann-Müller, Kristina Sahling, and Jennifer Haase. 2025. Let's Be Realistic: AI-Recommender Use in a Complex Management Setting. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3706599.3720131

## 1 Introduction

Decision-making lies at the heart of managerial tasks, especially in complex organizational environments where data-driven insights play an increasingly important role. As the variety, velocity, and amount of data increase, Recommender Systems (RS) support decision quality by generating task-specific recommendations [9, 11]. A RS aims to decrease the potentially great number of alternatives [32] and deal with the information overload problem common for management tasks in real-life [9, 11, 40].

Within Human-Computer Interaction (HCI), RSs are studied from different perspectives while mainly taking on a technical focus, including different RS algorithms [1, 14] or concepts like RS fairness [8]. We also know from adjacent areas of research that human-related factors and psychological concepts are important when engaging with RSs in practice. Especially the user's mindset

and expertise influence their ability to engage with, trust, and critically evaluate RS outcome [4, 15, 17, 22, 27]. Thus, how RSs are branded, as "AI-powered" or "algorithmic", can reshape user trust and reliance by activating distinct mental models–also referred to as the labeling-effect [25]. Experiments by Langer et al. [22] have shown that differing terminology used to describe algorithmic decision-making systems influences laypeople's perception (e.g., complexity) and evaluation (e.g., trust). Further, positive or negative beliefs about AI accuracy influence actual decision outcomes, irrespective of the system's actual performance, also known as the "AI-placebo" effect [18, 34]. These examples show the necessity of mindful labeling and considering additional aspects of mental models evoked by such labels to ensure (a) robust and replicable research outcomes and (b) the accurate and effective application of RS in practical contexts.

Prior research has established the existence of labeling effects [16, 25, 36] and provided valuable insights into the human role in RS use. So far, these experimental studies relied on simplified decision or task scenarios, such as binary choices [15], preference evaluations [13, 24] or puzzle-like tasks [18], showing increased internal validity. However, to manage the complexity of real-world decision-making, particularly in managerial contexts where users must integrate contextual knowledge and manage higher task demands [7, 15], high external validity is essential to transfer results from research into practice.

With a focus on external validity, this study builds on existing HCI research and develops an experimental setting with comparably high task complexity on the basis of a concrete management scenario. Arnold and Sutton define task complexity as "the degree to which task completion or resolution taxes the cognitive abilities of the decisionmaker" [3, p.180]. Thus, participants were asked to solve management tasks supported by a dashboard with either *AI* or *Human-Expert* recommendations. We investigate the roles of *RS labeling, user mindset, and expertise* to deepen the understanding of human factors influencing decision-making in our scenario (Figure 1). Through the use of a management dashboard with RS simulating real-world scenarios, we address the following research questions:

> *RQ 1: To what extent does recommender systems' labeling affect objective performance in complex tasks?*
> *RQ 2: How do mindset and expertise influence objective performance in complex tasks when using a dashboard with recommendations?*

Our findings provide insights into the impact of RS labeling and human factors on task performance, informed by *AI* and *Human-Expert* recommendations. Results reveal that expertise, rather than labeling or mindset, dominates task performance. These results underscore the importance of user proficiency in leveraging RSs effectively and call for more external valid experimental designs in HCI to capture the nuanced dynamics of human-AI interaction in complex decision environments.

## 2 Background

### 2.1 Influence of Labeling of Recommender Systems on Task Performance

The labeling of RS plays an important role in shaping user perceptions, trust, and reliance on algorithmic decision aids. Labels such as "AI-powered", "algorithm", or "computer program" activate distinct mental models that influence how users interpret system outputs and their roles in decision-making [22]. For example, systems labeled "AI" often create high competence and precision expectations, particularly in objective domains such as financial analysis [25]. However, in ethically sensitive or subjective tasks like hiring, users may perceive AI as lacking contextual understanding, leading to skepticism [26, 39]. Misaligned labels can undermine trust and reduce critical evaluation. Erroneous or unhelpful recommendations labeled as "AI" exacerbate skepticism, especially when expectations of precision are unmet [7]. Conversely, helpful recommendations foster trust and improve decision outcomes in complex tasks, highlighting the importance of aligning labels with system capabilities [41].

### 2.2 Influence of Human Factors on Task Performance

Prior research has found various human factors relevant for the perception and use of RSs. First, a positive attitude, such as openness to innovation, promotes trust and greater reliance on AI, especially in tasks where computational precision enhances decision quality [27]. In contrast, skepticism or algorithm aversion can lead to either disengagement or overreliance on human judgment, even when AI recommendations are superior [7, 26]. Trust is a key determinant of AI adoption, shaped by system transparency and past user experiences [2, 23]. Transparent systems, where recommendation processes are clearly explained, foster critical engagement and improve decision-making [2]. However, blind reliance on AI can lead to errors, emphasizing the need for informed trust that encourages critical evaluation and effective use of RS outputs [28].

