

HEC MONTRÉAL
École affiliée à l'Université de Montréal

**An Exploration of Human-Artificial Intelligence Interactions in Organizational
Decision-Making :
The Impact of Transparency, Explainability, and User Control**

**par
Burak Öz**

Thèse présentée en vue de l'obtention du grade de Ph. D. en administration
(spécialisation Technologies de l'information)

Février 2026

© Burak Öz, 2026

HEC MONTRÉAL
École affiliée à l'Université de Montréal

Cette thèse intitulée :

**An Exploration of Human-Artificial Intelligence Interactions in Organizational
Decision-Making :
The Impact of Transparency, Explainability, and User Control**

Présentée par :

Burak Öz

a été évaluée par un jury composé des personnes suivantes :

Wietske Van Osch
HEC Montréal
Présidente-rapportrice

Pierre-Majorique Léger
HEC Montréal
Directeur de recherche

Constantinos K. Coursaris
HEC Montréal
Codirecteur de recherche

Ryad Titah
HEC Montréal
Membre du jury

Luc Cassivi
Université du Québec à Montréal
Membre du jury

Alexander Mädche
Karlsruhe Institute of Technology
Examineur externe

Joé T. Martineau
HEC Montréal
Représentante du directeur de HEC Montréal

Résumé

Dans les organisations, les systèmes d'intelligence artificielle (IA) sont de plus en plus utilisés pour soutenir la prise de décision. Cependant, de nombreuses organisations rencontrent des difficultés à améliorer durablement la qualité des décisions et la performance. Ces difficultés ne s'expliquent pas uniquement par les limites techniques de l'IA, mais aussi par la manière dont les interactions entre les utilisateurs et l'IA, ainsi que l'interprétation et la gestion des recommandations de l'IA, évoluent au fil du temps. Les recherches existantes sur l'interaction humain-IA offrent toutefois des résultats empiriques contrastés, notamment quant au rôle de la conception des interfaces dans la compréhension, la confiance, l'attention et la performance des utilisateurs.

Cette thèse vise à comprendre comment les choix de conception des interfaces dans les systèmes décisionnels organisationnels soutenus par l'IA influencent la compréhension des utilisateurs, le calibrage de leur confiance, l'attention et la performance décisionnelle. Elle adopte une approche multiméthode ainsi qu'une perspective bidirectionnelle à propos de l'interaction humain-IA, en distinguant les mécanismes par lesquels les systèmes d'IA fournissent de l'information aux utilisateurs de ceux par lesquels les utilisateurs peuvent influencer ou ajuster le fonctionnement de l'IA.

La thèse est composée de trois essais empiriques. Le premier essai repose sur des entretiens semi-dirigés menés auprès de professionnels du domaine de la chaîne d'approvisionnement et de la logistique et examine comment les utilisateurs évaluent et gèrent les recommandations de l'IA en pratique. En s'appuyant sur la théorie de la conscience de la situation (Situation Awareness Theory), cet essai introduit une

perspective de méta-prise de décision pour décrire l'évolution du recours à l'IA dans les contextes organisationnels. Le deuxième essai examine de manière expérimentale l'effet de la transparence et de l'explicabilité, dans le contexte d'un assistant d'IA générative, sur la compréhension, la confiance, le traitement de l'information et la performance au fil d'interactions répétées. Le troisième essai étudie le contrôle utilisateur comme mécanisme de conception permettant de soutenir l'engagement attentionnel et la détection des erreurs lorsque les recommandations de l'IA sont peu fiables, en mobilisant des mesures comportementales, auto-rapportées et neurophysiologiques.

Dans l'ensemble, cette thèse contribue à une meilleure compréhension de la conception des interactions humain-IA dans les tâches de prise de décision organisationnelle, en identifiant les facteurs de conception pertinents, puis en définissant, opérationnalisant et testant les effets de la transparence de l'IA, de l'explicabilité de l'IA et du contrôle utilisateur sur la confiance, l'attention et la performance décisionnelle des utilisateurs.

Mots-clés

Interaction humain-IA, systèmes d'aide à la décision, conception des interfaces, transparence, explicabilité, contrôle utilisateur, conscience situationnelle, recours à l'IA, attention, NeuroIS

Méthodes de recherche

Recherche qualitative, expérimentations en laboratoire, oculométrie, mesures neurophysiologiques

Abstract

Artificial intelligence (AI) systems are increasingly used to support organizational decision-making, yet organizations frequently report difficulties in achieving sustained improvements in decision quality and performance. Prior research suggests that these difficulties are not solely attributable to limitations in AI accuracy, but also to how users interact with, interpret, and manage AI recommendations over time. However, existing work on human-AI interaction provides mixed and sometimes inconsistent empirical evidence regarding how interface design influences user understanding, reliance, attention, and performance.

This dissertation examines how interface-level design choices in AI-supported organizational decision systems shape users' understanding, confidence calibration, attention, and decision performance. The research adopts a multi-method approach and a bidirectional perspective on human-AI interaction, distinguishing between mechanisms through which AI systems communicate information to users and mechanisms through which users can influence or intervene in AI behavior.

The dissertation comprises three empirical essays. The first essay uses semi-structured interviews with professionals in supply chain and logistics contexts to examine how users assess, accept, reject, or override AI recommendations in practice. Drawing on Situation Awareness Theory, the study introduces a meta-decision-making perspective to describe how reliance on AI is continuously evaluated and revised in organizational settings. The second essay experimentally investigates how transparency and explainability, operationalized within a generative AI assistant context, influence users' understanding,

confidence, information processing, and performance across repeated interactions. The third essay examines user control as a design mechanism for supporting attentional engagement and error detection when AI systems produce unreliable recommendations, using behavioral, self-reported, and neurophysiological measures.

Overall, this dissertation contributes to the design of human-AI interactions in organizational decision-making tasks by identifying relevant design factors and then defining, operationalizing, and testing the impact of AI transparency, AI explainability, and user control in influencing users' understanding, confidence, information processing, and decision performance.

Keywords

Human-AI interaction, decision support systems, interface design, transparency, explainability, user control, situation awareness, reliance, attention, NeuroIS

Research Methods

Qualitative research, laboratory experiments, eye-tracking, neurophysiological measurement

Contents

Résumé	iii
Abstract	v
List of Tables	xiii
List of Figures	xvii
List of Acronyms	xxi
Acknowledgements	xxv
Preface	xxvii
Introduction	1
1 The Dynamics of Human-AI Interactions in Organizational Decision-Making	23
Abstract	23
1.1 Introduction	24
1.2 Background	26
1.2.1 The Adoption and Use of IS and AI	26
1.2.2 The Situation Awareness Model	26
1.3 Methods	28
1.3.1 Sampling and Participants	28
1.3.2 Data Collection	29

1.3.3	Analysis of Collected Data	30
1.4	Results	31
1.4.1	Decision-Making with AI Support and Meta Decision-Making	31
1.4.2	Meta-Decisions at the First SA Level	34
1.4.3	Meta-Decisions at the Second SA Level	38
1.4.4	Meta-Decisions at the Third SA Level	42
1.4.5	Meta-Decisions at the Decision/Action Selection Phase	44
1.5	An Extended Process Model of Decision-Making with AI	50
1.5.1	Attributed System SA	51
1.5.2	Bidirectional Adjustment of User and System SA	53
1.5.3	Meta-Decision Making	54
1.6	Discussion	57
1.6.1	Implications for Research	57
1.6.2	Implications for Practice	59
1.6.3	Limitations	60
1.7	Conclusion	61
	References	62

2	Exploring the Impact of Generative AI Transparency and Explainability on Decision Performance	65
	Abstract	65
2.1	Introduction	66
2.2	Background	70
2.2.1	Artificial Intelligence and Organizational Decision-Making	70
2.2.2	Human-AI Interaction and the Situation Awareness Model	71
2.2.3	Human-AI Interaction and Eye-Tracking Research	75
2.3	Hypothesis Development	78
2.3.1	Perceived Understanding and Confidence in the GenAI Assistant	78
2.3.2	Information Processing Efficiency	82

2.4	Methodology	86
2.4.1	Experimental Design	86
2.4.2	Experimental Setup	87
2.4.3	Experimental Procedure	97
2.4.4	Sample Description and Data Collection	100
2.4.5	Data Cleaning and Preprocessing	102
2.5	Results	103
2.5.1	Measurement Instruments	104
2.5.2	Descriptive Statistics	106
2.5.3	Hypothesis Testing	107
2.5.4	Post-Hoc Analysis of Interaction Effects	114
2.5.5	Post-Hoc Analysis of Temporal Decline in Attention Shift Rates	114
2.5.6	Post-Hoc Analysis of Long-Term Effects	116
2.6	Discussion	117
2.6.1	Key Findings	117
2.6.2	Theoretical Contributions	120
2.6.3	Methodological Contributions	122
2.6.4	Practical Implications	124
2.6.5	Limitations and Future Research	125
2.7	Conclusion	127
	References	128

3 User Control in Human-AI Interaction: How AI Configuration Shapes

	User Attention and Error Responses	145
	Abstract	145
3.1	Introduction	146
3.2	Background	150
3.2.1	NeuroIS Research	150
3.2.2	Human-AI Interaction in Organizational Decision-Making	151

3.3	Hypothesis Development	154
3.3.1	AI Errors and Confidence in the AI Assistant	154
3.3.2	AI Errors and Attentional Responses	155
3.3.3	AI Errors and Attentional Resource Allocation	156
3.3.4	User Autonomy in Human-AI Interactions	157
3.3.5	Confidence in the AI Assistant and Decision Performance	158
3.3.6	Attentional Reactions and Decision Performance	158
3.3.7	Attentional Resource Allocation and Decision Performance . . .	159
3.3.8	User Control and Perceived Understanding of AI	160
3.4	Methods	161
3.4.1	Experimental Design	161
3.4.2	Experimental Setup	162
3.4.3	Sample Description and Data Collection	174
3.5	Results	176
3.5.1	Experimental Design Checks	176
3.5.2	Manipulation Check: Configuration and Perceived Autonomy .	176
3.5.3	Manipulation Check: AI Errors	177
3.5.4	Data Measurement Model	178
3.5.5	Hypothesis Testing	180
3.5.6	Post-Hoc Analyses	187
3.6	Discussion	191
3.6.1	Key Findings	191
3.6.2	Theoretical and Methodological Implications	195
3.6.3	Practical Implications	196
3.6.4	Limitations and Future Research	197
3.7	Conclusion	199
	References	200

Conclusion	215
-------------------	------------

A	Supplementary Materials for Chapter 1	i
A.1	Methods	ii
A.1.1	Semi-Structured Interview Guide	ii
B	Supplementary Materials for Chapter 2	v
B.1	Methods	vi
B.1.1	Manipulation Check Study	vi
B.1.2	The Design of GenAI Assistants as Experimental Stimuli	viii
B.1.3	Measurement Model	xii
B.1.4	Experimental Task Instructions	xvi
B.2	Results	xxiii
B.2.1	Descriptive Statistics	xxiii
B.2.2	Hypothesis Testing	xxv
C	Supplementary Materials for Chapter 3	xxxiii
C.1	Methods	xxxiv
C.1.1	Measurement Model	xxxiv
C.1.2	Experimental Task Instructions	xxxviii
C.2	Results	1
C.2.1	Descriptive Statistics	1
C.2.2	Hypotheses 1 and 4a	li
C.2.3	Hypotheses 2 and 4b	li
C.2.4	Hypotheses 3 and 4c	lii
C.2.5	Hypothesis 5	lii
C.2.6	Hypothesis 6	lii
C.2.7	Hypothesis 7	liii
C.2.8	Decision Performance	liii
C.2.9	Hypothesis 8	liv

C.2.10	Post-Hoc Mediation Analyses	liv
C.2.11	Post-Hoc Analysis of Confidence Recovery Following AI Errors	lv

List of Tables

1	Keywords used in search queries. Search results must include at least one row from each column of this table. * indicates a wildcard character that represents any number of characters.	9
2	An illustrative subset of the reviewed studies, summarizing HCAI problems, proposed solutions, and empirical results.	12
1.1	The factors influencing meta-decisions at various stages of decision-making.	34
2.1	Hypothesis Testing Summary	113
3.1	Experimental design.	162
3.2	Used measures.	174
3.3	Comparison of the baseline performance levels of participants across experimental conditions	176
3.4	Hypothesis Testing Summary	187
B.1	Manipulation check items and t-tests results.	viii
B.2	Measurement items.	xii
B.3	Perceived Understanding item descriptive statistics	xiii
B.4	Perceived Understanding scale inter-item correlations.	xiii
B.5	Factor Loadings of Perceived Understanding Items	xiii
B.6	AICP item descriptive statistics. (r) indicates reversed items.	xiv
B.7	AICP scale inter-item correlations.	xiv

B.8	CFA Loadings for Workload Alleviation Subscale	xiv
B.9	CFA Loadings for Monitoring Subscale	xv
B.10	Computer Self-Efficacy item descriptive statistics	xv
B.12	Factor Loadings of Computer Self-Efficacy Items	xv
B.11	Computer Self-Efficacy scale inter-item correlations.	xvi
B.13	Block-Level Descriptive Statistics	xxiii
B.14	Trial-Level Behavioral Descriptive Statistics. Performance values are calculated using a loss function. Higher performance values are better, and the highest possible performance value is 0.	xxiv
B.15	Eye-Tracking Descriptive Statistics	xxiv
B.16	Attention Shift Saccades on First Trial	xxiv
B.17	Linear Mixed-Effects Model Predicting Perceived Understanding (H1a-H1c)	xxv
B.18	Linear Mixed-Effects Model Predicting Confidence in the GenAI Assistant (H2a-H2c)	xxvi
B.19	Linear Mixed-Effects Models Predicting Log(Decision Performance) at the Block Level (H3)	xxvi
B.20	Linear Mixed-Effects Models Predicting Log(Decision Performance) at the Trial Level (H4)	xxvii
B.21	Linear Mixed-Effects Model Predicting Log(Fixation Duration) (H5)	xxvii
B.22	Linear Mixed-Effects Model Predicting Log(Attention Shift Rate) (H6)	xxvii
B.23	Linear Mixed-Effects Models Predicting Log(Decision Performance) at the Trial Level (H7)	xxviii
B.24	Linear Mixed-Effects Models Predicting Log(Decision Performance) at the Trial Level (H8)	xxviii
B.25	Linear Mixed-Effects Model Predicting Log(Decision Performance)	xxviii
C.1	Measurement items.	xxxiv
C.2	AICP item descriptive statistics. (r) indicates reversed items.	xxxv

C.3	AICP scale inter-item correlations. (r) indicates reversed items. . . .	xxxv
C.4	Factor Loadings for the AICP - Workload Alleviation Subscale	xxxv
C.5	Factor Loadings for the AICP - Monitoring Subscale	xxxvi
C.6	Computer Self-Efficacy item descriptive statistics	xxxvi
C.7	Computer Self-Efficacy scale inter-item correlations.	xxxvi
C.8	Factor Loadings of Computer Self-Efficacy Items	xxxvii
C.9	Perceived Autonomy item descriptive statistics.	xxxvii
C.10	Perceived Autonomy scale inter-item correlations.	xxxvii
C.11	Factor Loadings of Perceived Autonomy items	xxxvii
C.12	Perceived Understanding item descriptive statistics.	xxxvii
C.13	Perceived Understanding scale inter-item correlations.	xxxviii
C.14	Factor Loadings of Explanation Satisfaction Items	xxxviii
C.15	Block-Level Descriptive Statistics for Perceived Understanding of AI	1
C.16	Descriptive Statistics for AI Errors on Correction Behavior	1
C.17	Linear Mixed-Effects Model Predicting Correction Behavior (Error Manipulation Check)	1
C.18	Trial-Level Descriptive Statistics by Configuration and Error States .	1
C.19	Trial-Level Descriptive Statistics by Configuration and Error States .	li
C.20	Linear Mixed-Effects Model Predicting Confidence in the AI Fore- casting System (H1 & H4a)	li
C.21	Linear Mixed-Effects Model Predicting P3a Amplitude (H2 & H4b) .	li
C.22	Linear Mixed-Effects Model Predicting Post-Fixation Duration (H3 & H4c)	lii
C.23	Linear Mixed-Effects Model Predicting Task Performance (H5) . . .	lii
C.24	Linear Mixed-Effects Model Predicting Task Performance (H6) . . .	lii
C.25	Linear Mixed-Effects Model Predicting Task Performance (H7) . . .	liii
C.26	Linear Mixed-Effects Model Predicting Log(Decision Performance) .	liii
C.27	Linear Model Predicting Perceived Understanding of the AI System (H8)	liv

C.28	Linear Mixed-Effects Models for Mediation via Confidence in AI in the "Configuration" Condition	liv
C.29	Linear Mixed-Effects Models for Mediation via Confidence in AI in the "No Configuration" Condition	liv
C.30	Linear Mixed-Effects Models for Mediation via Attention Allocation on Manual Task Elements in the "Configuration" Condition	lv
C.31	Linear Mixed-Effects Models for Mediation via Attentional Responses in the "No Configuration" Condition	lv
C.32	Summary of Censored and Uncensored Observations by Experimental Condition	lv
C.33	Cox proportional hazards model of confidence recovery	lv

List of Figures

1	PRISMA flow diagram of the primary study identification and screening process.	10
2	Dissertation structure	21
1.1	The process model of a user’s decision to automate a complex decision-making task.	56
2.1	Research model summarizing the hypothesized effects of transparency and explainability on user perceptions, information processing efficiency, and decision performance.	86
2.2	The system interface as used by participants during the experiment. .	91
2.3	An example GenAI assistant with high transparency & high explainability. The annotations highlight the components introduced through the transparency and explainability manipulations.	94
2.4	Experimental design and protocol.	100
2.5	Validated research model.	113
2.6	Kaplan-Meier fit plot corresponding to the change in attention shift rates over trials. This plot indicates that the conditions with high transparency reach stabilization, i.e., 40% of a participant’s initial attention shift rates, faster.	115
3.1	The role of physiological measures in IS research. Adapted from (Riedl and Léger, 2016).	150

3.2	Research model summarizing the hypotheses on cognitive responses to AI errors and the availability of AI configuration.	160
3.3	Screens shown before each trial in the configuration and no-configuration conditions.	165
3.4	Main forecasting interface used during the experiment. Highlighted elements show user interface elements that were revealed exactly three seconds after participants clicked "Generate AI Recommendation."	167
3.5	Experimental protocol.	170
3.6	Estimated marginal means of confidence in the AI recommendation system across error (erroneous vs. non-erroneous) and configuration conditions. Error bars represent 95% confidence intervals.	181
3.7	Grand average plots of waveforms at the Pz electrode channel.	182
3.8	Grand average plots of waveforms at the Oz electrode channel.	182
3.9	Topographic maps (250–280 ms) across configuration and accuracy conditions.	183
3.10	Validated research model.	186
3.11	Kaplan-Meier survival estimates for confidence recovery following AI errors. Participants in the configuration condition recovered confidence faster ($p = .043$). Risk at 10 trials is zero because no participant experienced an interval longer than 10 trials before encountering another AI error.	191
B.1	Screenshot of the GenAI assistant interaction shown to participants in Group 1 during the manipulation check study. This version represents the low transparency and high explainability condition.	vi
B.2	Screenshot of the GenAI assistant interaction shown to participants in Group 1 during the manipulation check study. This version represents the high transparency and no explainability condition.	vii

B.3	Tech stack used in the design of the experimental stimuli.	viii
B.4	An example conversation with the GenAI assistant in the low transparency & low explainability condition.	ix
B.5	An example conversation with the GenAI assistant in the low transparency & high explainability condition.	ix
B.6	An example conversation with the GenAI assistant in the high transparency & low explainability condition.	x
B.7	An example conversation with the GenAI assistant in the high transparency & high explainability condition.	xi
B.8	Experimental task instructions, page 1 visual supplement.	xvii
B.9	Experimental task instructions, page 2 visual supplement.	xviii
B.10	Experimental task instructions, page 3 visual supplement.	xix
B.11	Experimental task instructions, page 4 visual supplement.	xx
B.12	Experimental task instructions, visual supplement for page 1 after the trial block.	xxi
B.13	Experimental task instructions, visual supplement for page 1 after the trial block.	xxi
B.14	Experimental task instructions, visual supplement for page 1 after the trial block.	xxii
B.15	Experimental task instructions, visual supplement for page 2 after the trial block.	xxiii
B.16	Attention shift rates over trial numbers within an experimental block. The plot indicates that on average, participants reduced their attention shift rates by about 50% after having used the same GenAI assistant in similar cases for 10 times.	xxv
B.17	Kaplan-Meier fit plot corresponding to post-hoc analyses. This plot indicates that the conditions with high transparency reach stabilization, i.e., 30% of a participant's initial attention shift rates, faster.	xxix

B.18	Kaplan-Meier fit plot corresponding to post-hoc analyses. This plot indicates that the conditions with high transparency reach stabilization, i.e., 50% of a participant's initial attention shift rates, faster.	xxx
B.19	Kaplan-Meier fit plot corresponding to post-hoc analyses. This plot indicates that the conditions with high transparency reach stabilization, i.e., 55% of a participant's initial attention shift rates, faster.	xxxi
B.20	Kaplan-Meier fit plot corresponding to post-hoc analyses. This plot illustrates the test where the stabilization threshold is set at 60% of a participant's initial attention shift rates.	xxxii
C.1	Experimental task instructions, page 1 visual supplement.	xxxix
C.2	Experimental task instructions, page 2 visual supplement.	xl
C.3	Experimental task instructions, page 3 visual supplement.	xli
C.4	Experimental task instructions, page 4 visual supplement.	xlii
C.5	Experimental task instructions, visual supplement for page 1 after the trial block.	xliii
C.6	Experimental task instructions, visual supplement for page 1 after the trial block.	xliii
C.7	Experimental task instructions, visual supplement for page 1 after the trial block.	xliv
C.8	Experimental task instructions, visual supplement for page 2 after the trial block.	xlvi
C.9	Experimental task instructions, visual supplement for page 3 after the trial block.	xlvi
C.10	Experimental task instructions, visual supplement for page 4 after the trial block.	xlix

List of Acronyms

AI	Artificial Intelligence
AICP	Automation-Induced Complacency Potential
AOI	Areas of Interest
API	Application Programming Interface
CFA	Confirmatory Factor Analysis
DSA	Distributed Situation Awareness
EEG	Electroencephalography
ERP	Enterprise Resource Planning
GenAI	Generative Artificial Intelligence
GUI	Graphical User Interface
HCAI	Human-Centered Artificial Intelligence
HCI	Human-Computer Interaction
HITL	Human-in-the-loop
ICA	Independent Component Analysis
IS	Information System

IT	Information Technology
MGT	Multi-Grounded Theory
ML	Machine Learning
OData	Open Data Protocol
OODA	Observe-Orient-Decide-Act
OOTL	Out-of-the-loop
RQ	Research Question
SA	Situation Awareness
SD	Standard Deviation
SCM	Supply Chain Management
SOR	Stimulus-Organism-Response
TAM	Technology Acceptance Model
UI	User Interface
UX	User Experience

To my family

Acknowledgements

This dissertation would not have been possible without the support, guidance, and contributions of many individuals and institutions, to whom I express my sincere gratitude.

First and foremost, I would like to thank my supervisors Pierre-Majorique Léger and Constantinos K. Coursaris for their guidance, trust, and continued support throughout my doctoral studies. Their thoughtful feedback and commitment to excellence were instrumental in shaping this thesis and in supporting my development as an independent researcher.

I would also like to thank faculty members at HEC Montréal and across the Montreal joint program for their support, mentorship, and guidance, which meaningfully influenced my academic path. In particular, I thank Jacques Robert, Sylvain Sénécal, Ryad Titah, Luc Cassivi, Pierre Hadaya, Gilbert Babin, and Bogdan Negoita for their insightful discussions, encouragement, and availability, all of which helped refine my thinking and strengthened my interest in both research and teaching.

My heartfelt thanks go to my colleagues at the ERPsim Lab, whose collaboration and expertise were critical to the success of this research. Their contributions were central to the design and refinement of the experimental stimuli. I would also like to acknowledge my colleagues at Tech3Lab for their support throughout this project. The infrastructure, resources, and technical assistance they provided were invaluable for the execution of the experimental studies and data collection.

I extend my sincere thanks to my fellow PhD students at HEC Montréal and within the

Montréal Joint PhD Program for their intellectual engagement, feedback on early versions of this work, and collegial support. The exchange of ideas and shared experiences made this journey intellectually rewarding.

Finally, I am deeply grateful to my family and friends for their continued support. I especially thank my spouse for her patience and encouragement during this process.

I gratefully acknowledge the financial support of the Fonds de Recherche du Québec (grant number 305619), the Natural Sciences and Engineering Research Council of Canada (grant number ALLRP 575332 - 22), NSERC-Prompt Industrial Research Chair in User Experience, and IVADO (PFR3 HCAI), whose funding made the completion of this research possible.

Preface

This dissertation, authored by Burak Öz, represents an original and partially published research endeavor conducted under the guidance of Dr. Constantinos K. Coursaris and Dr. Pierre-Majorique Léger. This work was funded by the Fonds de Recherche du Québec (grant number 305619), the Natural Sciences and Engineering Research Council of Canada (grant number ALLRP 575332 - 22), NSERC-Prompt Industrial Research Chair in User Experience, and IVADO (PFR3 HCAI).

A preliminary version of Essay 1 was presented at the Human-Computer Interaction International conference in 2023, and the complete version was published as a chapter for the edited book titled "The Design of Human-Centered Artificial Intelligence for the Workplace."¹ The version included in this dissertation is a version prior to publication, which contains expanded methodological and analytical details. Burak Öz served as the primary author, responsible for the execution, data collection, analysis, and writing. Co-authors Dr. Joerg Beringer, Dr. Alexander J. Karran, Dr. Constantinos K. Coursaris, and Dr. Pierre-Majorique Léger have provided feedback at multiple stages of the study.

The experimental protocol and preliminary results of the study reported in Essay 2 were presented at the 2024 Pre-ICIS SIGHCI Workshop², and the final version is under preparation for submission as a journal article. Burak Öz served as the primary author, responsible for the design, execution, analysis, and writing. Dr. Alexander J. Karran

¹Öz, B., Karran, A. J., Beringer, J., Coursaris, C. K., & Léger, P. M. (2025). The Dynamics of Human-AI Interactions in Organizational Decision-Making. In *The Design of Human-Centered Artificial Intelligence for the Workplace*, pp. 29–46. Springer. https://doi.org/10.1007/978-3-031-83512-4_3

²Öz, B., Mussaferada, A., Coursaris, C. K., Senecal, S., and Léger, P. M. (2025). The Impact of Generative AI Transparency and Explainability on Decision Performance. SIGHCI 2024 Proceedings. 31. <https://aisel.aisnet.org/sighci2024/31>

contributed to the experimental design and reviewed the final version. Dr. Constantinos K. Coursaris and Dr. Pierre-Majorique Léger have provided feedback and guidance across all stages of the research project. Aliatta Mussaferada contributed to the experiment design by testing the initial generative AI assistant prototypes. This study used software applications designed by the ERPsim Lab at HEC Montreal as part of its experimental stimuli.

The experimental protocol and preliminary results of the study reported in Essay 3 were presented at the NeuroIS Retreat 2023³. The final version is under preparation for submission as a journal article. Burak Öz served as the primary author, responsible for the execution, data collection, analysis, and writing. Dr. Alexander J. Karran and Dr. Jared Boasen contributed to the experimental design and reviewed the final version. Dr. Constantinos K. Coursaris and Dr. Pierre-Majorique Léger have provided feedback and guidance across all stages of the research project. This study used software applications designed by the ERPsim Lab at HEC Montreal as part of its experimental stimuli.

³Öz, B., Karran, A. J., Boasen, J., Coursaris, C. K., & Léger, P. M. (2023). The Role of User Control in Enhancing Human-AI Collaboration Effectiveness: Insights from a Pilot Study. In *Information Systems and Neuroscience. NeuroIS 2023.*, pp. 185–193. Springer. https://doi.org/10.1007/978-3-031-58396-4_15

Introduction

In the last decade, organizational information systems (IS) have advanced alongside technological progress, enabling the collection and analysis of more data points across all business processes. With increased data visibility and availability, artificial intelligence (AI⁴) methods have become a central topic in enhancing organizational decision-making performance (Sebastian and Draper, 2024; Sharma et al., 2022). This increased attention on AI in organizational decision-making is not unwarranted, as these systems can support and replicate certain aspects of human intelligence by identifying patterns, interpreting information, and, in the case of generative AI (GenAI) models, creating new data points (Brynjolfsson et al., 2025; Faraj et al., 2018; Noy and Zhang, 2023). Consequently, the integration of AI across industries has been accelerating, with 42% of IT professionals reporting that their organizations are active users of AI technologies, while an additional 40% are in the exploration phase (Di Battista et al., 2023; IBM Newsroom, 2024).

AI technologies are now embedded across a wide range of organizational processes, extending far beyond the automation of routine and automated tasks into domains where decisions carry operational, financial, or societal consequences. In finance, AI supports credit risk assessment, fraud detection, and compliance monitoring (Zetzsche et al., 2020). In healthcare, it assists clinicians with diagnostic imaging, triage, and treatment planning (Erickson et al., 2017). In operations and logistics, AI optimizes routing, inventory management, and demand forecasting (Agrawal et al., 2024; Cain, 2022; Dubey et al.,

⁴This document employs the term "AI" to encompass a broad range of automation and autonomy capabilities, irrespective of the underlying technology, which could include learning agents like machine learning, deep learning, or artificial intelligence, as well as more static reasoning systems.

2020). Similarly, in human resources and talent management, AI screens candidates and identifies retention risks. Across these diverse contexts, AI systems increasingly participate in decision-making processes, while the accountability of outcomes in current implementations remains with humans (Kempton et al., 2023). This accountability aspect further reinforces the need for users to understand, monitor, and evaluate algorithmic outputs.

Despite the widespread adoption of AI technologies, evidence shows that organizations often struggle to extract the expected productivity improvements or decision-quality gains from these implementations. Cross-industry studies report that AI deployments frequently underperform relative to executive expectations, with many firms observing limited impact on core decision processes (Dwivedi et al., 2021; Endsley, 2023; Simkute et al., 2024). Large-scale surveys similarly document recurring challenges in realizing value from AI investments, even in data-rich functions such as forecasting, risk assessment, and operational planning, where algorithmic support should, in principle, deliver measurable improvements (Mediavilla et al., 2022; Nair and Saenz, 2024; Sharma et al., 2022). These patterns indicate a gap between the technical capabilities of AI systems and the actual benefits organizations obtain from their integration.

The difficulty of realizing AI's promised benefits is not specific to a single domain. In healthcare, diagnostic support systems require clinicians to verify algorithmic recommendations and investigate model errors, increasing workload rather than reducing it (Erickson et al., 2017). In human resources, biases in recruitment and screening algorithms necessitate human correction and ongoing oversight to prevent discriminatory outcomes (Vinopal, 2022). In supply chain and operations, planners often perceive a need to override or manually adjust AI-generated recommendations (Baryannis et al., 2019; Mediavilla et al., 2022; Nair and Saenz, 2024). Financial services face similar challenges, as regulatory requirements demand that AI-based risk or credit-scoring outputs be explainable and auditable, making human review a mandatory part of the workflow (Zetzsche et al., 2020). In summary, AI-supported organizational decision-making often requires continued human vigilance and intervention.

Supervisory roles in such human-AI interaction configurations place substantial cognitive demands on operators, particularly when their primary responsibility is to monitor system behavior and intervene only when necessary. Decades of research show that passive monitoring is among the most difficult forms of human work as vigilance deteriorates rapidly over time, infrequent events impair an operator's ability to detect anomalies, and reduced interaction with the underlying task diminishes the mental models needed to recognize subtle cues of system malfunction (Endsley et al., 2003; Miller and Parasuraman, 2003; Parasuraman et al., 1996). When individuals are required to oversee systems that generally function well but occasionally fail in rare or ambiguous ways, maintaining sustained attention and situation awareness becomes increasingly challenging (De Boer and Dekker, 2017). These cognitive limitations affect how well operators perceive relevant information, interpret system states, and predict future changes (Endsley, 1995). As a result, the effectiveness of supervisory control depends not only on system accuracy but on the degree to which users can remain engaged, attentive, and informed while performing inherently low-engagement, high-responsibility monitoring tasks.

The recurring difficulties observed across AI deployments mirror the classic ironies of automation: as more tasks become automated, the residual work left for humans becomes harder, less visible, less frequently practiced, and substantially more cognitively demanding (Bainbridge, 1983). When routine components of a task are delegated to an automated system, operators lose the ongoing interaction that once maintained their skills, leading to progressive deskilling and a diminished capacity to detect anomalies when they do arise (Endsley et al., 2003). Combined with the reduced engagement and vigilance issues, automation setups are argued to lead to configurations that are more complicated than manual task configurations (Bainbridge, 1983). These patterns from automation research remain relevant to modern AI implementations, forming a lens for understanding why technical improvements alone cannot prevent system-level breakdowns.

The patterns from automation research remain relevant in modern AI implementations, and they even further intensify these ironies because their core technical charac-

teristics introduce additional forms of complexities. Unlike traditional automation or IS, users cannot easily perceive what features an AI system attends to, how it weighs inputs, or why it arrives at particular recommendations (Czako et al., 2021; Hassenzahl and Tractinsky, 2006; Sharan and Romano, 2020). In addition, AI systems are adaptive. These systems constantly update their internal representations as environments shift (Pan, 2016). AI systems' propensity to update their reasoning process over time means that they lack functional consistency, i.e., their use may not lead to consistent outcomes (Schuetz and Venkatesh, 2020), making it difficult for operators to form stable mental models. Lastly, their probabilistic foundations generate outputs that vary across similar inputs, complicating predictability and challenging users' ability to anticipate system behavior. As a result, operators must continuously evaluate and correct AI behavior while maintaining awareness of the primary task. These conditions lead to fragile and interaction-heavy systems that require users to maintain engagement, interpret outputs, perform verification steps, supply additional inputs, or manage iterative exchanges that fragment attention and increase workload (Jiang et al., 2022; Prather et al., 2023; Simkute et al., 2024; Tankelevitch et al., 2024).

AI systems become actionable through the interfaces that mediate them, meaning that the user interface is the primary channel through which humans perceive, interpret, and act upon algorithmic outputs. As these systems increase in opacity, adaptiveness, and computational complexity, the user interface becomes the central determinant of whether users can maintain situation awareness, understand system behavior, and appropriately calibrate their confidence in AI recommendations (Endsley et al., 2003). In summary, as all cognitive interaction with AI is mediated through this layer, interface design critically determines whether an AI system reduces or increases cognitive load, clarifies or obscures system behavior, and enables or hinders effective human-AI interaction (Harbarth et al., 2025; Khosrowabadi et al., 2022; Shneiderman, 2020; Wang et al., 2021).

Taken together, these concerns suggest that realizing the benefits of AI requires not only technical advances but also a clearer understanding of how interface design shapes users' cognitive states and decision behaviors. Prior work shows that constructs

concerning human-AI interactions are defined inconsistently across disciplines, and empirical findings on reliance remain mixed, with studies documenting both underuse (Allen and Choudhury, 2022; Dietvorst et al., 2015) and overreliance (Anaraky et al., 2020; Merritt et al., 2019). These discrepancies indicate that the effects of design choices on users' attention, trust calibration, and situation awareness remain poorly understood. As organizations continue to report persistent gaps between anticipated and realized value from AI investments, there is a clear need for a research program that systematically examines how interface-level design choices shape cognition and performance in AI-assisted decision-making. Guided by this motivation, this dissertation addresses the following overarching research question:

How do interface-level design choices in AI-supported organizational decision systems shape users' attention, understanding, confidence calibration, and decision performance?

To answer this question, the dissertation follows a multi-method research program comprising three complementary empirical essays, which together examine human-AI interaction in organizational decision-making environments. The first essay explores how professionals experience and manage reliance on AI systems in real organizational settings. The second essay investigates how transparency and explainability influence users' understanding, confidence, and information processing efficiency when interacting with generative AI assistants. The third essay examines user control as a mechanism for sustaining effective decision-making when AI systems are prone to produce erroneous or unreliable recommendations. Together, these studies contribute to the understanding of how interface design shapes human cognition and performance in AI-supported decision-making contexts.

The first essay utilizes qualitative methods to explore how professionals experience AI-supported decision-making in practice. It investigates the factors that shape professionals' decisions to rely on or ignore AI systems in organizational contexts, with particular attention to how users form and maintain an understanding of AI recommendations over time. Grounded in the Situation Awareness Theory (Endsley,

1995), this study provides a framework and identifies factors that influence user confidence in AI.

The second essay focuses on two interface-level design constructs, transparency and explainability, and examines how they influence users' understanding, confidence calibration, and information processing efficiency when interacting with a generative AI assistant. Using a controlled, repeated-task experimental design, this study provides precise conceptual definitions and operationalizations of transparency and explainability in a generative AI assistant context. By utilizing behavioral, self-reported, and attentional measures, the essay evaluates how these design features shape user attention, cognition, and decision performance over time.

The third essay extends this investigation to situations in which AI systems are prone to produce erroneous or unreliable recommendations. It examines user control as a design mechanism for mitigating out-of-the-loop failures and sustaining effective decision-making under imperfect AI performance. Specifically, it investigates how allowing users to configure AI system parameters influences confidence calibration, attentional responses, and users' ability to detect and handle AI errors across repeated interactions. This essay uses behavioral, self-reported, and neurophysiological indicators to capture how user attention, cognition, and reliance evolve under varying levels of AI accuracy.

To situate and justify this research program, the following subsection reviews prior work on human-centered AI design in organizational decision contexts. This review aims to identify research gaps by mapping interface-level human-centered AI design challenges, proposed solutions to these challenges, and the empirical evidence supporting these solutions.

Scoping Review on Previous Work

The previous section illustrated that organizational struggles to realize the benefits of AI stem not only from technical limitations but also from the cognitive demands that current system designs impose on human operators. Building on these concerns, the

importance of user interface design in supporting effective human–AI interaction has been increasingly emphasized across information systems, human–computer interaction, and related fields (Harbarth et al., 2025; Shneiderman, 2020; Wang et al., 2021). Yet, because the study of human-centered AI in organizational decision-making is comparatively recent, relevant research has begun to emerge across a wide range of disciplines, from engineering to the humanities. Such breadth makes it difficult to obtain a consolidated view of how human-centered AI design is currently investigated, what problems have been identified, and which solutions have been proposed. These characteristics align with the conditions under which a scoping review is appropriate, e.g., when a topic is new, conceptually broad, and distributed across multiple research communities (Arksey and O’Malley, 2005).

In line with the field’s current state, this study adopted the scoping review methodology outlined by Arksey and O’Malley (2005) to systematically map and analyze the existing literature on AI-based decision-making tools implemented in workplace environments. In refining our methodology, we also followed additional recommendations pertaining to the scoping reviews and followed the PRISMA checklist for scoping reviews (Daudt et al., 2013; Levac et al., 2010; Tricco et al., 2018).

Identification of Relevant Studies

In this scoping review, we selected three databases for literature sourcing: Scopus, Web of Science, and the ACM Digital Library. The choice of these databases was informed by consultations with a librarian and an assessment of their relevance and comprehensiveness to our research focus. Scopus and Web of Science were chosen for their broad and multidisciplinary coverage of academic fields, while the ACM Digital Library was selected for its focus on high-quality journal publications in information technology.

Our keyword strategy encompassed four main categories: decision-making, AI, user interface (UI) design implications, and keywords associated with human-centered design

of AI. The research team adopted an iterative approach to refine the keyword selection, aiming to balance comprehensive coverage against the potential for irrelevant results. The initial keyword set, focused on decision-making, AI, and UI, was established through a team consensus. This was followed by an iterative process of refinement:

- Screening 30 randomly selected papers from the initial query results for relevance to our research questions.
- Extracting synonyms of query keywords from relevant papers' abstracts.
- Performing a backward search to gather additional references and keywords.
- Ending an iteration when no new synonyms emerged, followed by constructing a new query with the updated keywords.

This process underwent four iterations. However, due to the lack of standardization in naming AI techniques and algorithms, which led to significant noise in the query results, we conducted two additional iterations and held a consultative meeting with a researcher specializing in the field of responsible AI. The initial search was submitted to the databases on March 24, 2022, and preliminary results were charted and presented at a conference. Following the conference presentation, the search results and data analyses were expanded and updated to include recent publications, both on December 28, 2023; and March 3, 2025. This process identified 5,609 primary studies on Scopus, 1,226 on Web of Science, and 235 on the ACM Digital Library as of March 3, 2025. A list of search keywords is shown in Table 1, and a PRISMA flow chart indicating the screening process is shown in Figure 1.

Table 1: Keywords used in search queries. Search results must include at least one row from each column of this table. * indicates a wildcard character that represents any number of characters.

Decision-Making	AI	Design Implications	Human-Centeredness
decision science decision-making decision making knowledge work strategic thinking expert decisions decision support decisional guidance decision aids expert system advanced user support expert judgment dss business decision systems workplace decisions support system	artificial intelligence deep learning machine learning natural language processing recommendation agents robo-advisory ai algorithmic and human-based intelligence artificial general intelligence conversational agents intelligent assistants intelligent systems ml systems neural network nlp xai recommender system	design cues design evaluation design guidelines design heuristics design practices design principles design science ecological interface hci human computer interaction human-ai interaction interaction design interface characteristic* interface design interface evaluation man machine interaction message design screen design social interface user experience evaluation user interaction user interface design patterns human-agent interaction information ergonomics information visualization user experience user-computer interface anthropomorphism human-centered design human-centred design user-centered design user-centred design design implications	fair ethical moral reliab* safe* responsib* explainable XAI understandability interpretab* trust* privacy efficacy transparen* secur* user-cent* human-cent* people-cent* human-agent teaming human-in-the-loop human in the loop interaction human-ml interaction with ai interactive machine learning interpret* interpretable machine learning human expertise accessibility context aware human factors black box human involvement user-in-the-loop appropriate reliance user involvement

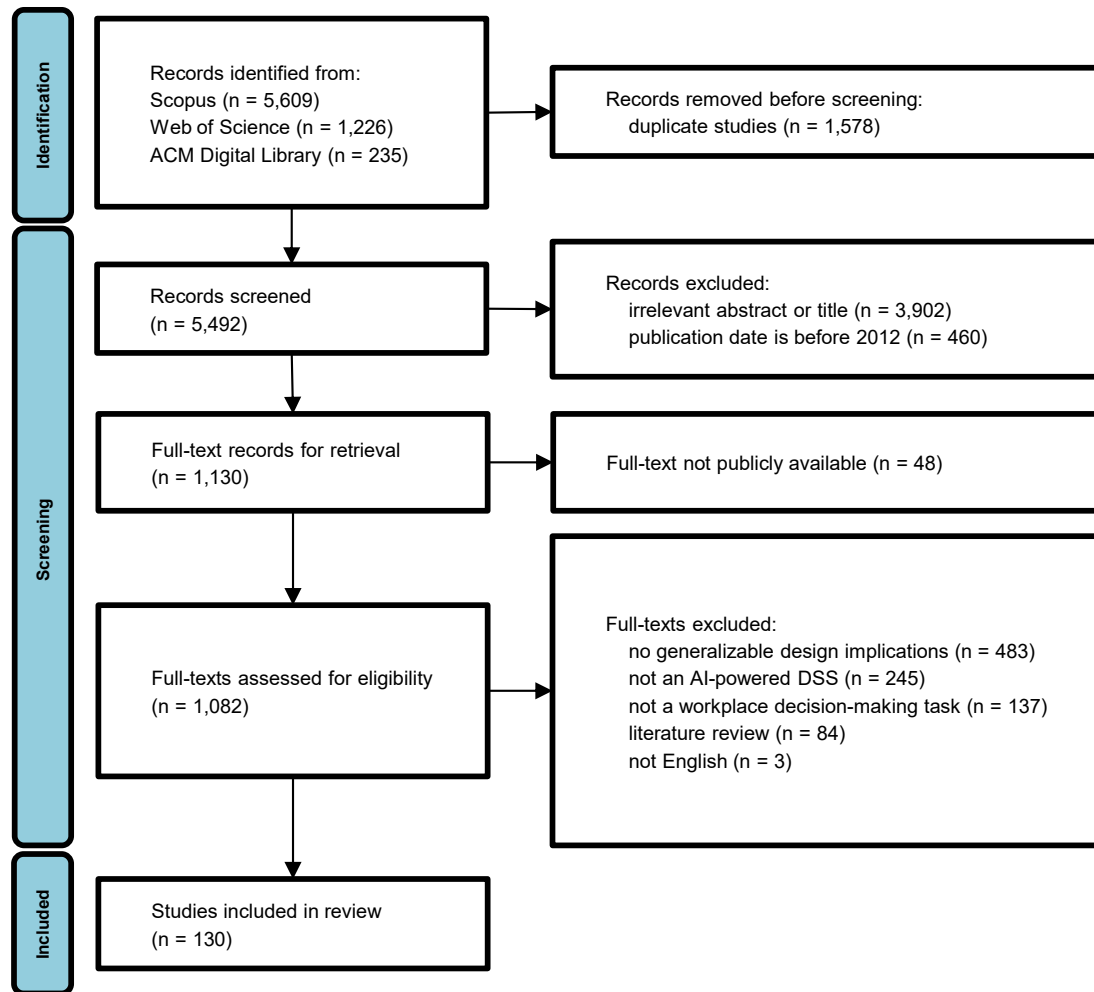


Figure 1: PRISMA flow diagram of the primary study identification and screening process.

Scoping Review Results

The review identified 130 studies examining interface-level mechanisms intended to support users' ability to understand, monitor, and regulate AI recommendations. Reflecting the developing nature of the field, this body of work is highly dispersed, with primary sources originating from information systems, computer science, human-computer interaction, ergonomics, and human factors. The methodological approaches used across these studies are similarly diverse, encompassing lab experiments, design science research, system design studies, and conceptual work. Table 2 presents a selection of the identified primary sources, from which we extracted (1) the domain in which AI support was applied, (2) the human-centered AI problem discussed, (3) the proposed solution, and (4) the empirical evidence, when available.

The identified primary sources address HCAI issues across a wide range of domains, including data analytics, healthcare, logistics and supply chain management, manufacturing, and other forms of knowledge work. This breadth aligns with industry reports indicating widespread interest in adopting AI to improve efficiency (Di Battista et al., 2023; IBM Newsroom, 2024). Moreover, the diversity of identified problems highlights differences in how decision-makers interact with AI systems across domains.

The majority of the primary sources focus on the lack of interpretability in AI systems (Bove et al., 2022; Madsen et al., 2025; Rebholz et al., 2024; Tsai et al., 2025). Another significant problem examined in these studies is the challenge of calibrating user confidence in the system (Chen et al., 2024; Karahasanovic et al., 2021; Mishler et al., 2025). In addition, studies both reported cases in which users exhibit a lack of confidence in AI (Bigras et al., 2018; Liu, 2021; Panigutti et al., 2023), as well as users' tendency to be excessively confident in AI's abilities (Brauner and Ziefle, 2019; Cecil et al., 2024; Dikmen and Burns, 2022). The out-of-the-loop problem (Bader and Kaiser, 2019), human cognitive biases (Wang et al., 2019), and the difficulty of integrating users' tacit knowledge into models (Dellermann et al., 2019; Kim et al., 2024; Mukhopadhyay et al., 2014) are among the other HCAI problems identified by the primary sources.

Table 2: An illustrative subset of the reviewed studies, summarizing HCAI problems, proposed solutions, and empirical results.

References	Domain	Study Type	HCAI Problem	Solution	Empirical Test
(Geuss et al., 2019)	Defense	Conceptual	Dangers of automating life-critical decisions	Ensuring explainability to augment human cognition	No empirical test
(Lu, 2024)	Human-AI interactions	Conceptual	Deskilling of users	Improving user agency and system explainability	No empirical test
(Micallef et al., 2017)	Human-AI interactions	System design	Integrating tacit knowledge	Interactive interfaces eliciting tacit knowledge	Including user knowledge within AI increases performance
(Mukhopadhyay et al., 2014)	Environmental planning	System design	Integrating tacit knowledge	Attention-tracking to capture expertise	No empirical test
(Kim et al., 2024)	Healthcare	System design	Integrating tacit knowledge	Eye-tracking human-in-the-loop model	Performance improves when user expertise is integrated within AI
(Dellermann et al., 2019)	Knowledge work	System design	Integrating tacit knowledge & Interpretability of black-box models	Interactive systems & graphical explanations	Qualitative evidence supports the solution
(Zhou et al., 2017)	Manufacturing	Lab experiment	Cognitive load from black-box systems	Explanations via data visualizations	Explanations reduce load, increase performance
(Robles-Flores and Roussinov, 2012)	Knowledge work	System design	Cognitive load from black-box systems	Natural-language explanations	No empirical test
(Bansal et al., 2019)	Human-AI interactions	Conceptual	Lack of functional consistency of AI	Ensuring consistent update behavior	No empirical test
(Ribeiro et al., 2016)	Data analytics	System design	Lack of interpretability	Local feature-based explanations	Explanations are useful

Continued on next page

Authors	Domain	Study Type	HCAI Problem	Solution	Empirical Test
(Morrison et al., 2018)	Healthcare	System design	Lack of interpretability	Visual data-relationship explanations	No empirical test
(Ehsan et al., 2019)	Human-AI interactions	System design	Lack of interpretability	Natural-language rationales	Detailed rationales preferred; intelligibility matters
(Nguyen et al., 2019)	Human-AI interactions	System design	Lack of interpretability	Interactive data exploration	No empirical test
(Hoque and Mueller, 2021)	Data analytics	System design	Lack of interpretability	Interactive causal-model explanations	Increased understanding; performance unchanged
(Rebholz et al., 2024)	Law	Lab experiment	Lack of interpretability	Post-hoc rationales & interactivity	Interactivity increases reliance
(Madsen et al., 2025)	Logistics	Lab experiment	Lack of interpretability	Transparency at situation awareness levels 1-3	Level 3 transparency, which helps understanding predictions, is preferred
(Tsai et al., 2025)	Healthcare	Lab experiment	Lack of interpretability	Displaying AI uncertainty	Uncertainty increases cognitive processing
(Krause et al., 2016)	Healthcare	System design	Lack of interpretability	Exploration dashboard	No empirical test
(Lamy et al., 2019)	Healthcare	System design	Lack of interpretability	Data exploration & plots	User test supports usefulness
(Schaffer et al., 2019)	Knowledge work	Lab experiment	Lack of interpretability	Local reliability explanations	Helps novices; risks overreliance
(Hepenstal et al., 2019)	Defense	Conceptual	Lack of interpretability	Natural-language explanations	No empirical test
(Kellner et al., 2021)	Manufacturing	Conceptual	Lack of interpretability	Data exploration & post-hoc explanations	No empirical test

Continued on next page

Authors	Domain	Study Type	HCAI Problem	Solution	Empirical Test
(van Capelleveen et al., 2021)	Supply chain management	System design	Lack of interpretability	Interactive exploration for understanding	Empirical results support the usefulness of interactive explorations
(Herm et al., 2022)	Human-AI interactions	Conceptual	Lack of interpretability	Global, local, and personalized explanations	No empirical test
(Bove et al., 2022)	Data analytics	Lab experiment	Lack of interpretability	Interactive data exploration	Improves contextual understanding
(Bigras et al., 2018)	Supply chain management	Lab experiment	Underreliance on AI	Rich explanations; easy access to information	Rich explanations improve adoption
(Panigutti et al., 2023)	Healthcare	System design	Underreliance on AI	Local explanations	Explanations increase adherence
(Dikmen and Burns, 2022)	Finance	Lab experiment	Overreliance on AI	Context-rich explanations	Improves error detection
(Cecil et al., 2024)	Human Resources	Lab experiment	Overreliance on AI	Heatmap or chart explanations	No reduction in overreliance
(Brauner and Ziefle, 2019)	Manufacturing	Lab experiment	Overreliance on AI	N/A	AI errors reduce performance
(Chen et al., 2024)	Healthcare	Lab experiment	Trust miscalibration (underreliance or overreliance)	Global vs local explanations	Explanations improve trust
(Mishler et al., 2025)	Cybersecurity	Lab experiment	Trust miscalibration (underreliance or overreliance)	Local transparency	No performance impact
(Wang et al., 2019)	Healthcare	Conceptual	Human biases	Explanations to mitigate bias	No empirical test
(Bader and Kaiser, 2019)	Human-AI interactions	Case study	Out-of-the-loop problem	Simple interactive interfaces	No empirical test

The most commonly proposed solution to the HCAI problem of the lack of interpretability of AI models is providing detailed system explanations in various forms, including sentences in natural language (Hepenstal et al., 2019; Robles-Flores and Roussinov, 2012), interfaces allowing interactive explorations (Bove et al., 2022; Hoque and Mueller, 2021; Kellner et al., 2021; Rebholz et al., 2024; van Capelleveen et al., 2021), or data visualizations and charts (Cecil et al., 2024; Dellermann et al., 2019). System explanations are also proposed as a potential solution to the problem of calibrating user confidence in AI, with suggestions ranging from explanations justifying a specific decision-making instance or recommendation (Cecil et al., 2024; Mishler et al., 2025; Panigutti et al., 2023) to explanations describing the overall logic and mechanism of the AI's recommendation-generation process (Chen et al., 2024). System explanations have also been proposed as a method of mitigating cognitive load caused by the complexities introduced by AI (Robles-Flores and Roussinov, 2012; Zhou et al., 2017). In addition, interactive user interfaces have been recommended to mitigate the out-of-the-loop problem that users experience when they must monitor an AI model's performance in a decision-making environment (Bader and Kaiser, 2019), or to address the challenges of integrating experts' tacit knowledge into AI models (Micallef et al., 2017).

While the proposed set of solutions covers a wide range of HCAI problems, the emphasis on system explanations and exploration interfaces is evident. However, when individual primary sources are examined in more detail, inconsistencies and differences in authors' conceptualizations of system explanations become apparent. For instance, the term 'explainability' has been used to refer to interactive interfaces for data exploration (Lamy et al., 2019), historical data visualizations (Zhou et al., 2017), explanations focusing on the justification of a specific decision-making recommendation (Mishler et al., 2025; Naiseh et al., 2021a; Ribeiro et al., 2016), a generic construct operationalized as information richness in the rationale provided by AI alongside its decision recommendations (Bigras et al., 2018), and as an encompassing term that covers both global explanations that can be valid across recommendation instances and local explanations that relate to a specific instance of generated recommendation (Naiseh et al.,

2021b; Schoonderwoerd et al., 2021). Similar variation exists in the use of the term ‘transparency,’ with studies conceptualizing it as explanations of system logic (Madsen et al., 2025), justifications of a specific decision-making recommendation (Nguyen et al., 2019), or design factors that make visible the source dataset used by the AI (Hepenstal et al., 2019; van Capelleveen et al., 2021).

The lack of standardization in conceptualization carries over into how primary studies operationalize, measure, and test these constructs in behavioral research. The identified studies yield conflicting results that could, in principle, help uncover the factors and potential moderator variables that lead to different outcomes across conditions. However, inconsistencies in conceptualization make it difficult to compare and synthesize findings across studies reporting behavioral evidence. For instance, the primary sources include reports of both increased cognitive load resulting from system explanations (Zhou et al., 2017) and reduced cognitive load resulting from system explanations (Tsai et al., 2025). Similar conflicting findings appear with respect to the impact of HCAI design factors on decision performance. Some studies report no significant performance gains from system transparency (Mishler et al., 2025), whereas others document performance improvements associated with increased transparency (van Capelleveen et al., 2021). Likewise, while several studies find no effect of explanations on performance (Cecil et al., 2024; Hoque and Mueller, 2021; Mishler et al., 2025), others report positive effects (Micallef et al., 2017; Zhou et al., 2017). Overall, these conflicting results are difficult to synthesize and to link to boundary conditions, largely because key constructs such as explainability and transparency are inconsistently conceptualized and frequently left undefined across studies.

The future research directions identified across the primary sources further underscore the underlying problem of conceptual definition. Several studies highlight the need to clarify which types of explanations are effective at different stages of interaction (Madsen et al., 2025; Tsai et al., 2025). Another recurring theme concerns the out-of-the-loop problem, with studies calling for research on HCAI design approaches that help maintain or improve decision performance when AI systems fail (Brauner and Ziefle,

2019; Panigutti et al., 2023; Rebholz et al., 2024; Xu et al., 2025).

Taken together, the scoping review indicates that research on human-AI interaction designs in organizational decision-making has grown rapidly over the past years across multiple domains. Prior work proposes a wide range of interface-level solutions to support human-AI interaction, with system explanations (e.g., Chen et al. 2024; Johnson et al. 2023), transparency mechanisms (e.g., Mishler et al. 2025; Vössing et al. 2022), and interactive interfaces (e.g., Katzburg et al. 2024; Sivaraman et al. 2023) as the most notable recommendations. However, it remains challenging to synthesize findings for several reasons. First, core constructs concerning these solutions are defined inconsistently and operationalized differently. For instance, the term explainability is used to refer to local explanations (e.g., Gentile et al. 2023), global explanations (e.g., Hauptman et al. 2024), both global and local explanations (e.g., Naiseh et al. 2023; Wulff and Finnestrand 2024), and interactive exploration interfaces (e.g., Sivaraman et al. 2023). A similar inconsistency persists in the use of the term transparency, which has been used to refer to both local (e.g., Mishler et al. 2025) and global explanations (e.g., Madsen et al. 2025) in prior studies.

Second, and reflecting these conceptual inconsistencies, there are conflicting findings regarding the effectiveness of these solutions in improving human-AI interaction. Several studies report benefits such as increased adherence to AI recommendations (e.g., Laxar et al. 2023; Silva et al. 2023) and increased confidence in AI recommendations (e.g., Sivaraman et al. 2023). However, empirical evidence also suggests that explainability can reduce confidence (e.g., Hauptman et al. 2024); increase the risk of overreliance (e.g., Naiseh et al. 2023), or have no significant effect on decision outcomes (e.g., Cecil et al. 2024; Gorre et al. 2023). Although these discrepancies may stem from contextual factors, the underlying inconsistencies in conceptualization and operationalization hinder the synthesis of results. As a result, it remains difficult to identify and distinguish the cognitive mechanisms through which design choices influence users' attention, confidence calibration, and decision performance under different task and domain characteristics.

Moreover, while existing research frequently acknowledges challenges such as miscalibration of confidence in AI (e.g., Chen et al. 2024; Mishler et al. 2025; Silva et al. 2023; Zerilli et al. 2022) and the out-of-the-loop problem (e.g., Bader and Kaiser 2019; Zhang et al. 2021), comparatively little work has examined how users' cognitive engagement and reliance evolve over repeated interactions, in ecologically valid settings where AI performance is prone to fluctuate.

In response to the identified research gaps, this dissertation contributes to the human-centered AI literature by examining how interface-level design choices shape users' cognitive states and decision behaviors in AI-assisted organizational decision-making. The following section outlines the organization of the dissertation and situates each empirical essay within this research program.

Organization of the Dissertation

To address the gaps identified in the preceding background section, this thesis sets out a research program aimed at improving our knowledge of the design principles that enable effective human-AI interaction in knowledge-intensive organizational decision-making. The scoping review highlighted that, although interest in human-centered AI design is growing across disciplines, existing work is dispersed, relies on inconsistent conceptualizations, and provides limited insight into how users develop understanding, confidence, and effective means of determining when to adhere to AI recommendations when interacting with AI systems. Consequently, the three empirical essays in this thesis pursue complementary objectives: identifying the real-world challenges professionals face when working with AI, defining and testing core design constructs that shape user understanding and performance, examining mechanisms that help users maintain effective decision-making when AI systems produce unreliable recommendations, and advancing methodological approaches by integrating physiological, behavioral, and self-reported measurements in ecologically valid repeated-task settings to obtain fine-grained insight into users' cognitive processes during interaction with AI systems. Across the

three essays, Situation Awareness (SA) theory (Endsley, 1995) serves as the guiding theoretical lens, which provides a process-oriented view of human-AI interaction in which perception of task-relevant information informs understanding and projection, which ultimately shape decision-making and subsequent behavioral adaptation through feedback over repeated interactions.

The first essay adopts an exploratory qualitative approach to examine how professionals experience AI-supported decision-making in practice. Drawing on semi-structured interviews with experts in supply chain and logistics, the study investigates how users assess, accept, reject, or override AI recommendations in real organizational settings. Grounded in Situation Awareness theory (Endsley, 1995), the analysis examines how professionals form, maintain, and revise their understanding of AI behavior over time while remaining accountable for decision outcomes.

The first essay develops an extended process model of human-AI interaction based on the Situation Awareness theory. The study explains how professionals form, update, and act upon situation awareness when interacting with AI systems, thereby clarifying how reliance and override decisions emerge over time. It introduces meta-decision-making as a secondary decision process through which users continuously evaluate whether AI recommendations should be followed, positioning this process as distinct from primary task execution. Within this process model, individual, task, and system factors operate as conditions shaping the outcome of meta-decisions. The study further extends the situation awareness theory by conceptualizing AI systems as entities whose behavior users interpret and monitor, enabling the application of distributed situation awareness perspectives to human-AI interaction.

The second essay focuses on AI-to-human mechanisms by examining how transparency and explainability influence user understanding and decision performance when interacting with a generative AI (GenAI) assistant. Using a controlled, repeated-task experimental design, the study provides clear conceptual definitions of transparency and explainability, distinguishes them, and operationalizes them within a conversational GenAI interface. It evaluates how these design features affect users' attention allocation to task-

relevant elements, perceived understanding, confidence calibration, and performance over time.

The second essay contributes to research by clarifying the difference between transparency and explainability and demonstrating how these AI-to-human design mechanisms influence attention, cognition, and performance. It shows that transparency plays a more prominent role in early interactions by supporting users' initial mental model formation, whereas explainability becomes increasingly important as interaction continues and users engage in deeper verification of AI outputs. Methodologically, the essay introduces the use of eye-tracking metrics, specifically saccade counts, as indicators of users' verification effort when evaluating AI recommendations. Together, this study contributes to the understanding of how information communicated from AI systems to users shapes cognitive effort, confidence calibration, and performance.

The third essay focuses on human-to-AI mechanisms by investigating AI configuration interfaces as a means of sustaining effective decision-making when AI systems are prone to produce erroneous or unreliable recommendations. The study examines how allowing users to configure AI system parameters influences attentional processes, confidence calibration, and users' ability to detect and manage AI errors across repeated interactions.

The third essay contributes to the literature on human-AI interaction by identifying AI configuration as a key mechanism through which users' attention to and perceptions of AI are shaped. It demonstrates that providing users with control interfaces enhances users' attentional engagement and improves their interaction with Level 1 situation awareness elements, supporting a more accurate perception of task-relevant information. By strengthening engagement at the input data level, asking for user configuration before providing them with AI recommendations enables improved understanding of AI, thereby shaping their attentional processes that are effective in error detection and intervention when AI behavior is imperfect. These findings extend prior work on out-of-the-loop failures by showing how human-to-AI interface mechanisms can support cognition and performance.

Together, the three essays examine how human-AI interaction evolves across the decision cycle: Essay 1 proposes a process model explaining how reliance decisions emerge during interaction with AI. Essays 2 and 3 then examine specific interface-level mechanisms operating within this process: Essay 2 investigates how AI-to-human information provision mechanisms influence perception and comprehension, and Essay 3 examines whether human-to-AI mechanisms can help sustain attention and effective intervention when AI performance deteriorates. This bidirectional perspective and the primary contributions of each essay are shown in Figure 2.