Furthermore, it was found that user expertise significantly moderates interactions with RSs. Experts with contextual knowledge critically evaluate AI outputs and align recommendations with task-specific goals, resulting in higher decision quality [7]. Conversely, users with limited expertise may struggle to assess recommendations, leading to overreliance or disengagement [16]. Familiarity with AI tools further influences trust and reliance. Experienced users develop a nuanced understanding of AI's strengths and limitations, enabling more balanced engagement [7], while inexperienced users may display skepticism or hesitation, especially in high-stakes scenarios [41]. Additionally, task complexity amplifies these dynamics. In cognitively demanding contexts, helpful recommendations enhance trust, whereas inconsistent outputs erode confidence and discourage system use [7, 37]. Designing adaptive RSs that accommodate varying levels of expertise is crucial to fostering effective decision-making and maintaining user trust.

## 3 Methods

To investigate our RQs, we created an experimental setup to examine the RS labeling effect and its impact on task performance

(RQ 1), along with the influence of individual factors (RQ 2). The experiment was developed with high external validity in mind; thus, we mimicked a realistic business scenario, including complex decision-making tasks. The study used a one (labeling of *AI* vs. *Human-Expert*) between-subject design accounting for the influence of human factors of expertise, mindset, and cognitive capacities. The study design was approved by the ethics committee of the authors' research institute.

## 3.1 Procedure

Participants were randomly assigned to the *AI* or the *Human-Expert* recommendation group. Both groups received a dashboard with identical information and completed the same set of tasks. At the bottom of the dashboard, task-specific recommendations were provided to assist participants in completing their decision-making tasks. These recommendations were designed to examine the labeling effect, where the same content was framed differently to be perceived as either AI-generated or Human-Expert-generated. The recommendations were tailored to the specific tasks but remained identical across both conditions, allowing for a controlled comparison of how labels influenced user trust, reliance, and task performance. The study has three parts (see Figure 2). In the first part of the experiment, participants were introduced to the cover story, in which they assumed the role of a supermarket manager using an accompanying management dashboard. They began with two comprehension checks to familiarize themselves with the dashboard's information and structure. Following these comprehension checks, participants completed the first block of six tasks. After every six questions, two additional comprehension checks were conducted before starting the next block of six tasks. Alongside task performance, participants indicated whether they relied on the information from the dashboard and the recommendations (see Appendix, Figure 3). After completing all tasks, participants proceeded to the second part of the study, where they answered questions about their cognitive load, maintaining the perspective of the supermarket manager role introduced earlier. In the final section, participants were asked to exit the manager role and complete questionnaires assessing general attitudes toward AI, trust in algorithmic decisions, AI literacy, graph literacy, and expertise with reporting systems.

## 3.2 Task and Dashboard Application

Participants assumed the role of a supermarket manager responsible for improving customer satisfaction to meet key performance indicators (KPIs) set by the parent company. The supermarket story contains the note that traditionally, managers relied on Excel spreadsheets and gut feeling to make decisions. Therefore, the dashboard was presented as a new analytics tool that has been introduced to enhance decision-making, aiming to optimize store performance and customer satisfaction, e.g., by reducing waiting times. The dashboard presents precise, visual KPI data through graphs, however, it is static in all the representations. The dashboard includes meta information like the date and time, as well as explanations for the top three KPIs. Visualizations include bar and line charts, with specific colour schemes and KPI boxes placed prominently at the top. The data presented in the dashboard consists of mock-up data, meaning that all figures were created to appear logical and feasible within a supermarket management context, rather than being derived from real datasets. This approach ensures experimental control while maintaining realism, preventing participants from being influenced by inconsistencies that might arise in real-world data. Figure 3 (see Appendix) illustrates the AI group's dashboard, showing both retrospective (blue) and predictive (orange) insights. Key KPIs include *Customer per day*, *Waiting time at the cash desk*, *Product Sales and Sales Forecast (in Euro)*, and *Product Variation List*. The dashboard remains the same across all decision-making tasks to simplify the experience and minimize potential cognitive overload. At the bottom of the dashboard, task-specific recommendations for complex tasks are provided, tailored to the tasks, but remain identical for both *AI* and *Human-Expert* group. This setup was designed to achieve high external validity, ensuring that the study reflects real-world managerial decision-making. The decision tasks were designed to reflect realistic managerial challenges requiring analytical reasoning, problem-solving, and strategic thinking. Participants were not merely asked to choose between predefined options; instead, they had to synthesize multiple data points to make decisions aligned with business goals. This complexity ensures that the study captures cognitive processing under realistic constraints, addressing a major limitation of previous experimental studies that used simplified decision scenarios with limited contextual depth [5, 7]. Therefore, the task design captures these key characteristics while maintaining experimental control. Participants were assigned to either receive *AI* and *Human-Expert* recommendations alongside the identical dashboard. The framings for both groups were:

> *AI: "the AI recommendation includes an AI-powered recommendation service that analyzes collected data using algorithms to generate actionable suggestions automatically."*
> *Human-Expert: "the expert recommendation includes a recommendation service powered by a subject matter expert (SME) who analyzes the data and generates suggestions based on their experience and domain knowledge."*

## 3.3 Measurements

We recorded task-related aspects, mindset-related, and expertise-related variables (for a complete overview, see Table 1 in Appendix). Demographics such as gender, age, work experience, and position were provided by the participant recruitment platform Prolific. The study focuses on a realistic management scenario, including only participants with prior management experience, to further enhance the external validity of the study.