	AI-to-human interface mechanisms	Human-to-AI interface mechanisms
Essay 1	<i>Identify the cognitive, system, and contextual factors that shape users' meta-decision-making regarding reliance on AI, including input data characteristics, transparency, explainability, interactivity, user control, and workload.</i>	
Essay 2	<i>Define and operationalize AI transparency and explainability, and examine how AI-to-human information provision influences users' understanding, confidence calibration, attention allocation, and decision performance in generative AI interactions.</i>	
Essay 3		<i>Investigate how user control as a human-to-AI interface mechanism shapes attentional processes, engagement with situation awareness cues, and users' ability to detect and handle AI errors over repeated interactions.</i>

Figure 2: Dissertation structure

Chapter 1

The Dynamics of Human-AI Interactions in Organizational Decision-Making

Abstract

Information systems (IS) are increasingly being used to support more complex and knowledge-intensive decision-making tasks. Contemporary systems often rely on machine learning and artificial intelligence (AI) methods that enable the processing of large datasets. While these technologies offer substantial potential, their use in organizational settings raises challenges related to how users understand, evaluate, and

This essay is a version of the following book chapter prior to publication: Öz, B., Karran, A. J., Beringer, J., Coursaris, C. K., & Léger, P. M. (2025). The Dynamics of Human-AI Interactions in Organizational Decision-Making. In *The Design of Human-Centered Artificial Intelligence for the Workplace*, pp. 29–46. Springer. The final authenticated version is available online at: https://doi.org/10.1007/978-3-031-83512-4_3

The version embedded in this essay is the author's manuscript prior to peer review and acceptance for publication. This version includes expanded details regarding the methodology and results, which were abbreviated in the version of record to align with the book's specific context and target audience.

regulate AI-supported decisions. Drawing on semi-structured interviews, this study explores how users assess, accept, reject, or override AI recommendations in practice. The findings indicate that users form and continuously revise interpretations of what the AI system takes into account, how it represents the decision environment, and what it anticipates. Perceived misalignment between users' situation awareness and their interpretation of the system's awareness shapes how decision-making tasks are selectively delegated to, or reclaimed from, the AI. The study proposes an extended process model on human-AI decision-making that highlights ongoing evaluation of AI performance and meta-decision-making as central features of AI use in organizational contexts.

1.1 Introduction

Information systems (IS) play a critical role in enabling employees to make better decisions in today's organizational environments. With continuous technological improvements in data processing capabilities, decision support has evolved from simple computations to providing recommendations in complex problem-solving scenarios. Consequently, this shift has transformed IS from simple tools to artificially intelligent agents that support, and in some cases, work together with, domain experts. This shift highlights the importance of collaborative decision-making models between humans and Artificial Intelligence (AI), which are becoming more common in organizational processes (National Academies of Sciences, Engineering, and Medicine, 2021).

As the extent of decision support from IS extends beyond simple calculations to autonomous decision-making, the relationship between the user and IS becomes more complex. Users may either rely excessively on the IS, potentially losing their ability to understand how future events impact their decision-making environment, or distrust the IS, thereby not fully benefiting from its capabilities in processing large volumes of data and deriving relevant insights (Chiang and Yin, 2022). Balancing these two tendencies is critical for optimal operational outcomes.

Previous research has extensively studied the impact of IS investments on organi-

zational outcomes (Brynjolfsson, 1993; Sabherwal and Jeyaraj, 2015). Extant theories and models, particularly the Technology Acceptance Model (TAM) and its extensions, highlight factors such as perceived usefulness and ease of use as drivers of IS adoption (Davis, 1989; Venkatesh and Bala, 2008). While TAM has been used extensively to explain user interactions with traditional IS, the dynamic, responsive, and often creative nature of AI-powered systems necessitates further research on the topic. Unlike traditional IS that function within defined parameters, AI systems possess the capability to continuously learn and modify their operational patterns based on new data inputs (Schuetz and Venkatesh, 2020). While the benefits of using AI in complex decision-making tasks are clear, this continuously evolving and often opaque nature of AI presents unique challenges (Endsley, 2023). One such challenge is understanding how users perceive and adapt to a dynamic AI-driven decision support system and how they make ongoing decisions regarding task delegation to AI. Therefore, in this chapter, we aim to answer the following research question:

RQ: How do professionals determine whether to rely on AI in organizational decision-making processes?

To understand users' decisions about delegating tasks to AI, Situation Awareness Theory emerges as a promising lens (Endsley, 1995). Situation awareness (SA) is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995, p. 1). While the SA concept was initially defined to understand and explain human decision-making with automation tools in aviation, it can be extended to the context of human-AI interactions (Endsley, 2023).

This chapter presents a study conducted to answer our research question by considering the perspectives of professionals shared through expert interviews. The rest of the chapter is organized as follows: first, we provide an overview of the SA concept and its relevance to human-AI interactions. We then describe our methods for collecting and analyzing stakeholder interview data within a supply chain management (SCM) context. Results are discussed in two parts: first, we introduce six contextual factors

influencing users' meta-decisions whether to delegate decision-making to AI; then, we incorporate these factors into the SA model and propose a conceptual extension. The chapter concludes with key considerations for successful AI implementation projects.

1.2 Background

1.2.1 The Adoption and Use of IS and AI

Information systems have been integral to organizational decision-making environments for several decades. As the success of an IS depends on its use, the factors leading to acceptance have been extensively studied in the literature. The Technology Acceptance Model (TAM) (Davis, 1989) is one of the well-known perspectives, which presents perceived usefulness and perceived ease of use as predictors of user acceptance. Over the years, the set of constructs predicting user acceptance has expanded, leading to models such as TAM2, the Unified Theory of Acceptance and Use of Technology, and TAM3 (Venkatesh and Davis, 2000; Venkatesh et al., 2003; Venkatesh and Bala, 2008).

While these attitude-based theories and other richer theoretical lenses provide valuable insights regarding the use of traditional IS, they assume that an IS is deterministic and behaves predictably according to the rules it was designed to follow (Schuetz and Venkatesh, 2020). AI, however, does not rely on a predetermined set of rules. Instead, AI-driven systems continuously learn from new observations and adapt their behavior accordingly (Czako et al., 2021). Therefore, to address the need to better understand human-AI interaction dynamics in organizational decision-making, we borrow the concept of situation awareness from the scientific discipline of human factors.

1.2.2 The Situation Awareness Model

Situation awareness (SA) is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995, p. 1). This definition of

SA is human-centric, focusing only on the human operators. According to this model, situation awareness is characterized by three levels:

1. Perception of elements in the current situation: This consists of the observations of data points relevant to decision-making.
2. Comprehension of the current situation: This level is about the use of the SA elements from the first level to create a meaningful synthesis that represents the current reality.
3. Projection of future status: This is achieved by integrating SA elements from the first two levels to provide a foundation for anticipating future changes in the environment.

Situation awareness is conceptualized as part of a cyclical decision-making loop: the state of the environment informs SA, which influences decision-making and action execution, which subsequently influences the decision outcomes that feed back to the situation awareness (Endsley, 1995, p. 35). Endsley's model identifies three determinants that influence this process: individual factors (e.g, the operator's goals, experience, and cognitive state), task factors (e.g., contextual complexity, workload, and environmental constraints), and system factors (e.g., interface design and automation capabilities). According to the process model illustrated by Endsley, these factors collectively shape the development of SA, action selection, and decision outcomes in the decision-making loop.

The SA theory initially aimed to explore the relationship between pilots and automation systems in aircraft (Endsley, 1995). The high volume of sensory data to be processed and the time sensitive nature of this context makes it important for pilots to maintain their SA. Similarly, modern organizational decision-making environments exhibit similar characteristics. With digital technologies increasingly involved in data generation and processing, the speed and accuracy expected from organizational decision-making processes have also risen (Danilova, 2019). Therefore, the conditions in modern

decision-making environments are parallel to those targeted by the initial SA theory (Endsley, 2023).

In summary, existing research and literature involving SA suggest that it is a promising theoretical lens through which to investigate human-AI interactions in complex decision-making scenarios. In this study, we explore the decision-making process of experts who have access to AI-driven decision support systems in their daily work. Through this exploration, we aim to extend our understanding of underlying factors that influence experts in either relying or ignoring AI systems during decision-making.

1.3 Methods

This study employed a qualitative research design, using semi-structured interviews to investigate human-AI interactions in an organizational decision-making context. The research protocol was reviewed and approved by the HEC Montréal Research Ethics Board (Certificate No. 2021-4436).

1.3.1 Sampling and Participants

A purposeful sampling method was used to recruit participants based on their roles and experience with an AI-powered SCM software solution (Palinkas et al., 2015; Seidman, 2006). Initial participants were recruited from a registry of the software solution's stakeholders, and additional participants were subsequently identified through snowball sampling.

The final sample ($N = 20$) comprised three primary stakeholder groups to ensure systematic coverage from system design to its use in practice. The following stakeholder groups were recruited in the following order: individuals from AI development teams ($N = 6$), consultants specializing in system implementation and end-user assistance ($N = 8$), and end-users ($N = 6$). Interviewing developers and consultants first allowed the researchers to establish the designers' view of the system, including what it is

intended to do, what it can and cannot do, and how its capabilities are configured, before examining how end users perceive, interpret, and act upon these capabilities in practice. In addition, developers and consultants were asked to share observations from their implementation and support experience regarding user needs, recurring requests, and human-AI interaction challenges, which informed the refinement of subsequent interview guides and supported more focused probing of the system's technical and contextual dimensions in interviews with end users.

Criteria for selection required participants to have active roles in configuring, managing, or utilizing AI systems in organizational SCM decision-making tasks.

1.3.2 Data Collection

This study employed a semi-structured interview method for data collection. The protocol was organized into three thematic sections to explore the nature of human-AI interactions: respondents' roles and tasks, human-AI interaction factors, and end-user perceptions of AI.

In addition to the thematic sections on human-AI interactions, the interviews used laddering techniques and targeted probes to elicit further insights into respondents' accounts of human-AI interaction dynamics, while adhering to the main interview structure (Reynolds and Gutman, 1988). This semi-structured format allowed participants to provide detailed accounts of their lived experiences while ensuring consistency across the sample (Xiao et al., 2017). To ensure content validity and technical accuracy, the guide underwent iterative refinement within the research team and was finalized following two review cycles with an industry expert specializing in the specific type of organizational IS under study. The complete interview guide is provided in the Appendix, Section A.1.1.

Interviews were conducted via videoconferencing from April 2021 to November 2021. Sessions were moderated by one member of the research team, with at least one additional member attending to assist in facilitating flow, ensuring adherence to the guide, and posing probing questions as needed. On average, interviews lasted 75 minutes (min: 49 minutes;

max: 98 minutes). Recordings were captured via the videoconferencing platform, with participants' written consent prior to the session and verbal consent at the beginning of their interview.

1.3.3 Analysis of Collected Data

Audio recordings of interview sessions were initially processed using automated transcription software (Zoom Communications, Inc., San Jose, CA, US). These transcripts were then imported to NVivo (QSR International, Burlington, MA, US) for correction and refinement. The resulting dataset comprised approximately 140,000 words.

In the early stages of this study's design, Situation Awareness (SA) Theory was identified as a suitable framework for investigating the dynamics between human decision-makers and AI agents in organizations. However, as SA Theory has primarily been applied in other fields, unique insights may emerge from applying it in our study's specific context. We adopted the Multi-Grounded Theory (MGT) method, which provides a balanced approach between inductive and deductive reasoning in an iterative process, making it possible for additional findings to emerge (Goldkuhl and Cronholm, 2010).

In line with MGT, our analysis included three grounding phases: empirical, theoretical, and internal. To ensure a common understanding of coding requirements, two researchers from the research team independently applied an inductive coding approach to a subset of approximately 28,000 words (20% of the total dataset). Disagreements were resolved through collaborative review to refine the coding rules. Following this phase, the primary author continued with the iterative transcription and coding of the remaining data using NVivo (QSR International, Burlington, MA, US).

The coding process followed the three steps of MGT:

1. Empirical Grounding (Step 1): Initial inductive coding was performed on the dataset without reference to existing theories.
2. Theoretical Grounding (Step 2): The research team held periodic meetings to review emergent codes and map them into the conceptual framework of SA theory.

3. Internal Grounding (Step 3): This phase involved theory condensation, where emergent codes that did not align with existing categories were evaluated to identify potential theory expansions.

These three phases were conducted iteratively as data collection progressed.

1.4 Results

As this study’s research focus was on understanding the dynamics between human and AI-systems during decision-making, we structure the findings in accordance with the decision-making process model described by Endsley (1995, p. 35). More specifically, we discuss the results concerning the three levels of SA (e.g., perception, comprehension, prediction), and the subsequent decision stage within the decision-making process (Endsley, 1995). Before addressing these factors, we first examine the emergent concepts of continued evaluation of AI and meta-decision-making.

1.4.1 Decision-Making with AI Support and Meta Decision-Making

Participants described the AI system not merely as a tool, but as a counterpart within the decision-making loop. Multiple participants explicitly referred to their interaction with the system as a form of collaboration, describing it as performing part of the decision task alongside them and characterizing the relationship as an active partnership:

‘I do not have time to spend all day trying to write schedules, so I am extremely thankful that it (the AI-powered SCM software solution) does the most part of it right. I do love a system that does about 75% (of the work). That collaboration is something that I am very thankful for.’ — *Participant 8*

Participants described the AI system using language that framed it as an active collaborator in the decision-making process. In these accounts, participants referred to what the system “knows,” how it “perceives” changes in the environment, and the limits

of what it takes into account, articulating expectations about the system's handling of information and its responsiveness to situational changes:

'We know that the system is intelligent enough to know when we need people (at) what times... I know it keeps all those factors in mind.' — *Participant 8*

'Sometimes the planner does a prediction override because there is something that the system cannot know, or the system needs some time to adjust. COVID is a good example.' — *Participant 11*

'Sometimes we had to do an override, but it was because the model did not see the right data. Then, we could see (that) over time, the system adjusted and now we are doing more long-term overrides (rather than short-term).' — *Participant 9*

'It fills the gaps based on the trends of our sales. For instance, back-to-school sale...(or) Christmas, I know it (the system) keeps all that in mind, and that is great.' — *Participant 8*

In addition, participants described an ongoing process of evaluating how AI outputs should be used in their decision-making activities. In these accounts, participants referred to repeatedly assessing the system's performance and reliability in order to decide when to rely on AI-generated recommendations and when to intervene. This process aligns with what we refer to as meta-decision-making, that is, a continuous process of "deciding how to decide" (Kottemann, 1986).

Participants described using various methods to monitor AI performance over longer time horizons, including retrospective comparisons between system outputs and realized outcomes:

'We have a table where we keep track of these forecast predictions so later on, in time, we can actually compare with actuals...' — *Participant 9*

'We use an acceptance panel to review (AI) recommendations...It is not (a) fully automated (process); each price proposal is verified by end users.' — *Participant 3*

'We constantly look at the model quality, (by comparing) the forecasted (sales) and the real (sales). We also analyze the data, maybe there was a promotion, or (maybe) there were additional events like (an) external event that the system does not know about.' — *Participant 3*

Participants most often described adjusting their engagement with the AI system in situations where system outputs differed from their own contextual knowledge or expectations. In these situations, participants reported making decisions about whether to modify, suspend, or override the system's involvement in the decision process:

'When we deliver (between) two countries with customs... I need to consider when they (customs offices) are operating, modify the settings, and change the status of my load to override (the AI recommendation).' — *Participant 18*

'...The system usually increases the prices before (Black Friday), because it knows about the (upcoming) promotion. We switch it off for two weeks because we do not want to increase the price right before Black Friday.' — *Participant 3*

'Sometimes we have to cheat. We have to create a longer transit time, a longer duration, or a shorter one.' — *Participant 13*

Participants described considering different aspects of the AI system when deciding whether and how to delegate decision-making tasks, with these considerations varying across stages of the decision-making process. To reflect this structure, the following results sections are organized by the stages of the decision-making process and the corresponding levels of Situation Awareness, as described by Endsley. Throughout the

following sections, delegation refers not to the transfer of the entire decision-making process, but to users' meta-decisions to rely on the AI system for specific sub-decision-making tasks associated with distinct levels of Situation Awareness or action selection. The factors described by participants at each stage are summarized in Table 1.1.

Table 1.1: The factors influencing meta-decisions at various stages of decision-making.

Situation Awareness			Decide
Perceive (Level 1)	Comprehend (Level 2)	Predict (Level 3)	
-Input data -User control -Data volume	-Data volume -AI transparency -AI configurability	-AI explainability	-AI interactivity
-Task complexity -User workload -User perceptions of AI -Task importance -Presence of an exception in the environment			

1.4.2 Meta-Decisions at the First SA Level

This section presents evidence from participants' accounts of meta-decisions about delegating perception of the environment to the AI system. In these accounts, participants referred to three recurring aspects when discussing whether they relied on the system for the initial perception of environmental elements: their assessment of the system's input data, the volume of data involved, and the extent to which they could control or adjust the input data used by the system. The following subsections describe how participants discussed these aspects in relation to delegation at the perception stage.

Input Data

Participants frequently referred to input data when describing whether they relied on the AI system. Across interviews, participants discussed assessing the system's perceptual capability with respect to three aspects: the completeness and relevance of the data available to the system, the alignment between the system's input variables and their

own contextual knowledge, and the extent to which the system could incorporate or accommodate external, non-quantifiable information.

In these accounts, participants described modifying or withdrawing reliance on the system at the perception stage when they perceived mismatches between the system's inputs and situational conditions. Participants referred to discrepancies involving tacit knowledge, exceptional events, or external constraints as prompting closer scrutiny of the system's outputs. In such situations, participants described supplementing the system's inputs, overriding its recommendations, or bypassing the system in order to reconstruct the decision process themselves:

'(Decision-makers) want to see every price because they are used to doing that. They know they spend a lot of time, but they still want to continue. They say—I want to look at each and every price, and if I am okay (with that), I will go to the next step.' — *Participant 4*

'Sometimes there are also information that cannot be quantified, so the system does not know about them... For example, we know about a special event that will take place in the future... If the system recommends price decrease for this item, we would reject such (a) proposal.' — *Participant 3*

In summary, considerations related to input data emerged as a recurring theme in participants' accounts of delegating perceptual sub-decision-making tasks to the AI system, particularly regarding the alignment between system inputs and situational conditions.

User Control on Input Data

Participants frequently referred to user control over input configuration when discussing whether they relied on the AI system. In addition to the data available to the system, participants emphasized the ability to adjust or supplement variables that shape how the system represents the decision environment. In these accounts, limited configurability

was described as making it more difficult to align the system's representation with contextual conditions known to users.

Participants described situations in which the system's input configuration did not reflect contextually relevant factors, such as employees' experience levels, which affect task duration and performance. In such situations, participants reported limiting reliance on the system's perception of the situation and taking a more active role in shaping the decision process:

'I just heard ten new associates (just started), and they are not experienced. So, I know that I can not look at the freight planning tool and put that exact amount of people (based on what the planning tool says). I am going to need a little bit more (associates) because they are going to be slower.' — *Participant 8*

Participants also described disengaging from AI-supported decision-making in contexts where known exceptions or strategic considerations could not be reflected through input configuration. In these situations, users suspended reliance on the system for the decision at hand, citing misalignment between the system's perception of the environment and their own understanding:

'There are some circumstances that we make a business decision (to apply) aggressive promotions due to market conditions, and we switch this (AI forecasting) system off for this period.' — *Participant 3*

In summary, user control over input configuration emerged as a recurring theme in participants' accounts of delegating perceptual sub-decision-making tasks, particularly in relation to aligning the system's representation of the environment with contextual knowledge.

Data Volume

Participants frequently referred to data volume when describing whether they relied on the AI system. Participants described situations in which the quantity of data required to form an initial understanding of the decision environment exceeded what could reasonably be processed through manual inspection. In these contexts, participants discussed relying on the AI system to support perceptual awareness across large-scale datasets:

'I would never ask anyone to write schedules for (more than) 300 associates without having a system like this.' — *Participant 8*

Participants further described AI support as particularly relevant when perception involved aggregating and scanning large numbers of data points to identify patterns or baseline conditions. Rather than processing individual data elements one by one, participants referred to using the system to construct an initial representation of the environment from high-volume inputs:

'...the granularity of the planning and the complexity of that...you are trying to schedule physical trucks, loads, and drivers, and all that is very difficult to do it manually at a large scale.' — *Participant 12*

'Having millions of items,... it becomes very difficult to do it manually. That is where it (AI-driven decision support system) comes in. It automates the manual task, and it is highly scalable.' — *Participant 4*

When contrasting current AI-supported practices with earlier manual approaches, participants described limits in human processing capacity when data volumes were large:

'In the past, the product team responsible for pricing [was] able to analyze thoroughly 100–200 items a few times a season... and they calculated the best prices for other countries directly based on exchange rates. However, the competition in other markets is different... so it was sub-optimal... Right now,

they receive price proposals (from the system) for all items... in all markets, and it takes them one hour a week to analyze... and make those decisions.’

— *Participant 3*

In summary, data volume emerged as a recurring theme in participants’ accounts of delegating perceptual sub-decision-making tasks, particularly regarding the feasibility of manually developing an initial understanding of the decision environment.

1.4.3 Meta-Decisions at the Second SA Level

This section presents participants’ accounts related to Level 2 Situation Awareness (comprehension) elements during meta-decision-making. Across interviews, participants referred to three recurring aspects related to the comprehension stage when discussing reliance on the AI: the volume of information to be synthesized, the transparency of the system’s internal logic, and the extent to which they could configure how recommendations were generated. The following subsections describe how participants discussed these aspects in relation to retaining or delegating responsibility for interpreting AI-generated information.

Data Volume

As in the perception stage, participants described situations in which the amount of information exceeded what could reasonably be processed manually. At the comprehension stage, these accounts focused on integrating and interpreting large numbers of system-generated recommendations rather than on perceiving individual data elements.

Participants described situations in which the volume of recommendations constrained their ability to systematically review and make sense of information within the available time. In these contexts, participants referred to relying on the system to organize and structure outputs in ways that supported situational understanding:

'We do not analyze product behavior in time because... there are thousands of proposals each week and (we have) only one hour... so we go through those proposals, and usually we do it on product groups... (for example) we set price decreases... then we see if there are any price increases and those can be analyzed in more detail.' — *Participant 3*

Rather than engaging with each recommendation individually, participants described a need for higher-level aggregation to guide interpretation and attention. When such aggregation was unavailable, participants reported relying on their own heuristics to decide where to focus their analysis and expressed a preference for summary views that would facilitate understanding of the system's representation of the decision environment:

'I see recommendations, for instance, a certain product category in (a) store is having a low accuracy... (but) we have too many stores to actually bother about... I think having an initial higher(-level) aggregated view on those recommendations would help guide us about where we need to look. Now, we just think ourselves what we want to check and whether we like what we see.' — *Participant 5*

Participants also described operational contexts in which comprehension required integrating multiple constraints and entities simultaneously. In these situations, participants referred to the system's ability to synthesize information across orders as supporting their understanding of feasible solutions:

'(The system) allows the person in charge of the planning... to deal with many, many orders and to be able to optimize the truck(s). (...) If I compare it to another system... I will have to pick every order, one by one, and then calculate myself what I will be able to put in each (truck)... (The strength of the system is) that it can, if I have... 100 orders, it will pick it all, and then give me... five different (trucks) all (filled) with 100 (percent).' — *Participant*

18

In summary, data volume emerged as a recurring theme in participants' accounts of delegating comprehension-related sub-decision-making tasks, particularly in the context of integrating and interpreting large numbers of system-generated recommendations.

System Transparency

System transparency refers to "the understandability and predictability of the system" (Endsley et al., 2003, p. 146). Prior work describes transparent systems as those that allow users to understand the data sources and internal logic underlying system outputs (Chen et al., 2014). In participants' accounts, transparency was discussed in relation to delegation to the AI system. These descriptions extended beyond perceiving information to making sense of how AI-generated outputs were produced and how different factors were combined in recommendations.

Participants referred to limited insight into the system's internal logic when describing challenges in interpreting AI outputs, particularly when recommendations diverged from expectations or required justification. In these situations, participants described the system as operating as a "black box," which made it difficult to develop a coherent understanding of how decisions were generated:

'So, from the point of view of the planner, I doubt they have this kind of understanding, honestly. Ideally, the whole thing is happening in a black box, as far as the decisions that are made on how the goods are transferred, which goods are transferred first, and how trucks are packed.' — *Participant 16*

Participants also referred to transparency when discussing situations in which AI-generated recommendations needed to be evaluated, explained, or defended. In these accounts, requests for explanations were framed as a need for accessible descriptions of how the system arrived at specific outcomes:

'...it was (the system had) very high-level (explanations). All it said was these (are) your demand hours based on your trends, and these are the hours... I like

to understand how things work. It helps me buy into the system and want to work for the system and not push back.’ — *Participant 8*

’I would say the decision-maker gets the knowledge of the system to a quite low extent, because there are so many factors affecting the final price proposal. And sometimes it is difficult to explain why the system recommended a particular price.’ — *Participant 3*

’They (end-users) want us to tell them in very plain language "this is the problem." Currently, when we are generating exceptions, they want to understand what happens. They do not want a cryptic error.’ — *Participant 16*

In contrast, participants also described situations in which the system’s logic was sufficiently clear for them to follow how recommendations were constructed. In these cases, participants referred to being able to integrate AI outputs into their own understanding of the decision environment:

’...for all of these parameters, it (the system) tries to find solutions to load on time, to deliver on time, to have food trucks, to find capacity. And, in the morning, we will check if we have non-eligible nodes and try to find solutions according to what we see.’ — *Participant 13*

In summary, system transparency emerged as a recurring theme in participants’ accounts of delegating comprehension-related sub-decision-making tasks, particularly regarding understanding how recommendations were generated and how multiple factors were combined.

System Configurability

Participants referred to system configurability when discussing reliance on the AI system. In these accounts, participants described not only the need to understand how the system

interpreted information, but also the ability to adjust parameters or test alternative scenarios in order to engage with how recommendations were produced. Participants described using configurability to explore how changes in inputs were reflected in system outputs.

Participants described the ability to modify model parameters as allowing them to experiment with different scenarios and observe how recommendations changed in response:

'The biggest value that the (AI-driven decision support system) brings is quickly modeling it (a scenario) and then playing around with it, such as changing parameters quickly and running it once again.' — *Participant 4*

Participants also discussed configurability in relation to transparency. They described situations in which limited configurability constrained their ability to interpret system outputs:

'We are not really configuring it because that is a given for us. In a way, it is a black box.' — *Participant 5*

In summary, system configurability emerged as a recurring theme in participants' accounts of delegating comprehension-related sub-decision-making tasks, particularly in relation to engaging with how recommendations were produced through parameter adjustment and scenario exploration.

1.4.4 Meta-Decisions at the Third SA Level

This section presents participants' accounts related to the delegation of Level 3 Situation Awareness (prediction) to the AI system. At this stage, participants described evaluating how the system anticipated future states and outcomes, and they frequently referred to system explainability when discussing these evaluations.

System Explainability

Explainability refers to "the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model" (Vilone and Longo, 2021, p. 91). Unlike transparency, which relates primarily to the broader understanding of the system and its underlying model, explainability focuses specifically on a given prediction and the factors influencing that outcome.

Participants described a need to understand what triggered particular forecasts, especially when AI-generated predictions diverged from their expectations. In these situations, participants referred to seeking explanations that would allow them to assess whether the system's anticipated future states were reasonable:

'We still need to understand, in what scenarios, what is triggering this forecast, which feels off (compared to our expectations).' — *Participant 9*

Participants also referred to explainability when discussing concrete recommendations, such as prices or resource allocations, that reflected the system's predictions about future conditions:

'It would be very useful to have some explanation on why a given product has a specific price (recommendation).' — *Participant 3*

'...(I) definitely needed some more training on the understanding as to why the system gave me 95 hours on a Monday.' — *Participant 8*

'...all these parameters say (to me), do not look at (the profit) margin, just get rid of this item. And we see that the system still does not recommend any price decrease... We do not understand why the system does not want to decrease the price...' — *Participant 3*

In summary, system explainability emerged as a recurring theme in participants' accounts of delegating prediction-related sub-decision-making tasks, particularly regarding the understanding and evaluation of AI-generated anticipations of future states.

1.4.5 Meta-Decisions at the Decision/Action Selection Phase

The decision-making process described by Endsley (1995, p. 35) concludes with the decision-action selection phase, in which a commitment is made to implement a chosen course of action. Unlike earlier stages of situation awareness, which focus on perceiving, interpreting, and anticipating aspects of the decision environment, this phase centers on committing to an action.

In participants' accounts, meta-decisions at this stage concerned whether to delegate the final action to the AI system or to retain manual control over action selection. The following subsections present participant accounts related to these meta-decisions, including references to task importance, task complexity, user workload, exceptional cases, system interactivity, and perceptions of AI decision performance.

Task Importance

Participants referred to task importance when discussing decisions about whether to intervene in or delegate action selection to the AI system. In these accounts, participants described considering the potential organizational impact and perceived risk associated with a decision when evaluating AI-generated recommendations. Tasks perceived as more consequential were discussed in relation to closer scrutiny of system outputs prior to committing to action:

'Planners do not want to take the risk, especially with important products, especially if the order (suggested by AI) is way lower (than our anticipations).'

— *Participant 9*

Participants also described handling decisions perceived as less critical differently. In these situations, participants referred to relying on AI-generated recommendations more broadly, often evaluating them in aggregate rather than examining each case individually:

'Some products, less important products,... products with lower quantity, they are accepted or rejected as a group of items. We spend more time analyzing

the high-value products, newer products.’ — *Participant 3*

‘It is not that time-consuming to filter those less important items and decide whether we accept (AI recommendations) or not.’ — *Participant 3*

In summary, task importance emerged as a recurring theme in participants’ accounts of final action selection with AI, particularly in relation to perceived risk and organizational impact.

Task Complexity

Participants referred to task complexity when discussing decisions about whether to intervene in or delegate final action selection to the AI system. In these accounts, participants described their decision-making context as involving multiple interdependent variables, requiring the construction of a comprehensive mental model of the decision environment. This complexity was described as challenging for users’ own sensemaking as well as for interpreting AI-generated recommendations:

‘Modeling (the) complex reality is not always easy, and there is typically a lot of discovery that has to happen.’ — *Participant 2*

Participants also described situations in which decision contexts involved many interacting factors, making it difficult to fully understand how specific recommendations were produced. In these situations, participants referred to limitations in their ability to develop a clear understanding of system outputs, even when collaborating closely with system developers:

‘I would say the decision-maker gets the knowledge of the system to a quite low extent, because there are so many factors affecting the final price proposal. And sometimes it is difficult to explain why the system recommended a particular price.’ — *Participant 3*

Task complexity was also described as interacting with system design in ways that constrained users' ability to safely intervene or configure the system:

'It is complex to run. If you open one of those files and you change one parameter without knowing what the impact is, you are likely to break it.' —

Participant 1

In summary, task complexity emerged as a recurring theme in participants' accounts of final action selection with AI, particularly in relation to assessing the consequences of intervention.

User Workload

Participants referred to workload when discussing decisions about whether to delegate final action selection to the AI system. In these accounts, participants described situations in which the time and effort required to perform tasks manually exceeded what they considered feasible given available resources and operational constraints. In these contexts, participants discussed delegation to the AI system as a practical way of managing workload demands:

'If (we) start doing this (task) manually, that is a couple of planners doing that and creating multiple stops to make sure about the distribution of loads across drivers... It can take (us) hours, or days.' — *Participant 2*

Participants also described situations in which high workloads constrained their ability to manually execute or closely review decisions. In these accounts, participants referred to accepting AI-generated recommendations when time constraints were salient:

'I do not have time to spend all day trying to write schedules, so I am extremely thankful that it (the AI-powered SCM software solution) does the most part of it right.... that eliminates like (about) five hours of work I have to do.' — *Participant 8*

In addition, participants described certain tasks as effectively unmanageable without AI support due to their scale or required responsiveness. In these cases, delegation was described as necessary to maintain operational continuity:

'I would never ask anyone to write schedules for 320 associates without having a system like this.' — *Participant 8*

Some participants also referred to the system's responsiveness in time-sensitive contexts when discussing reliance on AI-generated outputs:

'For the short-term forecast, the responsiveness to what is actually happening in the stores is really strong in that (AI recommender). That is something that we would never have achieved manually.' — *Participant 5*

In summary, workload emerged as a recurring theme in participants' accounts of final action selection with AI, particularly in relation to managing time and resource constraints.

Exceptional Cases

Participants referred to exceptional cases when discussing decisions about whether to retain manual control over final action selection rather than delegating execution to the AI system. In these accounts, participants described intervening manually when situations fell outside the system's expected scope or standard operating assumptions:

'We work mostly in automatic (mode), but sometimes the supply team sends us some emergencies that are standby manual internal order forms. We work in manual (in these cases).' — *Participant 13*

Participants described such cases as involving heightened specificity, urgency, or constraints that the system did not readily accommodate. In these situations, participants referred to manual intervention when the requirements of the task were not represented within the system's capabilities:

'Another example, if I have some emergencies, I have to deliver less tires than one full truck from one point to another. The system does not do that for me, so in such cases, I do this load setting manually.' — *Participant 18*

Participants also discussed exceptional cases in relation to limitations in system flexibility or access to relevant information. In these accounts, participants referred to overriding AI-generated recommendations when decision contexts involved factors that were not represented in the system's inputs or fell outside its configurable parameters:

'The overrides we typically do are because the commercial department has another idea than (the one) the system comes up with. It could be information not represented in the input data, or it could be a forecast for over six months, but we need to place orders now.' — *Participant 7*

In summary, participants' accounts indicate that exceptional cases were frequently discussed in discussions of final action selection, particularly when situations fell outside the AI system's standard operating conditions.

System Interactivity

Participants referred to system interactivity when discussing their engagement with the AI system during the action selection phase. In these accounts, participants described interactive interface features such as warnings, visual feedback, and adjustable recommendations in relation to evaluating and refining AI-generated actions prior to implementation:

'If I accidentally schedule someone seven days in a row, it is going to flag... I do love those warnings and those reminders that it gives us.' — *Participant 8*

Participants also described interactive elements for assessing AI-generated recommendations in real time. In these accounts, visual cues and feedback mechanisms were

discussed as providing information about potential misalignments with operational constraints:

'It drops (schedules) as a suggestion and there is a graph on the bottom that shows you based on trends of customers' demand, and then it shows you if you are either under the demand or over the demand for that day. You can go along with that graph...' — *Participant 8*

Participants further referred to interactivity when describing opportunities to adjust system settings and observe resulting changes before finalizing actions:

'As experiments, we also change some settings for selected stores or items and then we decide whether we keep those new settings or we come with the previous ones.' — *Participant 3*

In summary, system interactivity emerged as a recurring theme in participants' accounts of final action selection with AI, particularly regarding the review, adjustment, and confirmation of AI-generated actions.

User Perceptions of AI Decision Performance

Participants referred to their perceptions of AI decision performance when discussing whether to delegate or intervene at the final action selection stage. In these accounts, participants described monitoring system outcomes over time and reflecting on how AI-generated decisions compared with actual results:

'...the system is providing recommendations for us to evaluate... We need to have more cases where we also build trust to actually think the other way around (having the task fully automated).' — *Participant 7*

Participants also discussed the importance of retrospectively evaluating AI decisions by comparing system recommendations with realized outcomes. In these accounts,

participants referred to the absence of structured mechanisms for such comparisons when describing challenges in assessing system performance:

'(I) need to have a good, clever way of judging afterwards, who was right. I think that would really help building the trust in the system.' — *Participant 5*

Participants further described situations in which repeated comparisons between their own judgments and AI-generated decisions informed how they approached subsequent action selection:

'And if I see afterwards that I was totally wrong and the system was right, after two or three times I would say, 'Okay, I will not bother and just let the system do its work.'" — *Participant 7*

Participants also referred to maintaining informal or self-developed tracking mechanisms to support these comparisons over time:

'We do have a table where we keep track of these forecasts, so later on we can compare with actuals and calculate our own measures.' — *Participant 9*

In summary, perceptions of AI decision performance emerged as a recurring theme in participants' accounts of final action selection with AI, particularly in their reflections on past outcomes and in comparisons of AI-generated decisions with realized results.

1.5 An Extended Process Model of Decision-Making with AI

Our analysis enabled us to map participants' accounts of human-AI interaction onto the stages of the decision-making process. Using SA as an organizing lens, we examined how users described their perceptions, comprehensions, and predictions of the decision

environment when working with AI-driven decision support systems. At the same time, applying this lens revealed limits of the existing SA decision-making model (Endsley, 1995, p. 35), which is primarily centered on a single human decision-maker and does not explicitly account for the ongoing involvement of AI systems in these processes.

Mapping the empirical data into the SA framework highlighted three interrelated theoretical insights that refine how decision-making unfolds in AI-supported contexts. First, participants' accounts indicate that users reason about AI systems by forming working assumptions about what the system is aware of at different stages of the decision process. Second, the analysis revealed that situation awareness in human-AI decision-making is bidirectional between the user and the system. Users continuously adjust their own understanding based on AI outputs while also reshaping the system's contextual awareness through input data manipulation, model configuration, and decision intervention. Third, these interactions are governed by an ongoing layer of meta-decision-making through which users assess the alignment between their own situation awareness and their understanding of the system's awareness, and regulate the delegation of sub-decision tasks accordingly.

Together, these insights suggest that AI-supported decision-making can be conceptualized as a dynamic process in which situation awareness is distributed, compared, and updated over time between the human and AI, rather than as a static allocation of responsibilities. To capture this dynamic, we develop an extended process model that incorporates both user situation awareness and users' interpreted representations of system situation awareness, and the meta-decision-making processes that connect them. In the following subsections, we elaborate on these insights and clarify how they extend the model of situation awareness in dynamic decision-making (Endsley, 1995, p 35).

1.5.1 Attributed System SA

Our analysis reveals that users engage with AI-supported decision-making by forming working assumptions about what the AI system is aware of at different stages of the

decision process. When mapping participants' accounts onto SA theory, we observed that users repeatedly described the system in terms that implied assessments of what it "knows," what it "takes into account," and what it can or cannot anticipate, for example when explaining why they trusted a recommendation in routine cases but intervened when contextual factors were, in their view, not "known" by the system. Participants also referred to the system as "learning" shifts in contextual factors that affect their decisions, particularly when describing changes in system behavior over time. These descriptions do not indicate that users view AI systems as autonomous decision-makers, but rather that they construct interpretive representations of the system's awareness in relation to the task at hand.

This observation suggests extensions to the application of SA theory in AI-supported contexts. In its original formulation, SA denotes a cognitive state achieved by a human decision-maker through the perception, comprehension, and projection of environmental elements (Endsley, 1995). In the context of AI-supported decision-making, however, users appear to reason not only about their own situation awareness but also about the system's perceived awareness, inferred through its inputs, outputs, and behavior. Prior research in psychology and human-AI interaction has noted that users develop mental models of intelligent systems to interpret and anticipate their behavior (Bansal et al., 2021; Norman, 1988). Our analysis suggests that such mental models can be represented along the same dimensions that organize human SA, allowing users to reason about the system's perceived perceptual scope, its interpretive logic, and its anticipatory capacity.

Importantly, this attributed system situation awareness is not treated by users as fixed or complete. Instead, it is continuously evaluated against contextual conditions, task demands, and observed system performance. When users perceive gaps between their own understanding of the situation and their interpretation of the system's awareness, they adjust their interaction with the AI, including whether they rely on, build upon, or override its recommendations. This insight extends SA theory by highlighting that, in AI-supported decision-making, situation awareness moves beyond being an individual cognitive achievement and also serves as an interpretive reference point against which

users assess the appropriateness of delegating specific sub-decision tasks to the system.

1.5.2 Bidirectional Adjustment of User and System SA

Our analysis indicates that ongoing, bidirectional information transfers and adjustments characterize the interaction between users and AI systems during decision-making. Users did not describe their engagement with the system as simply receiving and acting upon AI-generated outputs. Instead, they referred to repeatedly interpreting system recommendations, modifying inputs or parameters, and intervening in the system's operation in response to contextual changes or perceived misalignments. For example, participants described revising their own understanding of demand patterns or operational constraints after reviewing system-generated forecasts, and reported instances in which they altered input data, adjusted parameters, or temporarily disabled automated functions when the system's recommendations conflicted with situational knowledge. These accounts suggest that human-AI interactions involve a series of adjustments to both users' situation awareness and the system's representation of the decision environment.

From a theoretical perspective, this pattern highlights a possible extension to how SA operates in AI-supported decision-making. In the SA theory, situation awareness is updated as new information is perceived and integrated by the decision-maker (Endsley, 1995). We found that in AI-supported contexts, users similarly revise their situation awareness through exposure to AI-generated summaries, recommendations, and projections, for instance, by reassessing their interpretation of trends or future conditions in light of system outputs. At the same time, users actively reshape the system's contextual awareness through operational mechanisms such as input modification, parameter configuration, recommendation overrides, or temporary suspension of system functions, particularly when they believe contextual factors are not adequately reflected in the system's inputs. By altering the scope and structure of the information the system processes, users described influencing the system's subsequent outputs.

This bidirectional adjustment process has implications for theorizing decision-making

with AI. Rather than treating situation awareness as a state achieved solely through the perception of external cues, the findings suggest that awareness is also updated through continued evaluation of, and interaction with, the AI system itself. As users revise their understanding based on system outputs while simultaneously intervening in the system's operation, situation awareness becomes relational and dynamic. When users perceive a misalignment between their own understanding of the situation and the system's representations, this misalignment creates tension that motivates reassessment and intervention (Jiang et al., 2021). This ongoing adjustment creates the conditions under which users engage in higher-order evaluations of when and how to rely on the AI, thereby motivating the meta-decision-making processes discussed in the following subsection.

1.5.3 Meta-Decision Making

Meta-decision making, or the act of "deciding how to decide" (Kottemann, 1986), emerges in this study as a key regulatory mechanism in AI-supported decision-making. While prior work conceptualizes meta-decision-making as the selection of decision strategies, information sources, or analytical techniques, our analysis indicates that, in AI-supported contexts, meta-decision-making extends to the ongoing evaluation of the AI system's role within the decision process itself. Users continuously assess whether, when, and to what extent specific sub-decision tasks should be delegated to the AI, retained under human control, or reclaimed following intervention.

When examined through the SA lens, meta-decision-making operates by regulating the distribution of SA sub-tasks across perception, comprehension, and prediction stages, as well as the final sub-task of decision or action selection. Users maintain their own situation awareness while simultaneously forming and updating an interpretation of the system's awareness. For example, participants described delegating perceptual sub-tasks to the AI when large volumes of data made manual scanning infeasible, relying on the system to surface relevant signals or baseline conditions. On the other hand, users also

described cases where they rejected the system's comprehension when its representations conflicted with contextual knowledge, leading them to disregard the system's framing of the situation and reconstruct their own understanding before proceeding.

Meta-decisions occur as users assess the alignment between these two forms of awareness and determine how responsibility for specific stages of the decision process should be allocated. These assessments are shaped by users' ability to interpret how the system represents the decision environment, to understand the basis of its recommendations, and to intervene when relevant contextual elements are not reflected in the system's inputs. As a result, meta-decision-making happens continuously alongside the primary decision process rather than as a discrete step. The notion of meta-decision-making reframes delegation in AI-supported decision-making at a more granular level and in an ongoing manner, rather than as a binary, one-time decision. Users do not decide once whether to rely on the AI; instead, they repeatedly recalibrate delegation in response to contextual changes, perceived system limitations, and prior outcomes. The process model shown in Figure 1.1 extends the model of situation awareness in dynamic decision-making (Endsley, 1995, p. 35) into human-AI interaction by representing users' attributed system SA, bidirectional adjustments, and the continuous meta-decision-making processes.

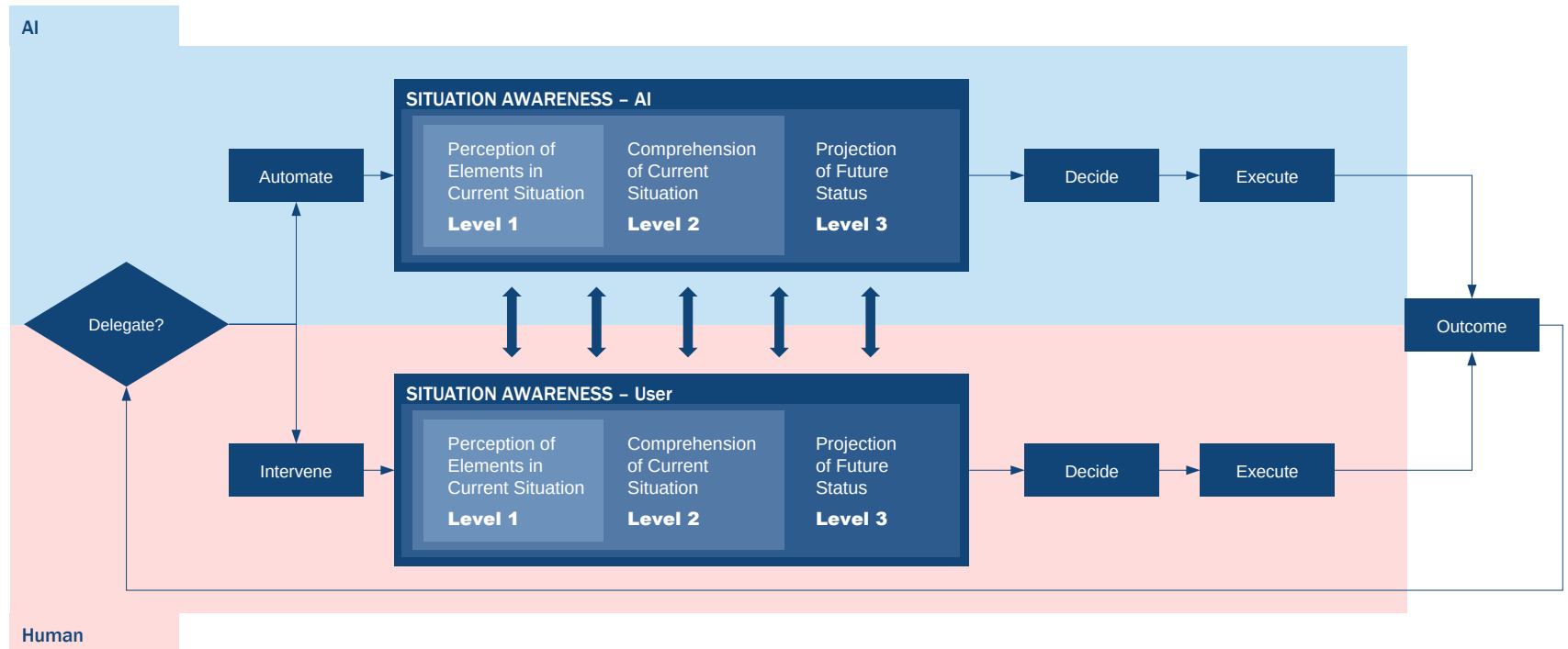


Figure 1.1: The process model of a user's decision to automate a complex decision-making task.

1.6 Discussion

In this study, we examined human-AI decision-making in complex organizational contexts through the lens of SA. Drawing on qualitative evidence from semi-structured interviews with professionals engaged in AI-supported decision-making, we proposed an extended process model that conceptualizes decision-making as an interactional process involving users' situation awareness, their interpretation of system awareness, and a continuous meta-decision-making regarding task delegation. In this section, we discuss the implications for research and practice on human-AI interaction in organizational decision-making.

1.6.1 Implications for Research

This study extends the model of SA in dynamic decision-making to AI-supported decision-making by moving beyond its focus on a single human decision-maker. In its original form, SA is conceptualized as an internal cognitive state achieved through perception, comprehension, and projection of environmental elements. Our findings suggest that, in AI-supported contexts, users reason not only about their own situation awareness but also about what they interpret the AI system to know, understand, and anticipate. We conceptualize this interpretive process as attributed system situation awareness, which is grounded in users' accounts of how they assess the system's inputs, representations, and outputs. This extension makes it possible to account for decision-making contexts in which intelligent systems participate in sensemaking without being treated as autonomous cognitive agents.

Another contribution of this study is the conceptualization of meta-decision-making as a continuous regulatory mechanism in human-AI decision-making. Rather than framing meta-decision-making solely as the selection of decision strategies or analytical techniques, our findings show that users engage in ongoing evaluations of how responsibility for specific sub-decision tasks should be distributed between themselves and the AI system. Through this regulatory process, users selectively delegate, reclaim,

or retain control over the perception, comprehension, prediction of the context, and the subsequent action-selection sub-task. Delegation is therefore not binary or static, but a dynamic outcome of a continuous evaluation process.

The findings further highlight the dynamic, interactive nature of situation awareness in AI-supported decision-making. Our analysis shows that users continuously revise their situation awareness not only through the perception of external cues, but also through their interaction with the AI system itself. Users update their understanding of the decision environment by engaging with system recommendations and explanations. At the same time, they intervene either by reshaping the system's awareness through modifying inputs or adjusting parameters, or by adjusting the system's involvement by overriding recommendations or suspending system functions when contextual elements are not adequately reflected. This bidirectional adjustment process suggests theorizing reliance on AI as an interactive, process-oriented aspect of AI-supported decision-making, in which awareness is continuously co-constructed through ongoing human-AI interaction.

The extended process model developed in this study highlights several future research opportunities on human-AI interactions. First, future research can build on the concept of attributed system SA by examining how users form, update, and revise their understanding of what an AI system knows, understands, and anticipates. Experimental studies could manipulate the visibility of system inputs, intermediate representations, or predictive outputs to assess how such cues shape users' assessments of system awareness. Through these designs, researchers could test how variations in perceived system SA influence delegation decisions and subsequent human-AI interaction behaviors.

Second, the notion of bidirectional adjustment highlights the potential in studying human-AI interaction as a coupled system. Experimental approaches could systematically vary users' ability to intervene in system operation, such as modifying inputs or overriding recommendations, to observe how these actions affect subsequent user behavior and interpretations of system awareness. Such experimental designs would make it possible to identify key AI configuration features that are beneficial in shaping decision outcomes in human-AI interactions.

Finally, the notion of meta-decision-making in human-AI interaction suggests that future studies could operationalize delegation as a sequence of decisions observed across evolving task conditions, rather than a single outcome. Experiments could vary contextual stability, task complexity, or the availability of intervention mechanisms to examine how users recalibrate delegation over time and how these recalibrations relate to changes in attributed system SA. Such designs would help clarify the factors influencing the decision between delegating to AI or intervention.

1.6.2 Implications for Practice

By conceptualizing decision-making as an interactional process involving attributed system situation awareness and continuous meta-decision-making, the model developed in this study suggests that effective human-AI interaction depends not only on improving algorithmic performance but also on supporting users' ability to perceive, comprehend, and predict system states and regulate delegation over time.

From a system design perspective, this implies that AI systems should be designed to accommodate selective and reversible delegation of decision-making sub-tasks, rather than assuming stable patterns of reliance. In practice, users require opportunities to assess how the system represents the decision environment at different stages, including what information it uses, how it models and interprets external reality, and how it generates predictions regarding future states. Interfaces that make these aspects visible and support users in evaluating whether the system's awareness aligns with situational conditions and task demands.

The model also highlights the importance of supporting user configuration of AI during decision-making. In organizational settings, users may encounter contextual factors that evolve over time or fall outside the system's initial scope. Systems that allow users to intervene, adjust, and configure AI-generated recommendations enable users to reshape the system's awareness when needed, rather than forcing either acceptance or complete rejection of system outputs. Allowing users to configure AI with respect to the

input data used, the selected model of reality, and the generation of predictions would support ongoing regulation of delegation and reduce the need for manual takeovers.

Finally, the continuous nature of meta-decision-making highlights the value of treating human-AI interaction as an ongoing process rather than a sequence of isolated decisions. In practice, this suggests that organizations may benefit from monitoring how users engage with AI systems over time, including patterns of intervention, reliance, and recalibration. Viewing these interactions as signals of alignment or misalignment between user and system awareness can inform when users need detailed explanations or when AI systems need to be updated or reconfigured. However, such monitoring should rely on optional functionality that gives users complete control over their personal data and privacy settings.

1.6.3 Limitations

This study has several limitations that define the scope within which its findings should be interpreted. First, the analysis is based on qualitative data from a single application domain, namely AI-supported decision-making in SCM. While this context provides a rich setting for examining various dynamic decision-making processes, the specific tasks, organizational structures, and forms of AI use may differ from those in other domains. Accordingly, the extended process model should be understood as empirically grounded in this context and therefore, future evaluations in other contexts may be needed for ensuring generalizability.

Second, the study relies on participants' retrospective accounts of their interactions with AI systems rather than direct observation of decision behavior. Although these accounts provide valuable insight into how users interpret system behavior and reason about delegation, future research could complement this perspective with observational or behavioral data to further examine the phenomenon identified in this study.

Finally, this study focuses on experienced users who regularly engage with AI-supported decision systems. The dynamics of attributed system situation awareness and

meta-decision-making may differ for novice users or those in early stages of system adoption. Accordingly, the findings may be less applicable to settings where users have limited familiarity with the system or where decision-making practices are still emerging.

1.7 Conclusion

This study contributes to research on human-AI interaction in organizational decision-making by extending the situation awareness theory to account for continuous interactions with intelligent systems. By identifying the concepts of attributed system situation awareness and meta-decision-making, the study conceptualizes decision-making with AI as an interactional process in which users actively regulate task delegation across stages of perception, comprehension, prediction, and action.

The proposed process model reframes delegation to AI as a dynamic process, highlighting how users continuously assess and recalibrate their interaction with AI systems in response to contextual conditions and system behavior. In doing so, the study advances a process-oriented perspective on human-AI interaction that complements existing cross-sectional views of AI use. Together, these contributions provide a theoretical lens for future research examining how situation awareness is constructed and managed across human and system boundaries during decision-making with AI.

References

- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.
- Brynjolfsson, E. (1993). The productivity paradox of information technology. *Communications of the ACM*, 36(12):66–77.
- Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M. (2014). Situation awareness-based agent transparency. Technical report, Defense Technical Information Center, Fort Belvoir, VA.
- Chiang, C.-W. and Yin, M. (2022). Exploring the effects of machine learning literacy interventions on laypeople’s reliance on machine learning models. In *27th International Conference on Intelligent User Interfaces*, pages 148–161, Helsinki, Finland. Association for Computing Machinery.
- Czako, Z., Sebestyen, G., and Hangan, A. (2021). Automaticai—A hybrid approach for automatic artificial intelligence algorithm selection and hyperparameter tuning. *Expert Systems with Applications*, 182:115225.
- Danilova, O. V. (2019). The place and role of decision making in the process of modern enterprise management. In G. Popkova, E., V. Chesnokova, A., and A. Morozova, I., editors, *Specifics of Decision Making in Modern Business Systems*, pages 57–67. Emerald Publishing Limited.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.

- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64.
- Endsley, M. R. (2023). Supporting Human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140:107574.
- Endsley, M. R., Bolté, B., and Jones, D. G. (2003). *Designing for Situation Awareness: an Approach to User-Centered Design*. CRC press.
- Goldkuhl, G. and Cronholm, S. (2010). Adding theoretical grounding to grounded theory: Toward multi-grounded theory. *International Journal of Qualitative Methods*, 9(2):187–205.
- Jiang, J., Karran, A., Coursaris, C. K., Léger, P. M., and Beringer, J. (2021). A situation awareness perspective on human-agent collaboration: Tensions and opportunities. <http://rgdoi.net/10.13140/RG.2.2.24937.62562>.
- Kottemann, J. E. (1986). Some requirements and design aspects for the next generation of decision support systems. In *Proceedings of the 19th Annual Hawaii International Conference on Systems Sciences*, pages 339–343.
- National Academies of Sciences, Engineering, and Medicine (2021). *Human-AI Teaming: State of the Art and Research Needs*. National Academies Press, Washington, D.C.
- Norman, D. A. (1988). *The Psychology of Everyday Things*. Basic books.
- Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., and Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research*, 42(5):533–544.
- Reynolds, T. J. and Gutman, J. (1988). Laddering theory, method, analysis, and interpretation. *Journal of Advertising Research*, 28(1):11–31.

- Sabherwal, R. and Jeyaraj, A. (2015). Information technology impacts on firm performance: an extension of Kohli and Devaraj (2003). *MIS Quarterly*, 39(4):809–836.
- Schuetz, S. and Venkatesh, V. (2020). The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction. *Journal of the Association for Information Systems*, 21(2):460–482.
- Seidman, I. (2006). *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*. Teachers College Press.
- Venkatesh, V. and Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences*, 39(2):273–315.
- Venkatesh, V. and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2):186–204.
- Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3):425.
- Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.
- Xiao, L., Guo, Z., and D’Ambra, J. (2017). Analyzing consumer goal structure in online group buying: a means-end chain approach. *Information & Management*, 54(8):1097–1119.

Chapter 2

Exploring the Impact of Generative AI Transparency and Explainability on Decision Performance

Abstract

As organizations continue to explore the integration of generative AI (GenAI) assistants into their enterprise decision support systems, understanding their impact on users' attentional processes and decision performance has become essential. Drawing on the Situation Awareness (SA) model and from eye-tracking methodology, the paper investigates how transparency and explainability in GenAI-generated outputs affect users' perceived understanding, confidence in the GenAI assistant, and information

This chapter includes an original and unpublished research work conducted by Burak Öz; with contributions from Dr. Alexander J. Karran, Dr. Constantinos K. Coursaris, and Dr. Pierre-Majorique Léger.

Burak Öz served as the principal author, responsible for the design, execution, analysis, and writing. Dr. Alexander J. Karran contributed to the experimental design and reviewed the manuscript. Dr. Constantinos K. Coursaris and Dr. Pierre-Majorique Léger provided feedback and guidance throughout all stages of the research.

processing efficiency during decision-making tasks. A within-subjects experiment using a business decision-making simulation embedded within a real-life ERP interface was conducted with 57 participants. During the decision-making task, participants interacted with GenAI assistants with various levels of transparency and explainability across multiple, sequential decision cycles. Results show that both features improve perceived understanding; however, they act as imperfect substitutes, with no additional benefit when both are present. Eye-tracking analyses reveal that transparency and explainability significantly affect users' attention shift rates, operationalized as saccadic eye movements between the decision interface and the GenAI interface. The attention shift rate declines across trials, with transparency leading to a faster reduction than explainability. The paper's findings offer design implications for building GenAI assistants and represent an initial step for attention-aware GenAI systems that adapt to users' understanding and attentional states during iterative interactions.

2.1 Introduction

Recent generative AI (GenAI) implementations have showcased the potential of such systems in novel human-computer interaction paradigms based on natural language (Feuerriegel et al., 2024). This transformative potential has quickly extended into the business world, where organizations are increasingly exploring GenAI to support complex decision-making processes. A key sign of this trend is the emergence of GenAI assistants: conversational systems specifically designed to aid decision-makers by answering questions and offering recommendations or insights derived from task-specific training datasets. Since late 2022, major software providers have begun piloting GenAI assistants in business applications across diverse domains (Budhwar et al., 2023; Roy et al., 2023; Suda and LeHong, 2025; Sellen and Horvitz, 2024), with notable early implementations including GitHub Copilot for code generation (Dohmke, 2023), Microsoft Office Copilot for spreadsheet analysis (Microsoft, 2024), and SAP Joule and Salesforce X-Gen for analytics and reporting (SAP, 2024; Yavuz et al., 2024). Industry

forecasts suggest rapid growth, with one survey reporting GenAI assistants as the most commonly piloted GenAI use case in organizations in 2023 (Gabriele Rigon et al., 2024), and projections indicating that by 2027, more than 50% of enterprise GenAI deployments may fall into this category (Chandrasekaran et al., 2024). Despite this high enthusiasm and interest, the actual performance benefits and optimal ways to leverage these tools are yet to be fully known (Simkute et al., 2024).

Although GenAI assistants share some similarities with other AI tools in business settings, their interaction model introduces new challenges. Unlike structured systems with defined input-output flows, these assistants create a layer of engagement that includes secondary tasks such as prompt engineering, output verification (hereafter referred to as verification behavior), follow-up prompting, and deciding whether further clarification is needed (Jiang et al., 2022; Prather et al., 2023; Simkute et al., 2024; Tankelevitch et al., 2024). These secondary tasks effectively represent an additional cognitive layer, as they involve managing an agentic entity through natural language, and this engagement with an additional cognitive layer occurs whether the GenAI assistant functions as a complementary tool or is more deeply integrated into the user's primary task interface. Consequently, users of GenAI assistants must shift their attention not only across interface elements but also between different cognitive processes. When combined with GenAI's potential to produce factual errors, inconsistencies, and hallucinations (Banh and Strobel, 2023; Feuerriegel et al., 2024; Söllner et al., 2025), this iterative nature involving cognitive task switches between GenAI-related secondary tasks and the primary task can negatively impact decision performance. To understand how these challenges can be mitigated, we turn to the literature on automation in dynamic settings.

In dynamic organizational settings, decision-making often involves monitoring evolving task elements, updating mental models, and anticipating future states. The Situation Awareness (SA) model (Endsley, 1995), which has been widely used to explain performance in complex automation environments that demand sustained attention and adaptation, offers a useful lens for examining this context. In this literature, transparency refers to a system's "abilities to afford an operator's comprehension about an intelligent

agent's intent, performance, future plans, and reasoning process" (Chen et al., 2014, p. 2). Explainability, by contrast, is "the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model" (Vilone and Longo, 2021, p. 91). Prior research suggests that when users understand how and why recommendations are produced, they are more likely to make informed judgments, detect system errors, and calibrate their confidence in the model appropriately (Bansal et al., 2021). These factors are presumed to be important not only for user confidence and understanding but also for optimizing decision performance when using GenAI assistants. However, in the context of GenAI assistants, increasing transparency and explainability introduces a design trade-off as both tend to lengthen the GenAI responses. Although these features aim to improve understandability, they also place greater demands on users' attention by requiring them to parse and verify longer text sequences. As GenAI assistants are intended for use by decision-makers in continuously evolving environments involving similar decisions across changing contexts, investigating the immediate and longer-term effects of transparency and explainability on perceived understanding and decision performance becomes critical for informing GenAI assistant design. Such an investigation would also respond to recent calls for research on design characteristics that enable successful conversational systems (Janssen et al., 2020), explanations and interpretability techniques that shape user experience (Bauer et al., 2021; Maedche et al., 2019), and system features that optimize human-AI interaction in organizational decision-making contexts (Söllner et al., 2025).

Users' ability to understand AI outputs is dependent on system design (Kruschel et al., 2025; Pfeuffer et al., 2023); but it is also constrained by user characteristics, which include their past experience with the tool, available cognitive resources and attentional capacities (Gnewuch et al., 2022). In environments where attention is limited, the format and volume of AI-generated content may unintentionally hinder understanding. As shown in the case of dashboards that overwhelm users with visual complexity (Schulz and Knierim, 2024), lengthy textual outputs from GenAI assistants may lead to distraction, disengagement, or overlooked errors. Maintaining SA under these conditions requires users to continuously

balance the cognitive demands of interpreting GenAI outputs with monitoring the broader task context. However, capturing how attention and understanding evolve in real-time is methodologically challenging. Self-reports, though widely used, provide limited insight into moment-to-moment shifts in attention and are poorly suited for studies involving repeated, continuous interaction.