**Task-related Aspects.** Complex *task performance* was assessed using twelve tasks divided into two subsets: (a) tasks that could only be solved by utilizing the information provided by the *AI/Human-Expert* recommender–marked as important task– and (b) tasks that could be solved solely based on the dashboard. All recommendations we factually correct. As part of a larger research project involving the distinction of two subsets, this paper focuses on the statistical analysis of the important tasks only. We employed a "Wizard-of-Oz" RS [38], which delivered identical information, like additional facts to the specific situation relevant for the tasks for both groups. All tasks required participants to analyze dashboard

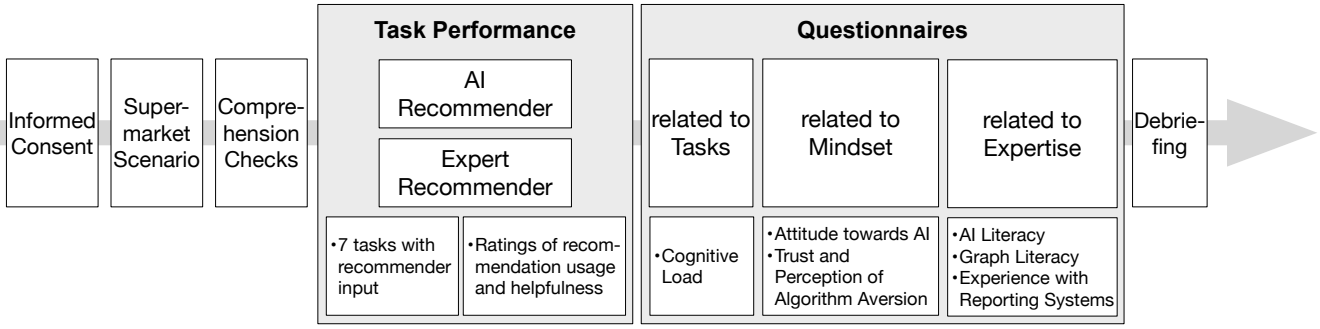


**Figure 2: Applied Experimental Design for the Survey**

**Flow diagram of the experimental design, beginning with informed consent, followed by a supermarket scenario, comprehension checks, task performance with AI or expert recommenders, questionnaires, and concluding with debriefing.**

information and make decisions aimed at improving customer satisfaction. Each task was presented as a multiple-choice question with one correct answer out of four options. The tasks were not time-constrained. One example task is: "How should you adjust your staffing and promotion strategy to accommodate the surge in customers?" the correct answer is: "Allocate additional staff to cash desks and launch a targeted promotion during the local event". The tasks were evaluated and refined through two rounds of pretesting to ensure validity and clarity[1].

To gain insights into user perception of recommendation and dashboard, we assessed the *Utilization and Evaluation of Recommendations' Helpfulness* with two items: (1) The first item evaluated the usefulness of the respective recommendation on a scale ranging from 1 "Not at all useful" to 5 "Totally useful", with an additional option of 0 "Did not use it". (2) The second item assessed the same for the usefulness of the dashboard.

**Expertise-related Aspects.** We used the *AI literacy* scale by [31] to capture the participants' self-rated perception of their AI knowledge and experience in usage and design. We assessed *Cognitive Load* with a questionnaire by [20] that distinguished the concept into *Mental load* and *Mental effort*. To collect participants' *Graph Literacy*, we used the short graph literacy by [29]. The scale presents four visualizations, each requiring participants to provide multiple-choice answers or open-ended responses based on the information displayed. To measure the *Experience with Reporting System*, we developed the following item: "When you think about a typical workday, how often do you use reporting systems?" we assessed experience with reporting systems on a scale from 1 "Never" to 5 "Regularly".

**Mindset-related Aspects.** To assess aspects of *Trust and Perceptions of Recommender*, four variables were measured using scales adapted from [12] measured on a 7-point Likert scale ranging from 1 "Strongly disagree" to 7 "Strongly agree". We tailored the items and instructions to align with the experimental condition, ensuring they addressed the corresponding recommender. Items include *AI Integrity* (e.g., "I think the AI Recommender makes unbiased decisions"), *AI Trust* (e.g., "I would heavily rely on the AI Recommender"), *AI Fairness* (e.g., "I think the AI Recommender makes fair recommendations"), and *AI Acceptance* (e.g., "I think I would accept the AI recommendation").