To address this challenge, we employ eye-tracking as a non-intrusive, real-time method for measuring users' information processing behavior. Eye-tracking data can reveal immediate shifts in gaze and attentional focus that occur as users interact with GenAI assistants, offering insight into how attention supports understanding and decision-making. By combining eye-tracking with self-reports and behavioral data, this study examines how transparency and explainability influence users' understanding and attention allocation over time. Based on these goals, we ask the following research questions:

- *RQ1. How do transparency and explainability of GenAI assistants influence users' perceived understanding and confidence in the GenAI assistant?*
- *RQ2. How do transparency and explainability influence users' information processing efficiency while interacting with GenAI assistants?*

To address these questions, a controlled multi-trial experiment was conducted in which participants repeatedly engaged with a business decision-making simulation scenario. This scenario involved a sequential multiple decision-making process, where each decision was a prediction and directly influenced subsequent outcomes. Participants observed the results of their decisions after each trial, advancing through a simulated timeline. The experimental task was designed to simulate a realistic, repeated decision-making environment within a real-world enterprise system. This setup mimicked the interconnected and time-dependent nature of decisions in organizational settings, allowing participants to observe the outcomes of their prior actions and adjust their subsequent behavior in a continuous flow. Given the persistent challenge of

achieving ecological validity in eye-tracking research (Balapour and Riedl, 2025), this business decision-making simulation environment provides a task framework that enabled continued participant engagement across dynamic decision cycles.

This study offers four key contributions. First, it defines and operationalizes transparency and explainability within the context of GenAI assistants embedded in enterprise decision support systems. Second, it demonstrates a methodological approach for evaluating human-AI interaction by embedding a GenAI assistant into a real-life ERP system to create an ecologically valid research setting using publicly available tools. Third, it empirically demonstrates how system design influences users' perceived understanding, information processing efficiency, and decision performance in a repeated-use setting that simulates the temporal dynamics of organizational decision-making. Finally, by identifying eye-tracking metrics that correlate with users' perceived understanding and confidence in the GenAI assistant, this study establishes a foundation for the development of GenAI systems that can dynamically adapt their output complexity based on real-time indicators of users' attentional states and cognitive needs.

2.2 Background

2.2.1 Artificial Intelligence and Organizational Decision-Making

Organizational decision-making has long relied on information systems (IS) to support planning, coordination, and operational efficiency. Early IS solutions focused on automating transactional functions such as order processing and record-keeping (Stevens, 1989). Over time, these systems evolved into integrated platforms such as Enterprise Resource Planning (ERP) systems, which synchronize planning, procurement, and resource allocation across functional units and organizational boundaries (Klaus et al., 2000; Lefebvre et al., 2003; Monden, 2011).

More recently, advances in data availability and computational power have enabled the application of artificial intelligence (AI) to support complex organizational decisions.

AI systems now assist with tasks such as forecasting, optimization, anomaly detection, and operational risk assessment (Agrawal et al., 2024; Sebastian and Draper, 2024; Sharma et al., 2022). These applications have demonstrated benefits in cost reduction, performance improvement, and increased transparency. For example, organizations use AI to optimize workflows (Cain, 2022), detect external disruptions (Wong et al., 2022), and anticipate resource needs using contextual information (Mediavilla et al., 2022; Nair and Saenz, 2024).

However, recent events have revealed critical limitations of AI in high-stakes decision contexts. Disruptions such as the global chip shortage and supply constraints in various industries have shown that AI models often underperform in the face of non-routine, ambiguous, or novel situations (Pettit et al., 2019). These failures emphasize the need for human flexibility and contextual reasoning in such contexts, and they point to a gap in current human-AI interaction models.

AI adoption has also reshaped the role of decision-makers. Instead of producing decisions autonomously, professionals are increasingly tasked with evaluating and refining AI-generated outputs (Faraj et al., 2018; Sharma et al., 2022; Toorajipour et al., 2021). This shift imposes new cognitive demands and highlights the importance of AI literacy, data fluency, and critical thinking as key organizational competencies (Di Battista et al., 2023). The integration of GenAI assistants further increases the need to understand how humans interact with AI in decision-making roles (Shareef and Picek, 2024).

Taken together, these developments highlight the importance of examining the design and use of GenAI systems from a human-centered perspective in organizational decision-making contexts and examine how they impact users' cognitive load, information processing efficiency, and decision performance.

2.2.2 Human-AI Interaction and the Situation Awareness Model

The increasing integration of AI into complex decision-making environments has led to growing interest in understanding how human and AI agents can collaborate effectively.

In organizational settings, this shift is not merely about automating routine tasks but about enabling decision-makers to interact meaningfully with intelligent systems that generate insights, forecasts, and recommendations (Faraj et al., 2018). As users are expected to assess the validity of AI outputs and determine when and how to act on them, fostering effective human-AI interactions becomes a critical design and research challenge (Agrawal et al., 2024). Research findings regarding the chatbots have made researchers realize that intelligent, conversational system designs are not only a technical challenge but also a design of an interaction process that is closer to a human-human communication than previous human-computer interaction paradigms (Gnewuch et al., 2022; Pfeuffer et al., 2019).

So far, human-AI interaction studies reveal conflicting evidence on how AI-powered decision support systems influence decision performance and how they are perceived by their users (Bansal et al., 2021; Lai and Tan, 2019; Rolon-Merette et al., 2025). While some highlight underreliance by expert decision-makers (Allen and Choudhury, 2022; Dietvorst et al., 2015; Kesavan and Kushwaha, 2020), others point to an overreliance (Anaraky et al., 2020; Merritt et al., 2019). These conflicting findings highlight the complexity of human-AI interactions and suggest a need for a better understanding of the factors influencing user reliance on AI.

A promising theoretical lens for examining these interactions is the Situation Awareness (SA) model, originally developed to explain human performance in high-stakes, dynamic environments such as aviation and military operations (Endsley, 1995). With its clear structure and domain-general applicability, the SA model has been useful in explaining human-automation interactions in various complex systems, including automation and AI (Endsley et al., 2003; Endsley, 2021). Its foundations in the human-automation interaction field make it a relevant framework for examining human-AI interactions in modern decision support settings (Endsley, 2023). SA is defined as “the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” (Endsley, 1995, p. 1). SA is characterized by three levels: (1) perception of relevant

contextual elements, (2) comprehension of the current situation based on those elements, and (3) projection of future events based on that understanding.

Although initially developed to explain pilots' interactions with flight automation systems, the SA model's relevance has expanded to socio-technical systems involving AI (Jiang et al., 2023; Endsley, 2023). In such systems, users must make sense of not only the decision context but also how and why the AI system behaves in a particular way. To reach Level 3 SA in these systems, users must first perceive and comprehend both the decision context and the AI's functioning, and then use that understanding to form expectations about the system's likely future actions. This view provides a comprehensive framework for identifying the types of information users may seek from AI-powered decision support systems to better understand the system.

SA shares conceptual alignment with broader theories of human information processing that describe how individuals perceive stimuli, allocate attention, and encode information into memory for reasoning and decision-making (Wickens et al., 2021; Card et al., 1986). Similar to SA's levels, these information processing models distinguish between pre-attentive perception, goal-directed attention, working memory constraints, and long-term knowledge retrieval (Healey and Enns, 2011). In both perspectives, attention is treated as a limited resource, and effective decision-making depends on how well users can filter and process relevant cues. The information processing model has been used in past eye-tracking research to study how visual design influences users' decision-making process (Toreini et al., 2022). Combined together, these cognitive models can help explain how GenAI interaction design might influence SA and reduce decision performance.

We now focus on two key system characteristics that have the potential to shape users' ability to understand AI: transparency and explainability. In accordance with the SA model and its three levels, we adopt the definition of transparency from the automation literature using the SA model as a theoretical lens. Accordingly, transparency refers to the system's "abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process" (Chen et al., 2014, p. 2). It

primarily supports Level 1 and Level 2 SA by helping users perceive what the system is doing and comprehend why it operates in a particular way. For example, a transparent GenAI assistant might disclose its data sources, modeling approach, or past prediction performance, all of which are details helping users to form a coherent mental model of the system. Transparency can also support Level 3 SA when it reveals stable reasoning patterns or predictive logic that helps users anticipate the system’s behavior under new conditions.

Explainability, by contrast, is defined as “the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model” (Vilone and Longo, 2021, p. 91). This definition, proposed following a comparative synthesis of the term’s usage across disciplines, provides a clear and generalizable conceptualization that helps distinguish explainability from transparency. According to this view, explainability focuses on individual outputs, contributing to Level 2 SA by helping users grasp why a particular recommendation was made and supporting Level 3 SA by identifying which input factors might influence future outputs. In highly transparent systems, explainability can be implemented by highlighting factors that led to a given recommendation (e.g., feature coefficients in a regression model). In less transparent systems based on complex models such as large language models or neural networks, explanation is often approximated using local, post-hoc interpretability techniques. Two widely used methods are LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) or SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017). These tools simulate local model behavior to generate explanations of how the assistants’ recommendations might change in response to variations in key inputs.

Both of these mechanisms aim to improve interpretability, a feature demanded by decision-makers using AI-powered decision support systems (Diederich et al., 2020; Landwehr et al., 2022). While these mechanisms are designed to improve users’ understanding of GenAI behavior, they also introduce a design trade-off. Providing both transparency and explainability information would result in longer, more complex GenAI responses. Although the additional details in GenAI responses may enhance user

confidence and decision performance, continued exposure to lengthy GenAI responses may also increase the attention demanded by the GenAI assistant. Attentional demands are already higher than human capacity in today's complex and time-constrained decision-making environments. Therefore, GenAI assistants create a tension between the increased text volume and the cognitive costs associated with processing more detailed outputs.

In decision-making contexts involving GenAI assistants, users must navigate both task-relevant information and system-generated outputs. Visual attention is central to developing SA, particularly in terms of perceiving system outputs (Level 1 SA) and forming a coherent mental model (Level 2 SA). As the workflow imposed by GenAI assistants force users to engage in a multitasking situation where they have two information systems to interact, i.e., the actual decision support system and the GenAI assistant, understanding how the design of a GenAI assistant impacts users' attention allocation becomes critical for supporting their situational awareness and decision performance.

2.2.3 Human-AI Interaction and Eye-Tracking Research

As indicated by the SA model, gathering and combining information from sensory systems is a critical first step in developing an understanding of both the environment and the AI-based decision support system in human-AI interaction scenarios. In information-rich digital decision-making contexts, examining user attention becomes important for understanding how information is gathered and processed by decision-makers. Attention enables individuals to select relevant information from a set of stimuli, allowing limited visual evaluation capacity to be used efficiently by filtering out irrelevant data and focusing on what is deemed important. The study of visual attention has expanded considerably over the past three decades, partly due to the increased availability of non-intrusive eye-tracking technologies (Carrasco, 2011). In human-computer interaction (HCI) research, eye tracking has become a widely used method for examining attentional dynamics in human-AI interactions (Vasseur et al., 2023).

The notion of selective attention stems from limitations in the neural capacity available to process visual information. The selective attention concept suggests that stimuli compete for an organism's limited resources. This view of competing stimuli has been supported by evidence from neuroimaging and behavioral studies (Beck and Kastner, 2009; Reynolds and Chelazzi, 2004). In complex decision-making scenarios, selective visual attention helps narrow cognitive resources to focal task elements, enhancing the perceptual representation of attended information relative to unattended information. Since gaze behaviors reflect users' selective attention strategies, eye-tracking enables real-time, non-intrusive observation of individuals' attentional patterns, providing an understanding of their cognitive focus and intentions (Riedl and Léger, 2016; Vasseur et al., 2023). Therefore, it is a promising and commonly used tool for examining how humans allocate attention in complex tasks.

In line with the selective visual attention concept, prior research has identified several eye-tracking measures associated with decision-making processes, task difficulty, and uncertainty (Brunyé and Gardony, 2017; Vasseur et al., 2023). These measures include fixations, which are defined as momentary pauses of eye movements targeted at a specific location for a minimum duration (Riedl and Léger, 2016). In tasks where humans need to search a visual environment for a stimulus, increasing task difficulty has been found to decrease fixation frequency and increase fixation duration (Hooge and Erkelens, 1996; Jacobs and O'Regan, 1987; Rayner, 1998). Similarly, saccades, rapid eye movements between fixations, have been shown to reflect decision-making strategies (Brunyé and Gardony, 2017). For example, the number of saccades between areas of interest (AOIs) tends to decrease when decision-making is supported or simplified (Toreini et al., 2022). Saccade amplitude also tends to decrease under high task difficulty (Phillips and Edelman, 2008), and saccade paths become less random as users gain task familiarity and understanding (Shiferaw et al., 2019).

These eye-tracking indicators are especially relevant for studying the attentional dynamics when users interact with GenAI assistants integrated into decision support environments. As stated in the introduction, engaging with these assistants introduces

an additional cognitive layer that is distinct from primary task execution and requires users to interpret, evaluate, and potentially act upon AI-generated recommendations. This layer of AI interaction necessitates cognitive effort and attentional shifts, regardless of whether the GenAI functionalities are presented as a visibly separate interface or are more deeply embedded within the primary task environment. In scenarios where information sources for the primary task and the GenAI output are spatially distinct, such as data from a database view or information accessed via other traditional interface elements, active search and verification behaviors become particularly observable, mainly through attention allocation as fixations, and attention shifts as saccadic transitions. These gaze behaviors, when the GenAI output and other information sources are defined as distinct AOIs, can be used to infer users' information processing as they manage their interaction with the AI.

However, to the best of our knowledge, research exploring the usefulness of eye tracking metrics in evaluating users' understanding of an AI model is limited, although several studies have noted this as a future research opportunity (Cao and Huang, 2022; Gupta et al., 2019; Nagendran et al., 2024).

Accordingly, this study draws on insights from the perceptual decision-making literature to investigate how GenAI assistant design—specifically transparency and explainability—affects attention allocation. We conceptualize saccadic transitions between the GenAI assistant and the decision interface as a behavioral indicator of users' understanding of the GenAI assistant and its outputs. A key objective is to identify eye-tracking metrics that correlate with perceived understanding, confidence in the GenAI assistant, and decision performance, thus informing the design of adaptive GenAI systems. Such systems could tailor content presentation based on real-time indicators of user understanding.

In summary, this study positions attention allocation and attention shift rates as critical variables concerning information processing in human-AI interaction and as a foundational step toward adaptive, attention-aware decision support systems.

2.3 Hypothesis Development

This section develops hypotheses to explain how transparency and explainability in GenAI assistant design influence users' attention allocation processes and decision outcomes. Guided by the Situation Awareness (SA) model and drawing on the methodological approaches from the eye-tracking research, the hypotheses address three interrelated research questions. First, the effects of transparency and explainability on users' perceptions are examined. Second, behavioral outcomes are explored through eye-tracking metrics that reflect attention allocation and shifts during interaction. Third, the temporal evolution of attention shifts is examined to capture the dynamic nature of human-AI interaction in repeated-use scenarios.

2.3.1 Perceived Understanding and Confidence in the GenAI Assistant

We begin by investigating the question *How do transparency and explainability of GenAI assistants influence users' perceived understanding and confidence in the GenAI assistant?* Transparency and explainability are theorized to support users in forming accurate mental models of AI systems, which are reflected as perceived understanding and confidence in the GenAI assistant. These concepts are expected to mediate the relationship between system design and decision performance.

Transparency is defined as the system's "abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process" (Chen et al., 2014, p.2), and is typically embedded in the system architecture and design documentation (e.g., model cards; Mitchell et al., 2019). Explainability, by contrast, is defined as "the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model" (Vilone and Longo, 2021, p.91). Unlike transparency, explainability pertains to specific outputs and is often operationalized through feature attribution techniques such as SHAP (Lundberg and Lee, 2017) or

LIME (Ribeiro et al., 2016). However, the literature frequently treats these concepts interchangeably or applies them inconsistently, often without clear differentiation. To maintain conceptual clarity while presenting empirical findings from this literature, we use the term explanations as an umbrella label for prior work. Within this framing, when referring to empirical findings, we interpret global explanations as analogous to transparency and local explanations as representative of explainability.

Perceived Understanding

According to Situation Awareness (SA) theory, a user’s understanding of system status and logic emerges from system design features that support the formation of Level 1 SA (perception of relevant information) and Level 2 SA (comprehension of the current situation) (Endsley, 1995). According to the SA theory, users’ perception of task-relevant elements (Level 1 SA) is a necessary condition for the Level 2 SA to be built. Features such as information displays, system cues, and interface feedback serve to clarify source data, agent rationale, and system behavior (Endsley et al., 2003). Within this framework, transparency and explainability serve as core system design attributes aimed at helping users build their Level 2 SA by making the AI’s goals, logic, and decision mechanisms more accessible to them (Coursaris, C., 2025).

We define perceived understanding as “the degree to which users feel that they sufficiently understand the AI system or process being explained to them” (Hoffman et al., 2023). While SA theory emphasizes actual comprehension of the situation, perceived understanding reflects a subjective belief that such comprehension has been achieved. Since the GenAI assistant presents outputs that describe a wide portion of the task environment, we treat perceived understanding of the assistant as closely aligned with users’ comprehension of the decision environment as a whole.

Past empirical studies have demonstrated that explanation modalities and design choices influence the perceived understanding of information systems used as decision support tools. Differences in explanation types have been associated with varying levels of user-reported understanding (Kaur et al., 2020). In conversational decision

support systems, improved explanation features have been found to reduce the number of follow-up questions, implying a higher level of understanding (Diederich et al., 2020). However, the effectiveness of these influences depends on how explanations are designed and presented. Evidence from a multi-study investigation shows that the contextual relevance and individual fit of explanations further moderate their impact, suggesting that explanation utility is not uniform across designs, tasks, or users (Stumpf et al., 2009).

H1a: Transparency increases users' perceived understanding of the GenAI assistant.

H1b: Explainability increases users' perceived understanding of the GenAI assistant.

Confidence in the GenAI Assistant

Confidence in a GenAI assistant refers to the “trustor’s perception of the trustee’s (in this case an AI) ability to perform a given task” (Chong et al., 2022, p. 2). Unlike trust, which is commonly conceptualized as a multidimensional construct incorporating elements such as ability, benevolence, and integrity (e.g., Barki et al. 2015), confidence is treated here as users’ evaluation of the AI system’s task-related capability without encompassing the broader relational or affective components associated with trust. In human-AI interaction, appropriate calibration of confidence to the system’s actual capabilities is essential (Lee and See, 2004). Excessively high confidence may lead to uncritical acceptance of flawed outputs, whereas insufficient confidence may result in the rejection of accurate recommendations. Consequently, confidence contributes positively to decision performance only when it aligns with system performance.

In the context of GenAI assistants, confidence emerges not only from outcome accuracy but also from system design features that contribute to the formation of situation awareness and mental models (Chen et al., 2014; Endsley et al., 2003). When a GenAI assistant is designed to reveal how it functions across different contexts (transparency), users are more likely to form coherent expectations about its behavior, reducing perceived randomness and fostering a sense of control. In parallel, when the system explains why it produced a particular output (explainability), users can assess its logic and determine whether the recommendation aligns with their expectations. These complementary

mechanisms both contribute to confidence formation but operate at different levels. More specifically, transparency fosters confidence in the GenAI assistant by clarifying its objectives and behavior in various contexts, thereby reducing perceived randomness. On the other hand, explainability would enhance confidence by providing a post-hoc rationale for specific predictions.

Empirical studies have shown that the presence, relevance, and clarity of explanations influence users' confidence in AI systems. For instance, in decision-support interfaces for data scientists, localized visual explanations (e.g., chart overlays) have been shown to increase users' confidence in system outputs (Kaur et al., 2020). In recommender systems, users' perceptions of explanation quality directly shaped their trust and acceptance of suggestions (Kunkel et al., 2019). Similarly, Stumpf et al. (2009) observed that users who received contextually aligned justifications reported higher confidence and required fewer follow-up interactions. These findings suggest that system explanations, whether global or local, play a central role in enabling users to develop their confidence in decision support systems.

Building on the past evidence and recognizing the need to disentangle how distinct explanation characteristics influence user perceptions, we focus on transparency and explainability as two specific system features. We hypothesize that each will independently enhance users' perceived understanding of and confidence in the GenAI assistant.

H2a: Transparency increases users' confidence in the GenAI assistant.

H2b: Explainability increases users' confidence in the GenAI assistant.

Perceived Understanding, Confidence in the GenAI Assistant, and Decision Performance

According to the SA model, effective decision-making in dynamic contexts requires forming accurate mental representations of the task environment and anticipating system behavior based on those models (Endsley, 1995). Transparency and explainability support this process by clarifying how the AI functions and how its outputs relate to task-relevant

input elements. When users better understand the system, they are more likely to integrate AI-generated recommendations appropriately into their decision-making (Lee and See, 2004).

Empirical research has shown that explanation modalities influence not only perceived understanding and confidence, but they also influence objective task performance. For example, different explanation formats such as SHAP and generalized additive models result in varying levels of confidence, which are further associated with differences in answer accuracy (Kaur et al., 2020). Other studies demonstrate that explanation quality affects not only understanding but also task efficiency, suggesting a connection between understanding and performance (Stumpf et al., 2009).

Explanations can also help users calibrate reliance on AI. Without them, users may default to their own judgment, even when the system is more accurate, potentially leading to suboptimal performance (Bussone et al., 2015). While explanations are often proposed as a means to promote appropriate reliance, findings also indicate that certain explanation types can lead people to rely too much on the AI's recommendation, even when the AI recommendations are incorrect (Bansal et al., 2021; Poursabzi-Sangdeh et al., 2021).

Together, these findings suggest that the influence of system explanations on decision performance, whether in the form of transparency or explainability, depends on whether these design characteristics successfully improve users' understanding of the system. Accordingly, the following hypotheses are proposed:

H3: Users' perceived understanding of the GenAI assistant leads to an increase in their decision performance.

H4: Users' confidence in the GenAI assistant leads to an increase in their decision performance.

2.3.2 Information Processing Efficiency

In this section, we investigate the second research question by drawing on empirical findings in the eye-tracking literature: *How do transparency and explainability influence*

users' information processing efficiency while interacting with GenAI assistants?

Confidence in an information source reduces the cognitive and physiological effort required during interaction (Lukyanenko et al., 2022). In the context of GenAI-supported decision-making, this reduction is expected to lower the need for sustained monitoring or cross-validation of GenAI assistant outputs. To capture such behaviors, we apply concepts from the scanpath analysis literature (Kosel et al., 2021), including the definition of areas of interest (AOIs) and the analysis of fixation transitions between them.

Attentional Resource Allocation

In this study, we define the GenAI assistant interface and the decision interface as AOIs, and we use the gaze behavior to understand attentional resource allocation patterns. Specifically, this section focuses on understanding how transparency and explainability influence users' distribution of attentional resources across the decision interface.

Both transparency and explainability are intended to make a GenAI assistant's behavior more interpretable and predictable. As users form a clearer understanding of the assistant's reasoning, they are expected to shift their attention away from manual input validation and allocate fewer cognitive resources to examining decision-relevant information. This shift reflects improved cognitive efficiency, as users increasingly delegate processing to the assistant and rely on its outputs.

Empirical findings on reading and problem-solving tasks support this expectation. Skilled readers and experienced problem-solvers allocate attention more selectively, exhibiting fewer and shorter engagements with task elements due to increased fluency and confidence (Dogusoy-Taylan and Cagiltay, 2014; Engbert et al., 2005; Rayner et al., 2011). Conversely, increased attentional allocation has been associated with task complexity and reduced familiarity (Jacob and Karn, 2003). In the context of GenAI assistant interactions, we hypothesize transparency and explainability to lower attentional demands by enhancing the user's understanding of the AI's internal logic.

H5a: Transparency reduces attention allocated on the decision interface.

H5b: Explainability reduces attention allocated on the decision interface.

Attention Shift Rate

We now focus on attention shifts to investigate how transparency and explainability influence users' information processing efficiency. In a GenAI assistant use context with AI-generated information and raw data on different AOIs, the transitions between these AOIs can serve as indicators of users' attempts to evaluate and confirm AI-generated recommendations. In the eye-tracking literature, similar behavioral patterns have been described as saccades, revisits or cross-checking behavior (Burch et al., 2011).

Empirical evidence shows that attention shifts reflect domain knowledge, understanding, and task confidence. In clinical decision-making, expert physicians shift attention between AOIs less frequently and exhibit more structured scanpaths than novices when interpreting X-ray scans or video recordings (Balslev et al., 2012; O'Neill et al., 2011; Tiersma et al., 2003). Similarly, in organizational settings, visually complex dashboards and difficult-to-interpret charts have been shown to elicit more frequent attention shifts, indicating lower understanding or confidence in the task environment (Burch et al., 2011; Toreini et al., 2022). Together, these findings demonstrate that visual scan behavior is sensitive to both user cognition and system design.

In our context, attention shift rates are expected to decline as users form a clearer mental model of the GenAI assistant's reasoning process. Transparency and explainability are expected to support this process by clarifying the assistant's logic and the relevance of its outputs.

H6a: Transparency decreases the rate of attention shifts between the GenAI assistant and the decision interface.

H6b: Explainability decreases the rate of attention shifts between the GenAI assistant and the decision interface.

Information Processing and Decision Performance

According to the SA model, the transition from comprehension to projection involves not only understanding the task environment but also directing attention to the most relevant

information (Endsley, 1995). As users internalize the GenAI assistant’s reasoning, they no longer need to revisit prior data or distribute attention across all task elements. Efficient information processing thus emerges as a behavioral manifestation of internalized system knowledge. In this framework, reduced attention allocation to the decision interface and fewer attention shifts are signs of effective mental model formation and reliance on the GenAI assistant’s support.

Eye-tracking studies confirm that such processing efficiency predicts improved decision performance. In clinical tasks, expert decision-makers exhibit fewer and more targeted gaze shifts, correlating with better diagnostic accuracy (O’Neill et al., 2011; Balslev et al., 2012). Similarly, in problem-solving contexts, reduced attentional dispersion has been associated with higher task performance (Dogusoy-Taylan and Cagiltay, 2014). These findings suggest that when individuals process information efficiently by concentrating their attention appropriately and minimizing unnecessary verification, they achieve higher performance outcomes.

In the context of GenAI-supported decisions, we argue that the benefits of transparency and explainability on decision performance are realized to the extent that these design features facilitate efficient information processing. By enabling users to construct reusable mental models and selectively allocate their attention, transparency and explainability promote appropriate reliance on GenAI assistant’s recommendations, leading to more accurate decisions.

H7: Attention allocated to the decision interface is associated with a lower decision performance.

H8: Attention shift rate is associated with a lower decision performance.

The hypotheses presented in this section are summarized in Figure 2.1.

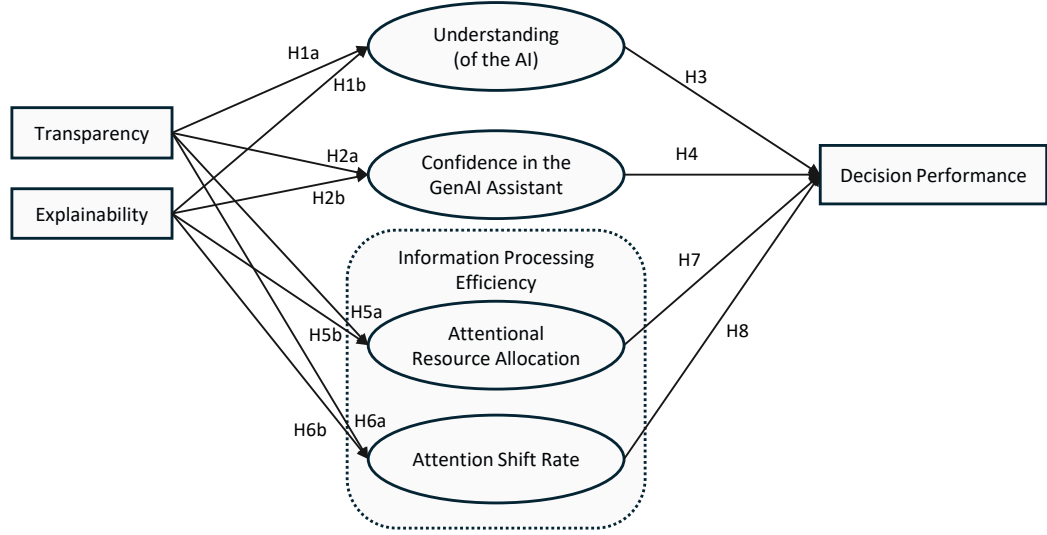


Figure 2.1: Research model summarizing the hypothesized effects of transparency and explainability on user perceptions, information processing efficiency, and decision performance.

2.4 Methodology

2.4.1 Experimental Design

To test the hypotheses, a within-subjects, 2x2 experimental design was used, where participants interacted with four GenAI assistant conditions, varying in the level of transparency (low or high), and in the level of explainability. To maintain a feasible session duration, each participant was exposed to two randomly assigned, counterbalanced conditions out of four conditions (Seltman, 2012). To eliminate within-condition carry-over effects on participants' perceptions of GenAI assistants, each GenAI assistant was implemented as a distinct system prototype with a separate backend, despite using the same user interface.

The experimental task was designed around the "newsvendor paradigm," a commonly-used decision-making scenario in operations management (Benzion et al., 2008; Kremer and Minner, 2008; Kremer et al., 2010; Wang et al., 2025). In this paradigm, a decision-maker determines the number of units, q , of a perishable good to be purchased at cost

c , selling at price p per unit and receiving a salvage price s for unsold inventory (where $s < p$). With uncertain demand d , the optimal order quantity under risk neutrality is given by the critical fractile solution $q = \Phi^{-1}((p - c)/p)^*$ when the demand distribution $\Phi(d)$ is known. However, in real-world settings, demand is typically unknown, making this problem an established framework for studying behavioral decision-making (Bolton and Katok, 2008; Schweitzer and Cachon, 2000; Wang et al., 2025). In this study, participants always engaged with a low-profit scenario, where the critical ratio was below 0.5 ($(p - c)/p < 1/2$) (Schweitzer and Cachon, 2000), with fixed cost, price, and salvage values across all trials at $c = 91.8$, $p = 126.25$, and $s = 50.5$, respectively. Excess inventory did not carry over the next weeks, and the demand distribution was not disclosed to participants. Instead, they were provided with a time series chart displaying demand observations from the last 20 periods. For consistency, all participants faced the same demand values at identical points in the experiment.

2.4.2 Experimental Setup

Simulated Decision-Making Environment

The experiment utilized ERPsim to instantiate the newsvendor paradigm as an ice-cream demand forecasting task. ERPsim is a business simulation platform, which simulates a competitive business environment where participants manage virtual companies (Léger, 2006; Léger et al., 2007; Oz and Léger, 2024). It is widely used in learning environments due to its realistic and responsive market dynamics and interactive decision-making processes. Consistent with prior research utilizing ERPsim, this study aimed to improve participant engagement and ecological validity by leveraging ERPsim's data generation capabilities for experimental task design (Loos et al., 2010).

ERPsim was used to generate sales data for dairy products over 100 simulated periods to create a realistic demand scenario. By default, ERPsim scenarios do not incorporate seasonality or trends, meaning that product demand values typically follow a stationary distribution. To better reflect real-world market conditions, a data transformation was

applied to introduce seasonality and trend effects, aligning with the experimental context where each period represented a weekly decision cycle. The seasonality pattern was determined after a detailed analysis of the ice cream production data (National Bureau of Economic Research, 2012), and a North American distributor's blog post on the seasonality of ice cream demand (FoodyDirect, 2014), while a slightly positive trend factor was introduced to reflect gradual market growth. Additionally, sporadic demand spikes were incorporated to simulate demand fluctuations observed in real-world settings.

In addition to the trend and seasonality factors embedded within the demand scenario, the dataset incorporated a controlled uncertainty component to reflect real-world decision-making environments. In this design, uncertainty refers to short-term variability in weekly demand relative to the immediately preceding period, representing forecasting uncertainty rather than structural market shocks. Demand uncertainty in the final simulation dataset was measured as percentage changes between consecutive simulated periods and was intentionally limited to ensure that the introduced trend and seasonality patterns remained perceptible to participants and could meaningfully inform forecasting decisions. Across the generated dataset, week-to-week demand changes exhibited a mean absolute variation of 3.67% (SD = 2.94%), with maximum observed changes below 10%. This moderate level of variability introduced realistic forecasting difficulty while preserving the interpretability of underlying temporal patterns within the decision-making task.

In summary, participants faced with a dataset that mirrored key real-world demand characteristics for this specific decision-making task, incorporating seasonality, trend, and variability while preserving experimental control. Participants were explicitly instructed to consider the season changes and these demand patterns when making their forecasts. Task instructions and supporting materials are available in Appendix, Section B.1.4.

Experimental Stimuli

During the experiment, participants interacted with two software tools: a custom-developed interface for SAP, an enterprise resource planning (ERP) system integrated

with ERPsim, and a GenAI assistant powered by OpenAI's Assistants Application Programming Interface (API). These tools were designed to ensure a controlled decision-making environment with a systematic variation in the GenAI assistant transparency and explainability.

The software solution for this experiment had the following requirements:

- The artifact needs to simulate the newsvendor paradigm within an ecologically valid IS environment, including both a decision interface and a GenAI assistant.
- The decision interface should provide all necessary information for decision-making, allowing task completion without a GenAI assistant.
- The GenAI assistant module of the software solution should be capable of providing reasonable recommendations to the planning decision, and it should be fully integrated with the simulated decision-making environment.
- The GenAI assistants with high transparency should provide the user with the source data information, past performance, and the reasoning process of the GenAI assistant while generating recommendations, while the ones with low transparency should provide only the past performance information.
- The GenAI assistants high explainability should complement each recommendation with a discussion of the factors influencing each recommendation.
- The GenAI assistants should provide unique responses for each participant while maintaining a consistent structure across and within experimental conditions.

The decision interface displayed key elements such as pricing information, a time-series chart of past demand (covering 20 virtual weeks), and an input field for entering forecasted order quantities for the upcoming week. Participants were restricted to modifying only the forecasted quantity, and the input field was designed to accept only numerical values to prevent errors. The decision-making process was structured into distinct steps: once a forecast was submitted, participants first completed a self-reported

confidence assessment, followed by a performance feedback report. The feedback report was designed to foster participant motivation and encourage continued effort throughout the experiment. Prior research indicates that social comparison and performance feedback can stimulate competitive engagement and goal-directed behavior (Garcia et al., 2013). In line with gamification research (Hamari et al., 2014), the report presented a comparison between participants' performance and that of an average decision maker and displayed the monetary value gained or lost in the previous decision period, together with summary performance indicators. This design aimed to sustain engagement and promote attentive decision-making through perceived competition and ongoing performance evaluation.

The GenAI assistant was developed as a chat-based decision support tool that is fully integrated into the decision-making scenario implemented using the ERPsim environment. The assistant's design involved a multi-step development process, including front-end and back-end logic programming, prompt engineering, and structured integration with the decision-making interface. The assistant was implemented using Python 3.9.2 and Streamlit 1.33.0, with data exchange facilitated through OData services published on the SAP interface. Before forwarding user queries to the OpenAI Assistants API (gpt-4o-2024-05-13), the system retrieved simulation state information from SAP, embedding relevant data into the user prompt. This ensured that AI-generated recommendations were contextually appropriate and aligned with the experimental conditions.

To maintain experimental control, four distinct GenAI assistants were configured using four different system prompts and slight changes in the visibility of front-end elements. These changes aimed to implement a different combination of transparency and explainability for each version. The design process began with the most detailed assistant (high transparency, explainability enabled), which was validated for response consistency and correctness. The remaining assistants were created by systematically modifying system prompts to reduce transparency and/or explainability while maintaining a consistent structure across conditions. The OpenAI API was configured with a temperature of 0.6 to limit response variability between participants while preserving

natural response generation. The code interpreter function was enabled, but code-like outputs were filtered or shown depending on the transparency levels set in the experimental conditions.

The web-based system was deployed on a Linux Debian server, utilizing NGINX reverse proxy to present the Streamlit application securely. Participants accessed the system through a web interface on laboratory computers, ensuring a standardized experimental setup. A visual summary of the complete tech stack is shown in Appendix, Section B.1.2, Figure B.3, while the complete experimental setup screen, as used by participants, is shown in Figure 2.2.

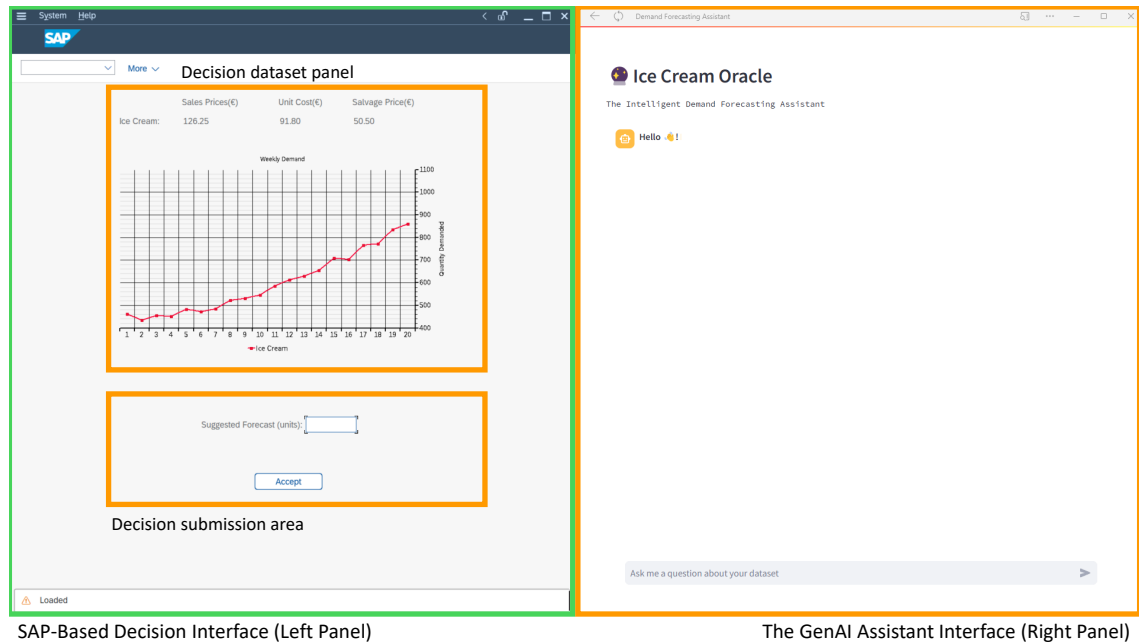


Figure 2.2: The system interface as used by participants during the experiment.

Manipulation of Transparency and Explainability

For each experimental condition, different GenAI assistants were designed to manipulate the levels of transparency and explainability. In the high transparency condition, the assistant displayed a welcome message presenting its goal, data sources, methods, and previous performance to address all four dimensions of our adopted transparency definition (Chen et al., 2014). It also provided a step-by-step description of how it derives

the recommendations and displays the code it generates to calculate its recommendations. In the low transparency condition, the assistant only presented a welcome message presenting its previous performance and did not provide any code or detailed description of its prediction process.

The explainability of GenAI assistants was manipulated by adjusting whether the GenAI assistant provided additional information after outputting its predicted sales for the subsequent period. This additional content focused on how weather, seasonality, and past demand in the simulation influenced the assistant's prediction. In case of low explainability, the assistant only provided the final prediction of sales without elaborating on its reasoning. An illustration depicting the GenAI assistants used in the main experiment and the specific pieces of information introduced under the transparency and explainability conditions is shown in Figure 2.3. Screenshots corresponding to each of the four experimental conditions are provided in Appendix, Section B.1.2.

An important step in designing an IS for the evaluation of theoretical concepts is ensuring instantiation validity, "the validity of IT artifacts as instantiations of theoretical constructs" (Lukyanenko et al., 2014, p. 322). To ensure the instantiation validity of transparency and explainability in the designed GenAI assistants, we conducted a between-subjects online manipulation check study. An online worker panel, Prolific, was used to recruit 224 participants for this manipulation check study. After excluding participants who failed attention checks, submitted the same response for all items, or had unusually fast completion times, 187 participants remained in the sample (93 in Group 1, 94 in Group 2).

In this manipulation check study, Group 1 participants were shown a screenshot of the GenAI assistant with high transparency and no explainability, where it answered a user query *What is the forecasted demand next week?* Group 2 participants viewed a similar screenshot depicting a GenAI assistant with low transparency and high explainability. Based on the interaction shown on the screenshot, participants rated the interaction using a five-item measurement tool on a 7-point Likert scale, where three items focused on perceived transparency and the remaining two focused on measuring perceived

explainability.

Composite scores were computed for perceived transparency and explainability. Independent samples t-tests revealed a statistically significant difference in perceived transparency between the two groups, with higher scores in Group 1 ($M = 5.16$) compared to Group 2 ($M = 4.71$), $t = 2.67$, $p = .0088$. Perceived explainability was significantly higher in Group 2 ($M = 4.99$) than in Group 1 ($M = 4.42$), $t = -2.47$, $p = .0147$. Item-level statistics and measurement details are provided in Appendix, Section B.1.1.


Ice Cream Oracle
 The Intelligent Demand Forecasting Assistant


Hello 🌟 I am your demand forecasting assistant.

I aim to help you make accurate demand forecasts based on historical sales data and weather forecasts. Here's how I do it:

Legend
 Present in "High Transparency" conditions only.
 Present in "High Explainability" conditions only.

- **Goal:**
 - Maximize profits.
- **Used Dataset:**
 - Sales data from the last two years (your company's database).
 - Pricing data from the last two years (your company's database).
 - Weather forecasts for the upcoming week (weather.com).
- **Methods:**
 - The data is used to estimate the baseline, trends, and patterns in sales.
 - The estimated parameters are then used in a forecasting method to get the next week's sales estimate.
- **Previous Performance:**
 - The model's confidence can vary between 64% and 92%.

Ask me to generate a demand forecast!


 What is your recommendation?



✓ Code

For the forecast, I will use the Holt-Winters exponential smoothing method, which is well-suited for data with seasonality, trends, and irregular components. Here is a step-by-step explanation of the process:

1. **Load the data:** We have loaded the historical data, which includes the week number, temperature, precipitation, wind speed, ice cream demand, and some other factors.
2. **Prepare the data:** We will focus on the demand data and use the past weeks' data to forecast the demand for week 32.
3. **Fit the Holt-Winters model:** We will fit the Holt-Winters model using the historical demand data. The model parameters (alpha, beta, gamma) will be estimated based on the data.
4. **Forecast for the next week:** Using the fitted model, we will forecast the demand for week 32.

 Let's implement this in Python.

✓ Code

Using the Holt-Winters exponential smoothing method, we have forecasted the ice cream demand for week 32.

Forecast Summary:

- Forecasted Demand for Week 32: 1016 units
- Confidence Level: 88%

Based on our objective to ensure stock availability, I recommend planning for 1016 units of ice cream for week 32.

Forecasted Demand:

1016 units

Discussion

The forecasted demand for week 32 is derived based on the following factors:

- **Seasonality:** The Holt-Winters method assumes yearly seasonality, capturing the annual demand patterns for ice cream.
- **Weather Conditions:** Temperature, precipitation, and wind speed are significant predictors of ice cream demand. These factors have been considered in the historical data used for forecasting.
- **Previous Week's Demand:** The demand for the previous week (1030 units) provides a reference point for the current forecast.
- **Smoothing Parameters:** The model uses the following smoothing parameters:
 - Alpha (α): 0.5 (high sensitivity to peaks)
 - Beta (β): 0.3 (moderate sensitivity to trends)
 - Gamma (γ): 0.2 (low sensitivity to seasonality)

Figure 2.3: An example GenAI assistant with high transparency & high explainability. The annotations highlight the components introduced through the transparency and explainability manipulations.

AI Assistant Prompts and Process

The experimental manipulations were applied by programmatically hiding parts of the response on our interface and engineering prompts for each condition's GenAI assistant using the chain-of-thought method (Wei et al., 2022). Each GenAI assistant was programmed with several blocks of instructions. The first block outlined the assistant's role and objective in supporting the user's decision-making process. The second block provided detailed instructions on generating decision recommendations for the planning decision, with a more complex step-by-step explanation for the high transparency conditions instead of direct instruction for the low transparency condition (Wei et al., 2022). The third block of the prompt included a set of constraints on the level of transparency and explainability that needed to be provided according to the experimental condition. The prompt concludes with an example code for the assistant to write and run if the user's query is unclear. This approach makes it possible to ensure recommendation consistency across the different versions of GenAI assistants. The prompting process involved an iterative testing process, where each iteration produced improvements to the prompts to ensure each GenAI assistant behaves according to the experimental conditions. This iterative process concluded when the tests did not yield further improvements.

Hardware Used

The experimental stimuli were presented on a 23.8 inch screen of 1920x1080 resolution at 60Hz, and participants' eye movements were captured using a Tobii Spectrum eye tracking device at 1200Hz. All participants used the same hardware in the same room under controlled environmental lighting conditions.

Measurements

Decision performance was measured using a loss function grounded in the planning problem definition, reflecting the cost implications of inventory decisions at each trial. Specifically, given the scenario parameters of cost ($c = 91.8$), price ($p = 126.25$), and

salvage price ($s = 50.5$), the cost of overestimating demand by one unit was $c - s = 41.3$, and the cost of underestimating it was $p - c = 34.45$. Based on these values, participant decisions were evaluated by comparing their order quantity to actual demand, with losses incurred for both over- and under-ordering. A perfect match between order quantity and demand resulted in a loss value of zero, while all deviations led to negative loss scores, indicating suboptimal performance.

Decision confidence and confidence in the GenAI assistant were recorded immediately after each trial using single-item, 7-point Likert scales (Davis et al., 1994). These measures were collected after participants submitted their order decisions but before viewing the performance report for that trial. Although single-item operationalizations do not permit internal consistency assessment and may be more sensitive to common method bias (De Guinea et al., 2013), prior research shows that single-item measures can provide reliable and valid assessments when constructs are unidimensional, concrete, and straightforward for respondents to evaluate (Ang and Eisend, 2018; Bergkvist and Rossiter, 2007; Wanous et al., 1997).

Because confidence was conceptualized as a unidimensional, trial-level assessment of perceived system ability, and multiple observations were collected repeatedly within each participant session, single-item measures were used to reduce response burden and preserve participant attention across trials, consistent with prior research employing repeated confidence assessments in dynamic decision-making contexts (Chong et al., 2022).

Perceived understanding of the GenAI assistant was measured at the end of each 10-trial experimental block using the explanation satisfaction scale, defined as “the degree to which users feel that they sufficiently understand the AI system or process being explained to them” (Hoffman et al., 2023, p. 3).

In addition, two control variables were measured before the experimental tasks: Automation-Induced Complacency Potential (AICP), administered before task-specific instructions (Merritt et al., 2019), and Computer Self-Efficacy, collected after the task and interface were introduced (Compeau and Higgins, 1995). The full list of measurement

items is provided in Appendix, Section B.1.3, Table B.2.

Participants' information processing activity was captured using two eye-tracking concepts grounded in scanpath analysis: attention shift rate and attentional resource allocation. Both of these concepts' operationalization relied on the definition of two AOIs: (1) the SAP-based decision interface on the left side of the screen, and (2) the GenAI assistant interface on the right. These AOIs are shown in Figure 2.2.

Attentional resource allocation was operationalized as the proportion of total fixation time spent within the decision interface AOI during each trial. This measure reflects the distribution of attentional resources across screen regions. Higher values suggest greater effort invested in interpreting contextual task data (e.g., demand history, pricing parameters), while lower values may reflect greater reliance on GenAI output.

Attention shift rate was measured by the number of rapid eye movements between fixations (saccades). In this study, saccades beginning in one interface and ending in the other were counted to operationalize attention shifts between the GenAI assistant and the decision interface. Higher attention shift rates may suggest uncertainty, effortful reasoning, or a lack of confidence in or understanding of the GenAI assistant.

2.4.3 Experimental Procedure

Each participant completed 25 trials as shown in Figure 2.4. Experiment sessions were structured as follows:

1. **Initial Instructions and Self-Reported Control Variable Measurements** Participants read instructions explaining the experimental task and decision interface. They also completed a short survey including control variable measurements. After these self-reported measurements are submitted, the business simulation was initialized on ERPsim with the parameters $n = 1$, $m = 20$.
2. **Training Block (5 trials, No GenAI assistant):** Participants completed five consecutive training trials ($n = 1$ to $n = 5$), where each trial represented consecutive virtual weeks $m = n + 19$. In the five trials within the training block, participants

did not have access to any GenAI assistant. Instead, they only interacted with the decision interface to familiarize themselves with the task and interface.

3. **Introductions and Self-Reported Control Variable Measurements:** Participants read instructions explaining the use of GenAI assistants and then completed a short survey including control variable measurements that probe their perceptions of the introduced GenAI assistant.
4. **Experimental Block 1 (10 trials, GenAI assistant):** Participants resumed the business simulation from where they left at the end of the training block ($n = 6$, $m = 25$). In this experimental block, consisting of 10 trials, participants continued interacting with the previously initiated business simulation while following the same virtual calendar. In addition to the training block, this experimental block featured a GenAI assistant, and participants could interact with it freely during the experimental block. This assistant's transparency and explainability levels were set according to the experimental conditions assigned to each participant.
5. **Survey Measurements:** Participants completed a survey on their perceived understanding of the GenAI assistant they used in Block 1.
6. **Experimental Block 2 (10 trials, GenAI assistant):** Participants resumed the business simulation from where they left off at the end of the first experimental block ($n = 16$, $m = 35$). As in the previous experimental block, participants continued to interact with the business simulation and followed the same virtual calendar. A different, randomly assigned GenAI assistant was available to participants during these 10 trials.
7. **Survey Measurements:** Participants completed a survey on their perceived understanding of the GenAI assistant they used in Block 2.

Each trial n represented a virtual week $m = n + 19$ in the simulation, and participants received performance feedback at the end of each trial to track the impact of their decisions and adjust their strategies accordingly.

Each trial n followed the sequence below:

1. **Forecast Submission:** Participants forecasted the demand for week $m + 1$.
2. **Confidence Self-Report Measurements:** Participants self-reported their decision confidence, and their confidence in the GenAI assistant, if applicable.
3. **Simulation Calendar Update:** A week in the simulation calendar has passed and m is updated as $m = m + 1$.
4. **Performance Feedback:** Participants were shown a report of their performance for week $m - 1$.

As an exception, during the first trial, participants were permitted to seek guidance from the experiment moderator on how to use the interface for submitting forecasts. If a participant requested an explanation, experiment moderators followed a predetermined script outlining the scenario and the user interface. Moderators were strictly required to adhere to this script and were not permitted to provide additional guidance beyond the predefined instructions.

A complete experiment session was structured as illustrated in Figure 2.4. Before the training block (trials 1 to 5), participants were asked to respond to a set of demographic questions and control variable measurements. Participants then received detailed instructions on both the decision-making scenario and the decision interface for data analysis and submission.

At the conclusion of the training block at the end of trial 5, participants were introduced to the GenAI assistant as a new feature within the simulated decision-making environment. After reading the instructions, participants were asked to report their perceptions of the GenAI assistant before proceeding with the first 10-trial block of the experimental phase (trials 6 to 25).

At the end of trial 15, the simulation calendar paused and participants were asked to complete a questionnaire assessing their perceptions of the GenAI assistant used in the previous 10 trials. Then, the second experimental block was configured with a different

GenAI assistant prototype, while the simulation calendar continued from where they left at the end of trial 15. At the end of trial 25, participants were asked to complete the same questionnaire for the previous 10 trials within the second experimental block. The experiment concluded with a debriefing on the study's objectives.

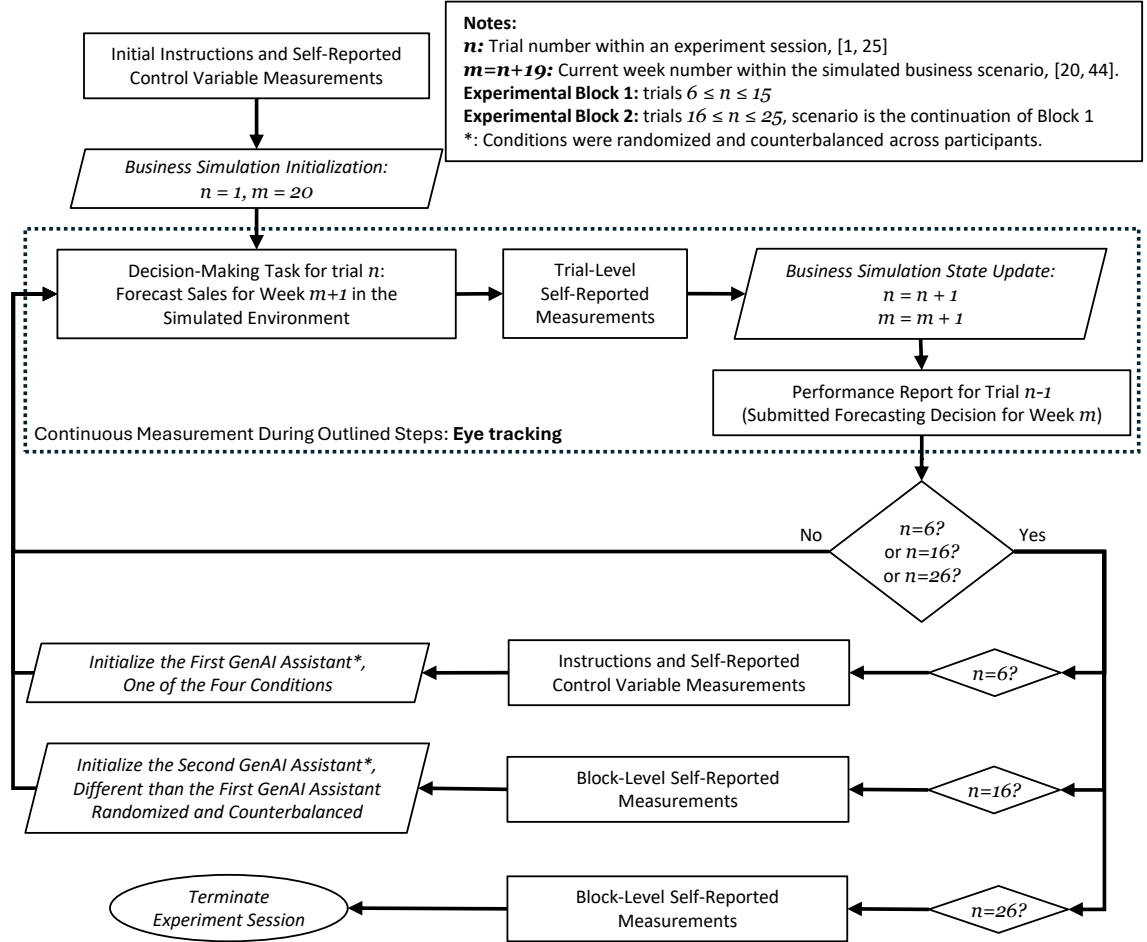


Figure 2.4: Experimental design and protocol.

2.4.4 Sample Description and Data Collection

To determine the required sample size, a power analysis was conducted using G*Power 3.1 (power = 0.90, medium effect size, $\alpha = 0.05$) (Faul et al., 2007). This analysis indicated a minimum of 48 participants to achieve sufficient statistical power. To account

for potential no-shows and technical failures, 78 participants were scheduled for data collection between 15 July 2024 and 27 September 2024.

Participants were recruited from a convenience sample through our institution's participant panel. Prior to registration, participants were screened based on three inclusion criteria: (1) prior experience with ERP systems, (2) prior experience with ERPsim, (3) completion of at least an introductory logistics and demand forecasting course. These criteria ensured a consistent level of domain knowledge by requiring all participants to have a fundamental understanding of the experimental task. Additionally, participants were screened for having normal or corrected-to-normal vision.

The study was reviewed and approved by our university's Ethics Committee (certificate number: 2023-5420). Subjects received a compensation of CAD 25 for their participation, and each session lasted approximately 90 minutes. Prior empirical studies using the newsvendor paradigm have employed both fixed (e.g., Wang et al. 2025) and performance-based (e.g., Benzion et al. 2008; Kremer et al. 2010) compensation schemes. Experimental evidence indicates that performance-based financial incentives do not consistently improve decision performance in judgment tasks (Camerer and Hogarth, 1999). Accordingly, a fixed compensation structure was adopted, while motivational engagement was supported through gamification elements, including performance feedback and social comparison presented at the end of each trial.

Of the 78 scheduled participants, 16 did not show up or canceled their session within 24 hours of their appointment. Additionally, two sessions were concluded without data collection due to external IT service outages (CrowdStrike failure on 19 July 2024, and OpenAI API failure on 28 August 2024). These participants were still compensated for their attendance. The final dataset consisted of 60 participants (32 females; mean age 29.1 ± 6.41 ; range, 19-47 years), with further exclusions applied in the data cleaning phase.

2.4.5 Data Cleaning and Preprocessing

Following data collection, all textual logs of GenAI interactions were parsed using Python. Each user-assistant message pair was matched and processed to identify the user prompt, assistant response, company code, trial number within the experiment, experimental condition (transparency and explainability), response time, and response length with and without code blocks. In addition, pattern-matching routines were used to extract the reported confidence level and recommended sales forecast by the GenAI assistant when such information was included. A total of 1,667 user queries remained across 1,117 experimental trials. This indicated that, on average, participants submitted 1.49 user queries per trial.

In cases where participants requested multiple forecasts within a single trial—for example, by asking the GenAI assistant to regenerate its recommendation using different parameters—the last recommendation was considered as the trial’s final GenAI assistant forecast. In cases where participants did not request any recommended decision from the GenAI assistant—for example, by only asking about expected changes in weather conditions without requesting a sales forecast—the lack of a recommended action by the GenAI assistant meant that the explainability condition could not manifest as tested in the manipulation check study. Based on this procedure, 269 trials were marked as invalid. Consequently, 848 experimental trials were included in the cleaned dataset.

As participants had full control over how the GenAI assistant generated its recommendations, they could also configure it in a way that led to very low-performing forecasts. To ensure consistent GenAI performance across participants and experimental conditions, we further excluded one trial with abnormally poor recommendation quality. In addition, two participant input errors (typos) that were noted during experiment sessions were manually corrected. This led to a final dataset of 847 trials across 102 experimental blocks and 57 participants. This cleaned dataset was used for all analyses involving behavioral and self-reported metrics.

2.5 Results

Prior to hypothesis testing, we evaluated whether the counterbalancing model used for experimental condition assignments remained valid after excluding participants and observations during data cleaning. A chi-squared test was conducted to examine the relationship between the four experimental conditions and their assigned order within experiment sessions. The result was not statistically significant ($\chi^2 = 1.66$, $p = 0.65$), indicating that the dataset did not contain a systematic error in terms of the ordering of experimental conditions within experiment sessions.

Individual characteristics that could affect decision-making performance (such as automation complacency potential, computer self-efficacy, and experience with ERP systems) and other demographic variables, such as age and gender, are controlled via randomization.

In addition, to ensure that our manipulation of transparency and explainability through system prompts did not inadvertently influence the quality of the recommendations generated by these GenAI assistants, we conducted an additional analysis of the recommendation performance of the four distinct GenAI assistants. GenAI assistants' recommendation quality was calculated by transforming the GenAI assistants' recommendations using the same loss function used to calculate participants' decision performance. A linear mixed-effects model was estimated using square-root transformed GenAI assistant performance as the dependent variable. The model included transparency, explainability, and their interaction as fixed effects, with trial week number included as a covariate to control for demand variation across weeks in the simulated environment. Random intercepts were specified for each participant-block combination to account for repeated measures within each GenAI assistant instance. The results indicated no significant main effects of transparency ($\beta = -0.28$, $p = .849$) or explainability ($\beta = 1.30$, $p = .371$), and no significant interaction effect ($\beta = 0.46$, $p = .836$). These results suggest that the GenAI assistant recommendation quality remained stable across experimental conditions.

2.5.1 Measurement Instruments

Measurement of Perceived Understanding of the GenAI Assistant

We assessed the measurement properties of the perceived understanding concept using confirmatory factor analysis (CFA). Initially, all eight items that were adapted from the original source were included in a single-factor model. In this round, the standardized loadings ranged from 0.503 to 0.905.

To improve measurement reliability, we re-evaluated the measurement items and refined the model by removing two items. The 7th item was excluded because it measured perceptions of the assistant's self-reported accuracy, which was not manipulated in this study and was always visible to participants in all conditions. The 8th item was also removed due to its lower factor loading value in the initial analysis. The resulting six-item model had standardized loadings ranging from 0.666 to 0.914, with Cronbach's $\alpha = 0.92$. Measurement items' factor loading values at each CFA run can be seen in Appendix, Section B.1.3, Table B.5.

Measurement of Confidence in the GenAI Assistant

In line with prior work utilizing single-item measures under time-constrained experiments (Davis et al., 1994; Chong et al., 2022), confidence in the GenAI assistant was measured at the end of each trial using a one-item scale. This assessment was collected before participants were shown any feedback about their performance, ensuring that the response captured anticipatory confidence rather than outcome-based satisfaction. Participants rated their confidence on a 7-point Likert scale, with higher values indicating greater confidence in the AI's recommendation for that specific trial.

Across all conditions and trials ($N = 847$), the average reported confidence in the GenAI assistant was 4.64 ($SD = 1.33$), with scores ranging from 1 to 7. Among the four experimental cells, the condition combining high transparency ($T=1$) with no additional explainability ($E=0$) yielded the highest average confidence rating ($M = 5.02$, $SD = 1.31$; $n = 221$). Descriptive statistics at the group level are reported in Table B.14, including

standard deviations and cell sizes.

Measurement of Control Variables

To account for individual differences, we measured several control variables at the participant level. Automation-Induced Complacency Potential was assessed prior to task instructions, and Computer Self-Efficacy was measured following the initial instructions.

The 10-item AICP scale was modeled according to its two theorized subdimensions: Workload Alleviation (AICP-W) (Items 1-4, 6) and Monitoring (AICP-M) (Items 5, 7-10) (Merritt et al., 2019). Confirmatory factor analyses (CFA) were conducted for each subscale separately. While the Monitoring subscale demonstrated low internal consistency ($\alpha = 0.58$) and was excluded from further analyses, the Workload Alleviation subscale showed acceptable reliability (Cronbach's $\alpha = 0.70$) and was retained as a control variable.

The Computer Self-Efficacy construct was modeled as a single latent factor using the full 10-item scale from Compeau and Higgins (1995). Following CFA, four items with loadings below 0.5 were excluded. The final model included six items, which led to a scale with all item factor loading values exceeding 0.57 and a Cronbach's $\alpha = 0.87$. Because computer self-efficacy represents an established construct and was included as a control variable measured prior to the experimental task rather than as a focal construct in hypothesis testing, the remaining items were retained based on their conceptual relevance despite two items with loadings below the preferred 0.708 threshold (Churchill, 1979; McKenzie et al., 2011).

As an additional behavioral control variable, baseline task performance was calculated using the final four trials of the five-trial training block, during which participants performed the task without any GenAI assistant. The first training trial was excluded, as participants could request assistance from the experiment moderator.

2.5.2 Descriptive Statistics

Descriptive statistics are discussed separately for three datasets used in the study: block-level self-reported data, trial-level behavioral data, and trial-level eye-tracking data. The number of observations differs across datasets due to the nature of data collection and analysis. Block-level data are based on post-block questionnaires, yielding two observations per participant, one per block. Trial-level behavioral data are collected for each decision task. While the eye-tracking data is based on the trial-level dataset, it was further filtered based on eye-tracking measurement validity, reducing the number of usable trials.

A summary of block-level data on perceived understanding is reported in Table B.13 in Appendix, Section B.2. Participants in the high transparency, low explainability condition reported the highest average perceived understanding (mean = 5.29, SD = 0.97), while those in the high transparency, high explainability condition reported lower understanding compared to high transparency, low explainability (mean = 4.97, SD = 1.28). On the other hand, participants in the low transparency, low explainability condition reported the lowest average perceived understanding (mean = 4.08, SD = 1.55).

At the trial level, results indicate that participants in the high transparency, low explainability condition reported the highest average confidence in the GenAI assistant (mean = 5.02, SD = 1.31) and the highest decision performance (mean = -978, SD = 897). In line with the patterns in perceived understanding and confidence in the GenAI assistant, decision performance in the high transparency, high explainability condition (mean = -1180, SD = 1008) was lower than in high transparency, low explainability. This consistent pattern across perceptions and decision outcomes suggest the presence of a non-additive relationship when both transparency and explainability are combined. Full descriptive statistics are reported in Table B.14 in Appendix B.2.

In the eye-tracking data, both fixation duration on the decision interface and total

task time were lowest in the high transparency, high explainability condition, suggesting more efficient processing in this group. In addition, a time-series plot of the attention shifts over ten trials shows that on average, participants reduced their attention shift rates by about 50% after having used the same GenAI assistant in similar cases for 10 times (See Figure B.16 in Appendix B.2). Descriptive statistics of extracted eye-tracking metrics can be seen in Table B.15 in Appendix Section B.2.

2.5.3 Hypothesis Testing

Perceived Understanding of the GenAI Assistant

To test the impact of transparency and explainability on users' perceived understanding of the GenAI assistant (H1), we estimated a linear mixed-effects model with participant-level random intercepts. The analysis was conducted at the block level, using 102 observations, as perceived understanding was measured at the end of each 10-trial block. The manipulations of transparency and explainability were included in the model through two dummy variables, as we are interested in interpreting the main effect of a variable while the other independent variable is at its baseline level of zero (Gelman, 2008).

Control variables were removed from the model as they were found to be insignificant, with the exception of AICP-W, which approached significance, and block number, which was retained due to its relevance for capturing fatigue or learning effects associated with experimental order.

The results support H1a and H1b, indicating that both transparency ($\beta = 0.24$, $p < .001$) and explainability ($\beta = 0.12$, $p = .018$) significantly increased perceived understanding. The interaction term between transparency and explainability was negative and significant ($\beta = -0.21$, $p = .009$), indicating that their combined presence yielded a smaller improvement than when each was presented alone. The fixed effects explained 18% of the variance in perceived understanding ($R_m^2 = .18$), ($R_c^2 = .61$).

Model estimates are reported in Appendix Section B.2, Table B.17.

Perceived Understanding and Decision Performance

To test whether users' perceived understanding of the GenAI assistant improves their decision performance (H3), we estimated a series of linear mixed-effects models at the block level ($N = 102$). In the full model including perceived understanding (Model 1), the coefficient for perceived understanding was positive but not statistically significant ($\beta = 0.08, p = .226$). When excluding perceived understanding (Model 2), the interaction effect of transparency and explainability was significant ($\beta = -0.113, p = .047$), indicating that the combined presence of these features reduced performance compared to their separate contributions. This effect was partially attenuated when perceived understanding was included in Model 1, although it remained marginally significant ($\beta = -0.1, p = .085$). To probe the indirect path, we estimated a third model where perceived understanding was regressed on the experimental manipulations (Model 3), replicating earlier results showing strong positive effects of transparency ($\beta = 0.243, p < .001$) and explainability ($\beta = 0.125, p = .02$), and a negative interaction ($\beta = -0.219, p = .008$). Collectively, these findings suggest that while perceived understanding was enhanced by transparency and explainability, it did not significantly mediate their effects on decision performance in this experimental setting. As the partial model tests ruled out a mediation relationship, no further analyses regarding mediation were done. Model details are reported in Appendix, Section B.2, Table B.19.

Confidence in the GenAI Assistant

To examine how transparency and explainability influence users' confidence in the GenAI assistant (H2), we estimated a linear mixed-effects model with participant-level and block-level random intercepts. A maximal model including random slopes was initially tested (Barr et al., 2013); however, high correlations between random factors suggested overfitting. Therefore, in line with methodological recommendations (Matuschek et al., 2017), we report results based on a more parsimonious statistical model with participant- and block-level random intercepts.

The model was fitted on trial-level data ($N = 847$) and included trial number and block number as control variables, as they relate to temporal structure and potential fatigue effects. Other control variables were removed due to them being insignificant.

The results support H2a: transparency had a significant positive effect on users' confidence in the GenAI assistant ($\beta = 0.122$, $p < .001$, one-tailed). However, the results do not support H2b, as explainability did not have a significant impact on confidence in the GenAI assistant ($\beta = 0.017$, $p = .32$, one-tailed). The interaction between transparency and explainability was significantly negative ($\beta = -0.098$, $p = .044$, one-tailed), indicating that the combined presence of both features led to a smaller increase in confidence than when each was presented alone.

In this model, the fixed effects explained 5% of the variance in confidence in the GenAI assistant ($R_m^2 = .05$), ($R_c^2 = .58$). Model details are provided in Appendix, Section B.2, Table B.18.

Confidence in the GenAI Assistant and Decision Performance

To test whether users' confidence in the GenAI assistant improves their decision performance (H4), we estimated a linear mixed-effects model with trial-level observations ($N = 847$) and participant-level random intercepts. In the full model including confidence (Model 1), users' confidence in the GenAI assistant was a significant predictor of decision performance ($\beta = 0.063$, $p = .015$). In the direct effects model (Model 2), transparency significantly improved decision performance ($\beta = 0.032$, $p = .031$), while explainability had no significant effect ($\beta = 0.006$, $p = .709$). Notably, the interaction between transparency and explainability was again negative and significant ($\beta = -0.068$, $p = .004$), once again indicating a sub-additive pattern. To validate the indirect path, we estimated a third model predicting confidence in the GenAI assistant using transparency and explainability as independent variables. Results indicated that transparency significantly increased confidence in the GenAI assistant ($\beta = 0.131$, $p < .001$), and the results of these partial models in Appendix, Section B.2, Table B.20 indicate a possible mediation effect of confidence in the GenAI assistant between

transparency and decision performance.

To test the mediation hypothesis without breaking the normality assumptions, we implemented a nonparametric bootstrap approach with 5,000 iterations using the `boot` package in R (Hayes, 2017). Specifically, at each bootstrap iteration, a resampled dataset was generated with replacement. Separate linear mixed-effects models were fit on each resample to estimate the effect of transparency on confidence (path a) and the effect of confidence on decision performance (path b). The product $a \times b$ was computed to obtain the indirect effect estimate for that iteration.

Based on 5,000 bootstrap samples, the 95% percentile-based confidence interval for the indirect effect was [0.0013, 0.0168], indicating that the mediation effect is statistically significant, as the confidence interval does not include zero. These results provide additional support confirming that transparency increases decision performance partially through its positive impact on users' confidence in the GenAI assistant.

Attentional Resource Allocation

To examine how transparency and explainability influence attention allocated on the decision interface (H5), we estimated a linear mixed-effects model using trial-level data with valid eye-tracking measurements ($N = 808$), with participant-level random intercepts. The dependent variable was the log-transformed fixation duration on the SAP-based decision interface, adjusted for trial duration.

The results support both H5a and H5b. Transparency significantly reduced attention allocation on the decision interface ($\beta = -0.061$, $p < .001$), and explainability also had a significant negative effect ($\beta = -0.038$, $p = .0096$). The model explained 6% of the variance in attention allocation ($R_m^2 = .06$), ($R_c^2 = .42$).

Control variables showed mixed effects. While trial number within block had a significant positive effect ($\beta = 0.008$, $p < .001$), block number was marginally significant ($\beta = -0.017$, $p = .060$), suggesting potential adaptation effects across blocks.

Attentional Resource Allocation and Decision Performance

To test whether attention allocation mediates the relationship between GenAI design features and decision performance (H7), we estimated a linear mixed-effects model using trial-level observations ($N = 799$) with random intercepts by participant. In the full model (Model 1), reduced attention allocation, measured as the percentage of time spent fixating on the decision interface, significantly predicted higher decision performance ($\beta = -0.109$, $p = .010$). In the direct effects model (Model 2), transparency significantly improved decision performance ($\beta = 0.034$, $p = .042$), whereas explainability had no significant ($\beta = 0.002$, $p = .906$). To test the mediating role of attention allocation, we estimated a third model predicting fixation duration. Transparency ($\beta = -0.060$, $p < .001$) and explainability ($\beta = -0.042$, $p = .004$) both significantly reduced attention allocation, supporting a potential indirect effect. Full model details are reported in Appendix, Section B.2, Table B.23.

To test the mediation hypothesis without assuming normality, we applied a nonparametric bootstrap procedure with 5,000 iterations using the `boot` package in R (Hayes, 2017). At each iteration, a resampled dataset was created with replacement, and separate linear mixed-effects models were fit to estimate the effect of transparency on fixation duration on the decision interface (path a) and the effect of fixation duration on decision performance (path b). The product $a \times b$ was computed to obtain an indirect effect estimate for that iteration.

The 95% percentile-based bootstrap confidence interval for the indirect effect was [0.0017, 0.0133], indicating statistical significance, as the interval excludes zero. These results support the presence of a significant indirect path from transparency to decision performance via fixation duration. Since the direct effect of transparency on decision performance remained significant after controlling for the mediator, the results indicate a partial mediation effect.

Attention Shift Rate

To examine how transparency and explainability influence participants' attention shift rates (H6), we estimated a linear mixed-effects model using trial-level data with valid eye-tracking measurements ($N = 799$), with participant-level random intercepts. The dependent variable was the natural logarithm of attention shift saccade counts per trial.

The model results provide support for H5a only. Transparency had a negative effect on attention shifts ($\beta = -0.13$, $p = .024$), suggesting that users exposed to higher transparency exhibited lower attention shift rates. Explainability did not have a significant effect on attention shift rates of participants ($\beta = -0.08$, $p = .155$). The model explained 8% of the variance in the rate of attention shifts between the GenAI assistant and the decision interface ($R_m^2 = .08$), ($R_c^2 = .46$).

Saccade frequency decreased across experimental blocks ($\beta = -0.203$, $p < .001$) and with increasing trial number within each block ($\beta = -0.054$, $p < .001$).

Attention Shift Rate and Decision Performance

To test whether attention shift rates mediate the effects of transparency and explainability on decision performance (H8), we estimated a series of linear mixed-effects models at the trial level ($N = 799$). In the full model including attention shifts (Model 1), the coefficient for log-transformed attention shift saccade count was negative but not significant ($\beta = -0.014$, $p = .171$). The interaction between transparency and explainability remained significant ($\beta = -0.069$, $p = .008$), with minimal attenuation compared to the model excluding the mediator (Model 2: $\beta = -0.070$, $p = .007$).

In the mediator model (Model 3), transparency significantly reduced attention shifts ($\beta = -0.133$, $p = .025$), while explainability and the interaction term had no significant effects. Overall, these results indicate that attention shift rates were influenced by transparency but did not significantly mediate its effect on performance. Full model details are reported in Appendix, Section B.2, Table B.24.

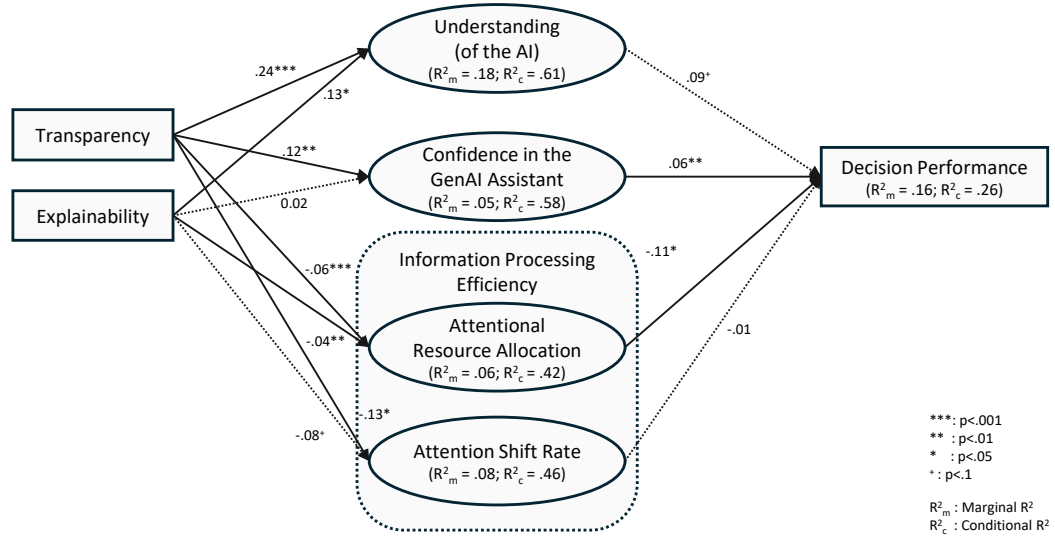


Figure 2.5: Validated research model.

Table 2.1: Hypothesis Testing Summary

Hypothesis	t-value	p-value	Significance
H1a: Transparency increases users' perceived understanding of the GenAI assistant.	4.77	<.001	***
H1b: Explainability increases users' perceived understanding of the GenAI assistant.	2.49	.008	**
H2a: Transparency increases users' confidence in the GenAI assistant.	3.33	0.001	**
H2b: Explainability increases users' confidence in the GenAI assistant.	0.47	0.319	n.s.
H3: Users' perceived understanding of the GenAI assistant leads to an increase in their decision performance.	1.36	0.089	n.s.
H4: Users' confidence in the GenAI assistant leads to an increase in their decision performance.	2.37	0.009	**
H5a: Transparency reduces attention allocated on the decision interface.	-4.06	<0.001	***
H5b: Explainability reduces attention allocated on the decision interface.	-2.88	0.002	**
H6a: Transparency reduces the rate of attention shifts between the GenAI assistant and the decision interface.	-2.25	0.013	*
H6b: Explainability reduces the rate of attention shifts between the GenAI assistant and the decision interface.	-1.42	0.078	n.s.
H7: Attention allocated to the decision interface is associated with a lower decision performance.	-2.59	0.005	**
H8: Attention shift rate is associated with a lower decision performance.	-1.37	0.171	n.s.

2.5.4 Post-Hoc Analysis of Interaction Effects

To further explore how transparency and explainability interact in shaping user outcomes, we examined the interaction term between these two design features in our models predicting perceived understanding, confidence in the GenAI assistant, attention allocation, and attention shift rates.

Across these tests, the interaction term consistently indicated a substitution effect between transparency and explainability. For perceived understanding ($\beta = -0.21$, $p = .009$) and confidence ($\beta = -0.098$, $p = .044$), the interaction was significantly negative, indicating that the combination of transparency and explainability did not yield additional benefits beyond those of each feature alone. In the model predicting saccade behavior, the interaction was also negative and significant ($\beta = -0.10$, $p = .004$). In the model predicting fixation duration, the interaction term was not statistically significant ($\beta = 0.013$, $p = .562$).

2.5.5 Post-Hoc Analysis of Temporal Decline in Attention Shift Rates

To explore how attention shift rates evolve over repeated interactions with GenAI assistants, we conducted a post-hoc survival analysis using data from the first experimental block. The objective was to determine whether participants in different experimental conditions showed varying rates of change in their gaze behavior, specifically attention shifts.

A participant was considered to have “stabilized” once their trial-level attention shifts dropped below 60% of their initial saccade count within the same block. This threshold captures a sustained reduction in attention shift rates, which can be interpreted as decreased verification behavior and increasing reliance on the GenAI assistant over time. The analysis was based on 421 trial-level observations from 57 participants.

Kaplan-Meier survival curves were estimated for the four experimental conditions defined by the two levels of transparency and explainability. A log-rank test

confirmed that stabilization rates significantly differed across conditions ($p = .038$). Participants exposed to high transparency reached stabilization earlier, while those in high explainability conditions showed a slower rate of decline in attention shift rates.

Robustness checks were conducted using alternative stabilization thresholds (55%, 50%, 40%, and 30% of the participant's initial saccade count), which yielded consistent patterns. In addition, descriptive statistics on initial attention shift rates indicate no substantial differences in baseline saccade counts in conditions with high transparency (Table B.16 in Appendix, Section B.2). This confirms that the faster decline observed in transparency conditions is not due to initially higher levels of attention shift rates.

The Kaplan-Meier curves for all experimental conditions are presented in Figure 2.6. According to the plot, participants across all conditions reach stabilization at around the eighth trial within an experimental block.

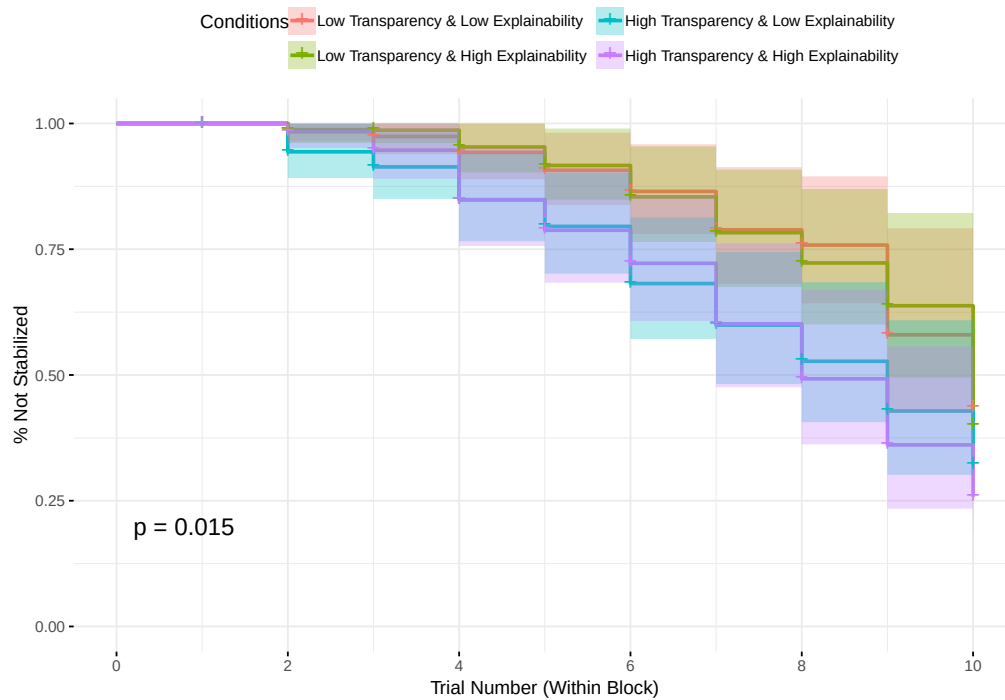


Figure 2.6: Kaplan-Meier fit plot corresponding to the change in attention shift rates over trials. This plot indicates that the conditions with high transparency reach stabilization, i.e., 40% of a participant's initial attention shift rates, faster.

2.5.6 Post-Hoc Analysis of Long-Term Effects

Motivated by the observed behavioral shifts over time across experimental conditions, we conducted an additional exploratory analysis to better understand how transparency and explainability influence users' confidence in the GenAI assistant and their decision performance after having exposure to it for a number of repetitions. Specifically, we re-estimated the models for Hypotheses 3c and 3d using only the final two trials within each experimental block. These trials correspond to later stages of the interaction, during which participants had gained experience with the GenAI assistant. As indicated by the post-hoc analysis results shown in Figure 2.6, attention shift rates tended to stabilize at this point, making it an appropriate window to assess long-term effects.

Three models were estimated using 171 trial-level observations from 55 participants. First, we examined whether participants' confidence in the GenAI assistant predicted decision performance (Model 1). The results revealed no statistically significant relationship between confidence in the GenAI assistant and performance ($\beta = 0.026$, $p = 0.565$), suggesting that confidence was not a significant performance driver in the later stages of interaction. Second, we estimated a model excluding confidence to evaluate the direct effects of transparency and explainability on performance (Model 2). Both transparency ($\beta = 0.052$, $p = 0.027$, one-tailed) and explainability ($\beta = 0.050$, $p = 0.032$, one-tailed) had a significant positive effect on performance. Moreover, their interaction remained negative and significant ($\beta = -0.099$, $p = 0.016$), confirming the sub-additive pattern observed in prior analyses. Third, we assessed the impact of transparency and explainability on confidence (Model 3). Transparency significantly increased confidence in the GenAI assistant ($\beta = 0.171$, $p < .001$), as well as explainability ($\beta = 0.074$, $p = 0.043$, one-tailed). Their interaction was not significant. Taken together, the findings indicate that transparency serves as an early enabler of decision performance, while explainability gains significance with continued interaction—highlighting its growing role of explainability as users become more familiar with the GenAI assistant.

2.6 Discussion

2.6.1 Key Findings

Drawing on the Situation Awareness (SA) model, this study investigated how transparency and explainability in GenAI assistant design influence users' cognitive states and decision outcomes during dynamic task settings. The SA model conceptualizes situation awareness as comprising perception, comprehension, and projection, and identifies system design as a primary external factor shaping these cognitive processes (Endsley, 1995; Endsley et al., 2003). Our results provide empirical support for this model in the context of GenAI assistant use in organizational decision-making, and highlight how specific design features influence user attention, confidence in the GenAI assistant, and performance over time. Below, we summarize our key findings and present the hypothesis tests in Table 2.1.

First, both transparency and explainability independently improved users' perceived understanding of and confidence in the GenAI assistant (H1a, H1b). These results align with the SA model's emphasis on system-level factors in supporting perception and comprehension (Endsley, 1995; Endsley et al., 2003). As Endsley notes, the acquisition and display of relevant environmental information are foundational for SA formation (Endsley, 1995, p. 50). Our findings involving self-reported perceived understanding measurements suggest that GenAI assistants, when designed to communicate intent and rationale clearly, can facilitate SA development.

Second, transparency significantly increased users' confidence in the GenAI assistant (H2a), whereas explainability did not have a statistically significant effect on this outcome (H2b). This suggests that transparency enables users to predict the assistant's behavior more effectively, leading to higher confidence in the GenAI assistant. On the other hand, given that participants had only 10 trials with each GenAI assistant, the lack of support for H2b may mean that, in the initial 10 interactions, an AI's justifications of individual recommendations are not as effective as detailed, step-by-step process information. As transparency maps more with Level 1 and 2 SA whereas explainability maps more with

Level 2 SA; these findings support the overall structure of the SA model where decision-makers build each level of their SA by building on top of each level.

Third, hypotheses H3a and H3b were not supported: perceived understanding did not mediate the relationship between transparency/explainability and decision performance. One plausible explanation relates to the timing of measurement. Perceived understanding was assessed after an experimental block of ten trials, while decision performance was captured across those ten trials and then averaged. If understanding improved gradually across trials, the end-of-block measurement may not have captured its early influence on behavior.

Fourth, H3c and H6a were supported: confidence in the GenAI assistant and attention allocation significantly mediated the relationship between transparency and decision performance. These mediation effects highlight the role of projection in facilitating effective human-AI interaction. By providing insight into system behavior, transparency facilitates users' confidence in the GenAI assistant, reduces the need to allocate attention on source data, and eventually improves performance. In contrast, H3d was not supported: explainability did not influence decision performance, neither directly nor through confidence, at least in the early phases of interaction. However, it is important to note that our experimental design included 10 repeated interactions, meaning that our findings were observed during the first 10 interactions. Although interaction over ten decision-making cycles does not only represent use immediately after adoption, it still reflects an early period in continued use. Yet post-hoc analyses revealed that explainability's direct effect on performance became significant in later trials, suggesting its influence may grow with repeated exposure.

Fifth, findings from H4a, H4b, H5a, and H5b indicate that when GenAI assistants lack transparency or explainability, users compensate by exhibiting more attention shifts, potentially in an attempt to verify or confirm the information generated by the GenAI assistant. Such verification behaviors are most likely an indication of participants' attempts to construct their SA and their understanding of the GenAI assistant's behavior.

The tested model explained approximately 16% of the variance in decision perfor-

mance, indicating that cognitive, attentional, and design factors examined in this study account for a meaningful portion of performance outcomes. However, prior research suggests that additional individual factors such as prior experience, task-relevant knowledge, risk aversion, and individual decision strategies also contribute to decision performance and may explain the remaining variability (Bolton and Katok, 2008; Christensen-Szalanski and Beach, 1982; De Vericourt et al., 2013; Ericsson and Lehmann, 1996; Oz et al., 2020).

In addition, a post-hoc survival analysis of the attention shift rates over time revealed that even though transparency does not immediately reduce attention shift rates in initial interactions, it accelerates the rate of decline over time. The analysis showed that participants across all conditions exhibited a general reduction in saccades between the GenAI assistant and decision interface by the eighth trial. More interestingly, those exposed to high transparency conditions stabilized significantly earlier than those exposed to explainability in the GenAI assistant they used. By revealing general system logic, transparency appears to facilitate faster development of reusable mental models of the GenAI assistant, thereby reducing the need for ongoing cross-checking. This difference in attention shift rates across transparency and explainability is in line with the SA model, which posits that adaptive projection is enhanced when comprehension is based on stable and reusable mental representations. This finding also aligns with prior hypotheses from the algorithm aversion literature, which indicate that user confidence in algorithmic systems is dynamic and sensitive to design (Prahl and Van Swol, 2017; Logg et al., 2019), and that effective explanations can facilitate better trust calibration over time (Tomsett et al., 2020; Zhang et al., 2020).

An additional insight emerged from our analysis of control variables. Neither AICP-W (automation induced complacency potential) (Merritt et al., 2019) nor computer self-efficacy (Compeau and Higgins, 1995) significantly predicted users' perceived understanding, confidence in the GenAI assistant, or decision performance. In contrast, transparency and explainability demonstrated statistically significant relationships with most of these outcomes. These findings highlight that effective human-AI interaction

is primarily a design challenge and user preconceptions about automation or individual differences in manual task efficacy are less relevant. In other words, effective SA in AI-supported decision-making depends less on who the user is and more on how the system is designed to support what the user sees and understands.

In post-hoc analyses, we also examined how the simultaneous presence of transparency and explainability influenced outcomes by testing the interaction term between them. Surprisingly, the interaction did not yield additive benefits, but instead exhibited a substitution effect. In other words, when both features were present, perceived understanding and confidence in the GenAI assistant did not increase beyond the levels observed when either was presented alone. Moreover, this substitution effect persisted in the manipulations' impact on participants' information processing efficiency. This pattern suggests that increasing the amount of information provided by the GenAI assistant does not necessarily improve user outcomes and may, under certain conditions, introduce cognitive demands that compromise the effectiveness of system design. This supports the SA model's argument that in high-automation environments, increases in perceived system complexity can degrade user SA (Endsley, 1995).

2.6.2 Theoretical Contributions

This study makes three primary theoretical contributions. First, it extends the behavioral operations and decision-making literature by examining how GenAI assistance influences decision outcomes in a common planning task in operations management (Schweitzer and Cachon, 2000; Kremer and Minner, 2008; Kremer et al., 2010; Benzion et al., 2008). While prior studies have focused on task difficulty and behavioral deviations from normative benchmarks, our work shifts attention to GenAI interaction design and its downstream effects on performance. Moreover, as highlighted by Wang et al. (2025), focusing on such an important paradigm for a large research field not only strengthens the practical utility of the study but also is a good starting point for making a theoretical contribution. As human-AI interaction is a rapidly evolving research field, a theoretical

contribution of such an initiative nature is especially important. The extant literature in human-AI interaction often provides pieces of empirical evidence for fragmented arguments, with little focus on an integrative theoretical understanding. Our study contributes to the existing discussion on human-AI interactions by exemplifying the use of an established theoretical lens, SA model, in understanding human-AI interaction dynamics both in initial use and continued use settings.

Regarding the second source of theoretical contributions, this study contributes conceptual clarity by disentangling and operationalizing two often conflated constructs: transparency and explainability. We provide precise definitions from existing frameworks (Chen et al., 2014; Vilone and Longo, 2021), and illustrate a clear method to instantiate each construct in a GenAI assistant context. Considering the increasing popularity and importance of GenAI use in organizations, this study follows Whetten's (1989) calls for introducing new constructs and situating them within a coherent nomological network.

Our third theoretical contribution stems from our theorizing of relationships between new constructs and testing these relationships (Whetten, 1989). Inspired by the SA model, we identify transparency and explainability as relevant design features in the context of human-AI interactions, theorize relationships involving them, and test these hypothesized relationships in an ecologically valid setting using non-intrusive real-time measurement methods. Our findings clarify when and how design attributes improve or hinder performance by highlighting GenAI design as an explanation of the inconsistent results in the literature on algorithmic advice-taking (Bansal et al., 2021; Dietvorst et al., 2015; Allen and Choudhury, 2022; Anaraky et al., 2020). In line with Whetten's (1989) view that theoretical contributions emerge when new constructs change causal maps, we show that design attributes influence user reliance and cognitive responses to GenAI assistants, rather than treating algorithmic use as a binary adoption decision.

Beyond these core contributions, the study introduces several methodological and empirical innovations that will support future theoretical development efforts. First, we embed the GenAI assistant into a realistic ERP simulation using publicly available tools, thus illustrate a new method to address concerns about ecological validity in eye-

tracking research (Balapour and Riedl, 2025). We use real-time physiological measures to capture attention allocation and attention shifts as measures of information processing behaviors, and propose them as behavioral proxies for user confidence in the GenAI assistant (Schulz and Knierim, 2024). Furthermore, we model temporal dynamics in user attention allocation and demonstrate that human-AI interaction evolves over time. By empirically validating this dynamic aspect, we highlight the importance of longitudinal effects in studying GenAI systems.

Collectively, our results empirically support our key suggestion that GenAI assistant design characteristics constitute important dimensions of human-AI interaction. We identify a design trade-off between improved understanding and increased cognitive demand, and empirically validate that more explanations are not always better in a GenAI assistant context. By introducing critical system design variables and illustrating their instantiation in an experimental design, this study significantly advances the human-AI interaction research.

2.6.3 Methodological Contributions

This study introduces several methodological innovations that can be used to advance the empirical investigation of human-AI interaction in information systems and decision sciences.

First, we present a theoretically grounded and independently manipulated experimental design that clearly distinguishes between transparency and explainability, two constructs often conflated in prior work. Drawing on established definitions from the literature (Chen et al., 2014; Vilone and Longo, 2021), we operationalize transparency as access to general information about the AI’s decision logic and explainability as post-hoc justifications for specific outputs. This conceptual distinction allows for precise hypothesis testing and provides a replicable framework for future studies.

Second, we extend the methodological scope of eye-tracking in information systems research by extending the use of two complementary attention metrics, attentional

allocation and attention shift rate, in a GenAI assistant context. Attentional allocation was measured by gaze fixation duration within predefined AOIs, corresponding to the GenAI assistant and the decision interface. Prior work in reading and problem-solving shows that selective, reduced fixation reflects fluency and familiarity (Rayner et al., 2011; Engbert et al., 2005; Dogusoy-Taylan and Cagiltay, 2014), while broader and longer fixations correlate with cognitive load and task novelty (Jacob and Karn, 2003). We adapt these findings to the GenAI context, treating reduced fixation on raw data sources as a signal of increased reliance on the assistant.

Attention shift rate captures the frequency of saccades between AOIs, serving as an index of verification and information-seeking behavior. This metric draws from research in diagnostic reasoning and decision support, where increased transitions indicate uncertainty (O'Neill et al., 2011; Tiersma et al., 2003; Burch et al., 2011). In our context, we interpret attention shifts as attempts to verify GenAI outputs, therefore consider fewer attention shifts as an indicator of a better understanding of the GenAI assistant.

Third, this study illustrates the use of business simulations such as ERPsim to conduct realistic, multi-trial experiments that expose participants to an IS use case that is continued in the same decision-making environment across the trials within the experiment. By modeling each trial as the follow-up of the previous trial, this experimental design kept the core task and the main data points consistent, while incorporating small changes across trials, such as ones happening in day-to-day real-life business scenarios. By introducing such an approach, we modeled human–AI interaction as a temporal process rather than a static decision event. Each participant completed ten sequential decision-making trials, allowing us to capture the evolution of their confidence, attention allocation, and performance. With our post-hoc analyses, we illustrated the possibility of conducting tests on the evolution of user perceptions and performance over time as a function of our experimental manipulations. Unlike typical cross-sectional decision studies, this approach may be expanded to mirror real-world continued usage where users repeatedly use an information system in the same decision-making environment and gradually form

calibrated responses.

2.6.4 Practical Implications

This study offers several actionable insights for the design of GenAI assistants in enterprise decision-support contexts.

First, minimalistic GenAI responses that omit transparency and explainability fail to enhance decision performance and may inadvertently increase cognitive workload by forcing users to verify system outputs externally. Users exposed to such designs lead to higher levels of attention shift rates, indicating uncertainty and reduced confidence in the GenAI assistant. These findings challenge the often-assumed efficiency of quick, direct GenAI outputs and call for a more deliberate approach to balancing short outputs with informative content.

Second, our results show that transparency and explainability function as imperfect substitutes. Simultaneously providing both can overwhelm users without yielding additive benefits. Thus, organizations embedding GenAI into their workflows should avoid indiscriminate inclusion of system-level and output-level explanations. Instead, system design should either prioritize transparency or explainability, depending on the individual user's past experience with that specific GenAI assistant.

On that note, transparency not only improves initial perceptions but also accelerates users' adaptation by reducing attention shift rates in later interactions. However, as user experience with the GenAI assistant increases, the importance of explainability in determining decision performance also increases. This finding highlights the value of designing GenAI assistants that apply memory-based adaptations to change the tone of conversations over time across repeated use. Our results indicate that developers can design their GenAI assistants to prioritize transparency in initial user interactions, with a gradual shift to increased explainability and reduced transparency as the user gains experience using the GenAI assistant in similar decision-making tasks.

Finally, our use of saccade-based attention metrics reveals that attention shift rate is a

meaningful proxy for user confidence in a GenAI assistant. As eye-tracking capabilities become increasingly integrated into consumer devices, there is an opportunity for GenAI systems to become more adaptive and responsive to real-time user states. Our findings illustrate the usefulness of two eye-tracking metrics that can be used in the design of such attention-aware GenAI assistants that dynamically adjust output complexity based on observed user states.

2.6.5 Limitations and Future Research

As with all empirical research, our findings should be interpreted within the boundaries of the study design. In addition, discussing these limitations is important as they also offer meaningful directions for future inquiry.

First, the responses produced by our GenAI assistant were based on a fixed configuration of model temperature and randomness parameters. Although we configured the assistant to produce consistent and appropriate output structures, different parameter settings in different decision-making contexts might yield more uniform outputs. Future research should systematically examine how controllable system hyperparameters, e.g., temperature, top-p sampling, can be used to influence user understanding, cognitive load, and confidence in the GenAI assistant over time.

Second, the decision task used in our study was grounded in a well-established experimental paradigm in the behavioral decision-making literature. While this setup streamlined our study's experiment design process, it reflects a single type of structured decision-making context. Further studies should extend this line of inquiry to organizational decisions involving other knowledge work tasks such as knowledge generation and knowledge integration.

Third, our experimental design required participants to submit decisions using a decision interface while receiving decision support recommendations from a separate GenAI assistant interface. This two-interface setup was a deliberate methodological choice to facilitate the clear measurement of user behavior, particularly attentional shifts

(via eye-tracking), between direct task engagement and interaction with the GenAI. This visual design choice mirrors several real-world GenAI implementations and interaction patterns where users consult an assistant alongside primary task interfaces or verify GenAI outputs against distinct information sources (e.g., GitHub Copilot's separate panel on Microsoft Visual Studio Code, SAP Joule's prototype illustrations where it shows links to source data, or features in Google Gemini and OpenAI ChatGPT that link to external information). We therefore believe our experimental environment has relevance for many GenAI interaction scenarios involving decision-making. However, it is important to note that because our specific eye-tracking metrics (such as saccade counts and fixation durations on distinct interface areas) depended on this two-interface design, their direct applicability to GenAI implementations that serve as a sole user interface for both information access and decision submission may be limited. Regardless of this measurement consideration, we maintain that the fundamental cognitive switches involved in managing an agentic entity through natural language as part of an additional cognitive layer are likely to persist even in fully integrated GenAI environments. Future research could explore how these cognitive switches manifest and can be effectively measured in single-interface GenAI decision support systems.

Fourth, confidence in the GenAI assistant was measured using a single-item indicator to enable repeated assessment across trials while minimizing participant fatigue. Although appropriate for capturing dynamic, trial-level evaluations of perceived system ability, this operationalization may not fully capture the broader range of cognitive evaluations users form when assessing AI systems. Future research should employ more comprehensive multi-item instruments to examine additional dimensions of users' assessment of system ability and investigate how richer confidence measurements evolve across extended human–AI interaction contexts. In addition, confidence was measured only prior to the performance report. Future work should examine differences between confidence assessed before observing objective outcomes and confidence measured after outcome feedback, to better understand how performance feedback influences user confidence in AI systems.

2.7 Conclusion

This study examined how a GenAI assistant’s transparency and explainability influence user understanding, attention, and decision performance in enterprise decision-support contexts. Grounded in the Situation Awareness (SA) model, we theorized and empirically tested how these design features affect users’ perceptions and behaviors across repeated interactions. By embedding a GenAI assistant into a realistic ERP simulation and using eye-tracking to monitor users’ information processing behaviors, we offer a framework for designing effective human-AI interaction in dynamic enterprise environments. Our findings show that transparency improves initial understanding and reduces verification effort over time, while explainability starts to positively influence decision performance in later interactions. Moreover, we find that transparency and explainability are imperfect substitutes; thus, effective GenAI assistant designs should prioritize transparency during early use and gradually shift toward explainability to optimize user outcomes over time.

References

- Agrawal, N., Cohen, M. A., Deshpande, R., and Deshpande, V. (2024). How machine learning will transform supply chain management. *Harvard Business Review*, 103(3-4):128–137.
- Allen, R. and Choudhury, P. R. (2022). Algorithm-augmented work and domain experience: the countervailing forces of ability and aversion. *Organization Science*, 33(1):149–169.
- Anaraky, R. G., Knijnenburg, B. P., and Risius, M. (2020). Exacerbating mindless compliance: the danger of justifications during privacy decision making in the context of facebook applications. *AIS Transactions on Human-Computer Interaction*, 12(2):70–95.
- Ang, L. and Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research*, 58(2):218–227.
- Balapour, A. and Riedl, R. (2025). Ecological validity in neuroIS research: Theory, evidence, and a roadmap for future studies. *Journal of the Association for Information Systems*, 26(1):9–65.
- Balslev, T., Jarodzka, H., Holmqvist, K., de Grave, W., Muijtens, A. M., Eika, B., van Merriënboer, J., and Scherpbier, A. J. (2012). Visual expertise in paediatric neurology. *European Journal of Paediatric Neurology*, 16(2):161–166.
- Banh, L. and Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1):63.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on

- complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.
- Barki, H., Robert, J., and Dulipovici, A. (2015). Reconceptualizing trust: A non-linear boolean model. *Information & Management*, 52(4):483–495.
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.
- Bauer, K., Hinz, O., van der Aalst, W., and Weinhardt, C. (2021). Explaining it to me—explainable AI and information systems research. *Business & Information Systems Engineering*, 63:79–82.
- Beck, D. M. and Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, 49(10):1154–1165.
- Benzion, U., Cohen, Y., Peled, R., and Shavit, T. (2008). Decision-making and the newsvendor problem: an experimental study. *Journal of the Operational Research Society*, 59(9):1281–1287.
- Bergkvist, L. and Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, 44(2):175–184.
- Bolton, G. E. and Katok, E. (2008). Learning by doing in the newsvendor problem: a laboratory investigation of the role of experience and feedback. *Manufacturing & Service Operations Management*, 10(3):519–538.
- Brunyé, T. T. and Gardony, A. L. (2017). Eye tracking measures of uncertainty during perceptual decision making. *International Journal of Psychophysiology*, 120:60–68.

- Budhwar, P., Chowdhury, S., Wood, G., Aguinis, H., Bamber, G. J., Beltran, J. R., Boselie, P., Lee Cooke, F., Decker, S., DeNisi, A., Dey, P. K., Guest, D., Knoblich, A. J., Malik, A., Paauwe, J., Papagiannidis, S., Patel, C., Pereira, V., Ren, S., Rogelberg, S., Saunders, M. N. K., Tung, R. L., and Varma, A. (2023). Human resource management in the age of generative artificial intelligence: Perspectives and research directions on chatGPT. *Human Resource Management Journal*, 33(3):606–659.
- Burch, M., Konevtsova, N., Heinrich, J., Höferlin, M., and Weiskopf, D. (2011). Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2440–2448.
- Bussone, A., Stumpf, S., and O’Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169. IEEE.
- Cain, S. (2022). A.I. is getting your holiday gifts to you more efficiently than ever. <https://fortune.com/2022/12/20/tech-forward-everyday-ai-holiday-shipping-ups-fedex-amazon/>. Accessed: 2024-01-02.
- Camerer, C. F. and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, 19(1):7–42.
- Cao, S. and Huang, C.-M. (2022). Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–23.
- Card, S., MORAN, T., and Newell, A. (1986). The model human processor- an engineering model of human performance. *Handbook of Perception and Human Performance.*, 2(45–1):1–35.
- Carrasco, M. (2011). Visual attention: the past 25 years. *Vision Research*, 51(13):1484–1525.

- Chandrasekaran, A., Mullen, A., Kune, L. F., Greene, N., Hare, J., Ramos, L., and Verma, A. (2024). Predicts 2024: the future of generative AI technologies. <https://www.gartner.com/document-reader/document/5236863>. Accessed: 2025-04-20.
- Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M. (2014). Situation awareness-based agent transparency. Technical report, Defense Technical Information Center, Fort Belvoir, VA.
- Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., and Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018.
- Christensen-Szalanski, J. J. and Beach, L. R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance*, 29(2):270–278.
- Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, 16(1):64–73.
- Compeau, D. R. and Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 19(2):189–211.
- Coursaris, C. (2025). The role of situation awareness in HCAI: an integrated perspective. <https://www.youtube.com/watch?v=hJNCs4o66mU>. Accessed: 2025-09-23.
- Davis, F. D., Lohse, G. L., and Kottemann, J. E. (1994). Harmful effects of seemingly helpful information on forecasts of stock earnings. *Journal of Economic Psychology*, 15(2):253–267.
- De Guinea, A. O., Titah, R., and Léger, P.-M. (2013). Measure for measure: A two study multi-trait multi-method investigation of construct validity in is research. *Computers in Human Behavior*, 29(3):833–844.

- De Vericourt, F., Jain, K., Bearden, J. N., and Filipowicz, A. (2013). Sex, risk and the newsvendor. *Journal of Operations Management*, 31(1-2):86–92.
- Di Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., and Zahidi, S. (2023). Future of jobs report 2023. World Economic Forum.
- Diederich, S., Brendel, A. B., and Kolbe, L. M. (2020). Designing anthropomorphic enterprise conversational agents. *Business & Information Systems Engineering*, 62:193–209.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.
- Dogusoy-Taylan, B. and Cagiltay, K. (2014). Cognitive analysis of experts’ and novices’ concept mapping processes: an eye tracking study. *Computers in Human Behavior*, 36:82–93.
- Dohmke, T. (2023). Github copilot for business is now available. <https://github.blog/2023-02-14-github-copilot-for-business-is-now-available/>. Accessed: 2023-11-25.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64.
- Endsley, M. R. (2021). Situation awareness in teams: Models and measures. In McNeese, M., Salas, E., and Endsley, M., editors, *Handbook of Distributed Team Cognition: Contemporary Research Models, Methodologies, and Measures in Distributed Team Cognition*, pages 1–28. CRC Press, Boca Raton, FL.
- Endsley, M. R. (2023). Supporting Human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140:107574.

- Endsley, M. R., Bolté, B., and Jones, D. G. (2003). *Designing for Situation Awareness: an Approach to User-centered Design*. CRC press.
- Engbert, R., Nuthmann, A., Richter, E. M., and Kliegl, R. (2005). Swift: a dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777.
- Ericsson, K. A. and Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, 47(1):273–305.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1):62–70.
- Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191.
- Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1):111–126.
- FoodyDirect (2014). Proof we like ice cream all year long. <https://web.archive.org/web/20140727192556/http://blog.foodydirect.com/index.php/proof-like-ice-cream-year-long/>. Accessed: 2023-03-12.
- Gabriele Rigon, Leinar Ramos, Bern Elliot, and Anthony Mullen (2024). Lessons from generative AI early adopters. <https://www.gartner.com/document-reader/document/5695719>. Accessed: 2025-04-20.
- Garcia, S. M., Tor, A., and Schiff, T. M. (2013). The psychology of competition: A social comparison perspective. *Perspectives on Psychological Science*, 8(6):634–650.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873.

- Gnewuch, U., Morana, S., Adam, M. T., and Maedche, A. (2022). Opposing effects of response time in human–chatbot interaction: the moderating role of prior experience. *Business & Information Systems Engineering*, 64(6):773–791.
- Gupta, K., Hajika, R., Pai, Y. S., Duenser, A., Lochner, M., and Billinghamurst, M. (2019). In AI we trust: Investigating the relationship between biosignals, trust and cognitive load in vr. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, pages 1–10.
- Hamari, J., Koivisto, J., and Sarsa, H. (2014). Does gamification work?—a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*, pages 3025–3034. Ieee.
- Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.
- Healey, C. and Enns, J. (2011). Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188.
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2023). Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5.
- Hooge, I. T. C. and Erkelens, C. J. (1996). Control of fixation duration in a simple search task. *Perception & Psychophysics*, 58:969–976.
- Jacob, R. J. and Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The Mind’s Eye*, pages 573–605. Elsevier.
- Jacobs, A. M. and O’Regan, J. K. (1987). Spatial and/or temporal adjustments of scanning behavior to visibility changes. *Acta Psychologica*, 65(2):133–146.

- Janssen, A., Passlick, J., Rodríguez Cardona, D., and Breitner, M. H. (2020). Virtual assistance in any context: a taxonomy of design elements for domain-specific chatbots. *Business & Information Systems Engineering*, 62:211–225.
- Jiang, E., Toh, E., Molina, A., Olson, K., Kayacik, C., Donsbach, A., Cai, C. J., and Terry, M. (2022). Discovering the syntax and strategies of natural language programming with generative language models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, New Orleans LA USA. ACM.
- Jiang, J., Karran, A. J., Coursaris, C. K., Léger, P.-M., and Beringer, J. (2023). A situation awareness perspective on human-AI interaction: Tensions and opportunities. *International Journal of Human–Computer Interaction*, 39(9):1789–1806.
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.
- Kesavan, S. and Kushwaha, T. (2020). Field experiment on the profit implications of merchants’ discretionary power to override data-driven decision-making tools. *Management Science*, 66(11):5182–5190.
- Klaus, H., Rosemann, M., and Gable, G. G. (2000). What is ERP? *Information Systems Frontiers*, 2:141–162.
- Kosel, C., Holzberger, D., and Seidel, T. (2021). Identifying expert and novice visual scanpath patterns and their relationship to assessing learning-relevant student characteristics. *Frontiers in Education*, 5:612175.
- Kremer, M. and Minner, S. (2008). The human element in inventory decision making under uncertainty-a review of experimental evidence in the newsvendor model. *Zeitschrift für Betriebswirtschaft*, pages 83–97.

- Kremer, M., Minner, S., and Van Wassenhove, L. N. (2010). Do random errors explain newsvendor behavior? *Manufacturing & Service Operations Management*, 12(4):673–681.
- Kruschel, S., Hambauer, N., Weinzierl, S., Zilker, S., Kraus, M., and Zschech, P. (2025). Challenging the performance-interpretability trade-off: an evaluation of interpretable machine learning models. *Business & Information Systems Engineering*, pages 1–25.
- Kunkel, J., Donkers, T., Michael, L., Barbu, C.-M., and Ziegler, J. (2019). Let me explain: Impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.
- Lai, V. and Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38.
- Landwehr, J. P., Kühl, N., Walk, J., and Gnädig, M. (2022). Design knowledge for deep-learning-enabled image-based decision support systems: Evidence from power line maintenance decision-making. *Business & Information Systems Engineering*, 64(6):707–728.
- Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.
- Lefebvre, E., Cassivi, L., Lefebvre, L. A., Léger, P.-M., and Hadaya, P. (2003). Supply chain management, electronic collaboration tools and organizational innovativeness. *Journal on Chain and Network Science*, 3(2):81–94.
- Léger, P., Robert, J., Babin, G., Pellerin, R., and Wagner, B. (2007). Erpsim. *ERPsim Lab* (erpsim.hec.ca), HEC Montreal, Montreal, Qc.
- Léger, P.-M. (2006). Using a simulation game approach to teach enterprise resource planning concepts. *Journal of Information Systems Education*, 17:441–447.

- Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.
- Loos, P., Riedl, R., Müller-Putz, G. R., Vom Brocke, J., Davis, F. D., Banker, R. D., and Léger, P.-M. (2010). neuroIS: Neuroscientific approaches in the investigation and development of information systems. *Business & Information Systems Engineering*, 2:395–401.
- Lukyanenko, R., Evermann, J., and Parsons, J. (2014). Instantiation validity in is design research. In *Advancing the Impact of Design Science: Moving From Theory to Practice: 9th International Conference, DESRIST 2014, Miami, FL, USA, May 22-24, 2014. Proceedings 9*, pages 321–328. Springer.
- Lukyanenko, R., Maass, W., and Storey, V. C. (2022). Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities. *Electronic Markets*, 32(4):1993–2020.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., and Söllner, M. (2019). AI-based digital assistants: Opportunities, threats, and research perspectives. *Business & Information Systems Engineering*, 61:535–544.
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.
- McKenzie, S., Podsakoff, P., and Podsakoff, N. (2011). Construct measurement and validation procedures in mis and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(2):293–334.

- Mediavilla, M. A., Dietrich, F., and Palm, D. (2022). Review and analysis of artificial intelligence methods for demand forecasting in supply chain management. *Procedia CIRP*, 107:1126–1131.
- Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., and Shirase, L. (2019). Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology*, 10:225.
- Microsoft (2024). Introducing Microsoft 365 Copilot – your copilot for work. <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>. Accessed: 2025-03-10.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229.
- Monden, Y. (2011). *Toyota Production System: an Integrated Approach to Just-in-time*. Productivity Press.
- Nagendran, M., Festor, P., Komorowski, M., Gordon, A. C., and Faisal, A. A. (2024). Eye tracking insights into physician behaviour with safe and unsafe explainable AI recommendations. *NPJ Digital Medicine*, 7(1):202.
- Nair, D. and Saenz, M. J. (2024). Pair people and AI for better product demand forecasting. *MIT Sloan Management Review*. URL: <https://sloanreview.mit.edu/article/pair-people-and-ai-for-better-product-demand-forecasting/>. Accessed: 2024-04-07.
- National Bureau of Economic Research (2012). Ice Cream Production for United States [M0174BUSM411NNBR]. <https://fred.stlouisfed.org/series/M0174BUSM411NNBR>. Accessed: March 14, 2025.
- O’Neill, E. C., Kong, Y. X. G., Connell, P. P., Ong, D. N., Haymes, S. A., Coote, M. A., and Crowston, J. G. (2011). Gaze behavior among experts and trainees during optic

- disc examination: Does how we look affect what we see? *Investigative Ophthalmology & Visual Science*, 52(7):3976–3983.
- Oz, B. and Léger, P.-M. (2024). *Conducting Scientific Research with ERPsims: Research Guidelines*. ERPsims Lab, HEC Montréal.
- Oz, B., Tran-Nguyen, K., Coursaris, C. K., Robert, J., and Léger, P.-M. (2020). Using digital nudges on analytics dashboards to reduce anchoring bias. In *SIGHCI 2020 Proceedings*, number 3.
- Pettit, T. J., Croxton, K. L., and Fiksel, J. (2019). The evolution of resilience in supply chain management: a retrospective on ensuring supply chain resilience. *Journal of Business Logistics*, 40(1):56–65.
- Pfeuffer, N., Baum, L., Stammer, W., Abdel-Karim, B. M., Schramowski, P., Bucher, A. M., Hügel, C., Rohde, G., Kersting, K., and Hinz, O. (2023). Explanatory interactive machine learning: Establishing an action design research process for machine learning projects. *Business & Information Systems Engineering*, 65(6):677–701.
- Pfeuffer, N., Benlian, A., Gimpel, H., and Hinz, O. (2019). Anthropomorphic information systems. *Business & Information Systems Engineering*, 61:523–533.
- Phillips, M. H. and Edelman, J. A. (2008). The dependence of visual scanning performance on search direction and difficulty. *Vision Research*, 48(21):2184–2192.
- Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52.
- Prahl, A. and Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6):691–702.

- Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., and Santos, E. A. (2023). “It’s weird that it knows what I want”: Usability and interactions with copilot for novice programmers. In *ACM Transactions on Computer-Human Interaction*, volume 31, pages 4:1–4:31.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372.
- Rayner, K., Slattery, T. J., Drieghe, D., and Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):514.
- Reynolds, J. H. and Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27(1):611–647.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.
- Riedl, R. and Léger, P.-M. (2016). *Fundamentals of NeuroIS: Information Systems and the Brain*. Springer, Germany.
- Rolon-Merette, T., Guo, E. Y., Poivré, F., Benboust, M., Amoros, S., Karran, A., Coursaris, C., Senecal, S., and Leger, P.-M. (2025). Generative AI summaries do not increase user engagement: A study of Google AI overviews. In *SIGHCI 2024 Proceedings*, number 29.
- Roy, A., Paquet, R., Petty, M., Hunter, E., and Roussel, A. (2023). Tech ceos: Generative AI can make sales and marketing more efficient. <https://www.gartner.com/document/4869731>. Accessed: 2024-09-07.
- SAP (2024). Joule AI agents | collaborative AI agents. <https://www.sap.com/products/artificial-intelligence/ai-agents.html>. Accessed: 2025-03-10.

- Schulz, T. and Knierim, M. T. (2024). Cognitive load dynamics in generative ai-assistance: a NeuroIS study. In *ICIS 2024 Proceedings*, number 12.
- Schweitzer, M. E. and Cachon, G. P. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science*, 46(3):404–420.
- Sebastian, S. and Draper, J. (2024). How AI can improve consumer product and retail experiences. <https://www.sdcexec.com/software-technology/ai-ar/article/22891372/capgemin-how-ai-can-improve-consumer-product-and-retail-experiences>. Accessed: 2024-04-07.
- Sellen, A. and Horvitz, E. (2024). The rise of the AI co-pilot: Lessons for design from aviation and beyond. *Communications of the ACM*, 67(7):18–23.
- Seltman, H. J. (2012). *Experimental Design and Analysis*. Carnegie Mellon University Pittsburgh.
- Shareef, A. and Picek, R. (2024). Enhancing business insights: AI based chat toolset for erp systems. In *2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–6.
- Sharma, R., Shishodia, A., Gunasekaran, A., Min, H., and Munim, Z. H. (2022). The role of artificial intelligence in supply chain management: Mapping the territory. *International Journal of Production Research*, 60(24):7527–7550.
- Shiferaw, B., Downey, L., and Crewther, D. (2019). A review of gaze entropy as a measure of visual scanning efficiency. *Neuroscience & Biobehavioral Reviews*, 96:353–366.
- Simkute, A., Tankelevitch, L., Kewenig, V., Scott, A. E., Sellen, A., and Rintel, S. (2024). Ironies of generative ai: Understanding and mitigating productivity loss in human-AI interactions. arXiv:2402.11364 [cs].

- Stevens, G. C. (1989). Integrating the supply chain. *International Journal of Physical Distribution & Materials Management*, 19(8):3–8. publisher: MCB UP Ltd.
- Stumpf, S., Rajaram, V., Li, L., Wong, W.-K., Burnett, M., Dietterich, T., Sullivan, E., and Herlocker, J. (2009). Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662.
- Suda, N. and LeHong, H. (2025). The 3 business cases of generative AI value. <https://www.gartner.com/document-reader/document/6055563>. Accessed: 2025-09-25.
- Söllner, M., Arnold, T., Benlian, A., Bretschneider, U., Knight, C., Ohly, S., Rudkowski, L., Schreiber, G., and Wendt, D. (2025). ChatGPT and beyond: Exploring the responsible use of generative AI in the workplace. *Business & Information Systems Engineering*, 67(2):289–303.
- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., and Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–24, New York, NY, USA. Association for Computing Machinery.
- Tiersma, E. S. M., Peters, A. A., Mooij, H. A., and Fleuren, G. J. (2003). Visualising scanning patterns of pathologists in the grading of cervical intraepithelial neoplasia. *Journal of Clinical Pathology*, 56(9):677–680.
- Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., Pearson, G., and Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, 1(4).
- Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., and Fischl, M. (2021). Artificial intelligence in supply chain management: a systematic literature review. *Journal of Business Research*, 122:502–517.

- Toreini, P., Langner, M., Maedche, A., Morana, S., and Vogel, T. (2022). Designing attentive information dashboards. *Journal of the Association for Information Systems*, 23(2):521–552.
- Vasseur, A., Passalacqua, M., Sénécal, S., and Léger, P.-M. (2023). The use of eye-tracking in information systems research: a literature review of the last decade. *AIS Transactions on Human-Computer Interaction*, 15(3):292–321.
- Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.
- Wang, Y., Xu, M., and Zhao, L. (2025). Exploring the cognitive mechanisms behind the adoption of algorithmic advice: A neuro-is study of newsvendor decision-making. *Business & Information Systems Engineering*.
- Wanous, J. P., Reichers, A. E., and Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, 82(2):247.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, 14(4):490–495.
- Wickens, C. D., Helton, W. S., Hollands, J. G., and Banbury, S. (2021). *Engineering Psychology and Human Performance*. Routledge.
- Wong, L.-W., Tan, G. W.-H., Ooi, K.-B., Lin, B., and Dwivedi, Y. K. (2022). Artificial intelligence-driven risk management for enhancing supply chain agility: a deep-learning-based dual-stage PLS-SEM-ANN analysis. *International Journal of Production Research*, pages 1–21.

Yavuz, S., Pérez, D., and Serdiukova, V. (2024). xGen-Sales: Next-Gen AI for sales cloud. <https://www.salesforce.com/blog/xgen-sales/>. Accessed: 2025-03-10.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305.

Chapter 3

User Control in Human-AI Interaction: How AI Configuration Shapes User Attention and Error Responses

Abstract

As artificial intelligence (AI) systems become increasingly integrated into knowledge-intensive decision tasks, it is critical to understand how interaction design shapes users' attentional processes and information-processing efficiency. Prior research has emphasized transparency and explainability as mechanisms to improve user comprehension of AI recommendations, yet their effectiveness depends on users' attentional allocation and remains empirically inconclusive. Rather than focusing on explanations, this study

This chapter includes an original and unpublished research work conducted by Burak Öz; with contributions from Dr. Alexander J. Karran, Dr. Jared Boasen, Dr. Constantinos K. Coursaris, and Dr. Pierre-Majorique Léger.

Burak Öz served as the principal author, responsible for the design, execution, analysis, and writing. Dr. Alexander J. Karran and Dr. Jared Boasen contributed to the experimental design and reviewed the manuscript. Dr. Constantinos K. Coursaris and Dr. Pierre-Majorique Léger provided feedback and guidance throughout all stages of the research.

examines whether providing access to AI configuration, a design affordance intended to give users a sense of influence over how the AI generates its recommendation, improves attention, information-processing efficiency, error detection, and decision performance. In a mixed-factorial NeuroIS experiment, 48 participants completed 76 consecutive AI-assisted demand-forecasting tasks in a business simulation. Participants were assigned to either a configuration condition, where they set parameters they believed shaped the forthcoming AI recommendation, or a no-configuration condition. From the user's perspective, configuring these parameters offered an opportunity to influence the AI's output. However, despite one group's configuration input, all participants within this study received identical AI-generated recommendations, ensuring that any between-condition differences reflected the perceptual impact of the configuration affordance rather than actual variation in model behavior. In addition, AI accuracy was manipulated within participants to examine responses to erroneous recommendations. Using eye-tracking, electroencephalography (EEG), behavioral data, and self-reports, the study finds that access to configuration increased perceived autonomy, strengthened attentional responses to AI errors, reduced visual verification durations, and enhanced users' confidence and understanding of the AI assistant. Participants with configuration access also recovered confidence more rapidly following erroneous recommendations. These findings highlight the importance of supporting users' autonomy by providing them with affordances to be involved into AI-assisted decision interfaces to sustain attention, improve processing efficiency, and stabilize confidence in AI.

3.1 Introduction

The capabilities of artificial intelligence (AI) systems continue to grow, enabling organizations to automate an increasing number of decision-making tasks across industries (Currie et al., 2019; Faraj et al., 2018; Fourcade and Khonsari, 2019; Giger, 2018). These systems are widely adopted to reduce costs, streamline operations, and enhance decision quality in business environments. As AI systems take on

roles traditionally held by human experts, human-AI interaction becomes central to many organizational workflows. Such dynamics are especially critical in high-stakes, knowledge-intensive domains like supply chain management, where AI recommendations increasingly shape operational decisions (Toorajipour et al., 2021).

Despite these advances, the notion of fully autonomous AI remains unrealistic in many complex domains due to ethical, social, and operational concerns regarding the use of algorithmic decision-making (Akhtar and Mian, 2018; Anderson, 2019; Ferrer et al., 2021). As a result, hybrid human-AI systems have emerged, combining the analytical power of AI with human judgment, particularly in tasks that require both structured analytics approaches and interpretation of context-specific unstructured cues (Endsley, 2023; Rai et al., 2019).

A key challenge in human-AI decision-making is calibrating user confidence in AI recommendations relative to these systems' actual accuracy. Past evidence indicates that users often miscalibrate their confidence in AI-generated recommendations. Some undervalue algorithmic input despite its accuracy, e.g., algorithm aversion (Dietvorst et al., 2015; Allen and Choudhury, 2022; Kesavan and Kushwaha, 2020), while others depend on AI outputs without sufficient scrutiny, e.g., algorithm complacency (Merritt et al., 2019; Anaraky et al., 2020; Liu, 2024; Parasuraman and Manzey, 2010). Both patterns of miscalibrated confidence in AI lead to reduced decision quality and undermine the potential of AI implementations in organizations (Dwivedi et al., 2021; Faraj et al., 2018).

To address this challenge of calibrating user confidence in AI assistants, two primary design strategies have been proposed to improve user perceptions and behaviors in human-AI interaction: increasing system transparency through explanations (Jiang et al., 2023), and enhancing user control and perceived autonomy in human-AI interactions (Haesevoets et al., 2021; Herrmann, 2025; Westphal et al., 2023). While explanation-based approaches aim to improve understanding by revealing how AI systems arrive at their recommendations, findings on their effectiveness are inconclusive. Some studies report that explanations improve user trust and compliance, while others find that they

increase task complexity, cognitive load, or even lead to negative side effects like longer decision times and reduced engagement (Bansal et al., 2021; Cecil et al., 2024; Giboney et al., 2015; Kizilcec, 2016; Zhang et al., 2020). These mixed results suggest that explanations alone may not be sufficient to improve user outcomes, especially when they are unclear, overly complex, or misaligned with user needs (Ahn et al., 2021; Kizilcec, 2016; Nunes et al., 2012).

On the other hand, user control in human–AI interaction has typically been operationalized as allowing users to edit or override an AI’s recommended action before execution (Westphal et al., 2023). Far less is known about forms of control that give users a greater sense of autonomy in the interaction and allow them to influence how the AI generates its recommendation, such as configuring parameters to shape the perceived relevance or quality of the output. Moreover, prior work on user control has largely examined tasks in which AI outputs are accurate (Burton et al., 2020; Dietvorst et al., 2018; Westphal et al., 2023). In practice, however, real-world AI systems inevitably produce errors due to shifting contexts, limited data, or imperfect reasoning. Little is known about how users respond when AI systems make such mistakes, or how giving users an increased perception of autonomy in shaping the AI’s behavior affects their reactions to erroneous recommendations.