To evaluate *General Attitudes toward AI* for participants in the AI group, a questionnaire with two scales from [33] was employed: *Positive General Attitudes toward AI* (e.g., "I am interested in using Artificial Intelligence systems in my daily life" and "Artificial Intelligence can provide new economic opportunities for this country") and *Negative General Attitudes toward AI* (e.g., "I think Artificial Intelligence is dangerous" and "Artificial Intelligence might take control of people"). Responses were rated on a 5-point Likert scale ranging from 1 "Strongly disagree" to 5 "Strongly agree".

## 3.4 Sample

We had 395 participants in the main experiment. We excluded 32 data sets due to less than 50% correct answers in comprehension checks. This left 363 participants (140 female, 221 male, 2 prefer not to say) aged 18 to 69 ($m$ = 38.69, $SD$ = 10.72). Participants were randomly assigned to either the *AI* group ($N$ = 183; 50.4%) or the *Human-Expert* recommendation group ($N$ = 180; 49.6%). The majority were managers with 76.6%, with senior managers making up 23.4%. Work experience varied, with 42.5% having over 16 years, 21.0% with 6–10 years, 16.3% with 3–5 years, 15.7% with 11–15 years, and 4.4% with 1–2 years. The average survey duration was approximately 28.95 minutes ($SD$ = 13.16 min). Table 2 (see Appendix) summarizes the characteristics of the sample.

# 4 Results

To address *RQ 1*, the research data is compared between the two groups receiving either *AI* or *Human-Expert* recommendations using an ANCOVA. The influence of additional factors such as mindset towards the *AI/Human-Expert* recommendation and expertise on task performance was analyzed using linear regression *(RQ 2)*. The data was tested for compliance with the statistical requirements of both measurements. Levene test was used to confirm the assumption of homogeneity of variances across groups for tasks with important input ($F(1, 361) = 1.25$, $p > .265$), which confirmed no deviations from the assumption. We used the Variance Inflation Factor (VIF) to check for multicollinearity in regression analysis. No multicollinearity was detected among the predictors. In addition to answering the RQs, we conducted exploratory analyses to examine

[1]We conducted a first qualitative pretest with N = 12 and a second quantitative pretest with N = 40

variations in participants' utilization of AI recommendations for task completion and their evaluations of the recommendations' perceived helpfulness.

### 4.1 Effects of the Labeling of Recommendation System on Complex Task Performance

The participants performed well for tasks with important recommendations ($M$ = 5.83, $SD$ = 1.25). A comparison between the two groups revealed no performance differences, with participants receiving *AI* ($M$ = 5.84, $SD$ = 1.21) and *Human-Expert* ($M$ = 5.81, $SD$ = 1.30) recommendations performing equally well. Table 3 (see Appendix) provides a detailed overview of the means, standard deviations, and F-test results. To address *RQ 1*, we performed an ANCOVA calculation, including the complex task performance as the dependent variable and the labeling as the independent variable. We add expertise-based variables, such as graph literacy, experience with reporting systems, AI literacy, and cognitive load, as control variables into the ANCOVA to account for variances in task performance due to individual differences. Against prior research findings that the labeling of a dashboard-integrated RS leads to differences in complex task performance, the ANCOVA results reveal that both groups performed equally in the tasks regardless of the recommendations' labeling ($F(1,357) = 0.99$, $p = .32$). About the control variables that we included into the ANCOVA, our results show an effect of graph literacy on task performance ($F(1,357) = 37.31$, $p < .001$). In addition, the self-reported cognitive load arising from the tasks significantly influenced task performance ($F(1,357) = 9.57$, $p < .001$), with higher task results for participants reporting less cognitive demand.

### 4.2 Effects of Expertise and Mindset on Complex Task Performance within a Dashboard Scenario

To address *RQ 2*, we analyze variables influencing task performance within each subgroup, performing linear regression. In addition to expertise-related variables –graph literacy, AI literacy, experience with a RS, and cognitive load–, we integrated mindset-related variables into the analysis, such as integrity, trust, fairness, and acceptance of the respective recommendation. For participants receiving the *AI recommendation*, we added their positive and negative attitudes towards AI. Calculations reveal a significant linear model ($F(9,173) = 3.43$, $p < .01$). The results show a significant coefficient for graph literacy ($\beta = .23$, $p < .01$) and cognitive load ($\beta = -.19$, $p < .05$) on task performance indicating that participants with higher graph literacy perform better when receiving *AI recommendations* and participants with higher cognitive load show reduced task performance. Similar effects of graph literacy ($\beta = .50$, $p < .001$) can be observed in the *Human-Expert recommendation* group. Again, the linear model was significant ($F(6,173) = 8.04$, $p < .01$). In both groups, we found no significant impact for variables related to participants' mindsets on task performance.