Drawing on the situation awareness (SA) model and dual-systems theory (Endsley, 1995; Kahneman, 2011), we examine whether giving users the ability to configure aspects of the AI’s recommendation process, a manipulation intended to create a greater sense of autonomy and involvement, shapes the attentional and cognitive processes that underlie confidence calibration in human–AI interaction. This perceived autonomy may prompt users to engage more actively with the task, sustain attention during verification, and respond more effectively when the AI produces erroneous recommendations. Accordingly, we investigate whether configuration affects information-processing efficiency, confidence calibration, and error detection. More specifically, we address the following research questions:

RQ1: How does giving users the ability to configure the AI system’s parameters

at the beginning of the decision process influence their calibration of confidence in AI recommendations across varying levels of accuracy?

RQ2: How does giving users the ability to configure the AI system's parameters influence their attentional responses during AI-assisted decision making?

RQ3: How does giving users the ability to configure the AI system's parameters affect users' detection and handling of erroneous AI recommendations?

We investigate these questions in a mixed-factorial laboratory experiment involving a demand forecasting task completed over 76 trials. Participants were randomly assigned to one of two groups: a configuration group given the ability to adjust AI parameters intended to create a sense of influence over the AI, and a no-configuration group. AI accuracy was manipulated within participants, so that all participants received identical numeric AI recommendations, except for the randomly assigned 15 trials with erroneous AI recommendations to simulate prediction errors. Throughout the experiment, we recorded performance outcomes and real-time physiological measures (electroencephalography (EEG) and eye-tracking) to capture users' attentional responses and processes.

Our work contributes to the empirical literature on human–AI interaction in three ways. First, we test whether giving users the ability to configure aspects of the AI, thereby shaping their perception of autonomy in their AI interaction, improves the calibration of their confidence as the system's accuracy varies. Second, we incorporate physiological measures that capture real-time attentional responses to AI recommendations, complementing self-reported perceptions. Third, we extend current research by examining user behavior under imperfect AI performance, responding to calls to study how users adapt when AI systems produce errors.

The remainder of this paper is structured as follows. Section 2 reviews prior research on human-AI interactions in organizational decision-making. Section 3 presents our experimental design and measures. Section 4 reports the results. Finally, Section 5 discusses theoretical and practical implications, along with future research directions.

3.2 Background

3.2.1 NeuroIS Research

NeuroIS is an interdisciplinary field that integrates neuroscience methods into information systems (IS) research to examine users' internal states and behaviors in response to IS stimuli. A commonly adopted theoretical framework in this field is the Stimulus-Organism-Response (SOR) model, which conceptualizes user behavior as a function of environmental inputs (stimuli), internal cognitive-affective processing (organism), and observable behavior (response) (Dimoka et al., 2012; Riedl et al., 2020).

In this framework, IS artifacts and system features serve as stimuli, triggering organismic responses such as changes in attention or emotions, which then influence attitudes and behavioral responses (see Figure 3.1). Neurophysiological methods, including eye-tracking, electroencephalography (EEG), fMRI, and skin conductance, are employed to capture the organismic layer in this process. These techniques offer access to unconscious and rapid user responses that are often inaccessible through self-report instruments (Dimoka et al., 2011; Kirwan et al., 2023).

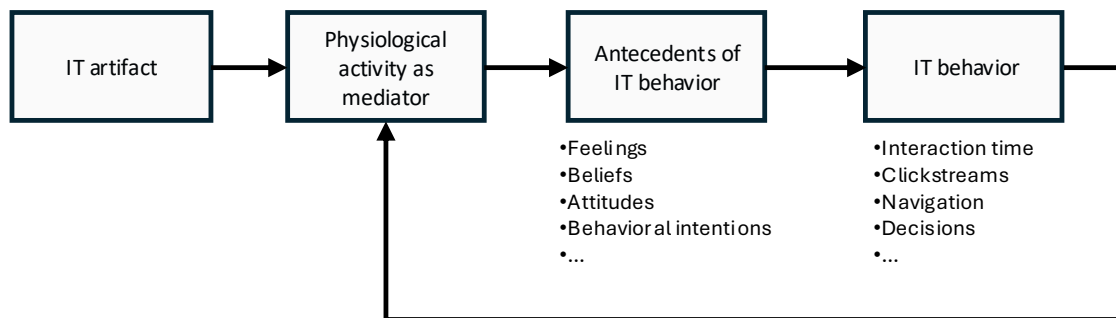


Figure 3.1: The role of physiological measures in IS research. Adapted from (Riedl and Léger, 2016).

Two widely used tools in this domain are eye-tracking and EEG (Riedl et al., 2020). Eye-tracking has been employed to examine attention allocation and information processing behaviors in response to IS interfaces (Lin et al., 2016; Ahn et al., 2018), while EEG has been used to monitor cognitive processes such as attentional engagement

and cognitive load (Anderson et al., 2016; Karran et al., 2019; Wang et al., 2025). The advantages of eye-tracking and EEG become particularly important in longitudinal or repeated-task settings, where participant responses may fluctuate as a function of trial-level within-participant manipulations (Lu and Sarter, 2019; Wang et al., 2024; Zhang et al., 2024). In such contexts, relying only on self-reported or behavioral data to assess how users processed AI errors and understood its boundary conditions is challenging, as users may not be consciously aware of their detection process, may follow different cognitive pathways to achieve similar behavioral outcomes, or may retrospectively rationalize their behavior in self-reports. Accordingly, this study employs physiological measures to gain deeper insight into users' attentional and cognitive processes when evaluating AI recommendations.

This study contributes to the NeuroIS literature by responding to recent calls for improved ecological validity in the NeuroIS literature (Balapour and Riedl, 2025). Although conducted in a controlled environment, it incorporates a relevant task paradigm from the operations research literature within an enterprise resource planning (ERP) system interface to create an event-related potentials experiment. With these design choices, this study aims to improve our understanding of cognitive and attentional processes in AI-assisted decision-making.

3.2.2 Human-AI Interaction in Organizational Decision-Making

Although AI systems have demonstrated remarkable capabilities in automating analytical and decision-support tasks, they bring significant challenges that arise from their complexity and opacity, a phenomenon often referred to as the "black box" problem of AI (Adadi and Berrada, 2018). As these systems are implemented across domains and areas of daily life, both AI experts and non-experts interact with them (Ribera and Lapedriza, 2019). However, AI models continue to face challenges such as hallucinations, context misinterpretation, and poor generalization to novel or ambiguous scenarios due to gaps in training data (Banh and Strobel, 2023; Faraj et al., 2018; Feuerriegel et al., 2024; Pettit

et al., 2019; Söllner et al., 2025).

In response to such limitations, the human-in-the-loop paradigm is often implemented, where AI is positioned as a decision-support assistant operating under human oversight rather than as a fully autonomous agent (Karran et al., 2019). This design approach assumes that human judgment provides critical inputs, such as contextual awareness, ethical reasoning, and adaptive expertise, that AI systems cannot yet fully replicate (Faraj et al., 2018). Nevertheless, empirical evidence reveals conflicting patterns in how users engage with AI recommendations. On one hand, users may undervalue algorithmic input, exhibiting algorithm aversion (Dietvorst et al., 2015; Allen and Choudhury, 2022; Kesavan and Kushwaha, 2020). On the other hand, especially as system performance improves, users may excessively rely on AI-generated decisions, a phenomenon known as algorithm complacency (Merritt et al., 2019; Anaraky et al., 2020; Liu, 2024; Parasuraman and Manzey, 2010).

The Situation Awareness (SA) model (Endsley, 1995; Endsley et al., 2003; Endsley, 2023) provides a suitable theoretical foundation for understanding human-AI interaction dynamics (Jiang et al., 2023). Originally developed to explain human performance in high-stakes environments such as aviation, the SA model conceptualizes human cognition in three hierarchical levels: perception of relevant cues (Level 1), comprehension of their meaning (Level 2), and projection of future system states (Level 3). In human-AI interaction, users must not only perceive and interpret the AI's outputs but also anticipate how the system will behave under changing conditions. Failures at any level of SA, whether due to insufficient perception of system output, poor comprehension of model logic, or inaccurate projection of outcomes, can lead to both algorithm aversion and complacency. Accordingly, understanding how AI design influences users' attentional processes and confidence is essential for sustaining appropriate reliance on AI systems.

Algorithm complacency has been linked to both individual and system-level factors. Individuals with limited domain knowledge, reduced situation awareness, or lower self-efficacy may feel less able to question or override AI outputs (Endsley, 2023). System designs that obscure model uncertainty or fail to communicate boundary conditions can

further reinforce blind trust in automation (Parasuraman and Manzey, 2010). At the same time, design features intended to enhance user understanding, such as detailed explanations in the form of system transparency, can also increase algorithm complacency by encouraging unwarranted confidence in the system (Anaraky et al., 2020; Bowden et al., 2023). Reduced human agency in AI-supported decision-making has also been linked to algorithm compliance challenges. As the level of automation increases in decision support workflows, the lack of user control in decision-making can cause humans to lose vigilance and agency, thereby increasing algorithm complacency (Ferraro and Mouloua, 2021; Merritt et al., 2019; Parasuraman and Manzey, 2010). This issue is further amplified as reduced autonomy and increased overreliance lead users to allocate less attention to the automated decision-making process, eventually increasing error rates (Onnasch et al., 2014; Parasuraman and Manzey, 2010). A potential strategy to increase users' attentional engagement and help maintain their SA is to design systems that enhance human agency in the decision-making process (Glikson and Woolley, 2020; Hancock et al., 2020; Jiang et al., 2023).

This study aims to address these calls for research on the impact of user autonomy on how users allocate attention, detect AI errors, and adjust their reliance on the system across repeated interactions. It does so by providing users with a configuration interface designed to create the perception that they can influence how the AI generates its recommendation. This configuration affordance is intended to increase users' perceived autonomy, thereby encouraging more active engagement with SA processes across perception, comprehension, and projection. By using neurophysiological, behavioral, and self-reported perceptual data, this study further aims to clarify how interaction design shapes attentional responses to AI recommendations and users' capacity to maintain effective oversight in organizational decision-making environments.

3.3 Hypothesis Development

3.3.1 AI Errors and Confidence in the AI Assistant

Confidence in an AI assistant refers to the "trustor's perception of the trustee's (in this case an AI) ability to perform a given task" (Chong et al., 2022, p. 2). In human-in-the-loop human-AI interaction settings, this perception must remain calibrated to the actual capabilities and limitations of the system to support effective decision-making (Lee and See, 2004). Unlike trust, which is commonly conceptualized as a multidimensional construct encompassing dimensions such as ability, benevolence, and integrity (e.g., Barki et al. 2015), our conceptualization of confidence focuses specifically on users' assessment of system ability.

According to the SA model, users continuously perceive, comprehend, and project information to maintain an understanding of their operational environment (Endsley, 1995). When interacting with AI systems, users form a mental model of the system's reliability that supports this awareness process. Errors generated by the AI act as discrepant feedback signals that violate user expectations, prompting re-evaluation and reallocation of attention toward diagnostic or corrective actions (Bansal et al., 2021; Zhou and Chen, 2019). This process aligns with the negative feedback loop conceptualized in the threat avoidance model, in which individuals detect deviations from goal-congruent performance and adjust their behavior to minimize risk or restore control (Liang and Xue, 2009). As the AI's output diverges from task objectives, users interpret the discrepancy as a potential threat to goal attainment, triggering a regulatory response that includes lowered adherence to AI recommendations and reduced confidence in the AI assistant.

Empirical findings from the human-AI interaction literature demonstrate that users' confidence in an AI assistant is highly sensitive to system performance (De Visser et al., 2020; Dzindolet et al., 2003; Hancock et al., 2011; Manzey et al., 2012; Mayer et al., 1995). Even minor deviations or errors in AI recommendations can undermine perceptions of reliability, as users infer incompetence or instability in the underlying

algorithm. Evidence from human-robot interaction experiments shows that frequent or salient errors significantly reduce confidence in and willingness to rely on automation (De Visser et al., 2020; Xu et al., 2025). Moreover, performance-based factors, including decision accuracy and failure rate, exert the strongest influence on the build-up and maintenance of confidence in automation within human-machine interactions (Hancock et al., 2011). Accordingly, in AI-assisted decision-making contexts, erroneous recommendations are expected to lead to a measurable decline in users' confidence in the AI recommendation system.

H1: Erroneous AI recommendations will lead to lower user confidence in the AI assistant than non-erroneous recommendations.

3.3.2 AI Errors and Attentional Responses

Errors in AI recommendations not only undermine users' confidence but also elicit distinct attentional responses within the task environment. When system outputs deviate from expectations, they act as salient prediction errors that disrupt users' ability to perceive, comprehend, and project task-relevant information (Endsley, 1995), prompting a reorientation of attention toward the source of the discrepancy. Neuroscientific evidence shows that deviant or infrequent events automatically evoke orienting responses and heightened neural activity associated with attention capture and stimulus evaluation (Coull, 1998; Luck, 2014; Müller-Putz et al., 2015; Polich, 2003). Event-related potential studies further demonstrate that unexpected events elicit components such as P300, reflecting early conflict detection and subsequent attentional updating (Polich, 2003). Similarly, presentation of conflicting information was also shown to elicit a P300 event-related potential component (Lakhiwal et al., 2023). In AI-assisted decision-making, erroneous recommendations function as such deviant events with conflicting information, thus we expect them to lead to amplified neural responses in the form of event-related potential components indicative of attentional reorientation.

H2: Erroneous AI recommendations will trigger stronger attentional responses than

non-erroneous recommendations.

3.3.3 AI Errors and Attentional Resource Allocation

Attention is a limited cognitive resource that must be selectively distributed across competing information sources (Healey and Enns, 2011). Eye-tracking research grounded in information processing theories describes how individuals perceive stimuli, allocate attention, and encode information into memory to support reasoning and decision-making (Wickens et al., 2021; Card et al., 1986). In line with the selective visual attention model, prior work using eye-tracking methods has identified several metrics associated with decision-making, task difficulty, and uncertainty (Brunyé and Gardony, 2017; Vasseur et al., 2023). Fixations, which can be defined as momentary pauses of eye movements directed toward a specific location for a minimum duration (Riedl and Léger, 2016), are among the most widely used of these measures. In visual search tasks, increasing difficulty has been shown to reduce fixation frequency while increasing fixation duration, reflecting deeper visual processing and diagnostic effort (Hooge and Erkelens, 1996; Inhoff and Radach, 1998; Jacobs and O'Regan, 1987; Rayner, 1998).

In human-AI interaction, attentional allocation to decision-making elements that support independent judgment can provide insights into users' perceptions of AI recommendation accuracy and their verification efforts. When automated systems perform reliably, users often operate with minimal attentional involvement. The occurrence of AI errors, however, abruptly increases task difficulty, as users must perform diagnostic and decision processes previously delegated to the AI. Studies on human-automation interaction show that when systems become unreliable, users exhibit a longer gaze duration on non-automated information sources (Bruder and Hasse, 2020; Dzindolet et al., 2003). In line with the empirical evidence on automation accuracy and gaze durations, we hypothesize:

H3: Erroneous AI recommendations will lead to greater attentional resource allocation on the manual task interface than non-erroneous recommendations.

3.3.4 User Autonomy in Human-AI Interactions

When AI systems request parameter configuration input before generating a recommendation, such requests function as an affordance that makes users believe they have influence over how the AI operates. This type of affordance differs from simple prompted input in natural language systems. In a configuration context, users adjust parameters they perceive as shaping the AI's recommendation logic and the model's behavior, thereby contributing to the quality or relevance of its output.

This perceived influence connects to two complementary mechanisms. First, providing input before the AI's analysis resembles traditional forms of user participation in information system development. Consistent with models of user participation and involvement, having autonomy and input in a system's functioning increases users' sense of ownership, importance, and comprehension of the system (Barki and Hartwick, 1994; Hartwick and Barki, 1994). Similarly, recent work proposes that enhancing user agency in automated decision-support systems may increase engagement and help sustain the attentional involvement needed to mitigate automation-induced complacency (Glikson and Woolley, 2020; Hancock et al., 2020; Jiang et al., 2023). Second, periodic prompts for configuration input serve as cognitive nudges, directing attention to task-relevant information. As suggested by the SA model, increased engagement with Level 1 SA elements enhances perception of the task environment and supports the development of Level 2 SA, which involves understanding how those cues relate to the AI's reasoning and potential limitations (Endsley, 1995). Strengthening Level 2 SA is expected to improve users' ability to detect AI errors, calibrate confidence, and allocate cognitive resources more efficiently during verification.

Empirical studies further show that user participation or the inclusion of deliberate, non-automated steps in AI-supported decision processes enhances engagement and understanding of system boundaries, leading to better-calibrated confidence and more appropriate recommendation adherence behavior (Passalacqua et al., 2025; Westphal et al., 2023; Malaguti et al., 2025). In line with this evidence, we hypothesize that access

to AI configuration moderates the effects of AI errors on users' confidence in the AI assistant and their information processing:

H4a: The ability to configure the AI assistant will moderate the effect of erroneous AI recommendations on confidence in the AI assistant, such that the error-related decrease in confidence will be smaller when configuration is available.

H4b: The ability to configure the AI assistant will moderate the effect of erroneous AI recommendations on attentional responses, such that the error-related increase in attentional responses will be greater when configuration is available.

H4c: The ability to configure the AI assistant will moderate the effect of erroneous AI recommendations on attentional resource allocation on the manual task interface, such that the error-related increase in allocation will be smaller when configuration is available.

3.3.5 Confidence in the AI Assistant and Decision Performance

Confidence in the AI assistant plays a central role in how users incorporate its recommendations into their decisions. When users have sufficient confidence in the AI's output, they are more likely to consider its recommendations effectively, thereby enhancing overall decision performance. Prior research on human-automation interaction shows that users who maintain higher confidence in reliable systems tend to achieve better performance outcomes (Dzindolet et al., 2003; Lee and See, 2004).

H5: Greater confidence in the AI assistant will be associated with higher decision performance.

3.3.6 Attentional Reactions and Decision Performance

Attention serves as a central mechanism linking error detection to decision performance. When users encounter unexpected or conflicting information, attentional reorientation facilitates the evaluation of task-relevant cues and the updating of situation awareness. Stronger attentional responses, reflected in neural activity, indicate more effective engagement of cognitive control processes. Empirical research shows that enhanced

amplitudes of event-related potential components such as N200 and P300 correspond to improved conflict detection, stimulus evaluation, and decision accuracy (Coull, 1998; Luck, 2014; Polich, 2003). These findings suggest that users who exhibit stronger attentional responses to AI errors are better able to assess discrepancies, integrate information, and maintain performance in AI-assisted decision-making.

H6: Greater attentional responses to erroneous AI recommendations will be associated with higher decision performance.

3.3.7 Attentional Resource Allocation and Decision Performance

The SA model suggests that effective decision-making depends on how well users allocate attention across information sources (Endsley, 1995). After encountering AI errors, users with accurate mental models of the system are expected to monitor task-relevant elements efficiently rather than relying on prolonged manual verification. As understanding improves, attention to the manual task interface decreases, reflecting a shift from compensatory checking to efficient information processing. In contrast, sustained fixation on manual interface elements may signal uncertainty or low confidence, increasing cognitive effort and reducing performance.

Empirical research links fixation duration to task efficiency and expertise by suggesting that shorter and more targeted fixations indicate more fluent information processing and higher performance, while longer fixations suggest uncertainty or difficulty integrating information (O'Neill et al., 2011; Dogusoy-Taylan and Cagiltay, 2014; Brunyé and Gardony, 2017). Accordingly, users showing longer total fixation duration on the manual task interface after AI errors are expected to demonstrate lower decision performance.

H7: Lower attentional resource allocation on the manual task interface following erroneous AI recommendations will be associated with higher decision performance.

The hypotheses are summarized as a conceptual model in Figure 3.2.

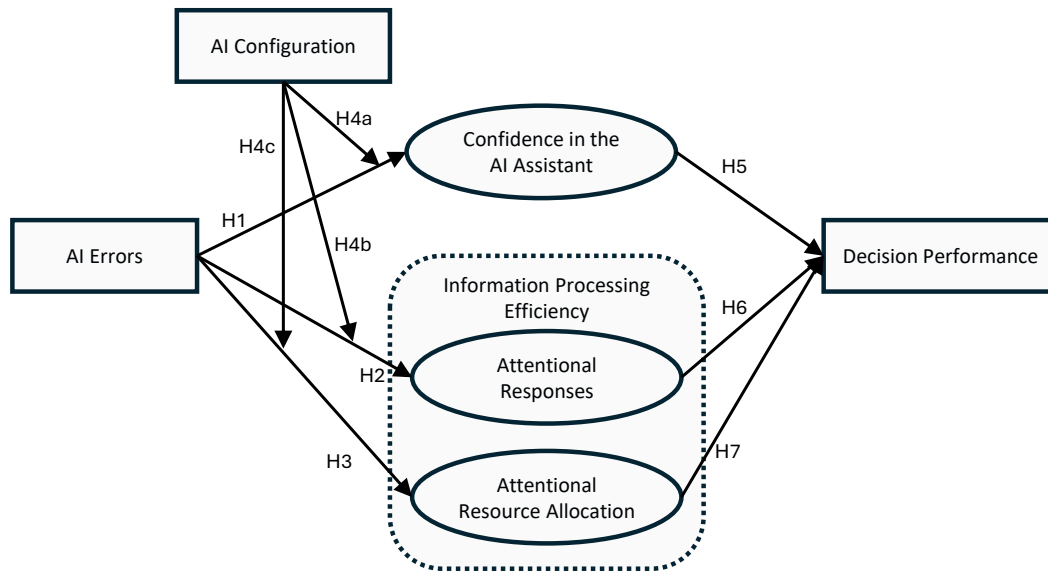


Figure 3.2: Research model summarizing the hypotheses on cognitive responses to AI errors and the availability of AI configuration.

3.3.8 User Control and Perceived Understanding of AI

Engaging with task-relevant parameters through AI configuration requires users to process the information that shapes the decision environment more actively. According to the SA model, interacting with Level 1 SA elements, the perception of task inputs, supports the development of Level 2 SA, which involves comprehension of relationships among these inputs, and Level 3 SA, which enables projection of system outcomes (Endsley, 1995). By prompting users to consider input variables and their implications, AI configuration promotes active engagement with the AI's decision logic rather than passive observation (Benke et al., 2022; Dietvorst et al., 2018; Westphal et al., 2023). This engagement helps users form a clearer mental model of how the AI operates and under what conditions its recommendations may be unreliable, thereby enhancing perceived understanding (Lee et al., 2019; Shneiderman and Plaisant, 2006; Westphal et al., 2023). As a result, users with access to AI configuration are expected to develop a stronger perceived

understanding of the AI system (Benke et al., 2022; Vandewaetere and Clarebout, 2013; Westphal et al., 2023).

H8: The ability to configure the AI assistant, as compared to having no configuration option, will increase users' perceived understanding of the AI assistant.

3.4 Methods

To test the hypotheses, this study employed a mixed-factorial, multi-trial experiment in which participants completed a sequence of consecutive demand-forecasting tasks within a simulated decision-making environment. This section provides a detailed description of the experimental design, the experimental task, the experimental protocol, and the measurement instruments.

3.4.1 Experimental Design

The experiment manipulated two variables. AI configuration availability served as a between-participants factor with two levels (configuration access vs. no configuration access), and AI accuracy served as a within-participants factor with two levels (15 trials with erroneous AI recommendations and 61 trials with non-erroneous recommendations). Table 3.1 summarizes the experiment design structure.

The experimental task was a modified version of the newsvendor problem, widely used to study human decision-making under uncertainty (Benzion et al., 2008; Kremer and Minner, 2008; Kremer et al., 2010). In this task, participants select an order quantity, q , for a perishable product with known cost c , selling price p , and salvage value s , where $s < p$. Theoretical work shows that, under risk neutrality and complete information, the optimal order quantity is given by $q = \Phi^{-1}((p - c)/p)$, where $\Phi(d)$ is the demand distribution. However, empirical research has consistently shown that individuals deviate from this optimal policy, especially when the demand distribution is unknown (Schweitzer and Cachon, 2000; Bolton and Katok, 2008). In addition to understanding human

behavior under uncertainty, we extend the use of the newsvendor paradigm by integrating AI recommendations into the scenario (Wang et al., 2025).

In our study, participants were not given access to $\Phi(d)$; instead, they observed the past 20 periods of realized demand in the form of a time series chart. For consistency, all participants faced the same set of demand values and fixed cost parameters ($c = 91.8$, $p = 126.25$, $s = 50.5$). Inventory did not carry over between trials. The decision environment featured a low-profit setting with a critical ratio $(p - c)/p$ of less than 0.5. This, combined with a salvage price that penalized overordering more severely than underordering (e.g., $p - c = 34.45$, while $c - s = 41.3$), created a context conducive to studying human-AI interaction and judgment under uncertainty (Schweitzer and Cachon, 2000).

Table 3.1: Experimental design.

Between subjects	Within subjects	AI Error-Free (60 trials)	AI Errors (15 trials)
AI without configuration access		No configuration available AI recommendations are accurate	No configuration available Erroneous AI recommendations
AI with configuration access		Configuration available AI recommendations are accurate	Configuration available Erroneous AI recommendations

3.4.2 Experimental Setup

Simulated Decision-Making Environment

To create a contextually situated and interactive newsvendor problem for the participants, we used the ERPsim business simulation platform (Léger, 2006; Léger et al., 2007). Specifically, we employed the Logistics version of ERPsim, which simulates operations within virtual dairy companies under competitive market conditions. This simulation environment is widely adopted in educational and experimental settings due to its ability to mimic realistic market dynamics and promote participant involvement in repeated decision-making processes (Karran et al., 2019; Oz et al., 2020). Building on prior work,

we leveraged ERPsim to increase both ecological validity and engagement in the context of AI-supported inventory decisions (Loos et al., 2010).

ERPsims was used to generate 100 periods of sales data for dairy products, which served as the base demand sequence for our task. In its default configuration, ERPsim generates demand values with no embedded seasonal or trending components. To better align with the experimental goal of simulating dynamic environments, we applied controlled transformations to the ERPsim output to introduce seasonality and trend components. The seasonal pattern was based on empirical analyses of U.S. ice cream production data (National Bureau of Economic Research, 2012) and insights from a food delivery company's blog post (FoodyDirect, 2014). An upward trend was added to simulate gradual growth in demand.

In addition to the introduced trend and seasonal components, the dataset incorporated controlled demand variability to represent forecasting uncertainty commonly observed in operational environments. In this study, uncertainty refers to short-term fluctuations in demand between consecutive simulated periods rather than large structural shocks. Demand uncertainty was operationalized as percentage changes between successive periods and was limited to a maximum value so that underlying trend and seasonal patterns remained detectable to participants during repeated decision-making. Across the generated dataset, consecutive-period demand changes exhibited a mean absolute variation of 3.09% (SD = 2.34%), with maximum observed changes of approximately 11%. This level of variability introduced realistic stochasticity while preserving the interpretability of systematic demand dynamics central to the experimental task.

The generated dataset aimed to reflect the stochastic characteristics of real-world demand while maintaining experimental control over its structure. Participants were explicitly instructed to attend to these features when configuring the AI system or making their own decisions. All instructions and task materials are documented in Appendix, Section C.1.2.

Experimental Stimuli

Participants completed the decision tasks using a custom-built interface designed to deliver a structured version of the newsvendor problem within a realistic organizational environment. The task interface was embedded within a working ERP system (SAP SE, Walldorf, Germany) to ensure contextual relevance while maintaining experimental control over the decision workflow.

Before accessing the main forecasting interface, participants encountered an initial step that differed between the two between-participants conditions: a configuration screen (configuration condition) or a waiting screen (no-configuration condition). This step operationalized the configuration manipulation intended to influence perceived autonomy, as described in Section 3.2.2. Only participants in the configuration condition interacted with parameters that appeared to influence the forthcoming AI recommendation. Participants in the no-configuration condition did not see any parameters and instead had to view a short waiting message for two seconds before continuing to the task. The 2-second delay duration was determined after analyzing the time spent on the configuration interface across 300 trials collected from 4 pre-test participants. Figures 3.3a and 3.3b illustrate the configuration and no-configuration screens shown at the beginning of each trial.

In the configuration condition, the three parameters were presented as adjustable inputs to create the perception that participants could influence how the AI generated its recommendation. However, these inputs did not affect the AI output. This design choice ensured that any differences between conditions resulted solely from the perceptual influence of the configuration affordance rather than differences in AI recommendations.

Considering the dairy market and ice cream sales, please select your suggestions for each of the following factors in determining the forecasting model parameters.

Smoothing parameter	Weak	☐	☐	☐	☒	☐	☐	☐	Strong
Trend parameter	Weak	☐	☐	☐	☒	☐	☐	☐	Strong
Seasonality parameter	Weak	☐	☐	☐	☒	☐	☐	☐	Strong

Apply

(a) Configuration

Please wait a few seconds for the AI model to prepare the calculation parameters.

Continue

(b) No-configuration - The highlighted button appeared after a 2-second delay.

Figure 3.3: Screens shown before each trial in the configuration and no-configuration conditions.

After completing the configuration or waiting step, all participants accessed the main forecasting interface (Figure 3.4). The interface displayed a time-series graph of historical demand, the product's pricing information, and a numeric input field for first viewing order quantities recommended by AI, and then editing the recommendation before submission, if participants deemed that necessary. Participants initiated each trial

by clicking the "Generate AI Recommendation" button ($t = 0$). The suggested forecast appeared exactly one second after the click ($t = 1$), and two seconds later ($t = 3$), additional interface elements became visible, including short textual explanations derived from the model parameters and two action buttons labeled "Accept" and "Edit Values." The suggested forecast could be edited only after clicking "Edit Values."

Although participants believed that the recommendations came from a functioning AI model, no machine learning models were embedded in the system. AI-generated recommendations were computed in advance using the Holt-Winters method implemented in the forecast package in R, based on the synthetic demand data described in Section 3.4.2. The recommendation explanations were presented as factor weights, which were generated using the Holt-Winters coefficients computed in advance for each trial. Trials designated as erroneous deviated from the system's typical performance by adding a positive or negative deviation that corresponds to 1.5 times the standard deviation of the model's historical forecast error.

Each trial followed a fixed sequence. After entering a forecast, the application prompted participants to self-report their confidence in the AI assistant and subsequently displayed a performance feedback report. The performance feedback report was designed to motivate participants and act as an incentive to help them perform better in tasks. Prior research shows that social comparison and performance feedback can elicit competitive engagement and goal-oriented behavior (Garcia et al., 2013). Consistent with findings from the gamification literature (Hamari et al., 2014), participants were shown a comparison between their performance and that of an average decision maker and were informed about the monetary value saved or lost in the previous trial, together with performance indicators summarizing forecasting outcomes. This design aimed to promote sustained task engagement and effort through perceived competition and performance evaluation.

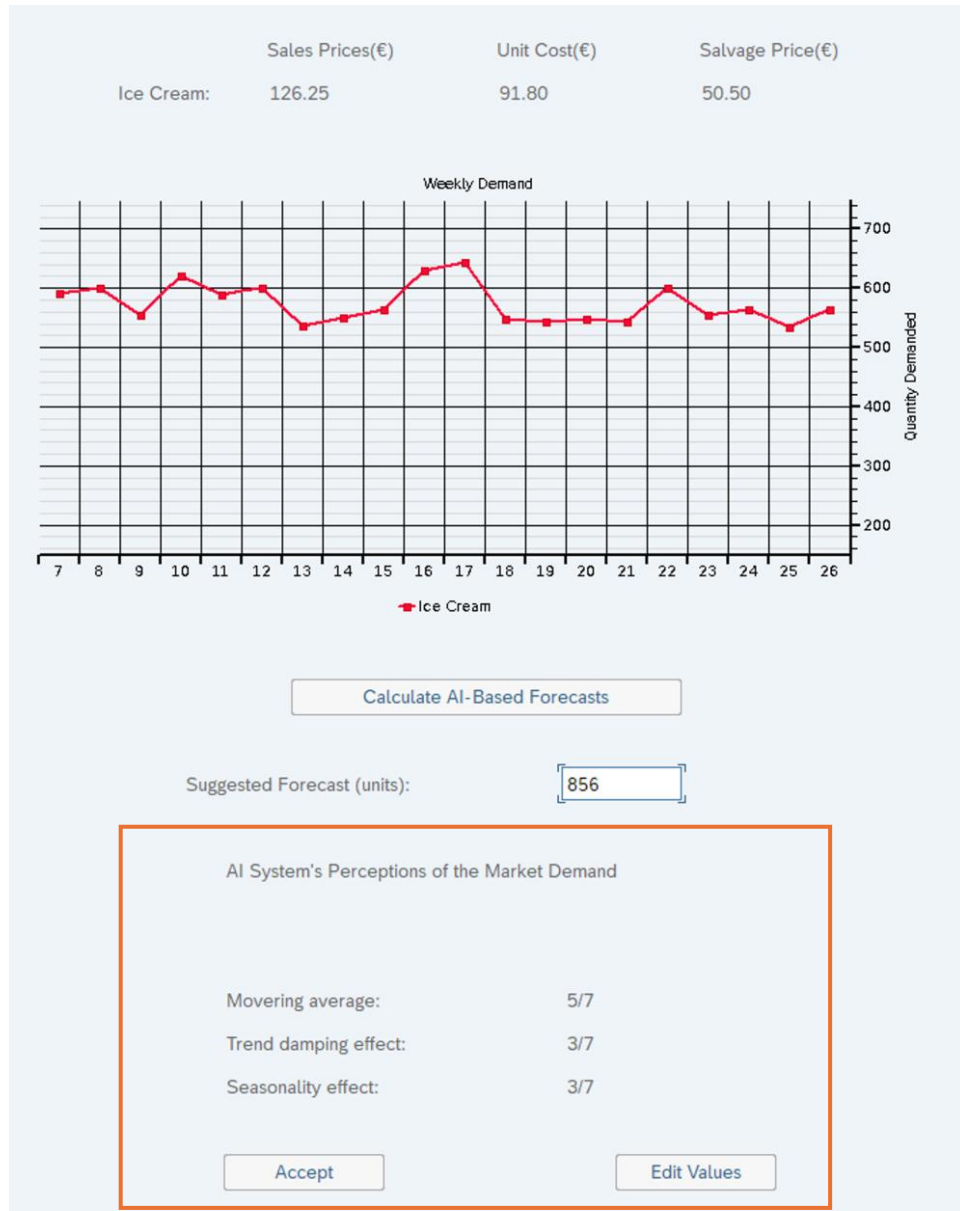


Figure 3.4: Main forecasting interface used during the experiment. Highlighted elements show user interface elements that were revealed exactly three seconds after participants clicked "Generate AI Recommendation."

Experimental Procedure

Each participant completed a total of 81 trials, organized as follows:

1. **Task Introduction and Pre-Task Survey:** Participants received written instructions describing the newsvendor scenario and completed a brief pre-task survey capturing demographic information and control variables.
2. **Training Block (5 trials, no AI support):** Participants completed five practice trials without AI assistance to become familiar with the decision-making interface and the task environment. At the end of each trial, they received a performance feedback report.
3. **Experimental Block (76 trials, with AI support):** Participants completed 76 main task trials with AI recommendations. AI support differed based on the between-participants condition (configuration vs. no configuration). Within this block, 15 trials between trials 15 and 70 were randomly assigned to contain erroneous AI recommendations. The first 14 trials served as a baseline experimental block to acclimate participants to using AI and to capture their baseline AI confidence values, with no influence of AI errors on their perceptions. Similarly, the last 6 trials did not include errors to ensure all participants experienced a similar number of non-erroneous trials before submitting their self-reported perceptions at the end of the session.
4. **Post-Task Survey:** After completing the experimental trials, participants responded to self-report items assessing their perceptions.

Each experimental trial n consisted of the following steps:

1. **Configuration or Waiting:**
 - **Configuration condition:** Participants were instructed to adjust smoothing, trend, and seasonality parameters, and told that these settings would influence

how the AI generated its recommendation. In reality, the configuration inputs had no effect on the forthcoming AI output. All participants, regardless of condition, received the same pool of AI recommendations; only the trial positions of erroneous recommendations varied randomly across participants.

- **No-configuration condition:** Participants were shown a brief waiting message for two seconds before being able to proceed.
2. **AI Recommendation Generation:** Participants were instructed to review the historical time-series chart of past sales data and click on "Generate AI Recommendation" when ready. The recommended order quantity appeared exactly 1 second after the click. Two seconds after the suggestion appeared, the system explanations and action buttons became visible.
 3. **Decision Submission:** Participants had the option to either accept the recommended forecast or edit the value before submission. When editing, the AI-recommended order quantity became an editable text field.
 4. **Performance Feedback:** After submitting their decision, participants received a performance report displaying realized demand and corresponding profit or loss.

Figure 3.5 outlines the experimental procedure.

Hardware Used

The experiment was conducted using standardized hardware across all participants. The decision-making interface was displayed on a 22-inch monitor with a 1920×1080 resolution and 60 Hz refresh rate in full-screen mode. Across trials, participants primarily used a mouse to navigate the task steps, with occasional keyboard input to modify AI-generated forecasts.

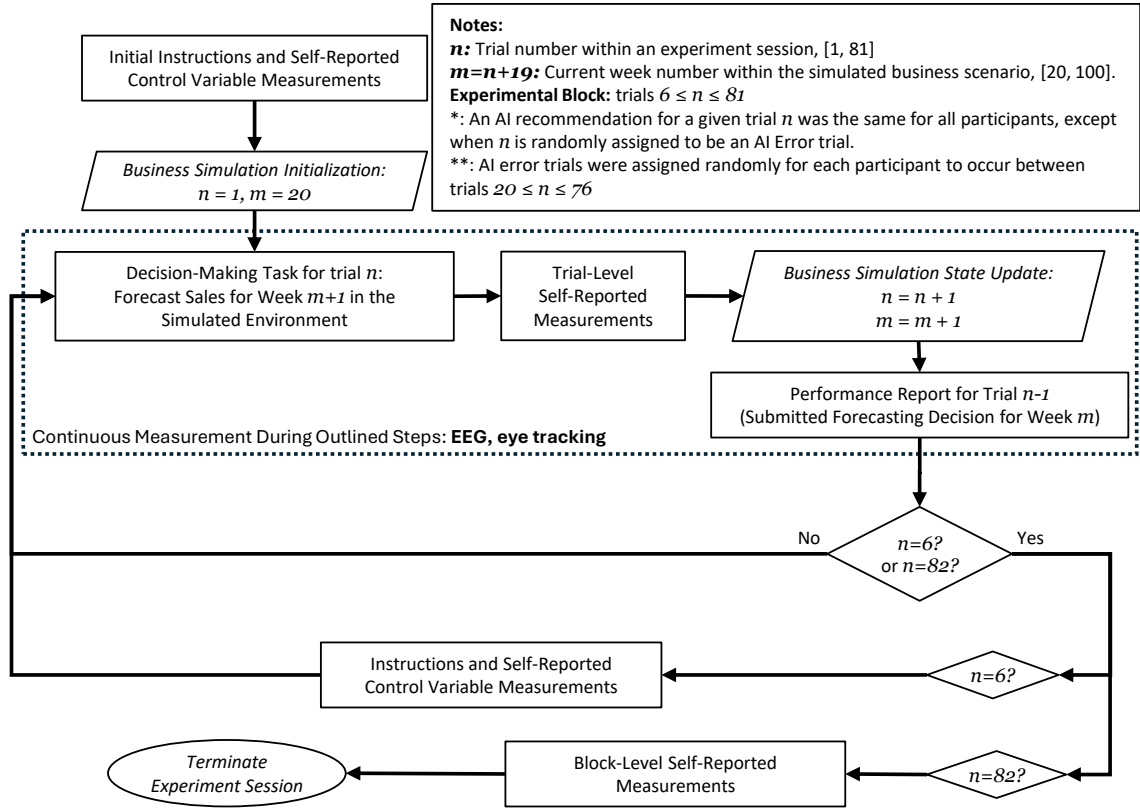


Figure 3.5: Experimental protocol.

Eye movements were recorded using a Tobii eye tracker (Tobii AB, Stockholm, Sweden), sampling at 60 Hz. EEG data were collected from 32 scalp electrodes arranged using the 32-channel Standard Cap layout for actiCAP (Brain Products GmbH, Munich, Germany). After cap placement, conductive gel was applied, and electrode impedance was checked to ensure proper signal quality. EEG signals were recorded raw at a 1000 Hz sampling rate. All sessions were conducted in the same laboratory room under controlled ambient lighting and noise conditions to maintain consistency across participants.

The decision-making interface was programmed to synchronize event timestamps with the continuous EEG and eye-tracking recordings. At the onset and offset of each trial, the interface transmitted two consecutive 10-ms square-wave auditory pulses, spaced 500 ms or 700 ms apart depending on whether the signal indicated trial onset or trial offset. These audio signals were routed through a stereo amplifier and into the TTL

input of the BrainVision trigger box (Brain Products GmbH). The amplifier increased the signal amplitude so that it exceeded the 5V threshold required to register a TTL trigger. These markers were subsequently used to extract the corresponding EEG and eye-tracking segments for each trial.

EEG Data Preprocessing

The collected EEG signals underwent a series of pre-processing steps using BrainStorm (Tadel et al., 2011) running in MATLAB (MathWorks, Natick, MA, USA), as follows. A 1–30 Hz band-pass filter was applied to attenuate slow drifts as well as higher-frequency environmental noise (e.g., line noise from AC power and monitor refresh rates) that commonly contaminates EEG recordings (Luck, 2014, p. 113). Independent component analysis (ICA) was then used to identify and remove components associated with ocular and muscle activity such as blinks. Data were epoched from -1000 ms to $+2000$ ms relative to the onset of the AI recommendation on each trial, yielding 76 epochs per participant.

All 3000 ms epochs were visually inspected, and trials with peak-to-peak amplitudes exceeding $100\mu\text{V}$ within the window were rejected. The included trials underwent baseline correction using the -200 ms to -1 ms prestimulus interval.

For event-related potential analysis, trial-wise time-averaged amplitudes were computed for each event-related potential component of interest. The P3a component was defined as the mean amplitude between 250–280 ms post-stimulus at the Pz electrode. These time-locked component averages were extracted on a per-trial basis and subsequently merged with the behavioral dataset for hypothesis testing using linear mixed-effects models (Frömer et al., 2018).

Measurements

Behavioral and Self-Reported Measures Behavioral data included decision performance, decision time, and AI suggestion adherence. Decision performance was measured

using a loss function grounded in the classic newsvendor paradigm, reflecting the cost implications of inventory decisions at each trial, in Euros. Specifically, given the scenario parameters of cost ($c = 91.8$), price ($p = 126.25$), and salvage price ($s = 50.5$), the cost of overestimating demand by one unit was $c - s = 41.3$, and the cost of underestimating it was $p - c = 34.45$. Based on these values, participant decisions were evaluated by comparing their order quantity to actual demand, with losses incurred for both over- and under-ordering. A zero difference between order quantity and demand resulted in a loss value of zero, while all deviations led to negative loss scores, indicating suboptimal performance. Decision time was computed as the time spent on the decision-making interface in Figure 3.4. AI suggestion adherence was computed in accordance with the appropriate reliance concept, which associates a negative score to each unit of forecast adjustment on non-erroneous AI recommendations, while associating a positive score to each unit of forecast adjustment on erroneous AI recommendations (Lee and See, 2004).

Confidence in the AI assistant was measured immediately after each trial using a single-item, 7-point Likert scale (Chong et al., 2022; Davis et al., 1994). The measure was administered after participants finalized their decision but prior to receiving performance feedback for that trial. Although single-item measures do not permit internal consistency assessment and may be more susceptible to common method bias (De Guinea et al., 2013), prior measurement research indicates that single-item indicators can provide reliable and valid assessments when constructs are unidimensional, concrete, and clearly understood by respondents (Ang and Eisend, 2018; Bergkvist and Rossiter, 2007; Wanous et al., 1997).

In this study, confidence in the AI assistant was conceptualized as a unidimensional, trial-level evaluation of the AI assistant's task-related capability (Chong et al., 2022). Accordingly, a single-item measure was employed to reduce participant burden and preserve attention, as this measurement was collected repeatedly across 76 consecutive trials within a single experimental session.

Perceived understanding of the AI assistant was measured at the end of each 10-trial experimental block using the explanation satisfaction scale, defined as "the degree to

which users feel that they sufficiently understand the AI system or process being explained to them" (Hoffman et al., 2023, p. 3).

In addition, three control variables were measured before the experimental tasks: Automation-Induced Complacency Potential (AICP), administered before task-specific instructions (Merritt et al., 2019), Computer Self-Efficacy, collected after the task and interface were introduced (Compeau and Higgins, 1995), and past experience with ERP systems, measured using a five-level single-item question at the beginning of each experiment session. In addition, participants' perceived autonomy was measured after the task completion as a manipulation check variable with the use of a four-item measure of perceived autonomy scale (De Vreede et al., 2021). The complete list of measurement items is provided in Appendix, Section C.1.1, Table C.1.

EEG-Based Measures of Attentional Responses to AI Recommendations Attentional responses to AI recommendations were investigated through event-related potential components that index distinct stages of cognitive processing. The P3a event-related potential component, recorded at Pz and Oz electrodes, was selected to operationalize attentional responses to AI recommendations. The P3a component reflects attentional orienting to salient or unexpected events, manifesting as a positive deflection in the 250–280 ms time window following stimulus onset, respectively (Coull, 1998; Luck, 2014; Polich, 2003).

Because single-trial EEG data are inherently noisy, the analysis followed a windowed event-related potential approach. For each trial, mean amplitudes were extracted within predefined temporal windows corresponding to the P3a component. These trial-level values were then entered into linear mixed-effects models with participant-level random intercepts to account for repeated measures (see (Frömer et al., 2018) for a detailed discussion of this procedure).

Eye-Tracking Measures of Information Processing Behaviors and Efficiency Participants' information processing behaviors were operationalized through gaze behavior

that was recorded using a Tobii eye tracking sampling at 60 Hz. Attentional resource allocation was measured as the fixation duration on screen areas where raw data relevant to decision-making were present. These areas of interest (AOIs) mainly focused on the raw data used to complete the decision, mainly in the form of a time series chart of the past 20 weeks' sales data, and the pricing information above that chart (see Figure 3.4 for an example view of the stimuli).

This fixation-based eye-tracking metric for attentional resource allocation was computed over the post-recommendation verification window, defined as the time period from the onset of the AI recommendation to the participant's decision submission in that trial.

A summary of operationalizations is shown in Table 3.2.

Table 3.2: Used measures.

Concept	Source	Operationalization	References
AI errors	Experimental manipulation	Binary variable	
AI configuration	Experimental manipulation	Binary variable	
Confidence in the AI assistant	Self-reported at each trial	One item, 7-level Likert scale	(Davis et al., 1994)
Perceived understanding of AI	Self-reported at the end of the experiment session	Four items, 7-level Likert scale	(Hoffman et al., 2023)
Perceived autonomy	Self-reported at the end of the experiment session	Three items, 7-level Likert scale	(De Vreede et al., 2021)
Propensity to trust automation	Self-reported at the beginning of the experiment session	Ten items, 7-level Likert scale	(Merritt et al., 2019)
Computer self-efficacy	Self-reported at the beginning of the experiment session	Ten items, 10-level Likert scale	(Compeau and Higgins, 1995)
Attentional responses	EEG	Event-related potential component P3a	(Polich, 2003)
Attentional resource allocation	Eye-tracking	Total fixation duration on the source data between AI recommendation onset and decision submission	(Toreini et al., 2022)
Decision performance	Behavioral	Losses from excess inventory or lost sales	(Benzion et al., 2008)

3.4.3 Sample Description and Data Collection

A power analysis was conducted using G*Power 3.1 to determine the required sample size, assuming a medium effect size, $\alpha = 0.05$, and statistical power of 0.90 (Faul et al.,

2007). The analysis indicated that at least 46 participants would be necessary to achieve adequate power. To allow for possible data loss, 59 participants were scheduled for data collection between March 23 and June 2, 2023.

Participants were recruited from the authors' institutional participant panel based on convenience sampling. Eligibility screening included three inclusion criteria: (1) prior use of ERP systems, (2) prior experience with ERPsim, and (3) successful completion of an introductory-level course in logistics and demand forecasting. These requirements ensured a uniform baseline of task-relevant knowledge. Additional criteria required participants to have normal or corrected-to-normal vision and be right-handed (to ensure uniformity of EEG signaling).

The study was reviewed and approved by the Ethics Committee of the authors' academic affiliation (certificate number: 2023-5420). Participants received a fixed compensation of CAD 50 for their participation, and each session lasted approximately two hours. Prior empirical studies using the newsvendor paradigm have employed both fixed (e.g., Wang et al. 2025) and performance-based (e.g., Benzion et al. 2008; Kremer et al. 2010) compensation schemes. Prior experimental evidence indicates that performance-based financial incentives do not consistently improve decision performance in judgment tasks, suggesting that participant engagement and motivation are the primary drivers of performance (Camerer and Hogarth 1999). Accordingly, a fixed compensation structure was adopted with the combination of gamification elements such as performance feedback and social comparison at the end of each trial.

Of the 59 individuals scheduled, seven canceled or failed to attend their session within 24 hours. Two sessions were discontinued due to EEG hardware malfunctions, and no data were collected from those participants, although they were still compensated. Additionally, two participants could not complete the number of trials required within the given time. To ensure consistency in participant fatigue and performance, these two participants were excluded from the dataset. Consequently, the final dataset included 48 participants (17 female; mean age 29.3 ± 6.82 ; age range 19-51), with additional exclusions applied during data preprocessing.

3.5 Results

3.5.1 Experimental Design Checks

Before testing the hypotheses, we examined whether the random assignment of participants to the two between-participant conditions was successful by comparing baseline performance and demographic characteristics across conditions. The chi-squared test indicated no significant differences between the two groups in gender, $\chi^2(1, N = 48) = 0.99$, $p = .320$. Two-tailed Fisher's exact test for past experience with enterprise resource planning systems also indicated no significant difference between conditions, $p = .186$. A Wilcoxon rank-sum test for baseline task performance indicated no significant difference between the two groups, $W = 266$, $p = .657$. These results suggest that participants in both conditions were not significantly different in terms of demographic and baseline performance variables, thereby supporting the success of the random assignment of between-subject conditions.

Table 3.3: Comparison of the baseline performance levels of participants across experimental conditions

Control variable	Conditions	Median	Mean	SD	W	P-value	R
Baseline Performance	Configuration	-1078.63	-1748.32	1687.60	266	0.657	0.06
	No configuration	-1057.88	-1640.87	1693.13			

3.5.2 Manipulation Check: Configuration and Perceived Autonomy

As a manipulation check, participants' perceived autonomy was measured using three adapted items from the Perceived Autonomy Scale (De Vreede et al., 2021). A confirmatory factor analysis indicated acceptable internal consistency ($\alpha = 0.80$), with standardized factor loadings ranging between 0.54 and 0.65. Because perceived autonomy served solely as a manipulation check rather than a focal construct in hypothesis testing or structural model estimation, item retention decisions emphasized conceptual coverage alongside empirical criteria. Therefore, indicators demonstrating acceptable loadings and

theoretical relevance were retained despite falling below the preferred loading threshold of 0.708 (McKenzie et al., 2011).

To examine whether the manipulation affected user perceptions, we regressed perceived autonomy scores on condition. Results showed a significant positive effect ($\beta = 0.19$, $SE = 0.09$, $t(46) = 2.20$, $p = 0.033$), indicating that participants in the configuration condition reported significantly higher perceived autonomy compared to those with no configuration access. These findings confirm that the manipulation was successful in shaping participants' perceptions of control in their interactions with the AI.

3.5.3 Manipulation Check: AI Errors

To verify that participants detected erroneous AI recommendations, we examined a direct behavioral indicator of error recognition: the magnitude of numerical adjustments participants made to the AI-recommended order quantity. The adjustment magnitude variable was computed as the absolute difference between the submitted order quantity and the AI-recommended order quantity for each trial. Using the log transformation of this variable as the dependent variable, we estimated a linear mixed-effects model with participant-level random intercepts. The model predicted the log-transformed adjustment magnitude as a function of whether the recommendation was erroneous, while controlling for configuration condition and computer self-efficacy.

We estimated a linear mixed-effects model with participant-level random intercepts to predict log-transformed adjustment magnitude from error presence (erroneous vs. accurate), configuration condition (configuration vs. no configuration), and their interaction, while controlling for computer self-efficacy. Error presence had a strong, significant effect on correction behavior ($\beta = 3.04$, $p < .001$), indicating that participants substantially increased their modifications when the AI was wrong. In contrast, configuration condition and its interaction with error presence were not significant (all $p > .20$), showing that error-related corrections did not differ between experimental groups. Computer self-efficacy also did not influence adjustment behavior ($\beta = -0.06$,

$p = .85$) ($R_m^2 = .543$, $R_c^2 = .576$).

These results confirm that participants successfully identified erroneous AI recommendations and modified their behavior to address inconsistencies. While subsequent sections examine the cognitive and performance consequences of these adaptations, this analysis establishes that the experimental manipulation of AI errors elicited a clear, measurable behavioral response.

It is also important to note that although both groups corrected erroneous recommendations to a similar degree, the hypotheses in this study focus on differences in the cognitive processes that lead to those corrections, such as attention allocation, error monitoring, and information-processing dynamics. These processes may still diverge even when the final behavioral adjustment is similar.

3.5.4 Data Measurement Model

Control Variables

The 10-item AICP scale was modeled in line with the two subdimensions theorized in the source study: Workload Alleviation (AICP-W) (Items 1-4, 6) and Monitoring (AICP-M) (Items 5, 7-10) (Merritt et al., 2019). Confirmatory factor analyses (CFA) were conducted for each subscale separately. Both the Workload Alleviation and Monitoring subscales demonstrated poor internal consistency levels ($\alpha = 0.67$ and $\alpha = 0.69$) and therefore, they were excluded from further analyses.

The Computer Self-Efficacy construct was modeled as a single latent factor using the 10-item scale adopted from Compeau and Higgins (1995). A CFA of the measurement yielded a Cronbach's $\alpha = 0.89$, with item loadings ranging between .6 and .89.

As an additional behavioral control variable at the participant level, baseline task performance values were computed with the use of participants' task performance in the last four trials of the initial five-trial training block. In these trials, participants performed the task without any AI recommendation. The first training trial was excluded from these baseline performance calculations, as participants were allowed to request assistance from

the experiment moderator.

Self-Reported Variables

Confidence in the AI assistant was measured using single-item instruments. Descriptive statistics by control and error conditions are reported in Appendix, Section C.2, Table C.18.

Perceived understanding of AI was modeled as a single latent factor measured with the use of seven items from the 8-item scale adopted from Hoffman et al. (2023). The measurement focused on the explanations provided by the AI recommendation system regarding the factor weights that influence the suggested forecast quantity. A CFA of the seven-item measurement instrument indicated a high reliability (Cronbach's $\alpha = 0.92$).

Behavioral Variables

During experiment sessions, participants' decision time and decision performance were captured automatically. In non-error trials, average performance was similar across configuration conditions, with a mean of -726 EUR in the no-configuration condition and -723 EUR in the configuration condition (values closer to 0 indicate better performance, with 0 denoting the optimal value). In error trials, both conditions exhibited lower performance overall, with -1188 EUR in the no-configuration condition and -1383 EUR in the configuration condition. Descriptive statistics across all experimental conditions are reported in Appendix, Section C.2, Table C.18.

Eye-Tracking Metrics

Fixation-based eye-tracking metrics were aggregated on the participant and trial basis and merged into a dataset as fixation duration and fixation rate per trial and participant. On erroneous trials, mean fixation duration on the manual-data chart averaged lower in the configuration condition ($M = 1.51$, $SD = 4.89$; $N = 303$), compared to the no configuration condition ($M = 2.03$, $SD = 5.59$; $N = 305$). On non-erroneous trials

(*error* = 0), mean fixation duration was similar across conditions (present: $M = 1.50$, $SD = 4.07$; $N = 1299$; absent: $M = 1.49$, $SD = 4.42$; $N = 1356$). Mean fixation rate averaged lower under the presence condition on non-erroneous trials (2.21 vs. 2.67) and was of similar magnitude across conditions on erroneous trials (2.21 vs. 2.15). Summary statistics for fixation duration and fixation rate are reported in Appendix, Section C.2, Table C.19.

3.5.5 Hypothesis Testing

Confidence in the AI Assistant (H1 & H4a)

To test the impact of erroneous AI recommendations on users' confidence in the AI assistant (H1) and the moderating role of configuration (H4a), we estimated a linear mixed-effects model with participant-level random intercepts. The model was fitted on trial-level data ($N = 2,602$) and included computer self-efficacy as a control variable.

The results provide strong support for H1. Erroneous AI recommendations significantly reduced user confidence in the AI recommendation system ($\beta = -0.289$, $p < .001$, one-tailed). The results also support H4b: the interaction between configuration and erroneous recommendations was significantly negative ($\beta = -0.082$, $p < .001$, one-tailed), indicating that the decrease in confidence following erroneous recommendations was larger when configuration was available. Neither configuration alone ($\beta = 0.006$, $p = .94$, one-tailed) nor computer self-efficacy ($\beta = 0.166$, $p = .083$, one-tailed) were significant predictors. The results are illustrated on the estimated means plot in Figure 3.6.

The model accounted for a modest proportion of variance in confidence ($R_m^2 = .31$), while participant-level random intercepts explained additional variance ($R_c^2 = .56$). Model details are provided in Appendix, Section C.2, Table C.20.

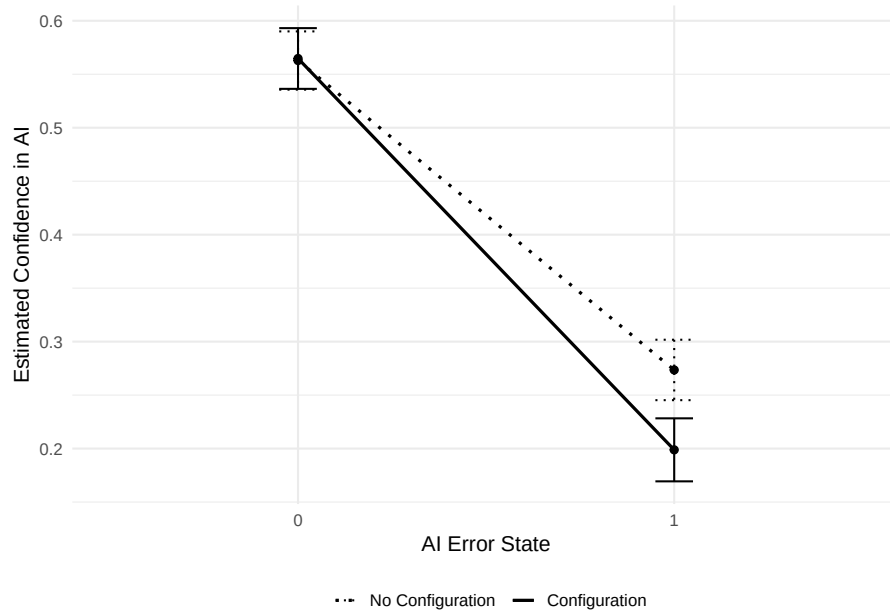


Figure 3.6: Estimated marginal means of confidence in the AI recommendation system across error (erroneous vs. non-erroneous) and configuration conditions. Error bars represent 95% confidence intervals.

Attentional Responses (H2 & H4b)

To test whether erroneous AI recommendations triggered attentional orienting as measured by the P3a component, we estimated a linear mixed-effects model with participant-level random intercepts ($N = 2,290$).

The results support H2: erroneous recommendations significantly reduced P3a amplitude ($\beta = -1.608$, $p < .001$, one-tailed), consistent with users reallocating attention under error conditions. The results also support H4b: the interaction between configuration and erroneous recommendations was significantly positive ($\beta = 1.99$, $p < .001$, one-tailed), suggesting that when configuration was available, attentional reallocation responses were more pronounced. A plot of the event-related potentials at posterior central electrodes is shown in Figures 3.7-3.8. The main effect of configuration was not significant ($\beta = 0.479$, $p = .16$, one-tailed). Model details are provided in Appendix, Section C.2 Table C.21.

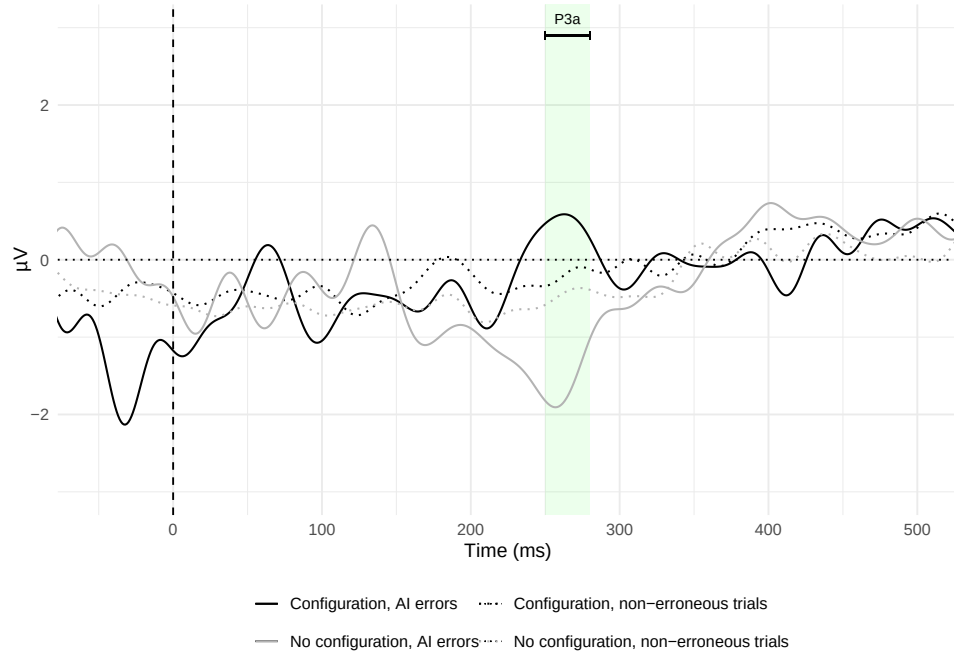


Figure 3.7: Grand average plots of waveforms at the Pz electrode channel.