### 4.3 Exploratory Analysis and Additional Observations

In addition to conducting the confirmatory analyses, we also explored the explanation of variance in participants' usage and perception of the *AI recommendation*. After each task, participants reported whether they used the dashboard and the *AI/Expert* recommendations for completing the tasks and rated the usefulness of the dashboard and recommendation in achieving task goals. We conducted linear regression, including perceived helpfulness or reported use as dependent variables and expertise-related factors, such as AI literacy, graph literacy, cognitive load, mindset-related variables, attitude toward AI, trust, acceptance, fairness, and integrity as independent variables. For participants receiving *Human-Expert* recommendation, the linear regression on perceived helpfulness of the *Human-Expert* recommendation revealed trust as a significant coefficient ($\beta = .2$, $p < .001$) on the helpfulness of the *Human-Expert* recommendation. In addition, we were also interested in whether participants utilized the AI recommendations for task completion. For participants who received *AI* recommendations, the findings indicate trust as a significant coefficient ($\beta = .38$, $p < .01$). Trust was also significant coefficient for those who received *Human-Expert* recommendations ($\beta = .26$, $p < .001$).

## 5 Discussion

Our study investigated the impact of RS labeling (*AI vs. Human-Expert*), mindset toward AI, and expertise on task performance within a complex management setting. Unlike prior research focusing on simplified decision tasks, our experiment involved actual managers working in a realistic, high-complexity decision environment. Regarding our *first research question*, task performance did not significantly differ between AI-labeled and Human-Expert-labeled recommendations. Instead, graph literacy and cognitive load emerged as the primary predictors of performance, supporting the view that expertise, rather than labeling effects, determines decision-making efficiency in data-driven environments [7, 30]. For our *second research question*, mindset factors, such as trust and acceptance of AI, did not significantly affect task performance. However, exploratory analysis revealed that AI-related attitudes positively influenced *subjective evaluations*—such as perceived helpfulness and reported usage of recommendations—indicating that while mindset affects how users engage with RSs, it does not necessarily translate into better decision accuracy.

These findings align with broader discussions in HCI and RS research, suggesting that once tasks reach a certain level of complexity, *cognitive abilities directly relevant to the task* become the primary determinant of performance, while factors such as mindset, RS labeling, and system explanations play a lesser role [16, 19]. The existing HCI literature often emphasizes technical RS aspects, including algorithmic efficiency, design patterns, and explanation mechanisms, while prioritizing internal validity in controlled experimental studies [1, 14, 21, 37]. However, our results suggest that integrating these factors into a high-complexity, externally valid setting shifts the primary performance drivers toward *comprehension of the dashboard and tasks*, reinforcing the importance of expertise in extracting meaningful insights from RSs.

Another key observation relates to the *AI placebo effect*, which has gained attention in recent RS and HCI research [19, 36]. While prior studies highlight the psychological influence of AI labels on perceived system effectiveness, our results suggest that in high-complexity decision-making contexts, managers' performance is not meaningfully influenced by these labels or their associated mental models. Instead, users relied on their expertise and task-related cognitive processing, with trust in AI affecting subjective evaluation metrics rather than objective performance outcomes. This finding calls into question the generalizability of AI placebo effects in structured managerial decision tasks, where domain knowledge and analytical skills take precedence over heuristic-driven biases.

Our study also underscores the methodological challenges of achieving external validity in RS experiments. While many prior studies focus on labeling effects, mental models, and system explanations in controlled settings, our findings emphasize the necessity of designing RS experiments that reflect real-world decision-making conditions [6, 35]. The interaction between user expertise, task complexity, and system design must be better understood to develop RSs that effectively support decision-makers navigating dynamic, ambiguous, and high-stakes environments. These results suggest that future research should move beyond isolated labeling effects to examine how expertise, task structure, and cognitive load jointly influence the effective use of AI-driven decision support systems.

### 5.1 Limitations

As with all experimental work, there are limitations. First, the dashboard design lacked interactivity, which might have reduced user engagement [10]. We do not believe this impacted task performance as the overall accuracy was high, and the reported cognitive load was relatively low. Second, the absence of explainable AI features, such as transparency regarding KPI and recommendation calculations, could have affected trust and decision behavior and thus mask the labeling effects commonly found. Third, our sample may not represent the diversity of real-world decision-makers. Expertise is task-related, and we had managers in our sample who had experience in dashboard use. Moreover, the exclusion of contextual factors, such as time pressure, organizational culture, and competing priorities, reduces the real-world applicability of our findings.

### 5.2 Future Research

We have identified three main areas for future research. First, designing the dashboard as an interactive dashboard –using filters and drop-downs– means allowing users to have personalized views and filter data to potentially further improve task performance. Through methods like click analysis and eye-tracking, recommendation usage, users' learning behavior, and cognitive load can be measured more explicitly, and the influence of more realistic dashboard-RS scenarios in complex task settings can be examined. Second, building RSs providing clear explanations of recommendation generation processes and accuracy levels might foster transparency and trust. Thus RSs could evolve into tools that not only provide suggestions but also help users build expertise over time, improving decision-making and performance in complex environments. Third, applying the dashboard-RS setup across diverse management domains would enhance claims of generalizability. The effects of mindset and cognitive load may vary across contexts, requiring RSs to adapt to specific needs and demands.

## 6 Conclusion

This research offers evidence-based insights into task performance with data-driven dashboards featuring AI- or Human-Expert-generated recommendations. Contrary to prior research findings, recommendation labeling did not significantly affect task performance. Instead, graph literacy and cognitive load emerged as key predictors of performance. Notably, a disconnect was identified: while participants with positive attitudes toward AI perceived recommendations as more helpful, these attitudes did not translate into improved performance. The findings suggest that the determinants of performance may shift significantly as task complexity increases. Specifically, when tasks reach a high level of complexity, cognitive abilities directly relevant to the task emerge as the primary predictors of performance. In contrast, human factors such as mindset and labeling effects, frequently emphasized in prior research, appear to have a diminished impact on objective performance. Instead, their influence is more pronounced in shaping subjective evaluations rather than objective performance outcomes. This shift underscores the importance of considering task complexity and real-life applicability when conducting experiments. It highlights the need for further research to explore the nuanced interplay between cognitive abilities, human factors, and task complexity in both experimental and applied settings.

## Acknowledgments

Jana Gonnermann-Müller's work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DII137 (Weizenbaum-Institute). Kristina Sahlings and Jennifer Haase's work was supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DII133 (Weizenbaum-Institute).

## References


[1] Marie Al-Ghossein, Talel Abdessalem, and Anthony BARRÉ. 2021. A Survey on Stream-Based Recommender Systems. *ACM Comput. Surv.* 54, 5 (2021), 104:1–104:36. doi:10.1145/3453443

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. doi:10.1145/3290605.3300233

[3] V. Arnold and Steve Sutton. 1998. The theory of technology dominance: Understanding the impact of intelligent decision aids on decision makers' judgments. *Advances in Accounting Behavioral Research* 1 (Jan. 1998), 175–194.

[4] Izak Benbasat and Weiquan Wang. 2005. Trust in and adoption of online recommendation agents. *Journal of the association for information systems* 6, 3 (2005), 4.

[5] Alex Bennet and David Bennet. 2008. The Decision-Making Process in a Complex Situation. In *Handbook on Decision Support Systems 1: Basic Themes*, Frada Burstein and Clyde W. Holsapple (Eds.). Springer, Berlin, Heidelberg, 3–20.

[6] Frada Burstein and Clyde W. Holsapple. 2008. Handbook on Decision Support Systems 2. (2008).

[7] Jason W. Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2020), 220–239. doi:10.1002/bdm.2155

[8] Simon Caton and Christian Haas. 2024. Fairness in Machine Learning: A Survey. *ACM Comput. Surv.* 56, 7 (April 2024), 166:1–166:38. doi:10.1145/3616865

[9] Li Chen, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, Francesco Ricci, and Giovanni Semeraro. 2013. Human Decision Making and Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 3, 3 (Oct. 2013), 17:1–17:7. doi:10.1145/2533670.2533675

[10] Juan Ignacio del Valle and Francisco Lara. 2024. AI-powered recommender systems and the preservation of personal autonomy. *AI & SOCIETY* 39, 5 (Oct. 2024), 2479–2491. doi:10.1007/s00146-023-01720-2

[11] Gerald Häubl and Valerie Trifts. 2000. Consumer Decision Making in Online Shopping Environments: The Effects of Interactive Decision Aids. *Marketing Science* 19, 1 (Feb. 2000), 4–21. doi:10.1287/mksc.19.1.4.15178

[12] Miriam Höddinghaus, Dominik Sondern, and Guido Hertel. 2021. The automation of leadership functions: Would people trust decision algorithms? *Computers in Human Behavior* 116 (March 2021), 106635. doi:10.1016/j.chb.2020.106635

[13] Nanami Ishizu, Wen Liang Yeoh, Hiroshi Okumura, and Osamu Fukuda. 2024. The Effect of Communicating AI Confidence on Human Decision Making When Performing a Binary Decision Task. *Applied Sciences* 14, 16 (Jan. 2024), 7192. doi:10.3390/app14167192

[14] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5 (2021), 105:1–105:36. doi:10.1145/3453154

[15] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2024. An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making. *MIS Quarterly* (2024). doi:10.25300/MISQ/2024/18512

[16] Agnes Mercedes Kloft, Robin Welsch, Thomas Kosch, and Steeven Villa. 2024. "AI enhances our performance, I have no doubt this one will do the same": The Placebo effect is robust to negative descriptions of AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–24. doi:10.1145/3613904.3642633

[17] Sherrie Y. X. Komiak and Izak Benbasat. 2006. The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents. *MIS Quarterly* 30, 4 (2006), 941–960. doi:10.2307/25148760 Publisher: Management Information Systems Research Center, University of Minnesota.