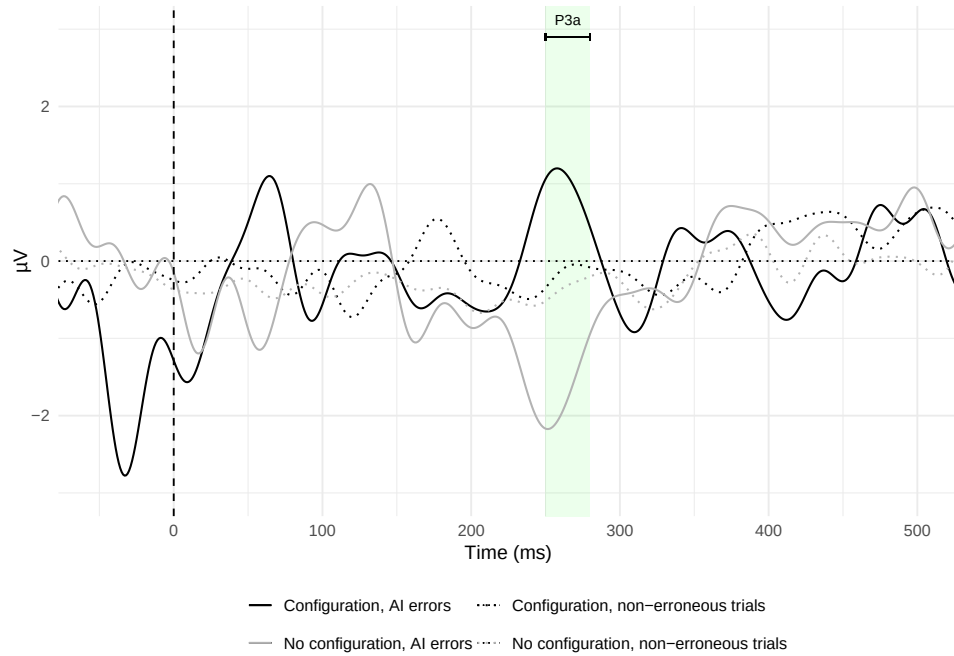


Figure 3.8: Grand average plots of waveforms at the Oz electrode channel.

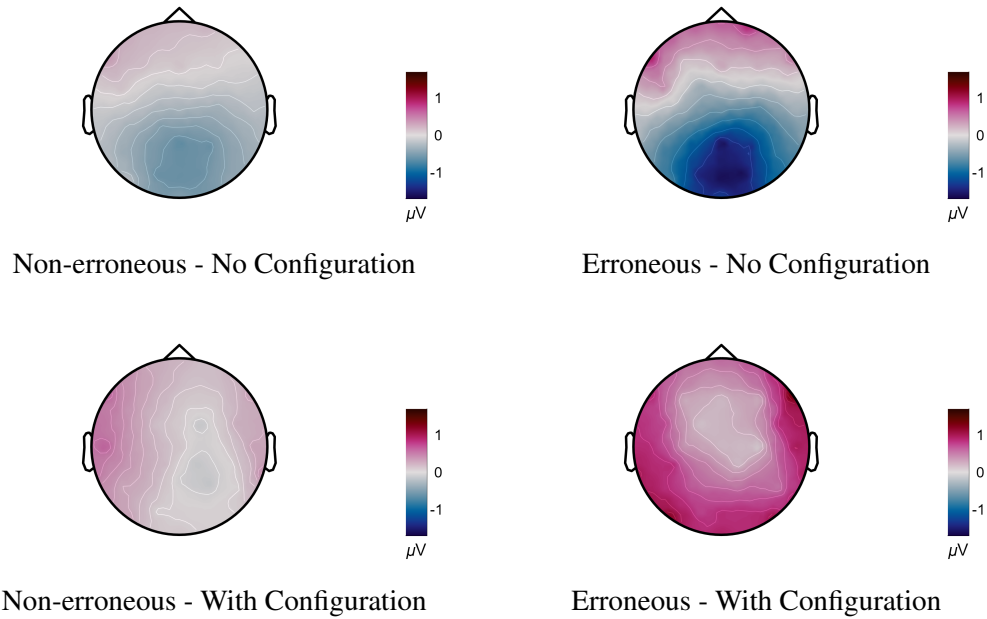


Figure 3.9: Topographic maps (250–280 ms) across configuration and accuracy conditions.

Attentional Resource Allocation (H3 & H4c)

To test whether erroneous AI recommendations increase users' visual attention allocated to manual task interface elements (H3) and whether configuration moderates this effect (H4c), we estimated a linear mixed-effects model with participant-level random intercepts. The model was fitted on trial-level data ($N = 2,602$) and included computer self-efficacy as a control variable.

The results support H3: erroneous AI recommendations significantly increased the total duration of fixations on task-relevant information between the onset of the AI recommendation and the submission of the final decision ($\beta = 0.075$, $p < .001$, one-tailed) ($R_m^2 \approx .04$, $R_c^2 \approx .47$). The interaction between configuration and erroneous AI recommendations was significant ($\beta = -0.052$, $p = .026$, one-tailed), indicating that configuration reduced the magnitude of the error-related increase in attentional allocation. Configuration alone had no significant main effect ($\beta = -0.031$, $p = .85$, one-tailed). Computer self-efficacy had a significant negative effect on fixation duration ($\beta = -0.478$,

$p = .026$, one-tailed), indicating that participants with higher self-efficacy spent less time fixating on task-relevant information. Further model details are provided in Appendix, Section C.2 Table C.22.

Confidence in the AI Assistant and Decision Performance (H5)

To test whether greater confidence in the AI assistant predicts higher task performance (H5), a linear mixed-effects model with random intercepts for participants was estimated. The model used trial-level data ($N = 2,602$) and predicted log-transformed performance scores using participants' self-reported confidence in the AI assistant as a predictor, controlling for computer self-efficacy.

The results indicated a significant positive association between confidence in the AI assistant and task performance ($\beta = 0.061$, $p < .001$), suggesting that participants who reported higher confidence achieved better performance across trials. This result supports the hypothesis that well-calibrated confidence in the AI facilitates effective integration of AI recommendations into decision-making, improving performance outcomes. Full model details are reported in Appendix, Section C.2, Table C.23.

Attentional Responses and Decision Performance (H6)

To evaluate whether attentional reactivity to AI errors improves performance (H6), we modeled task performance as a function of P3a amplitude (Pz, 250-280 ms) and computer self-efficacy, using a linear mixed-effects model with participant-level random intercepts ($N = 2,290$).

No significant effect was observed for P3a amplitude ($\beta = -0.0005$, $p = .16$) or for self-efficacy ($\beta = 0.011$, $p = .8$); and the explained variance was negligible ($R_m^2 = .001$, $R_c^2 = .008$), indicating no reliable relationship between neural attention responses and decision performance in this context. The model details are listed in Appendix, Section C.2, Table C.24.

Attentional Resource Allocation and Decision Performance (H7)

To examine whether increased visual verification behavior is associated with lower task performance (H7), we estimated a linear mixed-effects model with participant-level random intercepts. The trial-level model ($N = 2,602$) predicted log-transformed performance scores using log-transformed post-decision fixation duration on task-relevant information as a measure of visual verification effort, with computer self-efficacy included as a control.

The model revealed a significant negative relationship between visual verification effort and task performance ($\beta = -0.015$, $p = .03$, one-tailed), indicating that higher fixation duration following AI recommendation onset was associated with lower performance. Computer self-efficacy had no significant effect ($\beta = 0.0086$, $p = .85$). These findings are consistent with the interpretation that prolonged visual verification reflects lower task or AI understanding. The model results are summarized in Appendix, Section C.2, Table C.25.

To provide an aggregate assessment of predictors of decision performance, an additional mixed-effects model including all predictors of decision performance was estimated and is reported in Appendix, Section C.2, Table C.26. Variance inflation factor diagnostics indicated no multicollinearity concerns among predictors. This integrated model yielded a marginal R^2 of 0.05 and a conditional R^2 of 0.06, indicating that fixed predictors explain approximately 5% of variance in decision performance.

Perceived Understanding of the AI System (H8)

To test the impact of configuration on users' perceived understanding of the AI system (H8), we estimated a linear model at the block level, using 48 participants. Perceived understanding was measured only once at the end of the experiment session, resulting in a single observation per participant. The model included configuration as the main predictor while controlling for computer self-efficacy.

The results support H8. The configuration condition positively predicted perceived

understanding of the AI system ($\beta = 0.12$, $p = .043$, one-tailed), indicating that participants with AI configuration access reported higher perceived understanding than those with no configuration access. Computer self-efficacy was not a significant predictor ($\beta = 0.15$, $p = .22$, one-tailed). The model explained $R^2 = .08$ of the variance in perceived understanding, with an adjusted $R^2 = .04$.

Model estimates are reported in Appendix, Section C.2 Table C.27.

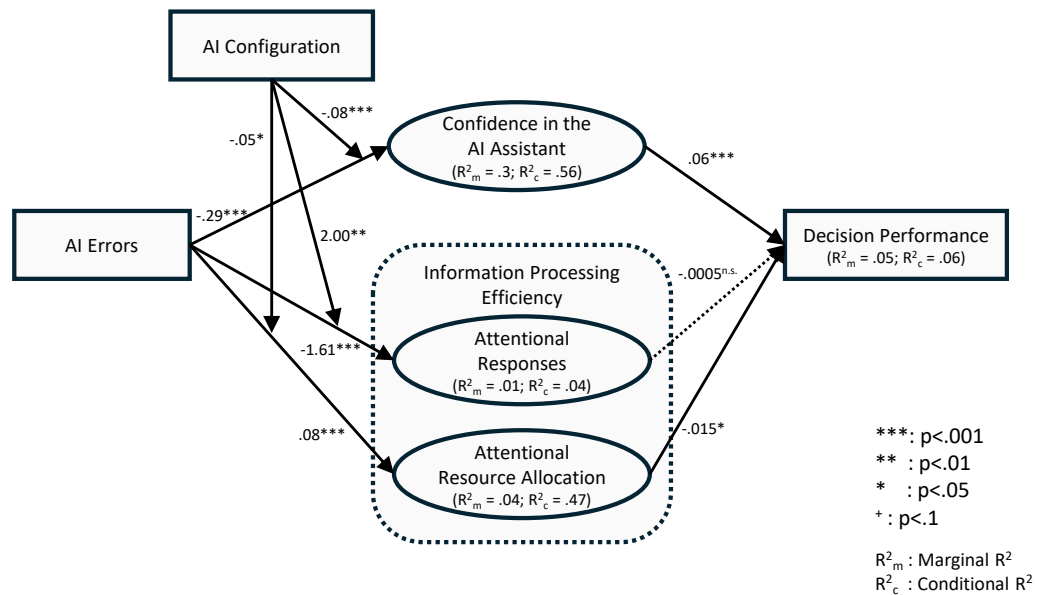


Figure 3.10: Validated research model.

Table 3.4: Hypothesis Testing Summary

Hypothesis	t-value	p-value	Significance
H1: Erroneous AI recommendations will lead to lower user confidence in the AI assistant than non-erroneous recommendations.	-26.56	< .001	***
H4a: The ability to configure the AI assistant will moderate the effect of erroneous AI recommendations on confidence in the AI assistant, such that the error-related decrease in confidence will be smaller when configuration is available.	-5.12	< .001	***
H2: Erroneous AI recommendations will trigger stronger attentional responses than non-erroneous recommendations.	-3.70	< .001	***
H4b: The ability to configure the AI assistant will moderate the effect of erroneous AI recommendations on attentional responses, such that the error-related increase in attentional responses will be greater when configuration is available.	3.06	< .001	***
H3: Erroneous AI recommendations will lead to greater attentional resource allocation on the manual task interface than non-erroneous recommendations.	4.09	< .001	***
H4c: The ability to configure the AI assistant will moderate the effect of erroneous AI recommendations on attentional resource allocation on the manual task interface, such that the error-related increase in allocation will be smaller when configuration is available.	-1.94	.026	*
H5: Greater confidence in the AI assistant will be associated with higher decision performance.	5.35	< .001	***
H6: Greater attentional responses to erroneous AI recommendations will be associated with higher decision performance.	-1.00	.16	n.s.
H7: Lower attentional resource allocation on the manual task interface following erroneous AI recommendations will be associated with higher decision performance.	-1.89	.030	*
H8: The ability to configure the AI assistant, as compared to having no configuration option, will increase users' perceived understanding of the AI assistant.	1.75	.043	*

3.5.6 Post-Hoc Analyses

Mediation Effects

Although not initially hypothesized, the conceptual relationships tested in Hypotheses 1-6 suggest potential mediation pathways linking AI errors to decision performance through confidence in the AI assistant and information processing efficiency, with configuration moderating the immediate impact of AI errors. This section reports post-hoc analyses examining these mediation relationships.

To account for the repeated-measures structure of the experiment while maintaining model parsimony, we first divided the dataset according to the two values of the binary moderator variable (configuration). A nonparametric bootstrap approach was then applied to test mediation effects without assuming normality. For each mediator variable, this approach required two separate analyses corresponding to each configuration condition. Within each condition, we estimated a series of linear mixed-effects models: a full model predicting decision performance from the independent variable and the mediator, a reduced model predicting decision performance from the independent variable only, and a model predicting the mediator from the independent variable. When these models indicated a possible mediation pattern, we conducted a nonparametric bootstrap analysis with 5,000 iterations. In this section, we report only significant mediation effects, while complete model specifications and detailed results are provided in Appendix, Section C.2.10.

Among the tested mediation pathways, only the confidence in the AI assistant under no-configuration condition, and attentional resource allocation on the manual task-relevant elements under the configuration condition suggested potential mediation effects. In these analyses, AI errors served as the independent variable, confidence in the AI assistant (or total fixation duration on the task-relevant information after the onset of AI recommendation) as the mediator, and decision performance as the dependent variable, with the configuration variable fixed to the value of 0 (or 1 in case of the mediation test involving attentional resource allocation on manual task elements).

In the analysis testing mediation through the confidence in the AI assistant, the mediator model showed that AI errors reduced users' confidence in the AI assistant ($\beta = -0.289, p < .001$). The outcome model, which included both AI errors and confidence as predictors of performance, revealed a positive association between confidence in the AI assistant and decision performance ($\beta = 0.033, p = .033$), while the direct effect of AI errors on performance remained negative and significant ($\beta = -0.035, p < .001$). A Monte Carlo simulation with 50,000 iterations was used to estimate the indirect effect. The average indirect effect was negative ($\bar{ab} = -0.0095$), and the 90% confidence interval

$[-0.0181, -0.0010]$ did not include zero, indicating a statistically reliable mediation effect at the 90% confidence level.

The mediator model indicated that AI errors were associated with increased fixation durations ($\beta = 0.431, p = .031$), suggesting stronger post-decision attentional engagement following erroneous AI recommendations. The outcome model, which included both AI errors and fixation duration as predictors of decision performance, revealed a significant negative association between fixation duration and performance ($\beta = -0.018, p = .038$), while the direct effect of AI errors remained negative and significant ($\beta = -0.057, p < .001$). A Monte Carlo simulation with 50,000 iterations was conducted to estimate the indirect effect. The average indirect effect was negative ($\bar{ab} = -0.001$), and the 90% confidence interval $[-0.018, -0.001]$ did not include zero, indicating a statistically significant mediation effect at the 90% confidence level.

These results suggest that when participants had access to AI configuration, AI errors reduced decision performance partly through their effect on attentional responses. Specifically, erroneous recommendations led to prolonged post-decision fixations, reflecting additional cognitive effort or verification behavior, which in turn corresponded to lower decision performance.

Recovery of Confidence in AI Following an AI Error

In this section, we report our post-hoc analysis examining participants' recovery from their initial loss of confidence following AI errors. The analysis aimed to determine whether configuration influenced how quickly participants restored their confidence in the AI assistant after encountering an erroneous AI recommendation.

To create a dataset suitable for survival analysis, we applied the following data processing steps. First, for each participant, we established a baseline level of confidence in the AI assistant by computing the mean of their reported confidence across the first fourteen trials with AI assistance but without any AI errors. We then compared participants' subsequent trial-level confidence ratings against their baseline confidence values and computed the number of trials it took for participants to return to or exceed

their baseline confidence following each AI error. In the survival analysis, the first observation where a participant's confidence in the AI assistant equaled or surpassed their baseline was treated as the stabilization event, and trials prior to that point represented periods during which the stabilization event had not yet occurred, i.e., confidence had not yet recovered. Because the analysis focused on the recovery of confidence in the AI assistant after the first AI error, the dataset excluded trials in which there were erroneous AI recommendations and all trials a participant completed before encountering their first erroneous AI recommendation. The dataset used in this analysis contained 1,946 trials, with approximately a 60% censoring rate in both experimental conditions of configuration. Descriptive statistics of the survival analysis are provided in Table C.32.

Using the trial-level dataset described above, we constructed Kaplan-Meier survival curves to estimate survival probabilities, defined as the probability that confidence remained below baseline as a function of the number of trials since the last AI error. Participants were grouped according to their assigned configuration condition, and a log-rank test indicated a significant difference between conditions ($p = .043$), suggesting that participants with AI configuration access recovered their confidence more rapidly after AI errors. As illustrated in Figure 3.11, both groups showed a gradual decline in survival probability (indicating increasing recovery of confidence), and recovery to the baseline AI confidence levels occurred earlier for those who had AI configuration access.

To complement the nonparametric comparison, we estimated a Cox proportional hazards model to quantify the effect of AI configuration on the rate of confidence recovery (Therneau and Grambsch, 2000). In this model, the hazard represents the instantaneous likelihood that a participant's confidence returned to or exceeded their baseline level at a given trial, given that recovery had not yet occurred. The model revealed a significant positive effect of AI configuration ($\beta = 0.15$, $p = .041$). The corresponding hazard ratio of 1.16 indicates that, at any given trial after an AI error, participants in the configuration condition were approximately 16% more likely to regain their confidence compared to those in the no configuration condition. The model's concordance index ($C = .52$) suggests moderate variability across participants in recovery

speed. These results reinforce the Kaplan-Meier findings by confirming that configuration access systematically accelerated participants’ recovery from confidence loss following AI errors. Details of the Cox model analysis are reported in Appendix, Section C.2, Table C.33.

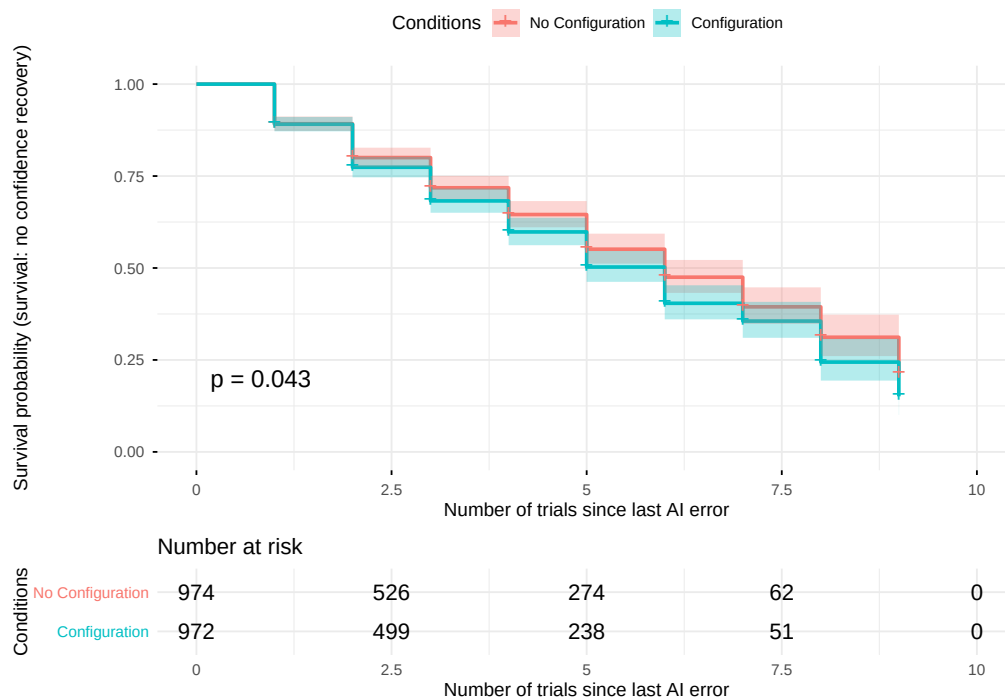


Figure 3.11: Kaplan-Meier survival estimates for confidence recovery following AI errors. Participants in the configuration condition recovered confidence faster ($p = .043$). Risk at 10 trials is zero because no participant experienced an interval longer than 10 trials before encountering another AI error.

3.6 Discussion

3.6.1 Key Findings

In addition to their promised productivity improvements, AI implementations introduce three key tensions from a human-centered AI perspective: human autonomy, user trust and confidence, and users’ perceived complexity (Jiang et al., 2023). This study aims to examine the effectiveness of AI configuration in mitigating the tension between

automation and user agency (Jiang et al., 2023). Drawing on the Situation Awareness (SA) model, it examines how a human-in-the-loop decision-making design, enabled by an AI configuration access, shapes users' confidence, attentional responses, and decision performance in a dynamic, AI-assisted decision-making task in which the AI is prone to errors at random points in time. The SA model conceptualizes user cognition as a multi-level process encompassing perception, comprehension, and projection, where system design features influence how users form, maintain, and recover their awareness of task context and system behavior (Endsley, 1995; Endsley et al., 2003; Endsley, 2023). Our findings provide empirical evidence for this framework in the context of imperfect AI assistance, demonstrating that design features enabling perceived autonomy can both facilitate user understanding and amplify cognitive responses to AI errors. Below, we summarize the main results, which are visually presented in Figure 3.10 and summarized in Table 3.4.

First, as hypothesized in H1, erroneous AI recommendations significantly reduced users' confidence in the AI assistant. This finding confirms that AI errors undermine users' belief in the reliability of algorithmic advice and aligns with prior work showing that users continuously evaluate information system outputs and their implications for decision performance throughout the decision-making process (e.g., the OODA loop (Endsley and Jones, 1997) and the cybernetics view of IS use behavior (Liang and Xue, 2009)). It also supports prior research demonstrating that early system failures are particularly detrimental to users' confidence calibration in AI systems (Dietvorst et al., 2015; Anaraky et al., 2020). Moreover, the moderating effect of configuration (H4a) was negative, indicating that participants with access to AI configuration experienced a larger decrease in confidence following AI errors. Although user control is often assumed to enhance perceived autonomy and trust (Haesevoets et al., 2021; Westphal et al., 2023), our results suggest that exposing system internals may also heighten users' sensitivity to errors and help mitigate the out-of-the-loop problem (Endsley et al., 2003). The faster initial confidence reduction in the configuration condition likely stems from users attributing responsibility for errors to their own configurations or perceiving the system

as less autonomous, thereby intensifying confidence loss when outcomes deviate from expectations. Although confidence in the AI assistant initially declined more sharply with configuration access, post-hoc analyses show that it also recovered more rapidly to baseline levels in this condition.

Regarding EEG measures, we observed a slightly unexpected finding: the two conditions showed opposite polarity around 250 ms in response to erroneous recommendations. Specifically, we observed a negative-going deflection in the no-configuration group, whereas the configuration group showed a positive-going response in the same time window. The negative-going activity can be explained by early attentional mechanisms linked to conflict monitoring or rapid error evaluation in visual tasks. Negative-going deflections in this latency range, often described as an N200, error-related negativity, or feedback-related negativity response, reflect an automatic, stimulus-driven allocation of attention toward unexpected or mismatching information (Folstein and Van Petten, 2008; Luck, 2014; Holroyd and Coles, 2002). In the absence of configuration access, erroneous recommendations appear to have been processed as externally generated anomalies, prompting a discrepancy-detection response rather than a deeper interpretive process. In contrast, the positive-going activity observed in the configuration condition aligns with an early P3a-like orienting component that reflects a higher-level, model-based allocation of attentional resources to events that violate an internal task representation (Polich, 2003). When users believed they had influenced the AI's output, errors triggered a context-updating process rather than a purely stimulus-driven mismatch reaction. This pattern supports the interpretation that configuration access strengthens attentional responses by engaging a more deliberate interpretive mechanism when AI errors occur and suggests that configuration activates distinct cognitive monitoring processes during error detection and correction, consistent with the SA model's proposition that system-level engagement influences users' perception and comprehension of task states.

Third, the results involving eye-tracking metrics (H3, H4c, and H8) indicate that a human-in-the-loop interaction design shapes how efficiently users allocate attention when processing and correcting AI errors. Erroneous recommendations increased fixation

duration on task-relevant information across all conditions, but this effect was weaker when users had access to AI configuration. Participants in the configuration condition spent less time examining task data, suggesting that they relied more on their internal understanding of how the AI operated and how its outputs related to the task, thereby reducing the need for prolonged visual verification. Consistent with the Situation Awareness (SA) model (Endsley, 1995; Endsley et al., 2003), these results imply that configuration access enhances comprehension (Level 2 SA) and projection (Level 3 SA) by prompting users to interpret and anticipate task-relevant contextual information more frequently. This stronger SA, in turn, reduces uncertainty during errors and supports a more efficient error correction process. This pattern aligns with the EEG evidence showing that configuration shifts early error processing from mismatch detection to a more deliberate attentional response that draws on users' internal task and system representations.

Fourth, the results for H5 and H7 revealed a consistent relationship between cognitive states and performance outcomes. Higher confidence in the AI assistant predicted better task performance (H5), indicating that well-calibrated confidence enables effective integration of AI recommendations into decision-making. In contrast, prolonged visual verification (H7) was negatively associated with performance, suggesting that sustained post-decision fixation reflects uncertainty and cognitive inefficiency. No significant relationship was observed between P3a amplitude and performance (H6), implying that neural attentional reactions alone do not directly translate into decision quality without corresponding improvements in confidence and comprehension.

Taken together, these findings highlight the importance of designing AI systems from a human-in-the-loop perspective. Providing users with configuration access not only enhances their understanding of how the AI operates but also supports the development of accurate confidence and more efficient attentional control during decision-making. By ensuring that users remain cognitively engaged and capable of managing AI errors effectively, configuration-based interaction helps mitigate the out-of-the-loop problem by improving users' comprehension of both the AI and the task context (Endsley et al.,

2003). Overall, the results confirm that enabling user involvement in AI-assisted decision processes fosters situational awareness, confidence calibration, and promotes a more adaptive and resilient decision-making process.

3.6.2 Theoretical and Methodological Implications

This study offers several theoretical contributions. A central contribution stems from extending behavioral operations and decision-making research on operations planning tasks by introducing AI decision support as a source of an additional meta-decision-making problem (Benzion et al., 2008; Kremer and Minner, 2008; Kremer et al., 2010; Schweitzer and Cachon, 2000). Rather than focusing on behavioral deviations from optimal benchmarks or on task-driven sources of difficulty, we redirect attention in the behavioral newsvendor literature to the design of the human–AI interaction itself and to the cognitive pathways through which AI assistance shapes decisions. Engaging with a well-established decision paradigm, as suggested by Wang et al. (2025), strengthens the theoretical grounding of our approach while maintaining relevance for a broad research community.

Second, this study extends the literature on the topic of user control over AI and understanding of AI by testing the impact of configuration under imperfect AI performance, a condition rarely explored in prior research. While prior studies have shown that user control enhances confidence in AI and decision performance when AI systems perform reliably (e.g., Westphal et al. 2023), our results demonstrate that user control also supports user attention and understanding, which also helps users when the AI produces errors. By situating these findings within the Situation Awareness (SA) theory (Endsley, 1995; Endsley et al., 2003), it illustrates that allowing users to configure how an AI generates its recommendations improves their attention to AI errors, their understanding, and their projection. These results strengthen the theoretical link between user control and dynamic confidence calibration, emphasizing that control mechanisms play a critical role not only in supporting initial confidence in AI but also in sustaining

effective attentional engagement in task input elements and faster attentional reactions to AI failures.

Methodologically, this study integrates two well-established experimental paradigms, the newsvendor decision task and the oddball paradigm, to examine event-related potentials in an ecologically valid decision-making environment. The findings demonstrate the feasibility of applying event-related potential methods to complex, dynamic decision-support contexts and reveal that users engage distinct cognitive processes when confronted with AI errors, depending on whether they had configuration access. By combining self-reports with eye-tracking data, we further show that fixation duration on task-relevant elements covaries with confidence in the AI assistant, suggesting that visual verification behavior may serve as an implicit indicator of confidence. Future research can build on this approach to evaluate fixation-based or behavioral measures as real-time indicators of user confidence or uncertainty. Finally, by employing a repeated-measures design, this study moves beyond the dominant focus on initial use and investigates how confidence evolves over extended human-AI interaction cycles.

3.6.3 Practical Implications

A key takeaway from this study is that the design of information systems should align with users' cognitive preferences while preserving the reliability and performance of the underlying AI. The results demonstrate that when users are meaningfully engaged in the decision loop, they exhibit stronger attentional responses to AI errors, form a deeper understanding of system behavior, and recover their confidence in AI more effectively after AI errors. Below, we discuss the practical implications of these findings from an enterprise system and human-AI interaction design perspective.

Our findings show that providing users with even limited opportunities to influence the AI's reasoning process can enhance their comprehension, attentional processing, and confidence calibration. Providing users with complete control over model internals may not always be feasible and safe in enterprise environments. Yet, designers can

introduce bounded forms of configuration such as adjustable parameter weights, scenario preferences, tolerance or confidence thresholds, or interpretable settings that shape the decision context without compromising model integrity. These mechanisms have the potential to satisfy users' need for autonomy, increase their engagement with task inputs, and foster more resilient confidence when the AI occasionally fails to produce reliable outcomes. In cases where configuration access cannot be granted to users, designers should consider nudging users to interact with relevant data before receiving AI assistance by using mandatory review steps, input data quality prompts, checklists, or other intentional friction elements.

The fixation results indicate that unexpected or erroneous AI outputs create heightened attentional demands. IS designers can translate this into adaptive support features that detect uncertainty and respond in real time. Although eye-tracking may not be feasible in typical enterprise deployments, similar insights can be inferred from cursor movement, scrolling patterns, extended verification time, or unusual navigation sequences (Grimes and Valacich, 2015). System designers should investigate methods for detecting such behavioral cues so that they can design their systems to provide additional clarification, highlight key variables, present a brief rationale for the recommendation, or deliver contextual examples.

Finally, because AI systems inevitably produce errors, interfaces should explicitly support users during such AI error events. This includes surfacing local uncertainty signals, providing concise justifications or warnings when outputs deviate from typical patterns, and proposing methods for correction. Systems that prepare users for occasional mistakes can help prevent algorithm compliance while nurturing user confidence in AI assistance.

3.6.4 Limitations and Future Research

As with all experimental research, the findings of this study should be interpreted within its methodological boundaries. First, the AI assistant generated identical

recommendations for all participants, regardless of whether they had configuration access or what configurations they submitted. This design choice ensured comparability across conditions but meant that the study primarily examined participants' perceived sense of control rather than the impact of actual control on decision outcomes. Future work should examine how granting users functional control over AI parameters influences performance, confidence calibration, and attention allocation during interaction.

Second, to prevent participants from anticipating AI errors, erroneous recommendations were distributed randomly across trials. As a result, the number of non-erroneous trials following each error varied across participants, limiting the interpretability of the post-hoc analysis of confidence recovery. Future studies should systematically manipulate the timing and frequency of AI errors to more precisely model the temporal dynamics of confidence recovery.

Third, confidence in the AI assistant was measured using a single-item indicator to allow repeated assessment across trials while minimizing participant effort. While suitable for capturing momentary, trial-level evaluations of perceived system ability, this operationalization may provide a narrower representation of users' evaluations than multi-item approaches. Future research should employ more comprehensive measurement instruments to capture additional facets of users' assessment of AI capability and examine how they evolve during longer-term human-AI interaction. In addition, future studies should compare confidence measured prior to observing objective task outcomes, as in this study's design, with confidence measured after objective outcome feedback, to better understand how performance feedback reshapes users' evaluations of AI capability.

Fourth, this study focused on system design factors shaping human-AI interaction; however, as reflected in the explained variance of 5% in decision performance, outcomes are likely influenced by a broader set of determinants. Future research should investigate additional individual- and task-level factors alongside system design characteristics, including users' prior experience, domain expertise, cognitive abilities, and task difficulty, to better understand how personal capabilities and adaptation processes jointly shape performance outcomes in AI-supported decision environments.

3.7 Conclusion

This study examines how user involvement in human-AI interaction influences confidence, attention, and performance in AI-assisted decision-making in cases where AI systems are prone to errors. Grounded in prior human-AI interaction research and the Situation Awareness (SA) model, it shows that involving users through a configuration interface enhances understanding of the AI system and supports confidence recovery after errors. By integrating behavioral, eye-tracking, and neurophysiological measures, the study finds that although AI errors increase attentional demands, users with configuration access manage these situations more efficiently and regain confidence more rapidly. These findings extend existing work by demonstrating that human-in-the-loop design patterns sustain engagement and comprehension even when AI performance is imperfect. Overall, the results highlight the importance of human-in-the-loop system design to preserve human attention, engagement, and understanding in dynamic, error-prone decision environments.

References

- Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160.
- Ahn, D., Almaatouq, A., Gulabani, M., and Hosanagar, K. (2021). Will we trust what we don't understand? Impact of model interpretability and outcome feedback on trust in AI. *arXiv Preprint arXiv:2111.08222*.
- Ahn, J.-H., Bae, Y.-S., Ju, J., and Oh, W. (2018). Attention adjustment, renewal, and equilibrium seeking in online search: an eye-tracking approach. *Journal of Management Information Systems*, 35(4):1218–1250.
- Akhtar, N. and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access*, 6:14410–14430.
- Allen, R. and Choudhury, P. R. (2022). Algorithm-augmented work and domain experience: the countervailing forces of ability and aversion. *Organization Science*, 33(1):149–169.
- Anaraky, R. G., Knijnenburg, B. P., and Risius, M. (2020). Exacerbating mindless compliance: the danger of justifications during privacy decision making in the context of facebook applications. *AIS Transactions on Human-Computer Interaction*, 12(2):70–95.
- Anderson, B. B., Vance, A., Kirwan, C. B., Jenkins, J. L., and Eargle, D. (2016). From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. *Journal of Management Information Systems*, 33(3):713–743.
- Anderson, P. L. (2019). Damages caused by AI errors and omissions: Management complicity, malware, and misuse of data. An-

- derson Economic Group https://www.andersoneconomicgroup.com/wp-content/uploads/06212019_AI_SpecRept_PtI_final.pdf. Accessed: 2022-12-03.
- Ang, L. and Eisend, M. (2018). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research*, 58(2):218–227.
- Balapour, A. and Riedl, R. (2025). Ecological validity in neuroIS research: Theory, evidence, and a roadmap for future studies. *Journal of the Association for Information Systems*, 26(1):9–65.
- Banh, L. and Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1):63.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.
- Barki, H. and Hartwick, J. (1994). Measuring user participation, user involvement, and user attitude. *MIS Quarterly*, 18:59–82.
- Barki, H., Robert, J., and Dulipovici, A. (2015). Reconceptualizing trust: A non-linear boolean model. *Information & Management*, 52(4):483–495.
- Benke, I., Gnewuch, U., and Maedche, A. (2022). Understanding the impact of control levels over emotion-aware chatbots. *Computers in Human Behavior*, 129:107122.
- Benzion, U., Cohen, Y., Peled, R., and Shavit, T. (2008). Decision-making and the newsvendor problem: an experimental study. *Journal of the Operational Research Society*, 59(9):1281–1287.

- Bergkvist, L. and Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, 44(2):175–184.
- Bolton, G. E. and Katok, E. (2008). Learning by doing in the newsvendor problem: a laboratory investigation of the role of experience and feedback. *Manufacturing & Service Operations Management*, 10(3):519–538.
- Bowden, V. K., Griffiths, N., Strickland, L., and Loft, S. (2023). Detecting a single automation failure: the impact of expected (but not experienced) automation reliability. *Human Factors*, 65(4):533–545.
- Bruder, C. and Hasse, C. (2020). What the eyes reveal: Investigating the detection of automation failures. *Applied Ergonomics*, 82:102967.
- Brunyé, T. T. and Gardony, A. L. (2017). Eye tracking measures of uncertainty during perceptual decision making. *International Journal of Psychophysiology*, 120:60–68.
- Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.
- Camerer, C. F. and Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of risk and uncertainty*, 19(1):7–42.
- Card, S. K., Moran, T. P., and Newell, A. (1986). The model human processor: An engineering model of human performance. In Boff, K. R., Kaufman, L., and P., T. J., editors, *Handbook of Perception and Human Performance*., volume 2, pages 1–35. John Wiley & Sons.
- Cecil, J., Lermer, E., Hudecek, M. F., Sauer, J., and Gaube, S. (2024). Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. *Scientific Reports*, 14(1):9736.

- Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., and Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018.
- Compeau, D. R. and Higgins, C. A. (1995). Computer self-efficacy: Development of a measure and initial test. *MIS Quarterly*, 19(2):189–211.
- Coull, J. T. (1998). Neural correlates of attention and arousal: Insights from electrophysiology, functional neuroimaging and psychopharmacology. *Progress in Neurobiology*, 55(4):343–361.
- Currie, G., Hawk, K. E., Rohren, E., Vial, A., and Klein, R. (2019). Machine learning and deep learning in medical imaging: Intelligent imaging. *Journal of Medical Imaging and Radiation Sciences*, 50(4):477–487.
- Davis, F. D., Lohse, G. L., and Kottemann, J. E. (1994). Harmful effects of seemingly helpful information on forecasts of stock earnings. *Journal of Economic Psychology*, 15(2):253–267.
- De Guinea, A. O., Titah, R., and Léger, P.-M. (2013). Measure for measure: A two study multi-trait multi-method investigation of construct validity in is research. *Computers in Human Behavior*, 29(3):833–844.
- De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., and Neerinx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, 12(2):459–478.
- De Vreede, T., Raghavan, M., and De Vreede, G.-J. (2021). Design foundations for AI assisted decision making: A self determination theory approach. In *Proceedings of the 54th Hawaii International Conference on System Sciences*.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.
- Dimoka, A., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Gefen, D., Gupta, A., Ischebeck, A., Kenning, P. H., Pavlou, P. A., Müller-Putz, G., Riedl, R., Vom Brocke, J., and Weber, B. (2012). On the use of neurophysiological tools in is research: developing a research agenda for neuroIS. *MIS Quarterly*, 36(3):679–702.
- Dimoka, A., Pavlou, P. A., and Davis, F. D. (2011). Research commentary—neuroIS: the potential of cognitive neuroscience for information systems research. *Information Systems Research*, 22(4):687–702.
- Dogusoy-Taylan, B. and Cagiltay, K. (2014). Cognitive analysis of experts’ and novices’ concept mapping processes: an eye tracking study. *Computers in Human Behavior*, 36:82–93.
- Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., and 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*, 57:101994.
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718.

- Endsley, M. and Jones, W. M. (1997). Situation awareness information dominance & information warfare. No. AL/CF-TR-1997-0156. United States Air Force Armstrong Laboratory, Wright-Patterson AFB, OH (1997).
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64.
- Endsley, M. R. (2023). Supporting Human-AI teams: Transparency, explainability, and situation awareness. *Computers in Human Behavior*, 140:107574.
- Endsley, M. R., Bolté, B., and Jones, D. G. (2003). *Designing for Situation Awareness: an Approach to User-centered Design*. CRC press.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1):62–70.
- Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G* power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191.
- Ferraro, J. C. and Mouloua, M. (2021). Effects of automation reliability on error detection and attention to auditory stimuli in a multi-tasking environment. *Applied Ergonomics*, 91:103303.
- Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., and Criado, N. (2021). Bias and discrimination in ai: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, 40(2):72–80.
- Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1):111–126.
- Folstein, J. R. and Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. *Psychophysiology*, 45(1):152–170.

- FoodyDirect (2014). Proof we like ice cream all year long. <https://web.archive.org/web/20140727192556/http://blog.foodydirect.com/index.php/proof-like-ice-cream-year-long/>. Accessed: 2023-03-12.
- Fourcade, A. and Khonsari, R. (2019). Deep learning in medical image analysis: a third eye for doctors. *Journal of Stomatology, Oral and Maxillofacial Surgery*, 120(4):279–288.
- Frömer, R., Maier, M., and Abdel Rahman, R. (2018). Group-level eeg-processing pipeline for flexible single trial-based analyses including linear mixed models. *Frontiers in Neuroscience*, 12:48.
- Garcia, S. M., Tor, A., and Schiff, T. M. (2013). The psychology of competition: A social comparison perspective. *Perspectives on Psychological Science*, 8(6):634–650.
- Giboney, J. S., Brown, S. A., Lowry, P. B., and Nunamaker Jr, J. F. (2015). User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems*, 72:1–10.
- Giger, M. L. (2018). Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3):512–520.
- Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.
- Grimes, M. and Valacich, J. (2015). Mind over mouse: the effect of cognitive load on mouse movement behavior. In *ICIS 2015 Proceedings*, number 9.
- Haesevoets, T., De Cremer, D., Dierckx, K., and Van Hiel, A. (2021). Human-machine collaboration in managerial decision making. *Computers in Human Behavior*, 119:106730.

- Hamari, J., Koivisto, J., and Sarsa, H. (2014). Does gamification work?—a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*, pages 3025–3034. Ieee.
- Hancock, J. T., Naaman, M., and Levy, K. (2020). Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100.
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527.
- Hartwick, J. and Barki, H. (1994). Explaining the role of user participation in information system use. *Management Science*, 40(4):440–465.
- Healey, C. and Enns, J. (2011). Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188.
- Herrmann, T. (2025). Intervenability as a design requirement for autonomy and oversight within human-centered AI. In *The Design of Human-Centered Artificial Intelligence for the Workplace*, pages 143–166. Springer.
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2023). Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5.
- Holroyd, C. B. and Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109(4):679.
- Hooge, I. T. C. and Erkelens, C. J. (1996). Control of fixation duration in a simple search task. *Perception & Psychophysics*, 58:969–976.

- Inhoff, A. W. and Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes. *Eye Guidance in Reading and Scene Perception*, pages 29–53.
- Jacobs, A. M. and O'Regan, J. K. (1987). Spatial and/or temporal adjustments of scanning behavior to visibility changes. *Acta Psychologica*, 65(2):133–146.
- Jiang, J., Karran, A. J., Coursaris, C. K., Léger, P.-M., and Beringer, J. (2023). A situation awareness perspective on human-ai interaction: Tensions and opportunities. *International Journal of Human-Computer Interaction*, 39(9):1789–1806.
- Kahneman, D. (2011). *Thinking, Fast and Slow*. Doubleday Canada, Toronto.
- Karran, A., Demazure, T., Léger, P.-M., Labonte-LeMoyne, E., Sénécal, S., Fredette, M., and Babin, G. (2019). Towards a hybrid passive BCI for the modulation of sustained attention using EEG and fNIRS. *Frontiers in Human Neuroscience*, 13(393).
- Kesavan, S. and Kushwaha, T. (2020). Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science*, 66(11):5182–5190.
- Kirwan, C. B., Vance, A., Jenkins, J. L., and Anderson, B. B. (2023). Embracing brain and behaviour: Designing programs of complementary neurophysiological and behavioural studies. *Information Systems Journal*, 33(2):324–349.
- Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395.
- Kremer, M. and Minner, S. (2008). The human element in inventory decision making under uncertainty-a review of experimental evidence in the newsvendor model. *Zeitschrift für Betriebswirtschaft*, pages 83–97.

- Kremer, M., Minner, S., and Van Wassenhove, L. N. (2010). Do random errors explain newsvendor behavior? *Manufacturing & Service Operations Management*, 12(4):673–681.
- Lakhiwal, A., Bala, H., and Léger, P.-M. (2023). Ambivalence is better than indifference: a behavioral and neurophysiological assessment of ambivalence in online environments. *MIS Quarterly*, 47(2):705–732.
- Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.
- Lee, M. K., Jain, A., Cha, H. J., Ojha, S., and Kusbit, D. (2019). Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.
- Léger, P., Robert, J., Babin, G., Pellerin, R., and Wagner, B. (2007). Erpsim. *ERPsim Lab (erpsim.hec.ca)*, HEC Montreal, Montreal, Qc.
- Léger, P.-M. (2006). Using a simulation game approach to teach enterprise resource planning concepts. *Journal of Information Systems Education*, 17:441–447.
- Liang and Xue (2009). Avoidance of information technology threats: a theoretical perspective. *MIS Quarterly*, 33(1):71.
- Lin, H., Hsieh, Y.-C., and Wu, F.-G. (2016). A study on the relationships between different presentation modes of graphical icons and users’ attention. *Computers in Human Behavior*, 63:218–228.
- Liu, P. (2024). Reflections on automation complacency. *International Journal of Human-Computer Interaction*, 40(22):7347–7363.
- Loos, P., Riedl, R., Müller-Putz, G. R., Vom Brocke, J., Davis, F. D., Banker, R. D., and Léger, P.-M. (2010). neuroIS: Neuroscientific approaches in the investigation and

- development of information systems. *Business & Information Systems Engineering*, 2:395–401.
- Lu, Y. and Sarter, N. (2019). Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems*, 49(6):560–568.
- Luck, S. J. (2014). *An Introduction to the Event-related Potential Technique*. MIT press.
- Malaguti, P., Karran, A. J., Le, D., Mortin, H., Coursaris, C. K., Sénécal, S., and Léger, P.-M. (2025). Investigating interaction friction in generative AI: Improving user experience and decision-making. In *SIGHCI 2024 Proceedings*, number 26.
- Manzey, D., Reichenbach, J., and Onnasch, L. (2012). Human performance consequences of automated decision aids: the impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6(1):57–87.
- Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.
- McKenzie, S., Podsakoff, P., and Podsakoff, N. (2011). Construct measurement and validation procedures in mis and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(2):293–334.
- Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., and Shirase, L. (2019). Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology*, 10:225.
- Müller-Putz, G. R., Riedl, R., and Wriessnegger, S. C. (2015). Electroencephalography (EEG) as a research tool in the information systems discipline: Foundations, measurement, and applications. *Communications of the Association for Information Systems*, 37.

- National Bureau of Economic Research (2012). Ice Cream Production for United States [M0174BUSM411NNBR]. <https://fred.stlouisfed.org/series/M0174BUSM411NNBR>. Accessed: March 14, 2025.
- Nunes, I., Miles, S., Luck, M., and De Lucena, C. J. (2012). Investigating explanations to justify choice. In *User Modeling, Adaptation, and Personalization: 20th International Conference, UMAP 2012, Montreal, Canada, July 16-20, 2012. Proceedings 20*, pages 212–224. Springer.
- O’Neill, E. C., Kong, Y. X. G., Connell, P. P., Ong, D. N., Haymes, S. A., Coote, M. A., and Crowston, J. G. (2011). Gaze behavior among experts and trainees during optic disc examination: Does how we look affect what we see? *Investigative Ophthalmology & Visual Science*, 52(7):3976–3983.
- Onnasch, L., Wickens, C. D., Li, H., and Manzey, D. (2014). Human performance consequences of stages and levels of automation: an integrated meta-analysis. *Human Factors*, 56(3):476–488.
- Oz, B., Tran-Nguyen, K., Coursaris, C. K., Robert, J., and Léger, P.-M. (2020). Using digital nudges on analytics dashboards to reduce anchoring bias. In *SIGHCI 2020 Proceedings*, number 3.
- Parasuraman, R. and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors*, 52(3):381–410.
- Passalacqua, M., Pellerin, R., Yahia, E., Magnani, F., Rosin, F., Joblot, L., and Léger, P.-M. (2025). Practice with less AI makes perfect: Partially automated AI during training leads to better worker motivation, engagement, and skill acquisition. *International Journal of Human-Computer Interaction*, 41(4):2268–2288.
- Pettit, T. J., Croxton, K. L., and Fiksel, J. (2019). The evolution of resilience in supply chain management: a retrospective on ensuring supply chain resilience. *Journal of Business Logistics*, 40(1):56–65.

- Polich, J. (2003). Theoretical overview of P3a and P3b. In *Detection of Change: Event-related Potential and fMRI Findings*, pages 83–98. Springer.
- Rai, A., Constantinides, P., and Sarker, S. (2019). Next generation digital platforms: Toward Human-AI hybrids. *MIS Quarterly*, 43(1):iii–ix.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372.
- Ribera, M. and Lapedriza, A. (2019). Can we do better explanations? A proposal of user-centered explainable AI. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019*. ACM New York, NY, USA.
- Riedl, R., Fischer, T., Léger, P.-M., and Davis, F. D. (2020). A decade of neuroIS research: Progress, challenges, and future directions. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 51(3):13–54.
- Riedl, R. and Léger, P.-M. (2016). *Fundamentals of NeuroIS: Information Systems and the Brain*. Springer, Germany.
- Schweitzer, M. E. and Cachon, G. P. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science*, 46(3):404–420.
- Shneiderman, B. and Plaisant, C. (2006). Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, page 1–7, New York, NY, USA. Association for Computing Machinery.
- Söllner, M., Arnold, T., Benlian, A., Bretschneider, U., Knight, C., Ohly, S., Rudkowski, L., Schreiber, G., and Wendt, D. (2025). ChatGPT and beyond: Exploring the responsible use of generative AI in the workplace. *Business & Information Systems Engineering*, 67(2):289–303.

- Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., and Leahy, R. M. (2011). Brainstorm: a user-friendly application for MEG/EEG analysis. *Computational Intelligence and Neuroscience*, 2011(1):879716.
- Therneau, T. M. and Grambsch, P. M. (2000). The cox model. In *Modeling survival data: extending the Cox model*, pages 39–77. Springer.
- Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., and Fischl, M. (2021). Artificial intelligence in supply chain management: a systematic literature review. *Journal of Business Research*, 122:502–517.
- Toreini, P., Langner, M., Maedche, A., Morana, S., and Vogel, T. (2022). Designing attentive information dashboards. *Journal of the Association for Information Systems*, 23(2):521–552.
- Vandewaetere, M. and Clarebout, G. (2013). Cognitive load of learner control: Extraneous or germane load? *Education Research International*, 2013(1):902809.
- Vasseur, A., Passalacqua, M., Sénécal, S., and Léger, P.-M. (2023). The use of eye-tracking in information systems research: a literature review of the last decade. *AIS Transactions on Human-Computer Interaction*, 15(3):292–321.
- Wang, K., Hou, W., Ma, H., and Hong, L. (2024). Eye-tracking characteristics: Unveiling trust calibration states in automated supervisory control tasks. *Sensors*, 24(24):7946.
- Wang, Y., Xu, M., and Zhao, L. (2025). Exploring the cognitive mechanisms behind the adoption of algorithmic advice: a neuro-is study of newsvendor decision-making. *Business & Information Systems Engineering*, pages 1–25.
- Wanous, J. P., Reichers, A. E., and Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology*, 82(2):247.

- Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., and Rafaeli, A. (2023). Decision control and explanations in human-ai collaboration: Improving user perceptions and compliance. *Computers in Human Behavior*, page 107714.
- Wickens, C. D., Helton, W. S., Hollands, J. G., and Banbury, S. (2021). *Engineering Psychology and Human Performance*. Routledge.
- Xu, S., Gong, Z., Li, Y., Zhao, Y., and Huang, S. (2025). The trust cost of AI errors: Examining cognitive and affective trust dynamics in Human-AI collaboration. *International Journal of Human-Computer Interaction*, pages 1–16.
- Zhang, T., Yang, J., Chen, M., Li, Z., Zang, J., and Qu, X. (2024). EEG-based assessment of driver trust in automated vehicles. *Expert Systems with Applications*, 246:123196.
- Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305.
- Zhou, J. and Chen, F. (2019). Towards trustworthy Human-AI teaming under uncertainty. In *IJCAI 2019 Workshop on Explainable AI (XAI)*.

Conclusion

This thesis aims to address the following research question: *How do interface-level design choices in AI-supported organizational decision systems shape users' attention, understanding, confidence calibration, and decision performance?* Rather than treating these outcomes as direct consequences of AI accuracy or adoption, the thesis examines how they emerge from users' ongoing interaction with AI systems and from the cognitive conditions created by interface design. Building on evidence from semi-structured interviews and quantitative analyses of experimental data, this section synthesizes the contributions of the three empirical studies in the thesis.

Taken together, the three studies indicate that these cognitive and behavioral outcomes depend on how system behavior is made visible, interpretable, and actionable to users. In particular, the findings suggest that interface design influences how users form and revise mental models of AI systems, how they allocate attention to verification and monitoring activities, and how they recalibrate their confidence in AI over repeated use.

The first study illustrates that user understanding of AI develops through continued interaction, as users assess AI behavior by inspecting outputs, questioning recommendations, and intervening when necessary. The second study shows that AI-to-human interface mechanisms, specifically, transparency and explainability, support users' understanding in different ways. The results also show a diminished benefit from the combination of those two mechanisms, highlighting that providing users with more information is not the optimal solution. On the other hand, the third study shows that users report a higher level of understanding of AI when they are provided with a human-to-AI interface, namely, an

AI configuration functionality.

The dissertation further shows that confidence calibration is closely tied to interface design and is dynamic, rather than being a stable attitudinal state. In the qualitative study, users describe their confidence in AI as being continuously reassessed through meta-decision-making about delegation. By including AI errors as a within-participants manipulation in its repeated-measures experimental design, the third study shows that confidence in AI responds dynamically to system performance. Further analyses show that temporal changes in confidence in AI are asymmetric: confidence in AI drops sharply following an AI error, and recovery requires several trials with non-erroneous AI recommendations. Analyses in the third study also show that AI configuration functionality improves confidence recovery, highlighting that confidence calibration depends on users' perceived ability to anticipate and manage AI behavior. In line with these findings, the second study shows that an AI-to-human interface, transparency, also improves confidence in AI.

Attentional measurements from the second and third studies provide additional evidence for how interface design shapes human-AI interactions. The second study shows that when transparency or explainability is absent, users increase attentional shifts between the AI assistant and task-relevant information, indicating greater verification effort. In repeated interactions, transparency is associated with a reduced need for verification efforts. The third study extends this result by showing that configuration access alters early attentional responses to AI errors, shifting processing toward more deliberate, model-based evaluation rather than purely stimulus-driven mismatch detection. These findings indicate that interface design influences not only where attention is directed, but also how users process information and respond to unexpected system behavior during AI-assisted decision-making.

In sum, the dissertation shows that interface-level design choices can influence decision performance indirectly, by shaping users' confidence in AI, their attention allocation to verification and monitoring, and their ability to recalibrate confidence in repeated human-AI interactions. Taken together, the thesis responds to calls for research

on design characteristics of successful conversational systems (Janssen et al., 2020), AI-based system explanation techniques (Bauer et al., 2021; Maedche et al., 2019), and human-centered AI applications in organizational decision-making (Söllner et al., 2025). The remainder of this conclusion discusses implications for research and practice, and directions for future research.

Theoretical Implications

This thesis makes three primary theoretical contributions to research on human-AI interaction and AI-supported organizational decision-making. First, it extends Situation Awareness (SA) theory to AI-supported decision contexts by showing that users' awareness encompasses not only their understanding of the task environment, but also their interpretation of how the AI system perceives, processes, and anticipates that environment. Across the studies, users are shown to continuously evaluate the alignment between their own situation awareness and their perception of the system's awareness. By introducing the notion of attributed system situation awareness, the thesis provides a theoretical basis for explaining how user confidence in AI develops through ongoing comparison rather than as a static attitude or one-time judgment (Endsley, 1995; Endsley et al., 2003).

Regarding the second source of theoretical contributions, this thesis clarifies the conceptual distinction between transparency and explainability by disentangling and operationalizing these two frequently conflated design constructs. Drawing on established definitions in the literature, transparency is conceptualized as system-level information that supports users' general understanding of AI behavior, whereas explainability is understood as output-level justifications tied to specific recommendations (Chen et al., 2014; Vilone and Longo, 2021). By independently manipulating these constructs within the same study and illustrating their instantiation in a contemporary GenAI assistant context, the thesis responds to Whetten's (1989) call for clearly defined constructs situated within a coherent nomological network.

The third theoretical contribution stems from theorizing and empirically testing relationships between these design constructs and user cognition and behavior. Building on the SA model, the thesis examines how transparency, explainability, and configuration shape users' confidence calibration, attention allocation, and decision performance over repeated interactions. The findings clarify when and how design attributes improve or hinder performance by showing that their effects depend on the interaction stage, task dynamics, and the cognitive demands introduced by the interface. In line with Whetten's (1989) view that theoretical contributions arise from changes to causal maps, the results demonstrate that interface-level design choices influence AI use through dynamic processes of understanding, verification, and delegation, rather than through binary adoption or reliance mechanisms (Bansal et al., 2021; Dietvorst et al., 2015; Allen and Choudhury, 2022; Anaraky et al., 2020).

Finally, this thesis extends research on behavioral operations and organizational decision-making by examining human-AI interaction within the well-established newsvendor problem, a central experimental paradigm in operations management (Benzion et al., 2008; Kremer and Minner, 2008; Kremer et al., 2010; Schweitzer and Cachon, 2000; Wang et al., 2025). Prior work in this literature has largely focused on task difficulty, behavioral biases, or deviations from normative benchmarks. By introducing AI decision support into this setting and systematically varying interface-level design features, the thesis highlights how interaction design shapes users' cognition and behavior during AI-supported decision-making. Using a well-known decision context as an analytical baseline strengthens the interpretability of the findings and illustrates how established paradigms can be leveraged to study contemporary human-AI interaction phenomena.

Methodological Contributions

This thesis makes several methodological contributions to the empirical study of human-AI interaction by demonstrating how cognitive and behavioral processes can be examined in ecologically valid, repeated-interaction settings. Rather than treating AI use as a

one-shot advice-taking event, the experimental studies modeled decision-making as an ongoing process in which users repeatedly interacted with the same AI system under evolving conditions. By involving multiple trials and repeated measures, the experimental designs captured how user confidence, attention allocation, and performance developed, deteriorated, and recovered over time. This approach enabled the investigation of adaptation in verification behavior and the calibration of confidence in AI.

Second, the thesis illustrates the value of integrating behavioral, attentional, and neurophysiological measures within a single research program. Eye-tracking was used to capture attention allocation and attention shift rates as indicators of verification effort and information-seeking behavior. Drawing on prior work in reading, problem-solving, and diagnostic reasoning (Rayner et al., 2011; Engbert et al., 2005; Jacob and Karn, 2003; O'Neill et al., 2011), fixation durations on task-relevant information and transitions between interface elements were treated as behavioral signals of uncertainty and reliance on AI support. Across studies, these measures covaried with self-reported confidence and performance, supporting their use as implicit indicators of user states during AI-assisted decision-making.

Third, the thesis contributes methodologically by addressing calls in NeuroIS for greater ecological validity in the study of cognition during information system use (Balapour and Riedl, 2025). By embedding AI assistants within a realistic ERP-based business simulation, the research demonstrates how controlled experimentation can be combined with complex, domain-relevant decision tasks. This design enabled the use of real-time physiological measures, including eye-tracking and EEG, to capture attention allocation, attention shifts, and early neural responses as indicators of information processing during AI-assisted decision-making. Moreover, by modeling human-AI interaction as a temporal process across repeated trials, the thesis empirically demonstrates that cognitive responses to AI evolve over time, highlighting the importance of longitudinal designs for studying AI-supported decision-making.

Finally, the thesis contributes methodologically by illustrating how the systematic design of GenAI assistants can be used as an experimental manipulation tool. To

maintain experimental control, multiple GenAI assistants were configured through carefully constructed system prompts and controlled interface-level visibility changes to manipulate transparency and explainability while preserving response consistency. With this approach, the second study demonstrates how GenAI behavior can be experimentally shaped in a replicable manner, providing guidance for future studies seeking to investigate interface-level AI design features. Together, these design and measurement choices offer methodological guidance for future experimental work on human-AI interaction in settings that reflect continued, real-world system use.

Practical Implications

Taken together, the three studies suggest that effective deployment of AI in organizational decision-making requires moving beyond a purely technical, performance-oriented view of AI systems toward a cognitive, human-centered perspective. The findings show that user outcomes are shaped not only by the accuracy of AI outputs, but also by how well systems help users understand what the AI is doing, how it reaches its recommendations, and how its behavior may change over time. From a practical standpoint, this implies that AI systems should be designed to support users' situation awareness and help them decide when and how to use AI recommendations during the decision-making process.

A central implication across the studies is the importance of giving users control over the level of AI involvement. Users do not rely on AI in a fixed or continuous manner; instead, they repeatedly decide whether to accept, question, or override AI recommendations as situations unfold. AI systems should therefore allow users to easily adjust how much support the AI provides at any point during the decision-making process. In practice, this means enabling users to influence inputs, adjust settings, or intervene in recommendations without friction, rather than forcing them to either fully accept AI outputs or abandon the system altogether. Such designs help users remain engaged while maintaining flexibility in how AI support is used.

The findings also indicate that presenting more information to users does not neces-

sarily improve decision outcomes. Providing extensive transparency and explainability features at the same time can increase cognitive demands without yielding additive benefits. Instead, the results point to an adaptive design approach in which different types of information are emphasized at different stages of use. Early interactions benefit from transparency that helps users form an initial understanding of how the AI works at an abstract level, while later interactions benefit more from targeted explanations that support verification, error handling, and confidence calibration.

Finally, the studies highlight the practical value of attending to how users interact with AI systems over time. Patterns of attention, verification, intervention, and responses to AI errors provide meaningful signals about whether users understand and trust the system. Even when advanced sensing technologies such as eye-tracking are not available, similar insights can be inferred from behavioral cues such as repeated overrides, interaction timing, or navigation behaviors (Grimes and Valacich, 2015). Using these signals to adapt system behavior and interface design may allow organizations to improve AI support over time while preserving user autonomy and trust.

Future Research Opportunities

The findings of this thesis highlight the relevance of longitudinal research designs for understanding human-AI interaction as an evolving process. While the present studies capture repeated interactions within controlled experimental sessions, they primarily reflect early-stage engagement rather than long-term use. Future research should therefore examine how attention allocation, confidence calibration, understanding, and decision performance develop over extended periods of interaction, particularly after stable usage patterns and habits emerge. Longitudinal field studies could investigate when reliance behaviors stabilize, how confidence evolves before and after exposure to objective performance feedback, and how users recalibrate delegation strategies following repeated AI successes and failures.

Future research should also further examine the cognitive mechanisms underlying

observed behavioral effects. Although this thesis interprets findings through the lens of situation awareness, alternative explanations related to cognitive load, information overload, or interruption effects may operate alongside attentional processes. Experimental designs that directly manipulate these mechanisms, or combine behavioral measures with richer physiological indicators, would help disentangle the specific cognitive pathways through which interface design influences human-AI interaction outcomes.

Another important direction concerns ecological realism and system design assumptions. The experimental studies relied on simulated decision environments and predefined configuration interfaces to ensure experimental control. Future work should investigate how the identified mechanisms operate in real organizational AI systems, where models continuously learn and evolve. Moreover, it should be noted that allowing users complete configuration of an AI system may not always be a feasible method to improve user attention. Future research should investigate adaptive or system-initiated personalization as alternative means of maintaining user attention.

At the measurement level, future studies should expand the operationalization of key constructs. Confidence in AI was measured using a single-item indicator to enable repeated assessment across trials; however, richer multi-item instruments could capture broader evaluative dimensions and distinguish confidence assessed before observing objective outcomes from confidence measured after performance feedback. Similarly, future work should further investigate performance outcomes using designs capable of capturing greater behavioral variance across longer time horizons, which may improve explanatory power and clarify how decision performance evolves beyond initial interaction stages.

Beyond the individual level, an important direction for future research lies in understanding how AI perceptions and interaction experiences propagate within teams and organizations. Individual users' confidence in AI, attributed system awareness, or willingness to rely on recommendations may shape other team members' perceptions through communication, observation, and shared decision processes. Future studies could examine how heterogeneous individual experiences with AI converge into shared

team-level understandings, how these collective perceptions influence coordination and reliance, and how organizational norms and policies regarding AI use emerge over time.

Finally, future research should investigate intervention mechanisms aimed at educating end users and increasing awareness of ethical and privacy-related issues surrounding AI use. As many AI systems are deployed as software-as-a-service solutions, user data are often stored and processed by external providers. Future work could examine how users understand these data practices, how privacy concerns influence interaction and adoption, and which design interventions or educational approaches effectively increase public awareness of data usage and privacy risks. Integrating these questions with cognitive and behavioral outcomes would support a more comprehensive understanding of responsible and sustainable human-AI interaction.