[18] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2022. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Transactions on Computer-Human Interaction* 29, 6 (Dec. 2022), 1–32. doi:10.1145/3529225

[19] Thomas Kosch, Robin Welsch, Lewis Chuang, and Albrecht Schmidt. 2023. The Placebo Effect of Artificial Intelligence in Human–Computer Interaction. *ACM Trans. Comput.-Hum. Interact.* 29, 6 (Jan. 2023), 56:1–56:32. doi:10.1145/3529225

[20] Moritz Krell. 2015. Evaluating an instrument to measure mental load and mental effort using Item Response Theory. *Science Education Review Letters* 2015, 1 (April 2015). doi:10.18452/8212

[21] Kislaya Kunjan, Bradley Doebbeling, and Tammy Toscos. 2019. Dashboards to Support Operational Decision Making in Health Centers: A Case for Role-Specific Design. *International Journal of Human–Computer Interaction* 35, 9 (May 2019), 742–750. doi:10.1080/10447318.2018.1488418

[22] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–28. doi:10.1145/3491102.3517527

[23] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (March 2004), 50–80. doi:10.1518/hfes.46.1.50_30392

[24] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (Jan. 2018), 205395171875668. doi:10.1177/2053951718756684

[25] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (March 2019), 90–103. doi:10.1016/j.obhdp.2018.12.005

[26] Hasan Mahmud, A. K. M. Najmul Islam, and Ranjan Kumar Mitra. 2023. What drives managers towards algorithm aversion and how to overcome it? Mitigating the impact of innovation resistance through technology readiness. *Technological Forecasting and Social Change* 193 (Aug. 2023), 122641. doi:10.1016/j.techfore.2023.122641

[27] Christian Montag, Johannes Kraus, Martin Baumann, and Dmitri Rozgonjuk. 2023. The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Computers in Human Behavior Reports* 11 (Aug. 2023), 100315. doi:10.1016/j.chbr.2023.100315

[28] Kathleen L. Mosier, Linda J. Skitka, Susan Heers, and Mark Burdick. 1998. Automation bias: Decision making and performance in high-tech cockpits. *The International journal of aviation psychology* 8, 1 (1998), 47–63.

[29] Yasmina Okan, Eva Janssen, Mirta Galesic, and Erika A. Waters. 2019. Using the Short Graph Literacy Scale to Predict Precursors of Health Behavior Change. *Medical Decision Making: An International Journal of the Society for Medical Decision Making* 39, 3 (April 2019), 183–195.

[30] John W. Payne. 1993. *The adaptive decision maker*. Cambridge University Press.

[31] Marc Pinski and Alexander Benlian. 2023. AI Literacy - Towards Measuring Human Competency in Artificial Intelligence. doi:10.24251/HICSS.2023.021

[32] Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). 2011. *Recommender Systems Handbook*. Springer US, Boston, MA.

[33] Astrid Schepman and Paul Rodway. 2020. Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports* 1 (Jan. 2020), 100014. doi:10.1016/j.chbr.2020.100014

[34] Alexander Skulmowski. 2024. Placebo or Assistant? Generative AI Between Externalization and Anthropomorphization. *Educational Psychology Review* 36, 2 (June 2024), 58. doi:10.1007/s10648-024-09894-x

[35] Van Hau Trieu, Andrew Burton-Jones, Peter Green, and Sophie Cockcroft. 2022. Applying and Extending the Theory of Effective Use in a Business Intelligence Context. *MIS Quarterly* 46, 1 (2022), 645–678.

[36] Steeven Villa, Robin Welsch, Alena Denisova, and Thomas Kosch. 2024. Evaluating Interactive AI: Understanding and Controlling Placebo Effects in Human-AI Interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–4. doi:10.1145/3613905.3636304

[37] Lucian L. Visinescu, Mary C. Jones, and Anna Sidorova. 2017. Improving Decision Quality: The Role of Business Intelligence. *Journal of Computer Information Systems* 57, 1 (Jan. 2017), 58–66. doi:10.1080/08874417.2016.1181494

[38] Sruthi Viswanathan, Behrooz Omidvar-Tehrani, Adrien Bruyat, Frédéric Roulland, and Antonietta Maria Grasso. 2020. Hybrid Wizard of Oz: Concept Testing a Recommender System. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3334480.3383097

[39] Chao Wang, Hengshu Zhu, Chen Zhu, Xi Zhang, Enhong Chen, and Hui Xiong. 2020. Personalized Employee Training Course Recommendation with Career Development Awareness. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1648–1659. doi:10.1145/3366423.3380236

[40] Bo Xiao and Izak Benbasat. 2007. E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly* 31, 1 (2007), 137–209. doi:10.2307/25148784

[41] Koji Yatani, Zefan Sramek, and Chi-Lan Yang. 2024. AI as Extraherics: Fostering Higher-order Thinking Skills in Human-AI Interaction. doi:10.48550/arXiv.2409.09218