Bibliography

- Agrawal, N., Cohen, M. A., Deshpande, R., and Deshpande, V. (2024). How machine learning will transform supply chain management. *Harvard Business Review*, 103(3-4):128–137.
- Allen, R. and Choudhury, P. R. (2022). Algorithm-augmented work and domain experience: the countervailing forces of ability and aversion. *Organization Science*, 33(1):149–169.
- Anaraky, R. G., Knijnenburg, B. P., and Risius, M. (2020). Exacerbating mindless compliance: the danger of justifications during privacy decision making in the context of facebook applications. *AIS Transactions on Human-Computer Interaction*, 12(2):70–95.
- Arksey, H. and O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1):19–32.
- Bader, V. and Kaiser, S. (2019). Algorithmic decision-making? the user interface and its role for human involvement in decisions supported by artificial intelligence. *Organization*, 26(5):655–672.
- Bainbridge, L. (1983). Ironies of automation. In *Analysis, Design and Evaluation of Man-machine Systems*, pages 129–135. Elsevier.
- Balapour, A. and Riedl, R. (2025). Ecological validity in neuroIS research: Theory,

- evidence, and a roadmap for future studies. *Journal of the Association for Information Systems*, 26(1):9–65.
- Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. (2019). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.
- Baryannis, G., Dani, S., and Antoniou, G. (2019). Predicting supply chain risks using machine learning: the trade-off between performance and interpretability. *Future Generation Computer Systems*, 101:993–1004.
- Bauer, K., Hinz, O., van der Aalst, W., and Weinhardt, C. (2021). Explaining it to me—explainable AI and information systems research. *Business & Information Systems Engineering*, 63:79–82.
- Benzion, U., Cohen, Y., Peled, R., and Shavit, T. (2008). Decision-making and the newsvendor problem: an experimental study. *Journal of the Operational Research Society*, 59(9):1281–1287.
- Bigras, E., Jutras, M.-A., Sénécal, S., Léger, P.-M., Black, C., Robitaille, N., Grande, K., and Hudon, C. (2018). In AI we trust: Characteristics influencing assortment planners’ perceptions of AI based recommendation agents. In *International Conference on HCI in Business, Government, and Organizations*, pages 3–16. Springer.
- Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., and Detyniecki, M. (2022). Contextualization and exploration of local feature importance explanations to improve understanding

- and satisfaction of non-expert users. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 807–819.
- Brauner, P. and Ziefle, M. (2019). Why consider the human-in-the-loop in automated cyber-physical production systems? Two cases from cross-company cooperation. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, volume 1, pages 861–866. IEEE.
- Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942.
- Cain, S. (2022). A.I. is getting your holiday gifts to you more efficiently than ever. <https://fortune.com/2022/12/20/tech-forward-everyday-ai-holiday-shipping-ups-fedex-amazon/>. Accessed: 2024-01-02.
- Cecil, J., Lermer, E., Hudecek, M. F., Sauer, J., and Gaube, S. (2024). Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. *Scientific Reports*, 14(1):9736.
- Chen, H., Ma, X., Rives, H., Serpedin, A., Yao, P., and Rameau, A. (2024). Trust in machine learning driven clinical decision support tools among otolaryngologists. *The Laryngoscope*, 134(6):2799–2804.
- Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., and Barnes, M. (2014). Situation awareness-based agent transparency. Technical report, Defense Technical Information Center, Fort Belvoir, VA.
- Czako, Z., Sebestyen, G., and Hangan, A. (2021). Automaticai—A hybrid approach for automatic artificial intelligence algorithm selection and hyperparameter tuning. *Expert Systems with Applications*, 182:115225.
- Daudt, H. M., van Mossel, C., and Scott, S. J. (2013). Enhancing the scoping study methodology: A large, inter-professional team’s experience with Arksey and O’Malley’s framework. *BMC Medical Research Methodology*, 13(1):48.

- De Boer, R. and Dekker, S. (2017). Models of automation surprise: Results of a field survey in aviation. *Safety*, 3(3):20.
- Dellermann, D., Lipusch, N., Ebel, P., and Leimeister, J. M. (2019). Design principles for a hybrid intelligence decision support system for business model validation. *Electronic Markets*, 29(3):423–441.
- Di Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., and Zahidi, S. (2023). Future of jobs report 2023. World Economic Forum.
- Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.
- Dikmen, M. and Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: a case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792.
- Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., Roubaud, D., and Hazen, B. T. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: a study of manufacturing organisations. *International Journal of Production Economics*, 226:107599.
- Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L. C., Misra, S., Mogaji, E., Sharma, S. K., Singh, J. B., Raghavan, V., Raman, R., Rana, N. P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., and 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*, 57:101994.
- Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64.
- Endsley, M. R. (2023). Ironies of artificial intelligence. *Ergonomics*, 66(11):1656–1668.
- Endsley, M. R., Bolté, B., and Jones, D. G. (2003). *Designing for Situation Awareness: an Approach to User-centered Design*. CRC press.
- Engbert, R., Nuthmann, A., Richter, E. M., and Kliegl, R. (2005). Swift: a dynamical model of saccade generation during reading. *Psychological Review*, 112(4):777.
- Erickson, B. J., Korfiatis, P., Akkus, Z., and Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics*, 37(2):505–515.
- Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1):62–70.
- Gentile, D., Donmez, B., and Jamieson, G. A. (2023). Human performance consequences of normative and contrastive explanations: an experiment in machine learning for reliability maintenance. *Artificial Intelligence*, 321:103945.
- Geuss, M. N., Larkin, G., Swoboda, J., Yu, A., Bakdash, J., White, T., Berthiaume, M., and Lance, B. (2019). Intelligent squad weapon: Challenges to displaying and interacting with artificial intelligence in small arms weapon systems. In *Artificial Intelligence and Machine Learning for Multi-domain Operations Applications*, volume 11006, pages 296–307. SPIE.

- Gorre, N., Fuhrman, J., Carranza, E., Li, H., Madduri, R. K., Giger, M., and El Naqa, I. (2023). User experience evaluation for midrc AI interface. In *Medical Imaging 2023: Imaging Informatics for Healthcare, Research, and Applications*, volume 12469, page 1246904. SPIE.
- Grimes, M. and Valacich, J. (2015). Mind over mouse: the effect of cognitive load on mouse movement behavior. In *ICIS 2015 Proceedings*, number 9.
- Harbarth, L., Gößwein, E., Bodemer, D., and Schnaubert, L. (2025). (Over)Trusting AI recommendations: How system and person variables affect dimensions of complacency. *International Journal of Human–Computer Interaction*, 41(1):391–410.
- Hassenzahl, M. and Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & Information Technology*, 25(2):91–97.
- Hauptman, A. I., Schelble, B. G., Duan, W., Flathmann, C., and McNeese, N. J. (2024). Understanding the influence of AI autonomy on AI explainability levels in Human-AI teams using a mixed methods approach. *Cognition, Technology & Work*, 26(3):435–455.
- Hepenstal, S., Kodagoda, N., Zhang, L., Paudyal, P., and Wong, B. W. (2019). Algorithmic transparency of conversational agents. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019*.
- Herm, L.-V., Steinbach, T., Wanner, J., and Janiesch, C. (2022). A nascent design theory for explainable intelligent systems. *Electronic Markets*, 32(4):2185–2205.
- Hoque, M. N. and Mueller, K. (2021). Outcome-explorer: a causality guided interactive visual interface for interpretable algorithmic decision making. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):4728–4740.
- IBM Newsroom (2024). Data suggests growth in enterprise adoption of AI is due to widespread deployment by early adopters. <https://newsroom.ibm.com/2024->

- 01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters. Accessed: 2024-04-07.
- Jacob, R. J. and Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The Mind's Eye*, pages 573–605. Elsevier.
- Janssen, A., Passlick, J., Rodríguez Cardona, D., and Breitner, M. H. (2020). Virtual assistance in any context: a taxonomy of design elements for domain-specific chatbots. *Business & Information Systems Engineering*, 62:211–225.
- Jiang, E., Toh, E., Molina, A., Olson, K., Kayacik, C., Donsbach, A., Cai, C. J., and Terry, M. (2022). Discovering the syntax and strategies of natural language programming with generative language models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, New Orleans LA USA. ACM.
- Johnson, M., Albizri, A., Harfouche, A., and Tutun, S. (2023). Digital transformation to mitigate emergency situations: Increasing opioid overdose survival rates through explainable artificial intelligence. *Industrial Management & Data Systems*, 123(1):324–344.
- Karahasanovic, A., Gausta Nilsson, E., Grani, G., Fredrik Pettersen, K., Karabeg, L., and Schittekat, P. (2021). User involvement in the design of ML-Infused systems. In *CHI Greece 2021: 1st International Conference of the ACM Greek SIGCHI Chapter*, CHI Greece 2021, New York, NY, USA. Association for Computing Machinery.
- Katzburg, O., Roimi, M., Frenkel, A., Ilan, R., Bitan, Y., et al. (2024). The impact of information relevancy and interactivity on intensivists’ trust in a machine learning-based bacteremia prediction system: Simulation study. *JMIR Human Factors*, 11(1):e56924.

- Kellner, D., Lowin, M., von Zahn, M., and Chen, J. (2021). Towards designing a user-centric decision support system for predictive maintenance in SMEs. In *Informatik 2021*, pages 1255–1260. Gesellschaft für Informatik, Bonn.
- Kempton, A., Parmiggiani, E., and Vassilakopoulou, P. (2023). Accountability in managing artificial intelligence: State of the art and a way forward for information systems research. *ECIS 2023 Research Papers*.
- Khosrowabadi, N., Hoberg, K., and Imdahl, C. (2022). Evaluating human behaviour in response to AI recommendations for judgemental forecasting. *European Journal of Operational Research*, 303(3):1151–1167.
- Kim, Y., Wu, J., Abdulle, Y., Gao, Y., and Wu, H. (2024). Human-in-the-loop chest x-ray diagnosis: Enhancing large multimodal models with eye fixation inputs. In Chen, H., Zhou, Y., Xu, D., and Vardhanabhuti, V. V., editors, *Trustworthy Artificial Intelligence for Healthcare*, pages 66–80, Cham. Springer Nature Switzerland.
- Krause, J., Perer, A., and Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697.
- Kremer, M. and Minner, S. (2008). The human element in inventory decision making under uncertainty-a review of experimental evidence in the newsvendor model. *Zeitschrift für Betriebswirtschaft*, pages 83–97.
- Kremer, M., Minner, S., and Van Wassenhove, L. N. (2010). Do random errors explain newsvendor behavior? *Manufacturing & Service Operations Management*, 12(4):673–681.
- Lamy, J.-B., Sekar, B., Guezennec, G., Bouaud, J., and Séroussi, B. (2019). Explainable artificial intelligence for breast cancer: a visual case-based reasoning approach. *Artificial Intelligence in Medicine*, 94:42–53.

- Laxar, D., Eitenberger, M., Maleczek, M., Kaider, A., Hammerle, F. P., and Kimberger, O. (2023). The influence of explainable vs non-explainable clinical decision support systems on rapid triage decisions: a mixed methods study. *BMC Medicine*, 21(1):359.
- Levac, D., Colquhoun, H., and O'Brien, K. K. (2010). Scoping studies: Advancing the methodology. *Implementation Science*, 5(1):69.
- Liu, B. (2021). In AI we trust? effects of agency locus and transparency on uncertainty reduction in human-AI interaction. *Journal of Computer-mediated Communication*, 26(6):384–402.
- Lu, W. (2024). Inevitable challenges of autonomy: Ethical concerns in personalized algorithmic decision-making. *Humanities and Social Sciences Communications*, 11(1):1–9.
- Madsen, A. N., Brandsæter, A., van de Merwe, K., and Park, J. (2025). Improving decision transparency in autonomous maritime collision avoidance. *Journal of Marine Science and Technology*, 30(1):134–152.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., and Söllner, M. (2019). AI-based digital assistants: Opportunities, threats, and research perspectives. *Business & Information Systems Engineering*, 61:535–544.
- Mediavilla, M. A., Dietrich, F., and Palm, D. (2022). Review and analysis of artificial intelligence methods for demand forecasting in supply chain management. *Procedia CIRP*, 107:1126–1131.
- Merritt, S. M., Ako-Brew, A., Bryant, W. J., Staley, A., McKenna, M., Leone, A., and Shirase, L. (2019). Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in Psychology*, 10:225.
- Micallef, L., Sundin, I., Marttinen, P., Ammad-Ud-Din, M., Peltola, T., Soare, M., Jacucci, G., and Kaski, S. (2017). Interactive elicitation of knowledge on feature

- relevance improves predictions in small data sets. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 547–552.
- Miller, C. A. and Parasuraman, R. (2003). Beyond levels of automation: an architecture for more flexible human-automation collaboration. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 47, pages 182–186. SAGE Publications Sage CA: Los Angeles, CA.
- Mishler, S., Parker, C., and Chen, J. (2025). Using anthropomorphism, transparency, and feedback to improve phishing detection aids. *Theoretical Issues in Ergonomics Science*, 26(1):45–65.
- Morrison, C., Huckvale, K., Corish, B., Banks, R., Grayson, M., Dorn, J., Sellen, A., and Lindley, S. (2018). Visualizing ubiquitously sensed measures of motor ability in multiple sclerosis: Reflections on communicating machine learning in practice. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–28.
- Mukhopadhyay, S., Singh, V. B., and Babbar-Sebens, M. (2014). User modeling with limited data: Application to stakeholder-driven watershed design. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3855–3860. IEEE.
- Nair, D. and Saenz, M. J. (2024). Pair people and AI for better product demand forecasting. *MIT Sloan Management Review*. URL: <https://sloanreview.mit.edu/article/pair-people-and-ai-for-better-product-demand-forecasting/>. Accessed: 2024-04-07.
- Naiseh, M., Al-Mansoori, R. S., Al-Thani, D., Jiang, N., and Ali, R. (2021a). Nudging through friction: an approach for calibrating trust in explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*, pages 1–5. IEEE.
- Naiseh, M., Al-Thani, D., Jiang, N., and Ali, R. (2021b). Explainable recommendation: When design meets trust calibration. *World Wide Web*, 24(5):1857–1884.

- Naiseh, M., Al-Thani, D., Jiang, N., and Ali, R. (2023). How the different explanation classes impact trust calibration: the case of clinical decision support systems. *International Journal of Human-Computer Studies*, 169:102941.
- Nguyen, A. T., Lease, M., and Wallace, B. C. (2019). Mash: Software tools for developing interactive and transparent machine learning systems. In *Joint Proceedings of the ACM IUI 2019 Workshops (IUI Workshops'19)*.
- Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.
- O'Neill, E. C., Kong, Y. X. G., Connell, P. P., Ong, D. N., Haymes, S. A., Coote, M. A., and Crowston, J. G. (2011). Gaze behavior among experts and trainees during optic disc examination: Does how we look affect what we see? *Investigative Ophthalmology & Visual Science*, 52(7):3976–3983.
- Pan, Y. (2016). Heading toward artificial intelligence 2.0. *Engineering*, 2(4):409–413.
- Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., and Rinzivillo, S. (2023). Co-design of human-centered, explainable AI for clinical decision support. *ACM Transactions on Interactive Intelligent Systems*, 13(4):1–35.
- Parasuraman, R., Mouloua, M., and Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors*, 38(4):665–679.
- Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., and Santos, E. A. (2023). “It’s weird that it knows what I want”: Usability and interactions with copilot for novice programmers. 31(1):4:1–4:31.
- Rayner, K., Slattery, T. J., Drieghe, D., and Liversedge, S. P. (2011). Eye movements and word skipping during reading: Effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):514.

- Rebholz, T. R., Koop, A., and Hütter, M. (2024). Conversational user interfaces: Explanations and interactivity positively influence advice taking from generative artificial intelligence.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.
- Robles-Flores, J. A. and Roussinov, D. (2012). Examining question-answering technology from the task technology fit perspective. *Communications of the Association for Information Systems*, 30(26):439–454.
- Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., and Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 240–251.
- Schoonderwoerd, T. A., Jorritsma, W., Neerincx, M. A., and Van Den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies*, 154:102684.
- Schuetz, S. and Venkatesh, V. (2020). The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction. *Journal of the Association for Information Systems*, 21(2):460–482.
- Schweitzer, M. E. and Cachon, G. P. (2000). Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science*, 46(3):404–420.
- Sebastian, S. and Draper, J. (2024). How AI can improve consumer product and retail experiences. <https://www.sdcexec.com/software-technology/ai-ar/article/22891372/capgemini-how-ai-can-improve-consumer-product-and-retail-experiences>. Accessed: 2024-04-07.

- Sharan, N. N. and Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8):e04572.
- Sharma, R., Shishodia, A., Gunasekaran, A., Min, H., and Munim, Z. H. (2022). The role of artificial intelligence in supply chain management: Mapping the territory. *International Journal of Production Research*, 60(24):7527–7550.
- Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31.
- Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., and Gombolay, M. (2023). Explainable artificial intelligence: Evaluating the objective and subjective impacts of XAI on human-agent interaction. *International Journal of Human-Computer Interaction*, 39(7):1390–1404.
- Simkute, A., Tankelevitch, L., Kewenig, V., Scott, A. E., Sellen, A., and Rintel, S. (2024). Ironies of generative ai: Understanding and mitigating productivity loss in human-AI interactions. arXiv:2402.11364 [cs].
- Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., and Perer, A. (2023). Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- Söllner, M., Arnold, T., Benlian, A., Bretschneider, U., Knight, C., Ohly, S., Rudkowski, L., Schreiber, G., and Wendt, D. (2025). ChatGPT and beyond: Exploring the responsible use of generative AI in the workplace. *Business & Information Systems Engineering*, 67(2):289–303.
- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., and Rintel, S. (2024). The metacognitive demands and opportunities of generative AI. In

- Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–24, New York, NY, USA. Association for Computing Machinery.
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D., Horsley, T., and Weeks, L. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Annals of Internal Medicine*, 169(7):467–473.
- Tsai, C.-C., Kim, J. Y., Chen, Q., Rowell, B., Yang, X. J., Kontar, R., Whitaker, M., and Lester, C. (2025). Effect of artificial intelligence helpfulness and uncertainty on cognitive interactions with pharmacists: Randomized controlled trial. *Journal of Medical Internet Research*, 27:e59946.
- van Capelleveen, G., van Wieren, J., Amrit, C., Yazan, D. M., and Zijm, H. (2021). Exploring recommendations for circular supply chain management through interactive visualisation. *Decision Support Systems*, 140:113431.
- Vilone, G. and Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.
- Vinopal, C. (2022). A growing reliance on AI in hiring is making regulators and lawmakers nervous. <https://observer.com/2022/12/a-growing-reliance-on-ai-in-hiring-is-making-regulators-and-lawmakers-nervous/>. Accessed: 2023-11-26.
- Vössing, M., Köhl, N., Lind, M., and Satzger, G. (2022). Designing transparency for effective Human-AI collaboration. *Information Systems Frontiers*, 24(3):877–895.
- Wang, D., Wang, L., Zhang, Z., Wang, D., Zhu, H., Gao, Y., Fan, X., and Tian, F. (2021). “Brilliant AI doctor” in rural clinics: Challenges in AI-Powered clinical decision support system deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, page Article 697, Yokohama, Japan. Association for Computing Machinery.

- Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–15.
- Wang, Y., Xu, M., and Zhao, L. (2025). Exploring the cognitive mechanisms behind the adoption of algorithmic advice: A neuro-is study of newsvendor decision-making. *Business & Information Systems Engineering*.
- Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, 14(4):490–495.
- Wulff, K. and Finnestrand, H. (2024). Creating meaningful work in the age of AI: explainable AI, explainability, and why it matters to organizational designers. *AI & Society*, 39(4):1843–1856.
- Xu, C., Xu, H., Sun, Y., and Xiong, W. (2025). The digital siren’s call: Accepting unethical AI advice. *International Journal of Human-Computer Interaction*, 41(12):7746–7762.
- Zerilli, J., Bhatt, U., and Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, 3(4).
- Zetsche, D. A., Arner, D. W., Buckley, R. P., and Tang, B. (2020). Artificial intelligence in finance: Putting the human in the loop. *CFTE Academic Paper Series: Centre for Finance, Technology and Entrepreneurship*, (1).
- Zhang, Z. T., Liu, Y., and Hussmann, H. (2021). Forward reasoning decision support: Toward a more complete view of the human-ai interaction design space. In *Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter*, pages 1–5.
- Zhou, J., Arshad, S. Z., Luo, S., and Chen, F. (2017). Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP Conference on Human-Computer Interaction*, pages 23–39. Springer.

Appendix A

Supplementary Materials for Chapter 1

A.1 Methods

A.1.1 Semi-Structured Interview Guide

All interviews were semi-structured, and the questions presented here served as prompts. When appropriate, respondents were asked to provide additional details. The probing suggestions were used by researchers to ask follow-up questions and elicit additional insights.

Part 1: Introduction and Overview

Goal: Understand the participant's professional task requirements, individual expertise, and situational awareness requirements to accomplish their tasks. Probing suggestions in this part target individual and task factors.

1. Can you describe your role at the company?

- *Probe for factors: Goals and roles; Experience; Abilities.*

2. In which contexts do you use (or have you used) this system?

- *Follow-up: Could you provide example tasks for which you use this system?*
- *Probe for factors: Task characteristics, Situation awareness requirements.*

Part 2: Human-AI Interaction Mechanisms within the Task Context

Goal: Understand the nature of human-AI interaction, users' task delegation to AI, and the perceived importance of the system for task execution. Probing suggestions in this part target individual, system, and task factors.

3. How would you describe your interaction with the system?

- *Follow-up: Does using the system help you understand what is occurring within your business?*
- *Probe for factors: Communication with AI; Collaboration with AI; Control over AI; Task repetition; Automation level; Task complexity; Presentation of AI predictions; AI predictions; Delegation of tasks to AI.*

4. Do you have the expertise to configure the business strategy used by the system?

- *If no:* **Who assists you with configuration?**
- *Follow-up:* **How frequently does this strategy need to be reconfigured?**
- *Probe for factors:* *Control over AI; Perceptions of AI (perceived AI reliability, confidence in AI, perceived influence in AI recommendations).*

5. Could you perform your job without this system?

- *If yes:* **How would you accomplish the tasks you currently use this system for?**
- *If yes:* **Is there anything the system does better?**
- *If no:* **What is the system doing that could not be obtained or performed without it?**
- *Probe for factors:* *Situation awareness requirements; Collaboration with AI; Interactions with AI; Workload; Task complexity; Procedures; User goals; System design.*

Part 3: Perceptions of AI

Goal: Evaluate perceptions concerning AI reliability, confidence in AI, participants' perceived understanding of AI.

6. Do you know how the system works to generate its suggestions?

- *Follow-up:* **Would knowing more about how information is processed by the system be useful to you?**
- *Follow-up:* **Do you think this system has all the necessary input attributes to make a good decision?**
- *Follow-up:* **Do you see the system as being rather static or as learning over time?**

- *Probe for factors: Perceived understanding of AI, Perceived transparency of AI, Perceived influence in AI recommendations.*

7. How confident are you in the solutions suggested by the system?

- *Follow-up: Do you evaluate the output of the system or use the results directly?*
- *Follow-up: What do you do when you feel the system's suggestions are out of calibration (e.g., ignore, override, reconfigure)?*
- *Follow-up: To what extent do you find the system and its outputs, plans, and recommendations to be reliable?*
- *Follow-up: Are you as confident in the system today as you were when you first started using it? If different, what has changed?*
- *Probe for factors: Attitudes concerning AI reliability, confidence in AI.*

8. Reflecting on when you were first tasked with using the system, describe that onboarding process and your initial experience.

- *Probe for factors: Perceived understanding of AI, user experience, user abilities.*

9. How could this system be improved to be more helpful for you and your business?

10. Would you recommend this system to someone performing your role?

Closing: Ask for additional remarks, appreciate their time and interest.

Appendix B

Supplementary Materials for Chapter 2

B.1 Methods

B.1.1 Manipulation Check Study

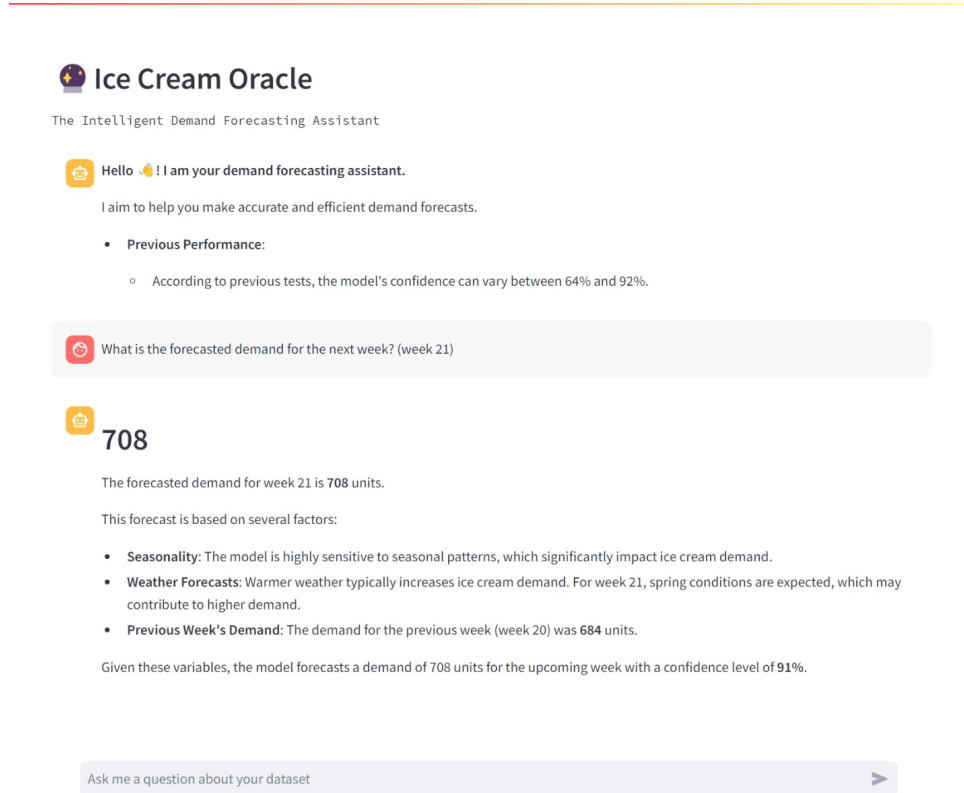


Figure B.1: Screenshot of the GenAI assistant interaction shown to participants in Group 1 during the manipulation check study. This version represents the low transparency and high explainability condition.

Ice Cream Oracle

The Intelligent Demand Forecasting Assistant

Hello 🌟 I am your demand forecasting assistant.

I aim to help you make accurate demand forecasts based on historical sales data and weather forecasts. Here's how I do it:

- **Goal:**
 - Maximize profits.
- **Used Dataset:**
 - Sales data from the last two years (your company's database).
 - Pricing data from the last two years (your company's database).
 - Weather forecasts for the upcoming week (weather.com).
- **Methods:**
 - The data is used to estimate the baseline, trends, and patterns in sales.
 - The estimated parameters are then used in a forecasting method to get the next week's sales estimate.
- **Previous Performance:**
 - The model's confidence can vary between 64% and 92%.

What is the forecasted demand for the next week?

Writing code 🧑‍💻 ...

I will now proceed with the following steps to forecast the demand for the next week using the Holt-Winters method:

1. **Load and Prepare Data:** The data has been loaded successfully. The dataset consists of weekly demand values.
2. **Set Parameters:** I will set the parameters for the Holt-Winters method, assuming yearly seasonality.
3. **Fit the Model:** I will fit the Holt-Winters model to the data.
4. **Forecast:** I will forecast the demand for the next week.

Let's proceed with the forecasting.

✓ Code

Using the Holt-Winters method with yearly seasonality, I've forecasted the demand for the next week. The process involved fitting the model to the historical data and then generating the forecast.

Confidence Level: 90%

Forecasted Demand for Next Week:

708

Ask me a question about your dataset ➡

Figure B.2: Screenshot of the GenAI assistant interaction shown to participants in Group 1 during the manipulation check study. This version represents the high transparency and no explainability condition.

Table B.1: Manipulation check items and t-tests results.

Item	Item Details	Group 1 Mean	Group 2 Mean	p-value	N
Tr1	It is clear to me what data sources are used to generate recommendations.	5.55	5.07	.013	187
Tr2	I can access a great deal of information that explains how the system works.	5.02	4.61	.056	187
Att	Please select "Somewhat agree" as a response to this question.	-	-	-	-
Tr3	I can see plenty of information about the system's inner logic.	4.91	4.45	.038	187
Ex1	The system provides information on the specific data values that led to the output forecast of 708 units.	4.75	5.14	.112	187
Ex2	I can see the specific values that produced the output of a 708 units forecast.	4.09	4.85	.004	187
Transparency	Composite of Tr1, Tr2, Tr3	5.16	4.71	.009	187
Explainability	Composite of Ex1, Ex2	4.42	4.99	.015	187

B.1.2 The Design of GenAI Assistants as Experimental Stimuli

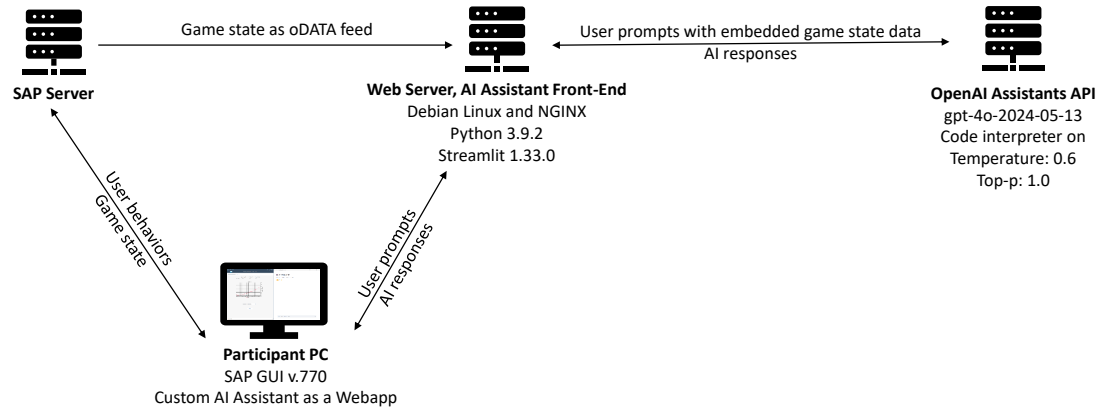


Figure B.3: Tech stack used in the design of the experimental stimuli.

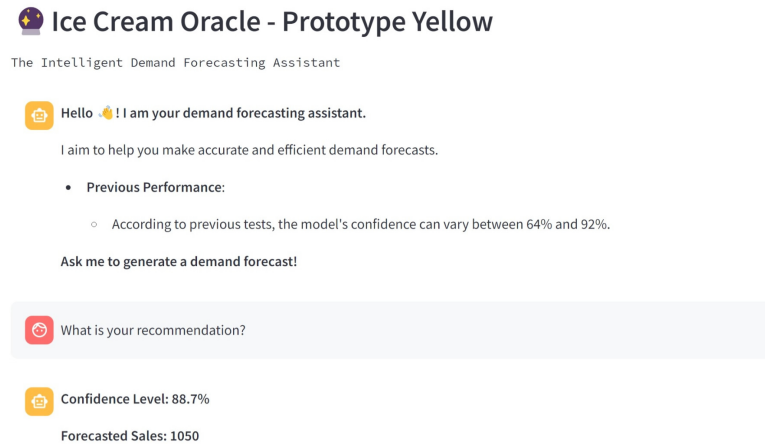


Figure B.4: An example conversation with the GenAI assistant in the low transparency & low explainability condition.

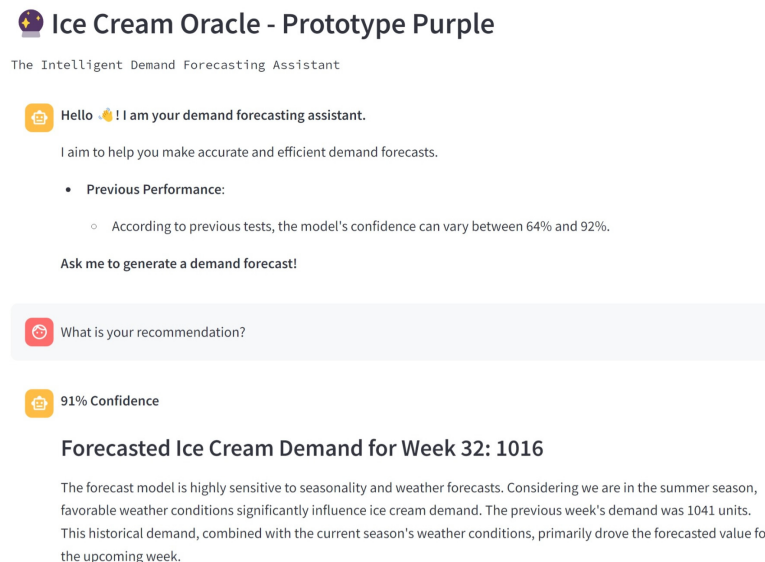


Figure B.5: An example conversation with the GenAI assistant in the low transparency & high explainability condition.

🍦 Ice Cream Oracle - Prototype Pink

The Intelligent Demand Forecasting Assistant

👋 Hello 🍦 I am your demand forecasting assistant.

I aim to help you make accurate demand forecasts based on historical sales data and weather forecasts. Here's how I do it:

- **Goal:**
 - Maximize profits.
- **Used Dataset:**
 - Sales data from the last two years (your company's database).
 - Pricing data from the last two years (your company's database).
 - Weather forecasts for the upcoming week (weather.com).
- **Methods:**
 - The data is used to estimate the baseline, trends, and patterns in sales.
 - The estimated parameters are then used in a forecasting method to get the next week's sales estimate.
- **Previous Performance:**
 - The model's confidence can vary between 64% and 92%.

Ask me to generate a demand forecast!

👋 What is your recommendation?

👋 ✓ Code

For the forecast, I will use the Holt-Winters exponential smoothing method, which is well-suited for data with seasonality, trends, and irregular components. Here is a step-by-step explanation of the process:

1. **Load the data:** We have loaded the historical data, which includes the week number, temperature, precipitation, wind speed, ice cream demand, and some other factors.
2. **Prepare the data:** We will focus on the demand data and use the past weeks' data to forecast the demand for week 32.
3. **Fit the Holt-Winters model:** We will fit the Holt-Winters model using the historical demand data. The model parameters (alpha, beta, gamma) will be estimated based on the data.
4. **Forecast for the next week:** Using the fitted model, we will forecast the demand for week 32.

Now, let's proceed with the forecasting.

Step-by-Step Forecasting Process:

1. **Load and prepare the data:**
 - We will use the historical demand data up to week 31.
2. **Fit the Holt-Winters model:**
 - The model will account for trends and seasonality in the data.
3. **Forecast for the next week (week 32):**
 - We will generate the forecast and calculate the confidence level.

Let's implement this in Python.

✓ Code

Using the Holt-Winters exponential smoothing method, we have forecasted the ice cream demand for week 32.

Forecast Summary:

- **Forecasted Demand for Week 32:** 1016 units
- **Confidence Level:** 88%

Based on our objective to ensure stock availability, I recommend planning for **1016 units** of ice cream for week 32.

Forecasted Demand:

1016 units

Figure B.6: An example conversation with the GenAI assistant in the high transparency & low explainability condition.

🧊 Ice Cream Oracle - Prototype Grey

The Intelligent Demand Forecasting Assistant

👋 Hello 🧊 I am your demand forecasting assistant.

I aim to help you make accurate demand forecasts based on historical sales data and weather forecasts. Here's how I do it:

- **Goal:**
 - Maximize profits.
- **Used Dataset:**
 - Sales data from the last two years (your company's database).
 - Pricing data from the last two years (your company's database).
 - Weather forecasts for the upcoming week (weather.com).
- **Methods:**
 - The data is used to estimate the baseline, trends, and patterns in sales.
 - The estimated parameters are then used in a forecasting method to get the next week's sales estimate.
- **Previous Performance:**
 - The model's confidence can vary between 64% and 92%.

Ask me to generate a demand forecast!

🗨️ What is your recommendation?

👋 ✓ Code

To provide a forecast recommendation for week 32, I will use the Holt-Winters method, which is suitable for capturing seasonality, trends, and level components in time series data. Here is a step-by-step explanation of the process:

Step-by-Step Process Explanation:

1. **Data Preparation:**
 - Load the data and focus on the columns `Week`, `Temperature`, `Precipitation`, `Wind_Speed`, and `Ice_Cream_Demand`.
 - Use historical demand data up to the current week (week 31) for forecasting purposes.
2. **Holt-Winters Parameters:**
 - **Alpha (α):** Smoothing level parameter, which controls the level component.
 - **Beta (β):** Smoothing trend parameter, which controls the trend component.
 - **Gamma (γ):** Smoothing seasonal parameter, which controls the seasonal component.
 - **Seasonal Periods:** Set to 52, assuming yearly seasonality based on weekly data.
3. **Model Fitting:**
 - Fit the Holt-Winters model to the historical demand data.
 - Use additive components for level, trend, and seasonality.
4. **Forecasting:**
 - Generate a forecast for week 32.
 - Provide a confidence level for the forecast recommendation.

Let's proceed with the calculations and generate the forecast for week 32.

✓ Code

Forecast Recommendation for Week 32

Using the Holt-Winters method, I have generated the forecast for week 32. Here's a detailed breakdown of the recommendation:

1. **Previous Week's Demand:**
 - The demand for week 31 was **1030 units**.
2. **Forecasted Demand for Week 32:**
 - The forecasted demand for week 32 is approximately **1024 units** (rounded up to the nearest integer).
3. **Confidence Level:**
 - The confidence level for this forecast is **90.1%**.

Discussion

The forecasted demand for week 32 is derived based on the following factors:

- **Seasonality:** The Holt-Winters method assumes yearly seasonality, capturing the annual demand patterns for ice cream.
- **Weather Conditions:** Temperature, precipitation, and wind speed are significant predictors of ice cream demand. These factors have been considered in the historical data used for forecasting.
- **Previous Week's Demand:** The demand for the previous week (1030 units) provides a reference point for the current forecast.

Figure B.7: An example conversation with the GenAI assistant in the high transparency & high explainability condition.

B.1.3 Measurement Model

Table B.2: Measurement items.

	Measures	Adapted from
Perceived Understanding	From the explanation, I understand how the AI model works. This explanation of how the AI model works is satisfying. This explanation of how the AI model works has sufficient detail. This explanation of how the AI model works seems complete. This explanation of how the AI model works tells me how to use it. This explanation of how the AI model works is useful to my goals. This explanation of the AI model shows me how accurate the AI model is. This explanation lets me judge when I should trust and not trust the AI model.	(Hoffman et al., 2023)
Automation Induced Complacency Potential	When I have a lot to do, it makes sense to delegate a task to automation. If life were busy, I would let an automated system handle some tasks for me. Automation should be used to ease people's workload. If automation is available to help me with something, it makes sense for me to pay more attention to my other tasks. Even if an automated aid can help me with a task, I should pay attention to its performance. Distractions and interruptions are less of a problem for me when I have an automated system to cover some of the work. Constantly monitoring an automated system's performance is a waste of time. Even when I have a lot to do, I am likely to watch automation carefully for errors. It's not usually necessary to pay much attention to automation when it is running. Carefully watching automation takes time away from more important or interesting things.	(Merritt et al., 2019)
Computer Self-Efficacy	<i>I could complete the job using SAP...</i> ...if there was no one around to tell me what to do as I go. ...if I had never used a package like it before. ...if I had only the software manuals for reference. ...if I had seen someone else using it before trying it myself. ...if I could call someone for help if I got stuck. ...if someone else had helped me get started. ...if I had a lot of time to complete the task. ...if I had just the built-in help facility for assistance. ...if someone showed me how to do it first. ...if I had used similar packages like this one before to do this task.	(Compeau and Higgins, 1995)
Confidence in GenAI Decision Confidence	Please rate your level of confidence in the AI forecasting system. Please rate your level of confidence in your decision.	(Chong et al., 2022)

Table B.3: Perceived Understanding item descriptive statistics

Item	Mean	SD	Min	Max	Skewness	Kurtosis	Item-Total r
Understand1	5.16	1.57	1.00	7.00	-1.12	0.55	0.76
Understand2	4.76	1.68	1.00	7.00	-0.71	-0.44	0.83
Understand3	4.52	1.77	1.00	7.00	-0.41	-0.97	0.83
Understand4	4.46	1.67	1.00	7.00	-0.38	-0.94	0.81
Understand5	4.70	1.70	1.00	7.00	-0.74	-0.47	0.72
Understand6	5.32	1.39	1.00	7.00	-0.91	0.33	0.70
Understand7	4.63	1.75	1.00	7.00	-0.38	-1.12	0.71
Understand8	4.51	1.72	1.00	7.00	-0.39	-0.78	0.52

Table B.4: Perceived Understanding scale inter-item correlations.

	Under-stand1	Under-stand2	Under-stand3	Under-stand4	Under-stand5	Under-stand6	Under-stand7	Under-stand8
Understand1	1.00							
Understand2	0.77	1.00						
Understand3	0.77	0.80	1.00					
Understand4	0.68	0.75	0.82	1.00				
Understand5	0.65	0.64	0.64	0.59	1.00			
Understand6	0.48	0.65	0.57	0.60	0.56	1.00		
Understand7	0.50	0.60	0.61	0.66	0.57	0.63	1.00	
Understand8	0.43	0.41	0.41	0.41	0.45	0.50	0.46	1.00

Table B.5: Factor Loadings of Perceived Understanding Items

Item	All Items, Standardized Loading	Selected Items, Standardized Loading
Understand1	0.821	0.831
Understand2	0.887	0.893
Understand3	0.905	0.914
Understand4	0.868	0.862
Understand5	0.732	0.720
Understand6	0.692	0.666
Understand7	0.710	-
Understand8	0.503	-

Table B.6: AICP item descriptive statistics. (r) indicates reversed items.

Item	Mean	SD	Min	Max	Skewness	Kurtosis	Item-Total <i>r</i>
AICP_1	4.02	0.74	2	5	-0.79	0.91	0.34
AICP_2	4.00	0.82	2	5	-0.75	0.28	0.30
AICP_3	4.51	0.54	3	5	-0.37	-1.19	0.34
AICP_4	4.18	0.78	2	5	-0.97	0.95	0.42
AICP_5(r)	3.65	0.72	3	6	0.89	0.43	0.18
AICP_6	3.21	0.90	2	5	0.16	-0.92	0.41
AICP_7	2.44	0.89	1	5	0.71	0.08	0.18
AICP_8(r)	4.93	1.07	3	7	0.22	-0.96	0.24
AICP_9	2.81	0.83	1	4	0.18	-1.22	0.46
AICP_10	3.18	0.78	2	4	-0.31	-1.34	0.42

Table B.7: AICP scale inter-item correlations.

	AICP_1	AICP_2	AICP_3	AICP_4	AICP_5(r)	AICP_6	AICP_7	AICP_8(r)	AICP_9	AICP_10
AICP_1	1.00**									
AICP_2	0.50**	1.00**								
AICP_3	0.38**	0.36*	1.00**							
AICP_4	0.33*	0.36*	0.34*	1.00**						
AICP_5(r)	0.01	-0.12	0.01	0.05	1.00**					
AICP_6	0.21	0.26	0.29*	0.28*	0.03	1.00**				
AICP_7	-0.07	-0.15	-0.10	-0.11	0.27*	0.42**	1.00**			
AICP_8(r)	0.05	0.00	0.22	0.19	0.11	0.02	0.09	1.00**		
AICP_9	0.15	0.29*	0.06	0.30*	0.27*	0.08	0.16	0.25	1.00**	
AICP_10	0.12	0.00	0.00	0.21	0.17	0.30*	0.25	0.25	0.46**	1.00**

* Represents significance at the 0.05 level. ** Represents significance at the 0.01 level. Items 5 and 8 were treated as reverse measures.

Table B.8: CFA Loadings for Workload Alleviation Subscale

Item	Standardized Loading
AICP_1	0.662
AICP_2	0.689
AICP_3	0.575
AICP_4	0.542
AICP_6	0.406

Table B.9: CFA Loadings for Monitoring Subscale

Item	Standardized Loading
AICP_5	0.363
AICP_7	0.335
AICP_8	0.356
AICP_9	0.682
AICP_10	0.666

Table B.10: Computer Self-Efficacy item descriptive statistics

Item	Mean	SD	Min	Max	Skewness	Kurtosis	Item-Total <i>r</i>
cse_1	6.21	2.15	1.00	10.00	-0.73	-0.27	0.66
cse_2	5.50	2.62	1.00	10.00	-0.06	-1.29	0.80
cse_3	6.14	2.20	1.00	10.00	-0.13	-0.86	0.52
cse_4	7.15	2.04	2.00	10.00	-0.54	-0.63	0.63
cse_5	8.17	1.75	2.00	10.00	-1.12	1.24	0.64
cse_6	8.38	1.75	2.00	10.00	-1.32	1.84	0.57
cse_7	7.42	1.90	3.00	10.00	-0.44	-0.61	0.59
cse_8	6.61	2.12	1.00	10.00	-0.84	0.34	0.66
cse_9	8.76	1.55	3.00	10.00	-1.71	2.97	0.59
cse_10	8.04	1.40	4.00	10.00	-0.47	0.05	0.48

Table B.12: Factor Loadings of Computer Self-Efficacy Items

Item	All Items, Standardized Loading	Selected Items, Standardized Loading
cse_1	0.420	-
cse_2	0.321	-
cse_3	0.420	-
cse_4	0.852	0.824
cse_5	0.544	0.575
cse_6	0.799	0.827
cse_7	0.652	0.614
cse_8	0.450	-
cse_9	0.833	0.860
cse_10	0.735	0.737

Table B.11: Computer Self-Efficacy scale inter-item correlations.

	cse_1	cse_2	cse_3	cse_4	cse_5	cse_6	cse_7	cse_8	cse_9	cse_10
cse_1	1.00**									
cse_2	0.52**	1.00**								
cse_3	0.28*	0.40**	1.00**							
cse_4	0.41**	0.43**	0.38**	1.00**						
cse_5	0.14	-0.07	0.18	0.40**	1.00**					
cse_6	0.11	0.09	0.42**	0.70**	0.57**	1.00**				
cse_7	0.39**	0.33*	0.31*	0.62**	0.23	0.44**	1.00**			
cse_8	0.32*	0.35*	0.28*	0.31*	0.25	0.29*	0.40**	1.00**		
cse_9	0.39**	0.11	0.26	0.68**	0.52**	0.70**	0.55**	0.29*	1.00**	
cse_10	0.25	0.17	0.18	0.60**	0.42**	0.60**	0.40**	0.52**	0.66**	1.00**

* Represents significance at the 0.05 level. ** Represents significance at the 0.01 level.

B.1.4 Experimental Task Instructions

Task Instructions, Page 1

The following text and the supplementing Figure B.8 were shown on the first page of instructions, before participants start engaging with the tasks in the trial block:

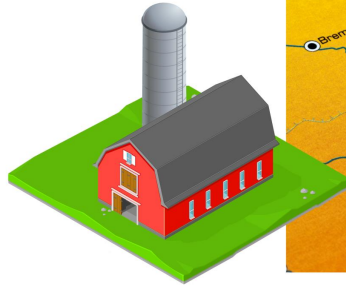
Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

Welcome to your new job at ERP Dairy! ERP Dairy is a natural ice cream distributor in the German market.

As a wholesaler, ERP Dairy purchases units of ice cream from dairy co-ops to sell them to grocery stores.

Welcome to Your New Job!



ERPSimLab
HEC MONTRÉAL | © ERPSim Lab, HEC Montréal

Commercial version
powered by BÂTON

Figure B.8: Experimental task instructions, page 1 visual supplement.

Task Instructions, Page 2

The following text and the supplementing Figure B.9 were shown on the second page of instructions, before participants start engaging with the tasks in the trial block:

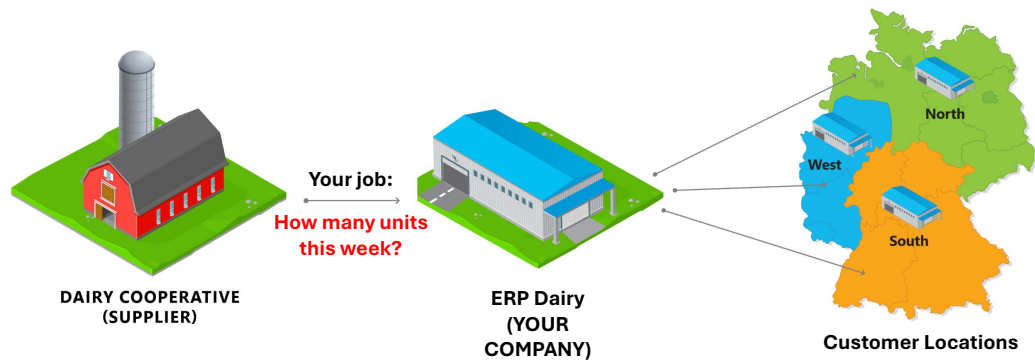
Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

As the demand forecaster, your job is to analyze the historical demand data, understand trends and seasonalities, and submit your sales forecasts to your suppliers.

Each week, you determine the number of units that you expect to sell, and then submit that sales estimate to your suppliers as a purchase order.

Logistics Game Design



ERPsimLab
HEC MONTRÉAL | © ERPsim Lab, HEC Montréal

Commercial version
powered by BATON

Figure B.9: Experimental task instructions, page 2 visual supplement.

Task Instructions, Page 3

The following text and the supplementing Figure B.10 were shown on the third page of instructions, before participants start engaging with the tasks in the trial block:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

Let's explore the tool you will be using: the SAP demand forecasting interface (see the screenshot below).

You will see the pricing information on top of the screen. **Sales price** is what your customers pay, **unit cost** is what you pay for each unit. Your product is a natural ice cream with a shelf life of one week only. Therefore, if you order excess ice cream, you sell them at a huge discount, 50 euros per unit. **Salvage price** indicates the discounted sales price of the excess units you ordered. The pricing information will be constant throughout the experiment.

Below the pricing information, you will see the weekly past demand chart. You are

expected to use this data to predict your sales for the next week. On this chart, week numbers start from January 1st, and your customers are located in Germany. Therefore, you can deduct your ice cream demand seasonality using the week numbers.

One year has 52 weeks, weeks from 20 to 35 are often summer months in the North.

Once you make your estimation, you must input that quantity into the textbox on the screen and click accept.

Example data chart:

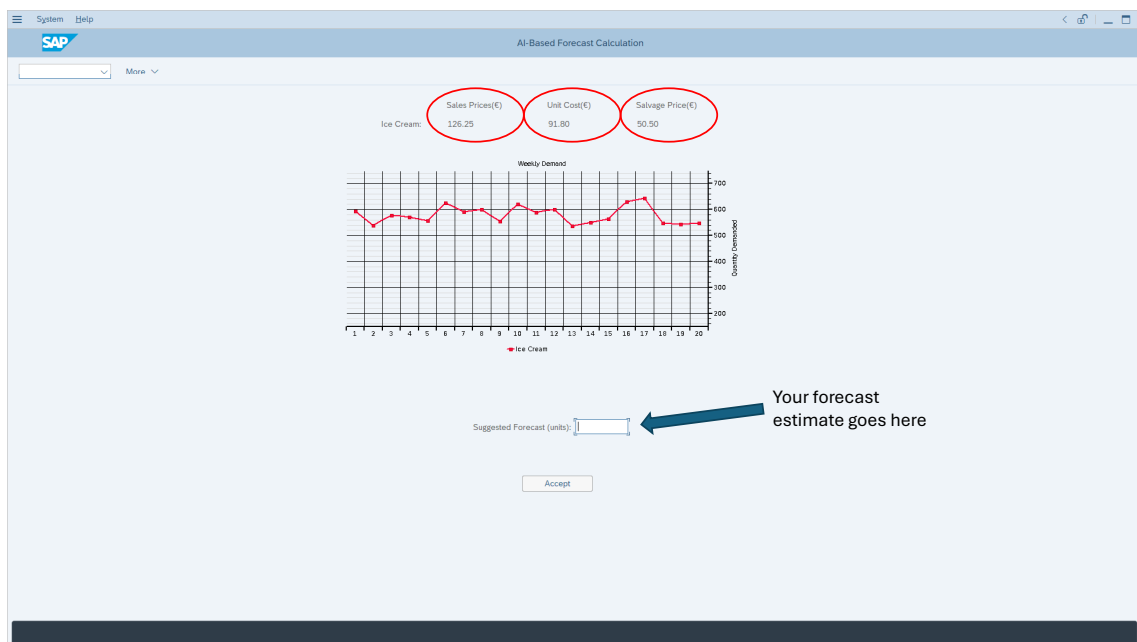


Figure B.10: Experimental task instructions, page 3 visual supplement.

Task Instructions, Page 4

The following text and the supplementing Figure B.11 were shown on the fourth page of instructions, before participants start engaging with the tasks in the trial block:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

When you submit your forecasts, the business simulation will advance by a week, and you will be presented with a performance report.

On this report, you will see how much you saved or lost compared to an average decision-maker.

There is also an overall forecasting error measure: MAPE. The lower the MAPE is, the better you are doing. A MAPE lower than 10% means you are doing an acceptable job.

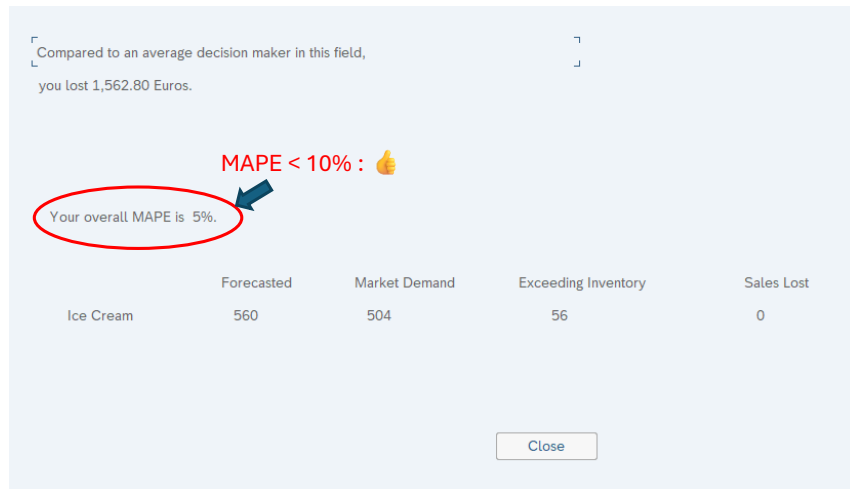


Figure B.11: Experimental task instructions, page 4 visual supplement.

Task Instructions After Trial Block, Page 1

The following text and the supplementing Figures B.12, B.13, B.14 were shown on the first page of instructions after the trial block, before participants start engaging with the experimental tasks:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

Now, you will learn about the three important parameters in a forecasting model. You do not need to know all details for today's task, but it's a good idea to have a basic understanding.

Smoothing is the weight of recent data points in the forecast. Its parameter, alpha, indicates how much importance the model will allocate to the most recent observation

compared to the demand history.

If the smoothing parameter is lower, forecasts are similarly impacted by older and recent sales data. If it's higher, forecasts consider recent data more, and forecasts might be impacted by irregular spikes.

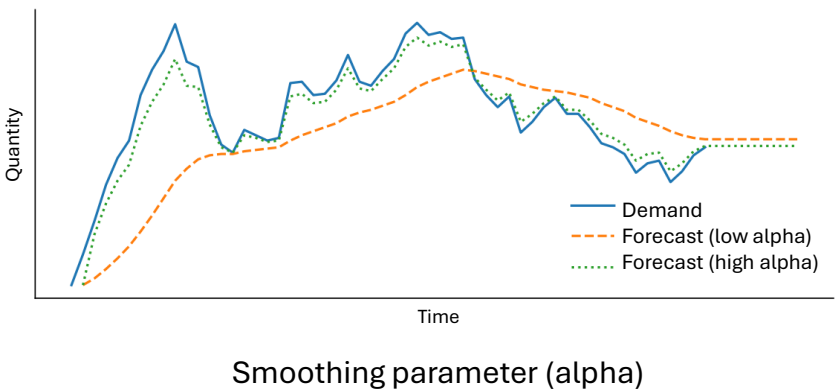


Figure B.12: Experimental task instructions, visual supplement for page 1 after the trial block.

Trend is the slope of the time series data. A higher value of the trend parameter gives more weight to the trend.

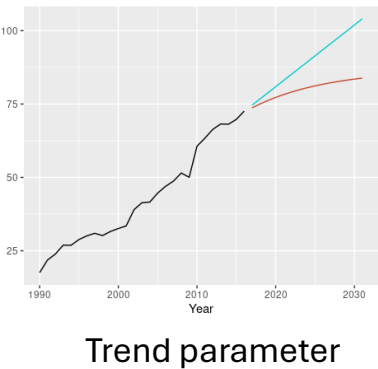
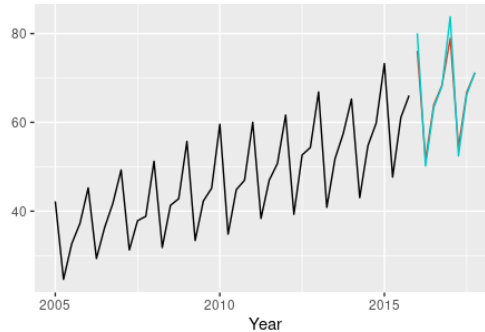


Figure B.13: Experimental task instructions, visual supplement for page 1 after the trial block.

Seasonality is the effect of periodic fluctuations in the market. A higher seasonality parameter gives more weight to the seasonal component, making the forecast more sensitive to changes in the seasonal pattern.



Seasonality parameter

Figure B.14: Experimental task instructions, visual supplement for page 1 after the trial block.

Task Instructions After Trial Block, Page 2

The following text and the supplementing Figures B.15 were shown on the second page of instructions after the trial block, before participants start engaging with the experimental tasks:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

Your employer has acquired an artificial intelligence (AI)-powered decision support agent to assist demand forecasting experts with their daily tasks.

This agent communicates using natural language, can read data, and runs customized forecasts for your virtual company.

In this experiment, we ask you to interact with this chat-based agent at least once each virtual week. You may ask as many follow-up questions as needed.

The agent will be displayed on the right side of your screen, while the decision interface will be on the left. Once you determine your demand forecasts, use the interface

on the left to submit your decision and advance the simulation time. You can see an example screenshot of the tools you will use below.

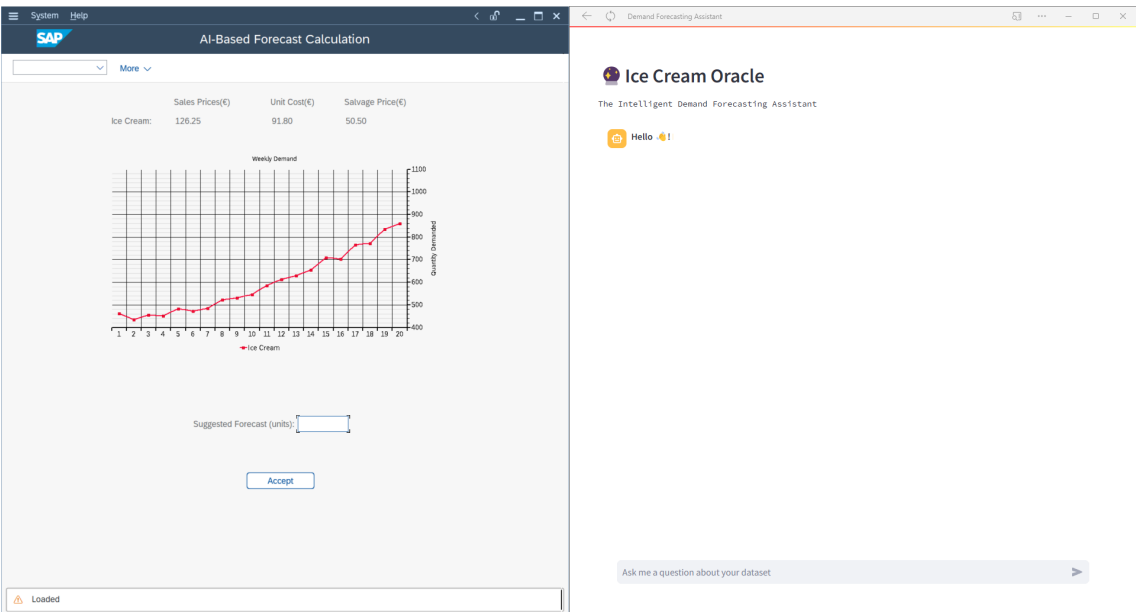


Figure B.15: Experimental task instructions, visual supplement for page 2 after the trial block.

B.2 Results

B.2.1 Descriptive Statistics

Table B.13: Block-Level Descriptive Statistics

Transparency	Explainability	Explanation Satisfaction (SD)	N
0	0	4.08 (1.55)	28
0	1	4.79 (1.19)	29
1	0	5.29 (0.97)	26
1	1	4.97 (1.28)	19

Table B.14: Trial-Level Behavioral Descriptive Statistics. Performance values are calculated using a loss function. Higher performance values are better, and the highest possible performance value is 0.

T	E	Conf. in GenAI (SD)	Decision Confidence (SD)	Performance (SD)	N
0	0	4.36 (1.45)	4.60 (1.34)	-1027 (1480)	257
0	1	4.53 (1.22)	4.94 (0.97)	-1012 (849)	218
1	0	5.02 (1.31)	5.05 (1.18)	-978 (897)	221
1	1	4.75 (1.15)	4.80 (0.90)	-1180 (1008)	151

Table B.15: Eye-Tracking Descriptive Statistics

T	E	Verif. Saccades (SD)	Fix. Duration (SD, ms)	Total Time (SD, ms)	N
0	0	23.7 (14.9)	15569 (13225)	68988 (49874)	245
0	1	21.2 (14.0)	16483 (12893)	86564 (57998)	207
1	0	23.2 (16.1)	13614 (9275)	76902 (53031)	204
1	1	21.0 (15.0)	12610 (7768)	88251 (57809)	143

Table B.16: Attention Shift Saccades on First Trial

T	E	Initial Saccades (SD)	N
0	0	40.0 (20.1)	24
0	1	34.9 (18.7)	24
1	0	35.4 (26.3)	17
1	1	31.0 (21.0)	15

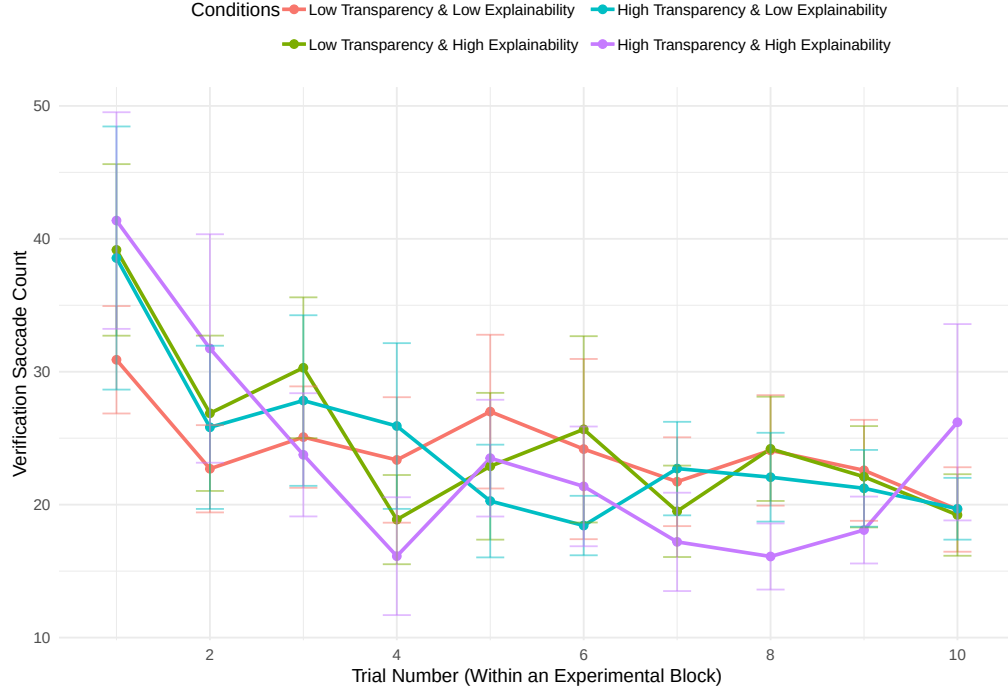


Figure B.16: Attention shift rates over trial numbers within an experimental block. The plot indicates that on average, participants reduced their attention shift rates by about 50% after having used the same GenAI assistant in similar cases for 10 times.