# A Appendix

## A.1 Dashboard Design


**Figure 3: Dashboard Design for the AI Recommender Experimental Condition**

## A.2 Measurements

**Table 1: Overview of Tasks, Variables, and Items used in the Experimental Study**

| Focus | Variable | Items | Response Format |
|---|---|---|---|
| **Task-related Aspects** | Task Performance | 12 tasks, e.g.,<br>*Task:* "How should you adjust your staffing and promotion strategy to accommodate the surge in customers?"<br>*Correct answer:* "Allocate additional staff to cash desks and launch a targeted promotion during the local event."<br>*Task:* "How should you manage the cash desk operations to reduce the waiting time to 6 minutes or less?"<br>*Correct answer:* "Open two cash desks and reallocate one colleague from the stockroom to the cash desk." | Multiple choice (1 correct answer out of 4 options) |
| | Utilization and Evaluation of Recommendations' Helpfulness | (1) Usefulness of the recommendation: "How helpful was the recommendation?"<br>(2) Usefulness of the dashboard: "How helpful was the dashboard?" | Scale: 1 "Not at all helpful" to 5 "Very helpful" (additional option 0 "Did not use it") |
| **Expertise-related Aspects** | AI Literacy [31] | Self-reported knowledge and experience in AI usage and design | Scale: 1 "Strongly disagree" to 7 "Strongly agree" |
| | Cognitive Load [20] | (1) *Mental load*: "The tasks were difficult to complete."<br>(2) *Mental effort*: "I put effort into the task." | Scale: 1 "Not at all" to 7 "Totally" |
| | Graph Literacy [29] | Four visualizations with multiple-choice or open-ended tasks | Multiple choice or open-ended responses |
| | Experience with Reporting System | "When you think about a typical workday, how often do you use reporting systems?" | Scale: 1 "Never" to 5 "Regularly" |
| **Mindset-related Aspects** | Trust and Perceptions of Recommender [12] | (1) *AI Integrity*: "I think the AI Recommender makes unbiased decisions."<br>(2) *AI Trust*: "I would heavily rely on the AI Recommender."<br>(3) *AI Fairness*: "I think the AI Recommender makes fair recommendations."<br>(4) *AI Acceptance*: "I think I would accept the AI Recommender's recommendations." | Scale: 1 "Strongly disagree" to 7 "Strongly agree" |
| | General Attitudes toward AI [33] | (1) *Positive General Attitudes toward AI*: "I am interested in using AI systems in my daily life."<br>(2) *Negative General Attitudes toward AI*: "I think AI is dangerous." | Scale: 1 "Strongly disagree" to 5 "Strongly agree" |

## A.3 Sample

**Table 2: Sample Description including Demographics and Control Variables**

| Demographics | Absolute | AI Group | Expert Group |
|---|---|---|---|
| **Age** mean (SD) | 38.69 (10.72) | 38.72 (11.2) | 38.65 (10.23) |
| **Gender** abs (%) | | | |
| Female | 140 (38.57) | 72 (39.34) | 68 (37.78) |
| Male | 221 (60.88) | 110 (60.11) | 111 (61.67) |
| No answer | 2 (0.55) | 1 (0.55) | 1 (0.56) |
| **Role** abs (%) | | | |
| Manager | 278 (76.58) | 140 (76.50) | 138 (76.67) |
| Senior Manager | 85 (23.42) | 43 (23.50) | 42 (23.33) |
| **Working Experience** abs (%) | | | |
| 1-2 years | 16 (4.42) | 5 (2.73) | 11 (6.15) |
| 3-5 years | 59 (16.30) | 34 (18.58) | 25 (13.97) |
| 6-10 years | 76 (21.00) | 36 (19.67) | 40 (22.35) |
| 11-15 years | 57 (15.75) | 27 (14.75) | 30 (16.76) |
| 16+ years | 154 (42.54) | 81 (44.26) | 73 (40.78) |
| **Experience with Reporting System** mean (SD) | 3.06 (1.10) | 3.05 (1.06) | 3.07 (1.15) |
| **Short Graph Literacy** mean (SD) | 2.30 (1.09) | 2.17 (1.09) | 2.42 (1.07) |
| **AI Literacy Overall Items** mean (SD) | 3.86 (1.65) | 3.83 (1.74) | 3.89 (1.56) |

## A.4 Outcomes

**Table 3: Results of Task Performance and Evaluation of Dashboard and Recommendation for the Full Sample, AI, and Expert Group**

| Variable | Full Sample | AI Group | Expert Group | *F* | *p* |
|---|---|---|---|---|---|
| Task Performance | 5.83 (1.25) | 5.84 (1.21) | 5.81 (1.30) | 0.99 | .32 |
| Dashboard Use | 6.63 (1.16) | 6.54 (1.32) | 6.72 (.97) | | |
| Recommendation Use | 6.7 (1.04) | 6.56 (1.26) | 6.84 (.75) | | |
| Recommendation Evaluation | 4.14 (.75) | 4.12 (.80) | 4.16 (.69) | | |
| Evaluation of Dashboard | 3.77 (.86) | 3.72 (.91) | 3.82 (.81) | | |

*Note:* Values are presented as M (SD). Dashboard Use and Recommendation Use are sum scores: 0–7 for tasks with necessary recommendations.