B.2.2 Hypothesis Testing

Table B.17: Linear Mixed-Effects Model Predicting Perceived Understanding (H1a-H1c)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.254	0.405	.532
Transparency	0.243	0.051	< .001***
Explainability	0.125	0.050	.016*
Log(Baseline Performance)	-0.035	0.052	.496
Exp. Block Number	-0.005	0.032	.866
AICP-W	0.061	0.033	.068
CSE	-0.026	0.126	.835
Transparency × Explainability	-0.219	0.080	.008**
<i>Random Effects:</i>			
Participant Intercept Variance	0.028 (SD = 0.167)		
Residual Variance	0.025 (SD = 0.157)		

Note: $N = 102$ observations, 57 participants. REML criterion: 0.5.

R^2 (Marginal / Conditional): 0.175 / 0.613

Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$

Table B.18: Linear Mixed-Effects Model Predicting Confidence in the GenAI Assistant (H2a-H2c)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.575	0.034	< .001***
Transparency	0.122	0.037	.0015**
Explainability	0.017	0.036	.639
Trial Number (within block)	-0.002	0.002	.357
Exp. Block Number	-0.008	0.023	.732
Transparency × Explainability	-0.098	0.057	.088
<i>Random Effects:</i>			
Participant Intercept Variance	0.017 (SD = 0.132)		
Participant:Block Intercept Variance	0.010 (SD = 0.099)		
Residual Variance	0.021 (SD = 0.145)		
<i>Note:</i> $N = 847$ observations, 103 blocks, 57 participants. REML criterion: -604.3.			
R^2 (Marginal / Conditional): 0.048 / 0.583			
Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, $p < .10$			

Table B.19: Linear Mixed-Effects Models Predicting Log(Decision Performance) at the Block Level (H3)

Fixed Effect	Model 1: With Perceived Understanding	Model 2: Direct Effects Only
	Estimate (SE)	Estimate (SE)
Intercept	0.163 (0.226)	0.184 (0.222)
Perceived Understanding	0.087 (0.064)	—
Transparency	0.012 (0.040)	0.029 (0.038)
Explainability	0.002 (0.038)	0.015 (0.037)
Log(Baseline Performance)	-0.028 (0.029)	-0.031 (0.028)
Exp. Block Number	-0.062 (0.025)*	-0.064 (0.026)*
AICP-W	0.005 (0.018)	0.010 (0.017)
Complacency Potential	-0.006 (0.068)	-0.009 (0.066)
Transparency × Explainability	-0.087 (0.058)	-0.099 (0.057)
<i>Random Effects:</i> Participant Intercept SD = 0.057 (M1), 0.045 (M2); Residual SD = 0.126 (M1), 0.130 (M2)		
<i>Model Fit:</i> REML = -79.3 (M1), -81.3 (M2); $N = 102$; 57 participants		
R^2 (Marginal / Conditional): 0.113 / 0.263 (M1), 0.093 / 0.191 (M2)		
Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, $p < .1$		

Table B.20: Linear Mixed-Effects Models Predicting Log(Decision Performance) at the Trial Level (H4)

Fixed Effect	Model 1: With Confidence in GenAI	Model 2: Total Effect Only
	Estimate (SE)	Estimate (SE)
Intercept	0.267 (0.021)***	0.302 (0.015)***
Confidence in GenAI	0.061 (0.026)*	—
Transparency	0.026 (0.015)*	0.033 (0.015)*
Explainability	0.003 (0.015)	0.005 (0.015)
Trial Number (within block)	0.017 (0.002)***	0.017 (0.002)***
Exp. Block Number	-0.037 (0.010)***	-0.037 (0.010)***
Transparency × Explainability	-0.061 (0.023)**	-0.067 (0.023)**
<i>Random Effects: Participant Intercept SD = 0.044 (M1), 0.043 (M2); Residual SD = 0.134</i>		
<i>Model Fit: REML = -898.8 (M1), -898.7 (M2); N = 847; 57 participants</i>		
<i>R² (Marginal / Conditional): 0.147 / 0.228 (M1), 0.141 / 0.220 (M2)</i>		
<i>Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, · $p < .1$</i>		

Table B.21: Linear Mixed-Effects Model Predicting Log(Fixation Duration) (H5)

Fixed Effect	Estimate (SE), p
Intercept	0.667 (0.023), $p < .001$ ***
Transparency	-0.060 (0.015), $p < .001$ ***
Explainability	-0.042 (0.015), $p = .004$ **
Exp. Block Number	-0.015 (0.009), $p = .103$
Trial Number (within block)	0.008 (0.002), $p < .001$ ***
Transparency × Explainability	0.015 (0.023), $p = .503$
<i>Random Effects: Participant Intercept SD = 0.092; Residual SD = 0.116</i>	
<i>Model Fit: REML = -1007; N = 799 trials; 57 participants</i>	
<i>R² (Marginal / Conditional): 0.063 / 0.423</i>	
<i>Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, · $p < .1$</i>	

Table B.22: Linear Mixed-Effects Model Predicting Log(Attention Shift Rate) (H6)

Fixed Effect	Estimate (SE), p
Intercept	3.630 (0.093), $p < .001$ ***
Transparency	-0.133 (0.059), $p = .024$ *
Explainability	-0.083 (0.058), $p = .155$
Exp. Block Number	-0.203 (0.037), $p < .001$ ***
Trial Number (within block)	-0.054 (0.006), $p < .001$ ***
Transparency × Explainability	0.097 (0.092), $p = .294$
<i>Random Effects: Participant Intercept SD = 0.393; Residual SD = 0.463</i>	
<i>Model Fit: REML = 1194.5; N = 799 trials; 57 participants</i>	
<i>R² (Marginal / Conditional): 0.076 / 0.463</i>	
<i>Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, · $p < .1$</i>	

Table B.23: Linear Mixed-Effects Models Predicting Log(Decision Performance) at the Trial Level (H7)

Fixed Effect	Model 1: With Attention Allocation	Model 2: Total Effect Only
	Estimate (SE)	Estimate (SE)
Intercept	0.340 (0.038)***	0.267 (0.025)***
Log(Fixation Duration)	-0.109 (0.042)**	—
Transparency	0.029 (0.017)	0.034 (0.017)*
Explainability	-0.002 (0.017)	0.002 (0.017)
Trial Number (within block)	0.020 (0.002)***	0.019 (0.002)***
Exp. Block Number	-0.044 (0.011)***	-0.042 (0.012)***
Transparency × Explainability	-0.071 (0.026)**	-0.070 (0.026)**
<i>Random Effects: Participant Intercept SD = 0.045 (M1), 0.044 (M2); Residual SD = 0.152</i>		
<i>Model Fit: REML = -658.5 (M1), -656.3 (M2); N = 799; 57 participants</i>		
<i>R² (Marginal / Conditional): 0.143 / 0.212 (M1), 0.136 / 0.203 (M2)</i>		
<i>Significance codes: ***p < .001, **p < .01, *p < .05, ·p < .1</i>		

Table B.24: Linear Mixed-Effects Models Predicting Log(Decision Performance) at the Trial Level (H8)

Fixed Effect	Model 1: With Attention Shift Rate	Model 2: Total Effect Only
	Estimate (SE)	Estimate (SE)
Intercept	0.319 (0.045)***	0.267 (0.025)***
Log(Attention Shift Rate)	-0.014 (0.010)	—
Transparency	0.033 (0.017)	0.034 (0.017)*
Explainability	0.000 (0.017)	0.002 (0.017)
Trial Number (within block)	0.018 (0.002)***	0.019 (0.002)***
Exp. Block Number	-0.045 (0.012)***	-0.042 (0.012)***
Transparency × Explainability	-0.069 (0.026)**	-0.070 (0.026)**
<i>Random Effects: Participant Intercept SD = 0.045 (M1), 0.044 (M2); Residual SD = 0.152</i>		
<i>Model Fit: REML = -650.8 (M1), -656.3 (M2); N = 799; 57 participants</i>		
<i>R² (Marginal / Conditional): 0.137 / 0.206 (M1), 0.136 / 0.203 (M2)</i>		
<i>Significance codes: ***p < .001, **p < .01, *p < .05, ·p < .1</i>		

Table B.25: Linear Mixed-Effects Model Predicting Log(Decision Performance)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.277	0.053	< .001***
Perceived Understanding	0.047	0.034	.166
Confidence in the GenAI	0.056	0.031	.068
Log(Fixation Duration)	-0.142	0.053	.008**
Log(Attention Shift Rate)	0.009	0.013	.509
Transparency	0.014	0.019	.452
Explainability	-0.003	0.017	.859
Exp. Block Number	-0.050	0.012	< .001***
Trial Number (within block)	0.018	0.002	< .001***
Transparency × Explainability	-0.060	0.027	.029*
<i>Random Effects:</i>			
Participant Intercept Variance	0.002 (SD = 0.050)		
Residual Variance	0.023 (SD = 0.151)		
<i>Note: N = 799 observations, 57 participants. REML criterion: -649.6.</i>			
<i>R² (Marginal / Conditional): 0.158 / 0.240</i>			
<i>Significance codes: ***p < .001, **p < .01, *p < .05, ·p < .10</i>			

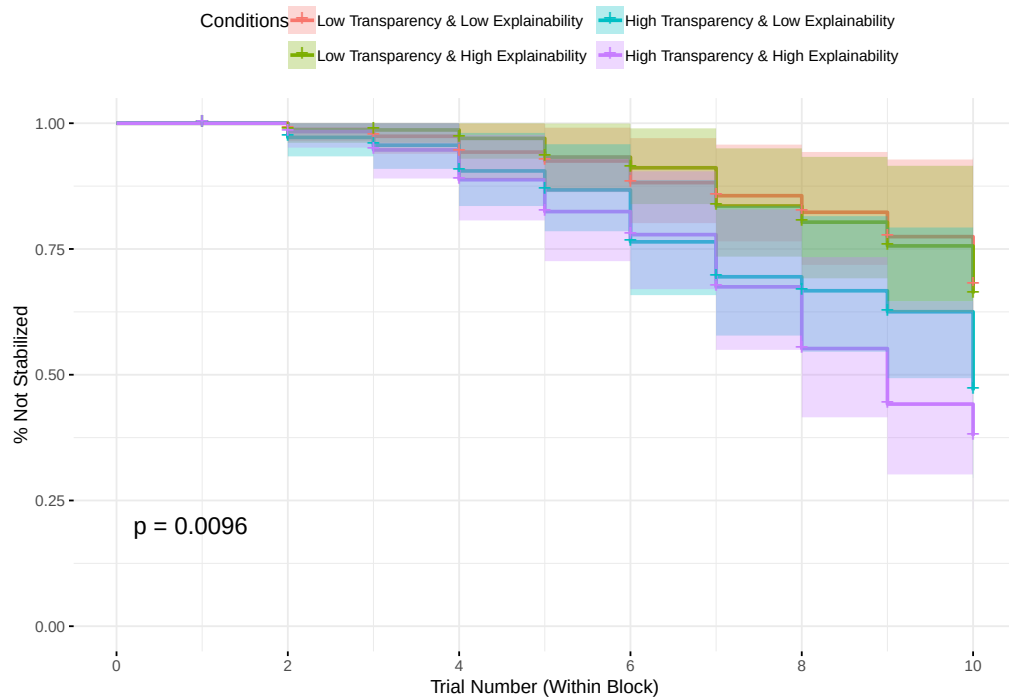


Figure B.17: Kaplan-Meier fit plot corresponding to post-hoc analyses. This plot indicates that the conditions with high transparency reach stabilization, i.e., 30% of a participant's initial attention shift rates, faster.

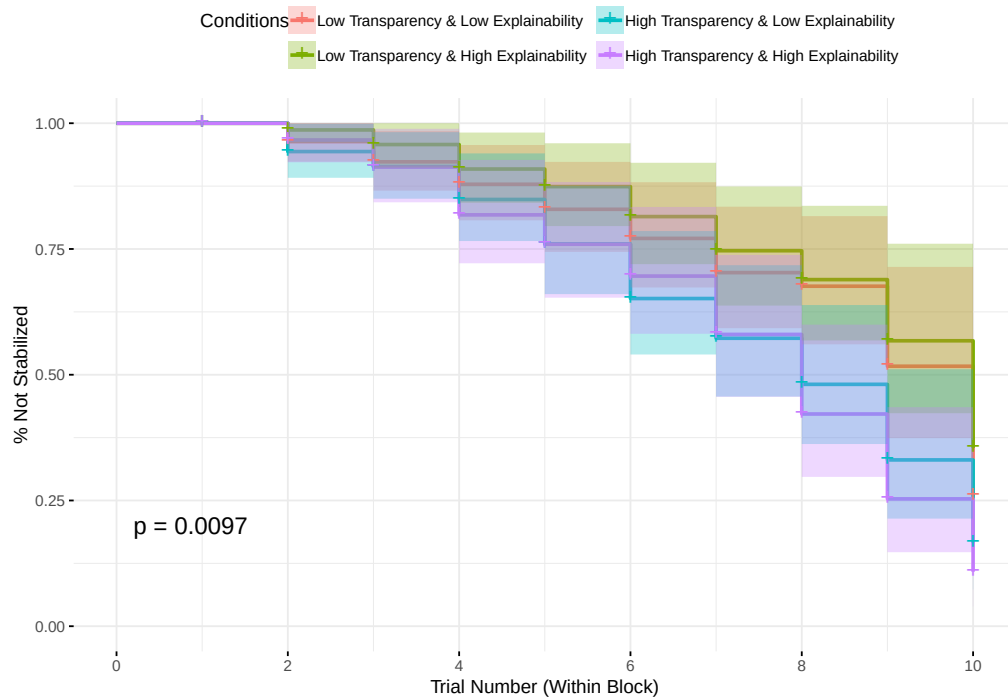


Figure B.18: Kaplan-Meier fit plot corresponding to post-hoc analyses. This plot indicates that the conditions with high transparency reach stabilization, i.e., 50% of a participant's initial attention shift rates, faster.

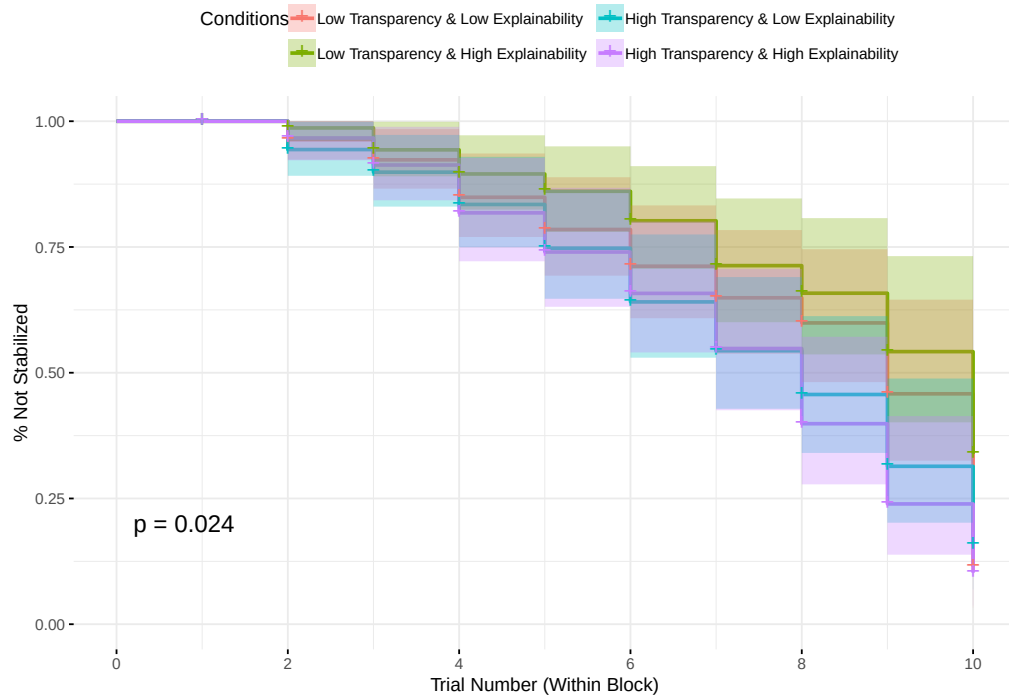


Figure B.19: Kaplan-Meier fit plot corresponding to post-hoc analyses. This plot indicates that the conditions with high transparency reach stabilization, i.e., 55% of a participant's initial attention shift rates, faster.

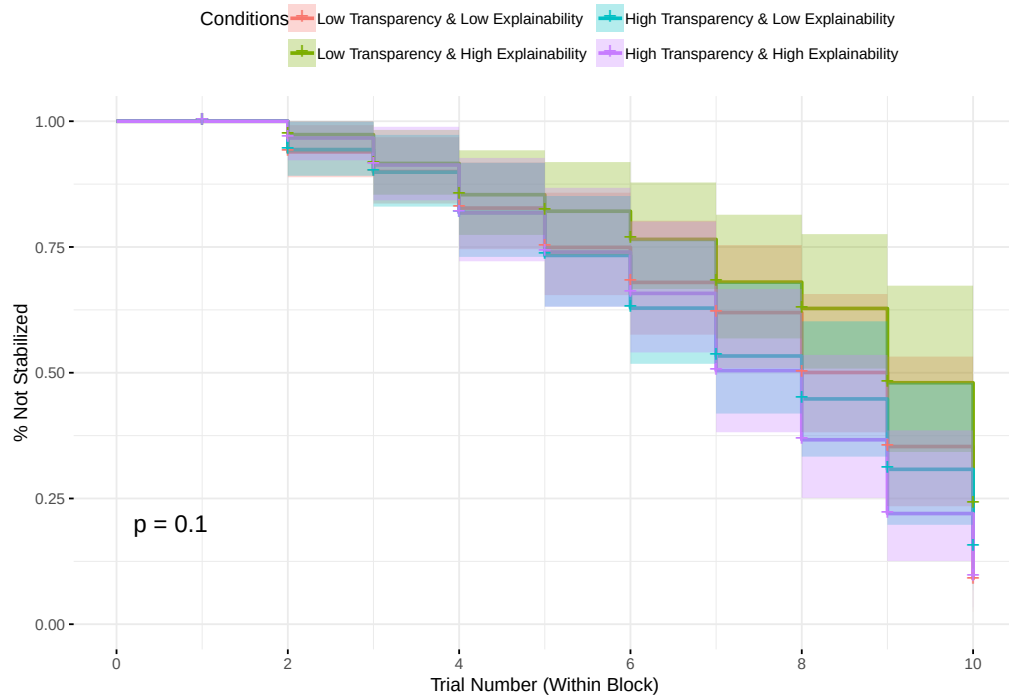


Figure B.20: Kaplan-Meier fit plot corresponding to post-hoc analyses. This plot illustrates the test where the stabilization threshold is set at 60% of a participant's initial attention shift rates.

Appendix C

Supplementary Materials for Chapter 3

C.1 Methods

C.1.1 Measurement Model

Table C.1: Measurement items.

	Measures	Adapted from
Perceived Autonomy	The AI forecasting system provided suggestions based on my true interests. The AI suggestion system let me do things my own way. The AI suggestion system allowed me to be in control of what I did.	(De Vreede et al., 2021)
Perceived Understanding	From the explanation, I understand how the AI model works. This explanation of how the AI model works is satisfying. This explanation of how the AI model works has sufficient detail. This explanation of how the AI model works seems complete. (excluded) This explanation of how the AI model works tells me how to use it. This explanation of how the AI model works is useful to my goals. This explanation of the AI model shows me how accurate the AI model is. This explanation lets me judge when I should trust and not trust the AI model.	(Hoffman et al., 2023)
Automation Induced Complacency Potential	When I have a lot to do, it makes sense to delegate a task to automation. If life were busy, I would let an automated system handle some tasks for me. Automation should be used to ease people's workload. If automation is available to help me with something, it makes sense for me to pay more attention to my other tasks. Even if an automated aid can help me with a task, I should pay attention to its performance. Distractions and interruptions are less of a problem for me when I have an automated system to cover some of the work. Constantly monitoring an automated system's performance is a waste of time. Even when I have a lot to do, I am likely to watch automation carefully for errors. It's not usually necessary to pay much attention to automation when it is running. Carefully watching automation takes time away from more important or interesting things.	(Merritt et al., 2019)
Computer Self-Efficacy	<i>I could complete the job using SAP..</i> ...if there was no one around to tell me what to do as I go. ...if I had never used a package like it before. ...if I had only the software manuals for reference. ...if I had seen someone else using it before trying it myself. ...if I could call someone for help if I got stuck. ...if someone else had helped me get started. ...if I had a lot of time to complete the task. ...if I had just the built-in help facility for assistance. ...if someone showed me how to do it first. ...if I had used similar packages like this one before to do this task.	(Compeau and Higgins, 1995)
Confidence in the AI Decision Confidence	Please rate your level of confidence in the AI forecasting system. Please rate your level of confidence in your decision.	(Chong et al., 2022)

Table C.2: AICP item descriptive statistics. (r) indicates reversed items.

Item	Mean	SD	Min	Max	Skewness	Kurtosis	Item-Total <i>r</i>
AICP_1	4.26	0.61	3	5	-0.19	-0.68	0.20
AICP_2	4.28	0.72	2	5	-0.80	0.49	0.20
AICP_3	4.57	0.58	3	5	-0.91	-0.25	0.19
AICP_4	4.15	0.87	2	5	-0.88	0.15	0.33
AICP_5	3.33	0.47	3	4	0.72	-1.52	-0.15
AICP_6	3.39	0.93	1	5	-0.34	-0.43	0.27
AICP_7	2.63	1.04	1	5	0.30	-0.98	0.31
AICP_8	4.72	1.03	3	7	0.08	-0.94	0.63
AICP_9	2.67	1.03	1	5	0.19	-0.98	0.33
AICP_10	3.20	0.98	1	5	-0.25	-1.13	0.42

Table C.3: AICP scale inter-item correlations. (r) indicates reversed items.

	AICP_1	AICP_2	AICP_3	AICP_4	AICP_5	AICP_6	AICP_7	AICP_8	AICP_9	AICP_10
AICP_1	1.00									
AICP_2	0.38**	1.00								
AICP_3	0.26	0.72**	1.00							
AICP_4	0.30*	0.32*	0.22	1.00						
AICP_5	-0.07	-0.28	-0.20	-0.29*	1.00					
AICP_6	0.09	0.26	0.20	0.34*	-0.09	1.00				
AICP_7	-0.16	-0.04	-0.09	0.28*	-0.25	0.18	1.00			
AICP_8	0.19	0.05	0.05	0.17	0.24	0.21	0.25	1.00		
AICP_9	-0.00	-0.23	-0.13	0.03	-0.05	-0.10	0.40**	0.60**	1.00	
AICP_10	0.02	-0.05	-0.04	-0.01	0.15	0.08	0.31*	0.56**	0.48**	1.00

* Represents significance at the 0.05 level. ** Represents significance at the 0.01 level.

Items 5 and 8 were treated as reverse measures.

Table C.4: Factor Loadings for the AICP - Workload Alleviation Subscale

Item	Standardized Loading
AICP_1	0.398
AICP_2	0.963
AICP_3	0.748
AICP_4	0.340
AICP_6	0.278

Table C.5: Factor Loadings for the AICP - Monitoring Subscale

Item	Standardized Loading
AICP_5	0.130
AICP_7	0.405
AICP_8	0.808
AICP_9	0.742
AICP_10	0.687

Table C.6: Computer Self-Efficacy item descriptive statistics

Item	Mean	SD	Min	Max	Skewness	Kurtosis	Item-Total <i>r</i>
cse_1	5.48	2.53	1	10	-0.31	-0.89	0.65
cse_2	4.69	2.63	1	9	-0.01	-1.23	0.63
cse_3	6.25	2.49	1	10	-0.67	-0.24	0.65
cse_4	7.67	2.23	1	10	-1.25	0.84	0.59
cse_5	8.29	2.03	1	10	-1.99	4.56	0.70
cse_6	8.42	1.64	1	10	-1.96	6.45	0.80
cse_7	7.60	2.35	1	10	-1.17	0.72	0.59
cse_8	6.65	2.52	1	10	-0.85	-0.02	0.58
cse_9	8.63	1.93	1	10	-2.55	7.35	0.61
cse_10	8.35	1.67	1	10	-1.85	5.69	0.62

Table C.7: Computer Self-Efficacy scale inter-item correlations.

	cse_1	cse_2	cse_3	cse_4	cse_5	cse_6	cse_7	cse_8	cse_9	cse_10
cse_1	1.00									
cse_2	0.81**	1.00								
cse_3	0.50**	0.55**	1.00							
cse_4	0.48**	0.50**	0.43**	1.00						
cse_5	0.57**	0.55**	0.55**	0.45**	1.00					
cse_6	0.46**	0.46**	0.53**	0.60**	0.64**	1.00				
cse_7	0.34*	0.32*	0.43**	0.21	0.47	0.65**	1.00			
cse_8	0.44**	0.31*	0.36**	0.33*	0.39**	0.51**	0.52**	1.00		
cse_9	0.23	0.19	0.33*	0.44**	0.51**	0.71**	0.63**	0.58**	1.00	
cse_10	0.30*	0.36**	0.53**	0.52**	0.42**	0.66**	0.40**	0.44**	0.49**	1.00

* Represents significance at the 0.05 level. ** Represents significance at the 0.01 level.

Table C.8: Factor Loadings of Computer Self-Efficacy Items

Item	Standardized Loading
cse_1	0.61
cse_2	0.60
cse_3	0.66
cse_4	0.65
cse_5	0.73
cse_6	0.89
cse_7	0.68
cse_8	0.61
cse_9	0.72
cse_10	0.69

Table C.9: Perceived Autonomy item descriptive statistics.

Item	Mean	SD	Min	Max	Skewness	Kurtosis	Item-Total <i>r</i>
PA_1	4.42	1.47	1	7	-0.49	-0.51	0.47
PA_2	5.19	1.62	1	7	-0.77	-0.38	0.41
PA_4	5.15	1.58	1	7	-0.99	0.25	0.46

Table C.10: Perceived Autonomy scale inter-item correlations.

	PA_1	PA_2	PA_4
PA_1	1.00		
PA_2	0.35**	1.00	
PA_4	0.42**	0.35**	1.00

* Represents significance at the 0.05 level. ** Represents significance at the 0.01 level.

Table C.11: Factor Loadings of Perceived Autonomy items

Item	Standardized Loading
PA_1	0.65
PA_2	0.54
PA_4	0.64

Table C.12: Perceived Understanding item descriptive statistics.

Item	Mean	SD	Min	Max	Skewness	Kurtosis	Item-Total <i>r</i>
exp_sat_1	4.50	1.69	1	7	-0.56	-0.73	0.67
exp_sat_6	4.50	1.69	1	7	-0.53	-0.73	0.60
exp_sat_7	3.42	1.88	1	7	0.44	-1.04	0.60
exp_sat_8	4.10	1.78	1	7	-0.24	-1.05	0.60

Table C.13: Perceived Understanding scale inter-item correlations.

	exp_sat_1	exp_sat_6	exp_sat_7	exp_sat_8
exp_sat_1	1.00			
exp_sat_6	0.49**	1.00		
exp_sat_7	0.49**	0.57**	1.00	
exp_sat_8	0.65**	0.41**	0.43**	1.00

* Represents significance at the 0.05 level. ** Represents significance at the 0.01 level.

Table C.14: Factor Loadings of Explanation Satisfaction Items

Item	Standardized Loading
exp_sat_1	0.81
exp_sat_6	0.64
exp_sat_7	0.65
exp_sat_8	0.74

C.1.2 Experimental Task Instructions

Task Instructions, Page 1

The following text and the supplementing Figure C.1 were shown on the first page of instructions, before participants start engaging with the tasks in the trial block:

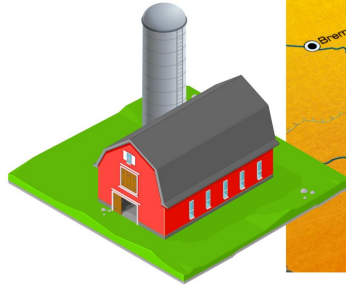
Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

Welcome to your new job at ERP Dairy! ERP Dairy is a natural ice cream distributor in the German market.

As a wholesaler, ERP Dairy purchases units of ice cream from dairy co-ops to sell them to grocery stores.

Welcome to Your New Job!



ERPsimLab
HEC MONTRÉAL | © ERPsim Lab, HEC Montréal

Commercial version
powered by BÂTON

Figure C.1: Experimental task instructions, page 1 visual supplement.

Task Instructions, Page 2

The following text and the supplementing Figure C.2 were shown on the second page of instructions, before participants start engaging with the tasks in the trial block:

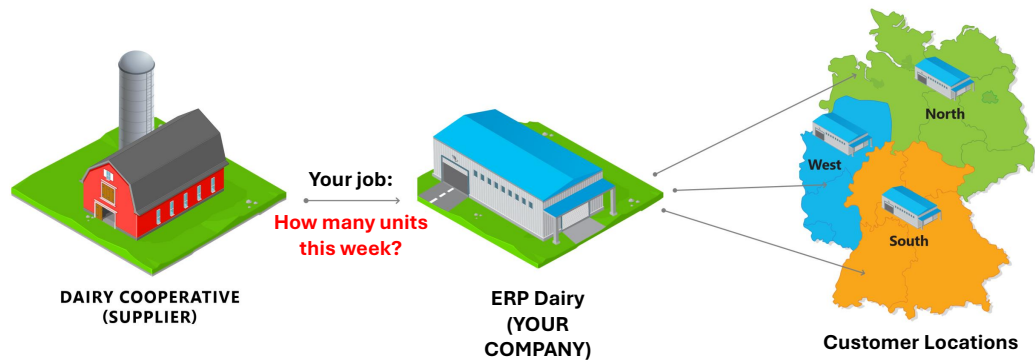
Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

As the demand forecaster, your job is to analyze the historical demand data, understand trends and seasonalities, and submit your sales forecasts to your suppliers.

Each week, you determine the number of units that you expect to sell, and then submit that sales estimate to your suppliers as a purchase order.

Logistics Game Design



ERPsimLab
HEC MONTRÉAL | © ERPsim Lab, HEC Montréal

Commercial version
powered by BATON

Figure C.2: Experimental task instructions, page 2 visual supplement.

Task Instructions, Page 3

The following text and the supplementing Figure C.3 were shown on the third page of instructions, before participants start engaging with the tasks in the trial block:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

Let's explore the tool you will be using: the SAP demand forecasting interface (see the screenshot below).

You will see the pricing information on top of the screen. **Sales price** is what your customers pay, **unit cost** is what you pay for each unit. Your product is a natural ice cream with a shelf life of one week only. Therefore, if you order excess ice cream, you sell them at a huge discount, 50 euros per unit. **Salvage price** indicates the discounted sales price of the excess units you ordered. The pricing information will be constant throughout the experiment.

Below the pricing information, you will see the weekly past demand chart. You are

expected to use this data to predict your sales for the next week. On this chart, week numbers start from January 1st, and your customers are located in Germany. Therefore, you can deduct your ice cream demand seasonality using the week numbers.

One year has 52 weeks, weeks from 20 to 35 are often summer months in the North.

Once you make your estimation, you must input that quantity into the textbox on the screen and click accept.

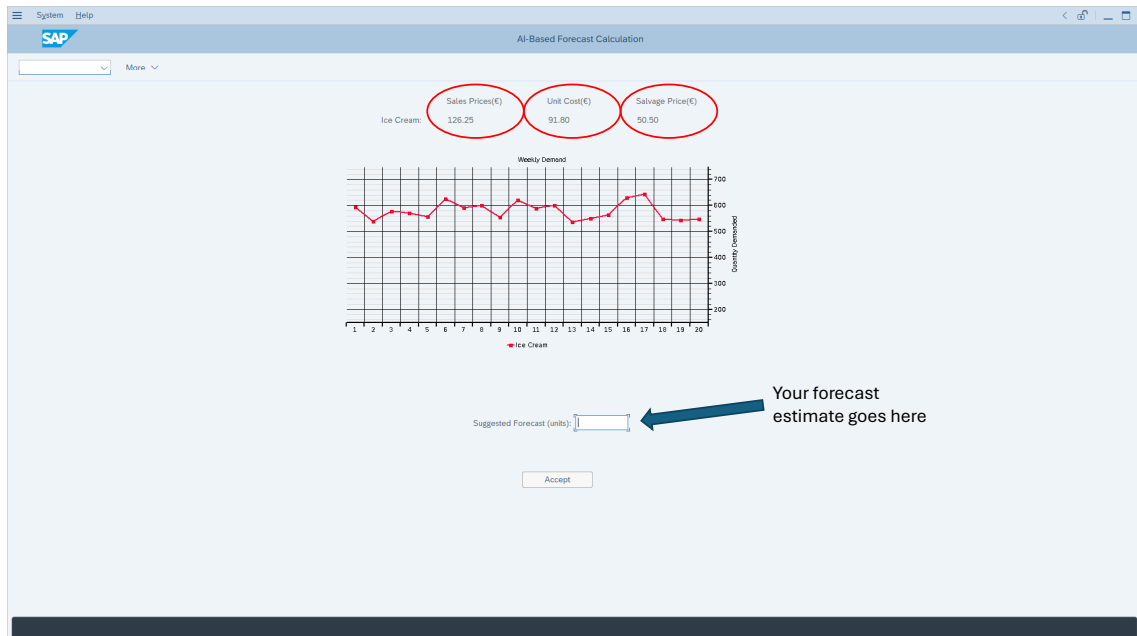


Figure C.3: Experimental task instructions, page 3 visual supplement.

Task Instructions, Page 4

The following text and the supplementing Figure C.4 were shown on the fourth page of instructions, before participants start engaging with the tasks in the trial block:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

When you submit your forecasts, the business simulation will advance by a week, and you will be presented with a performance report.

On this report, you will see how much you saved or lost compared to an average decision-maker.

There is also an overall forecasting error measure: MAPE. The lower the MAPE is, the better you are doing. A MAPE lower than 10% means you are doing an acceptable job.

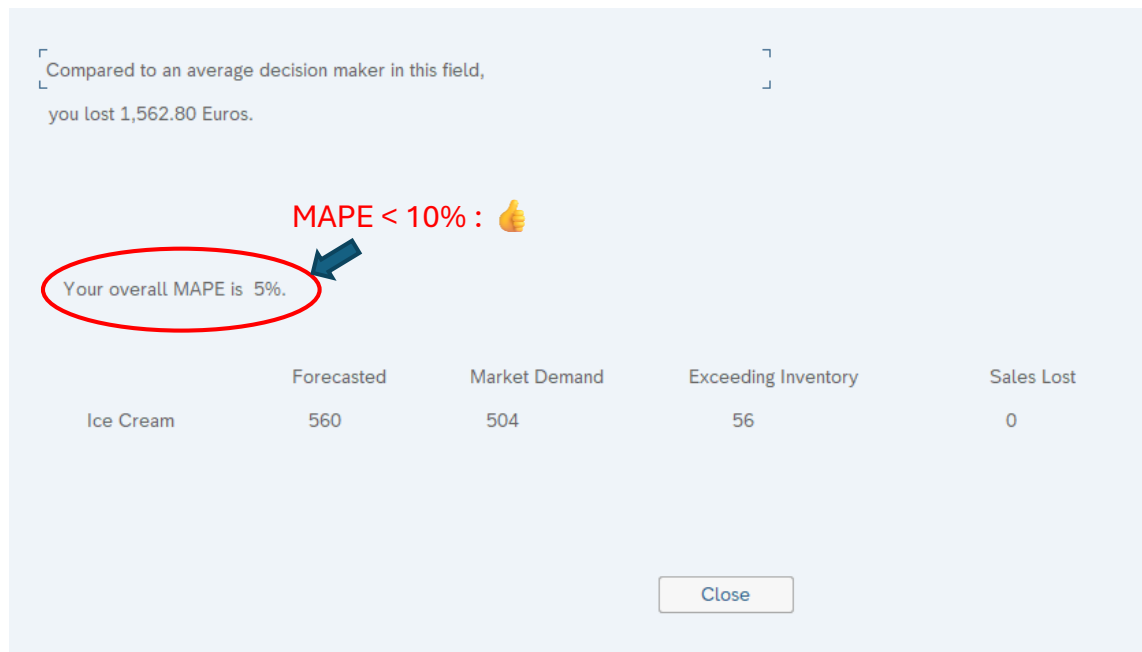


Figure C.4: Experimental task instructions, page 4 visual supplement.

Task Instructions After Trial Block, Page 1

The following text and the supplementing Figures C.5, C.6, C.7 were shown on the first page of instructions after the trial block, before participants started engaging with the experimental tasks:

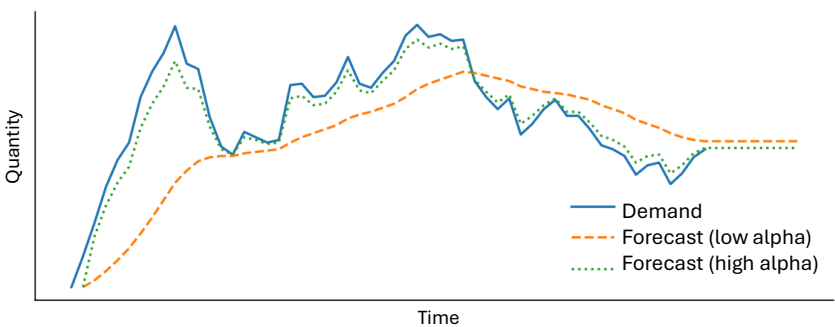
Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button anytime to continue to the next page of instructions.

Now, you will learn about the three important parameters in a forecasting model. You do not need to know all details for today's task, but it's a good idea to have a basic understanding.

Smoothing is the weight of recent data points in the forecast. Its parameter, alpha, indicates how much importance the model will allocate to the most recent observation compared to the demand history.

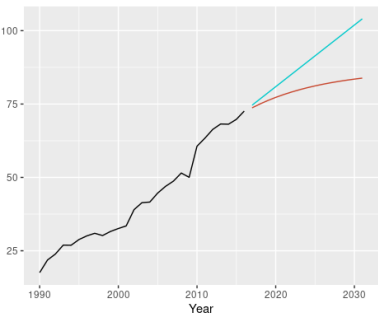
If the smoothing parameter is lower, forecasts are similarly impacted by older and recent sales data. If it's higher, forecasts consider recent data more, and forecasts might be impacted by irregular spikes.



Smoothing parameter (alpha)

Figure C.5: Experimental task instructions, visual supplement for page 1 after the trial block.

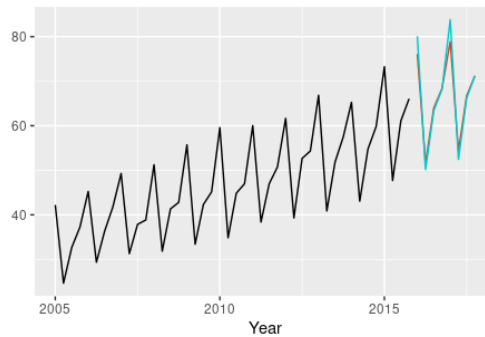
Trend is the slope of the time series data. A higher value of the trend parameter gives more weight to the trend.



Trend parameter

Figure C.6: Experimental task instructions, visual supplement for page 1 after the trial block.

Seasonality is the effect of periodic fluctuations in the market. A higher seasonality parameter gives more weight to the seasonal component, making the forecast more sensitive to changes in the seasonal pattern.



Seasonality parameter

Figure C.7: Experimental task instructions, visual supplement for page 1 after the trial block.

Task Instructions After Trial Block, Page 2

The following text and the supplementing Figure C.8 were shown on the second page of instructions after the trial block, before participants started engaging with the experimental tasks:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button when you are done.

Your employer purchased an artificial intelligence (AI) powered decision support system that helps demand forecasting experts in their daily tasks.

This system works on a different interface as shown on the screenshot below.

- First, it presents the information you previously used to make forecasting decisions.
- When you click the calculate suggestions button, it calculates demand forecasting estimates using a sophisticated algorithm and presents the results.

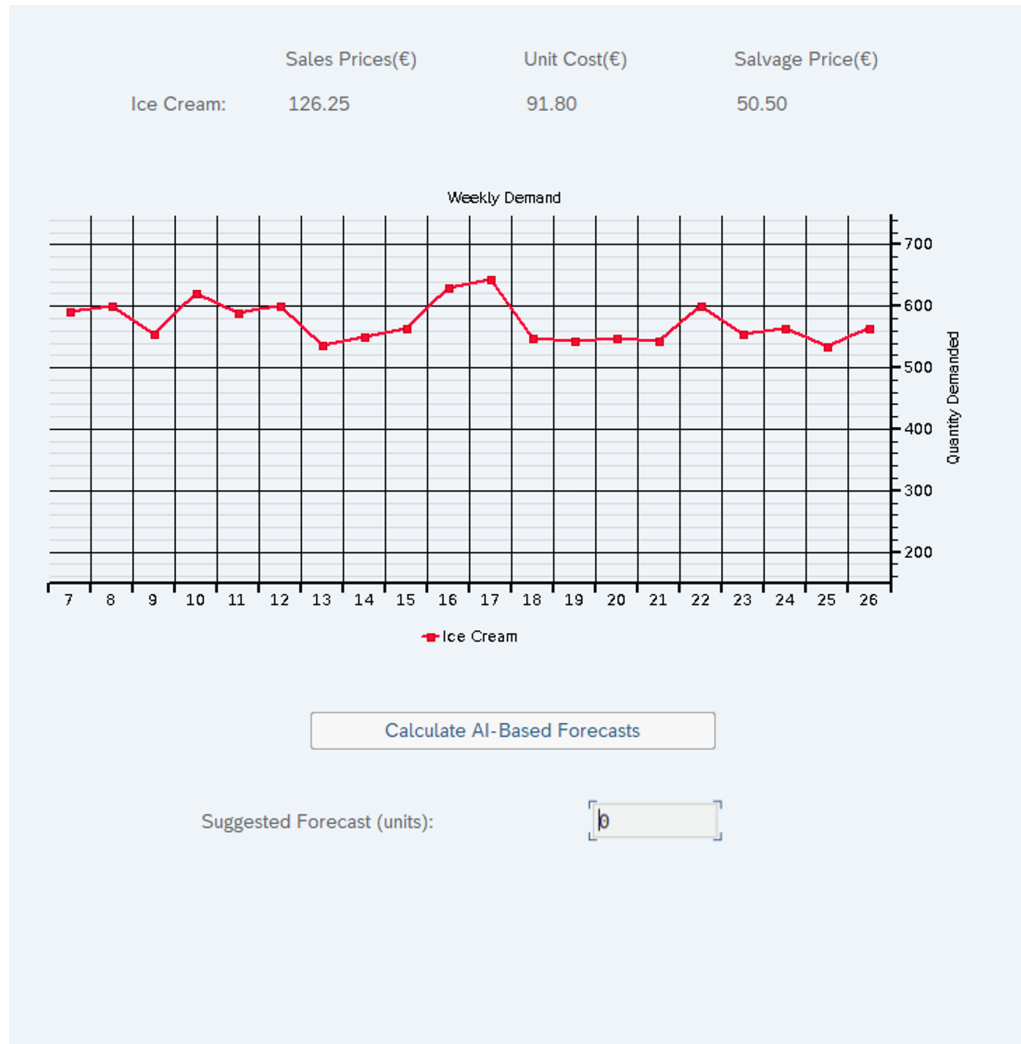


Figure C.8: Experimental task instructions, visual supplement for page 2 after the trial block.

Task Instructions After Trial Block, Page 3 for Participants Who Had Configuration Screen Access in Their Assigned Between-Subjects Condition

The following text and the supplementing Figure C.9 were shown as the third page of instructions after the trial block for participants assigned to the "configuration screen access" condition:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button when you are done.

Although this system is highly sophisticated, as other AI systems, it is not perfect. Therefore, you should evaluate its suggestions and explained factor weights, and make adjustments, if necessary.

Although the inner workings of this AI system are not completely transparent, this system will first ask for your input, and adjust its suggestions based on what you know about the dairy market.

The system will ask you for three parameters' weights, and then it will analyze the data to design its forecasting model for the next week.

Remember that the AI system will ask you these three parameters each week. You will control how the AI system responds to the demand data. So you will need to think about the market conditions and adjust the system according to what you think is the best condition for providing the best forecast.

After taking your input, the AI system will analyze the past data, and determine the most suitable set of parameters based on market conditions and your input.

- To accept AI suggestions as-is, you can click on the "Accept" button.
- To make adjustments on AI suggestions, you will first need to click on the "Edit Values" button first.
- When you are done editing, you can proceed to the upcoming week by using the "Accept" button.

Task Instructions After Trial Block, Page 3 for Participants in the AI without any configuration screen Condition

The following text and the supplementing Figure C.9 were shown as the third page of instructions after the trial block for participants assigned to the AI without any configuration screen condition:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button when you are done.

Although this system is highly sophisticated, as other AI systems, it is not perfect. Therefore, you should evaluate its suggestions and explained factor weights, and make adjustments, if necessary.

Although the inner workings of this AI system are not completely transparent, the application developers informed us that the system determines three parameters on its own, based on its perceptions of the demand. You have no control over these parameters, and we don't know how the system analyzes past data to identify them. However, the system explains the parameter weights as you see on the bottom half of the screen.

After taking your input, the AI system will analyze the past data, and determine the most suitable set of parameters based on market conditions and your input.

- To accept AI suggestions as-is, you can click on the "Accept" button.
- To make adjustments on AI suggestions, you will first need to click on the "Edit Values" button first.
- When you are done editing, you can proceed to the upcoming week by using the "Accept" button.

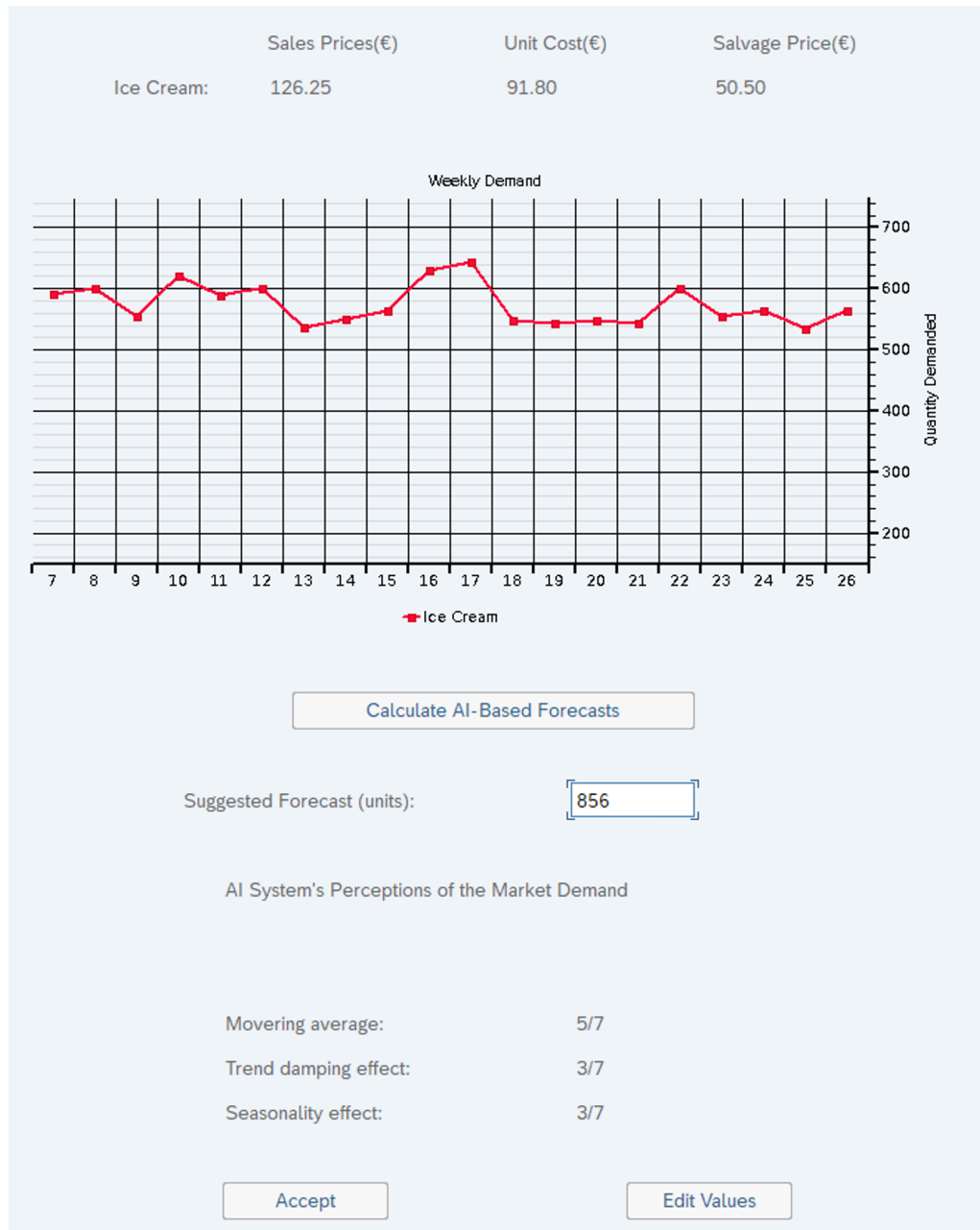


Figure C.9: Experimental task instructions, visual supplement for page 3 after the trial block.

Task Instructions After Trial Block, Page 4

The following text and the supplementing Figure C.10 were shown as the fourth page of instructions after the trial block, before participants started engaging with the

experimental tasks:

Please read the text explaining the scenario at your own pace and ask questions whenever needed.

You can press the next button when you are done.

Before we start, we have a request: For our brain measurements to work, we ask you to click the calculate button only after you analyzed the data chart and have an estimated forecast in your mind.

We also ask you to look at the AI suggestion textbox when you click the calculate button.



Figure C.10: Experimental task instructions, visual supplement for page 4 after the trial block.

C.2 Results

C.2.1 Descriptive Statistics

Table C.15: Block-Level Descriptive Statistics for Perceived Understanding of AI

Configuration	Mean	SD	N
0 (absent)	3.77	1.49	25
1 (present)	4.52	1.20	23

Table C.16: Descriptive Statistics for AI Errors on Correction Behavior

AI Error	Mean Correction Applied	SD	N
0	11.8	15.6	2609
1	155.0	55.9	711

Table C.17: Linear Mixed-Effects Model Predicting Correction Behavior (Error Manipulation Check)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	2.010	0.226	< .001***
Configuration	-0.033	0.098	.740
AI Error	3.039	0.063	< .001***
Computer Self-Efficacy (CSE)	-0.056	0.284	.846
Configuration × AI Error State	-0.116	0.093	.212
<i>Random Effects:</i>			
Participant Intercept Variance	0.091 (SD = 0.302)		
Residual Variance	1.173 (SD = 1.083)		

Note: $N = 3,252$ observations, 47 participants. REML criterion: 9,844.1.

R^2 (Marginal / Conditional): 0.543 / 0.576.

Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, · $p < .10$.

Table C.18: Trial-Level Descriptive Statistics by Configuration and Error States

Configuration	AI Error	Confidence in AI		Trial Duration		Performance		N
		Mean	SD	Mean	SD	Mean	SD	
0	0	0.579	0.218	22.4	17.0	-744	567	1355
0	1	0.272	0.226	24.3	12.7	-1202	1230	372
1	0	0.581	0.223	22.1	17.3	-748	618	1200
1	1	0.199	0.209	23.0	9.2	-1336	1394	325

Table C.19: Trial-Level Descriptive Statistics by Configuration and Error States

Configuration	AI Error	Post-click fixation duration		Post-click fixation rate		N
		Mean	SD	Mean	SD	
0	0	0.65	2.75	1.01	4.01	1355
0	1	1.20	4.30	1.59	5.11	372
1	0	0.46	2.03	0.57	2.45	1200
1	1	0.53	1.92	0.54	1.96	325

C.2.2 Hypotheses 1 and 4a

Table C.20: Linear Mixed-Effects Model Predicting Confidence in the AI Forecasting System (H1 & H4a)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.436	0.093	< .001***
Configuration	0.006	0.040	.939
AI Error	-0.289	0.011	< .001***
Computer Self-Efficacy (CSE)	0.166	0.117	.083
Configuration x Error	-0.082	0.016	< .001***
<i>Random Effects:</i>			
Participant Intercept Variance	0.018 (SD = 0.134)		
Residual Variance	0.032 (SD = 0.180)		

Note: $N = 2,602$ observations, 47 participants. REML criterion: -1365.

R^2 (Marginal / Conditional): 0.306 / 0.555 (approx.).

Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, $p < .10$

C.2.3 Hypotheses 2 and 4b

Table C.21: Linear Mixed-Effects Model Predicting P3a Amplitude (H2 & H4b)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.021	0.320	.98
Configuration	0.479	0.472	.16
AI Error	-1.608	0.435	< .001
Configuration x Error	1.993	0.651	.001
<i>Random Effects:</i>			
Participant Intercept Variance	1.358 (SD = 1.165)		
Residual Variance	44.759 (SD = 6.690)		

Note: $N = 2,290$ observations, 47 participants. REML criterion: 15242.5.

R^2 (Marginal / Conditional): 0.011 / 0.040

Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$

C.2.4 Hypotheses 3 and 4c

Table C.22: Linear Mixed-Effects Model Predicting Post-Fixation Duration (H3 & H4c)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	1.200	0.189	< .001***
Configuration	-0.031	0.082	.85
AI Error	0.075	0.018	< .001***
Computer Self-Efficacy	-0.478	0.238	.026*
Configuration x Error	-0.052	0.027	.026*
<i>Random Effects:</i>			
Participant Intercept Variance	0.075 (SD = 0.275)		
Residual Variance	0.092 (SD = 0.304)		
<i>Note:</i> N = 2,602 observations, 47 participants. REML criterion: 1385.4.			
<i>R</i> ² (Marginal / Conditional): 0.042 / 0.472 (approximate)			
Significance codes: *** <i>p</i> < .001, ** <i>p</i> < .01, * <i>p</i> < .05, <i>p</i> < .10			

C.2.5 Hypothesis 5

Table C.23: Linear Mixed-Effects Model Predicting Task Performance (H5)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.289	0.019	< .001***
Confidence in the AI Assistant	0.061	0.011	< .001***
Computer Self-Efficacy	0.006	0.02	.91
<i>Random Effects:</i>			
Participant Intercept Variance	0.0003 (SD = 0.018)		
Residual Variance	0.0215 (SD = 0.146)		
<i>Note:</i> N = 2,602 observations, 47 participants. REML criterion: -3580.9.			
<i>R</i> ² (Marginal / Conditional): 0.012 / 0.026.			
Significance codes: *** <i>p</i> < .001, ** <i>p</i> < .01, * <i>p</i> < .05, · <i>p</i> < .10.			

C.2.6 Hypothesis 6

Table C.24: Linear Mixed-Effects Model Predicting Task Performance (H6)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.313	0.0169	< .001***
P3a Amplitude (Pz)	-0.0005	0.00046	.16
Computer Self-Efficacy	0.011	0.022	.80
<i>Random Effects:</i>			
Participant Intercept Variance	0.00016 (SD = 0.013)		
Residual Variance	0.0217 (SD = 0.147)		
<i>Note:</i> N = 2,290 trials, 47 participants. REML = -2229.8.			
<i>R</i> ² (Marginal / Conditional): 0.001 / 0.008			
Significance codes: *** <i>p</i> < .001, ** <i>p</i> < .01, * <i>p</i> < .05, · <i>p</i> < .10			

C.2.7 Hypothesis 7

Table C.25: Linear Mixed-Effects Model Predicting Task Performance (H7)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.328	0.019	< .001***
Visual Verification (log(post fixation duration))	-0.015	0.0077	.03*
Computer Self-Efficacy	0.0086	0.021	.85
<i>Random Effects:</i>			
Participant Intercept Variance	0.00018 (SD = 0.013)		
Residual Variance	0.0217 (SD = 0.147)		
<i>Note:</i> N = 2,602 observations, 47 participants. REML criterion: -3607.7.			
<i>R</i> ² (Marginal / Conditional): 0.002 / 0.010 (approximate)			
Significance codes: *** <i>p</i> < .001, ** <i>p</i> < .01, * <i>p</i> < .05, · <i>p</i> < .10			

C.2.8 Decision Performance

Table C.26: Linear Mixed-Effects Model Predicting Log(Decision Performance)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.374	0.023	< .001***
P3a (Pz)	-0.0004	0.0004	.398
Confidence in the AI	0.058	0.012	< .001***
Log(Post-Fixation Duration)	-0.012	0.008	.132
Trial Number	-0.002	0.0002	< .001***
Computer Self-Efficacy (CSE)	-0.003	0.024	.896
<i>Random Effects:</i>			
Participant Intercept Variance	0.00028 (SD = 0.017)		
Residual Variance	0.02073 (SD = 0.144)		
<i>Model Fit:</i>			
REML Criterion	-2301.8		
Observations	$N = 2290$, 47 participants		
R ² (Marginal / Conditional)	0.045 / 0.058		
Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, [†] $p < .10$			

C.2.9 Hypothesis 8

Table C.27: Linear Model Predicting Perceived Understanding of the AI System (H8)

Fixed Effect	Estimate	Std. Error	p-value
Intercept	0.364	0.157	.012*
Configuration	0.117	0.067	.043*
Computer Self-Efficacy (CSE)	0.153	0.198	.220
<i>Model Fit:</i>			
Residual Std. Error	0.231 (df = 45)		
Multiple R^2	0.077		
Adjusted R^2	0.036		
F-statistic	1.88 on 2 and 45 df, $p = .165$ (two-tailed)		
<i>Note:</i> $N = 48$ observations.			
R^2 (raw / adjusted): 0.077 / 0.036			
Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, $^{\dagger} p < .10$			

C.2.10 Post-Hoc Mediation Analyses

Table C.28: Linear Mixed-Effects Models for Mediation via Confidence in AI in the "Configuration" Condition

Fixed Effect	Model 1: Outcome (DV) Estimate (SE)	Model 2: Total Effect Estimate (SE)	Model 3: Mediator (Confidence) Estimate (SE)
Intercept	0.297 (0.033)***	0.302 (0.033)***	0.294 (0.158)*
AI Error	-0.058 (0.010)***	-0.065 (0.012)***	-0.371 (0.012)***
Confidence in AI	—	-0.018 (0.021)	—
CSE	0.054 (0.042)	0.06 (0.042)	0.359 (0.202)
<i>Random Effects: Participant</i>			
<i>Model Fit:</i> REML = -1219 (M1), -1124.4 (M2), -586.4 (M3); $N = 1219$; 22 participants			
R^2 (Marginal / Conditional): 0.030 / 0.041 (M1), 0.031 / 0.040 (M2), 0.375 / 0.564 (M3)			
Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, $^{\cdot} p < .1$			

Table C.29: Linear Mixed-Effects Models for Mediation via Confidence in AI in the "No Configuration" Condition

Fixed Effect	Model 1: Outcome (DV) Estimate (SE)	Model 2: Total Effect Estimate (SE)	Model 3: Mediator (Confidence) Estimate (SE)
Intercept	0.334 (0.019)***	0.318 (0.021)***	0.492 (0.116)***
AI Error	-0.045 (0.009)***	-0.035 (0.010)***	-0.289 (0.011)***
Confidence in AI	—	0.033 (0.018)*	—
CSE	-0.0002 (0.023)	-0.004 (0.024)	0.092 (0.148)
<i>Random Effects: Participant</i>			
<i>Model Fit (REML):</i> -1457.4 (M1), -1454.6 (M2), -782.1 (M3); $N = 1383$ trials; 25 participants			
R^2 (Marginal / Conditional): 0.019 / 0.026 (M1), 0.022 / 0.029 (M2), 0.245 / 0.550 (M3)			
<i>Indirect Effect (Bootstrap, 5,000 simulations):</i> $\bar{a}\bar{b} = -0.01$, 95% CI [-0.0197, 0.0007]; 90% CI [-0.018, -0.001]			
Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, $^{\cdot} p < .1$			

Table C.30: Linear Mixed-Effects Models for Mediation via Attention Allocation on Manual Task Elements in the "Configuration" Condition

Fixed Effect	Model 1: Outcome (DV)	Model 2: Total Effect	Model 3: Mediator (Fixation)
	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept	0.297 (0.033)***	0.315 (0.033)***	10.328 (3.950)*
AI Error	-0.058 (0.010)***	-0.057 (0.010)***	0.431 (0.230)*
Log(Fixation Duration)	—	-0.0018 (0.0010)*	—
CSE	0.054 (0.042)	0.036 (0.042)	-10.175 (5.060)*

Random Effects (Participant Intercept SD / Residual SD):
M1: 0.016 / 0.150; M2: 0.014 / 0.150; M3: 3.053 / 3.545
Model Fit (REML): -1129.6 (M1), -1120.8 (M2), 6619.1 (M3)
Sample: N = 1219 trials; 22 participants
R² (Marginal / Conditional): 0.030 / 0.041 (M1), 0.033 / 0.042 (M2), 0.076 / 0.469 (M3)
Indirect Effect (Monte Carlo, 50,000 simulations): $\bar{a}\bar{b}$ = -0.0009, 95% CI [-0.0024, 0.0001]; 90% CI [-0.0021, -0.0000]
Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, · $p < .1$

Table C.31: Linear Mixed-Effects Models for Mediation via Attentional Responses in the "No Configuration" Condition

Fixed Effect	Model 1: Outcome (DV)	Model 2: Total Effect	Model 3: Mediator (Fixation)
	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept	0.334 (0.019)***	0.340 (0.020)***	6.377 (3.246)*
AI Error	-0.045 (0.009)***	-0.044 (0.009)***	0.719 (0.213)**
Log(Fixation Duration)	—	-0.00093 (0.00080)	—
CSE	-0.0002 (0.023)	-0.0048 (0.024)	-4.447 (4.152)

Random Effects (Participant Intercept SD / Residual SD):
M1: 0.012 / 0.141; M2: 0.013 / 0.141; M3: 4.085 / 3.509
Model Fit (REML): -1457.4 (M1), -1446.3 (M2), 7497.6 (M3)
Sample: N = 1383 trials; 25 participants
R² (Marginal / Conditional): 0.019 / 0.026 (M1), 0.020 / 0.028 (M2), 0.027 / 0.587 (M3)
Significance codes: *** $p < .001$, ** $p < .01$, * $p < .05$, · $p < .1$

C.2.11 Post-Hoc Analysis of Confidence Recovery Following AI Errors

Table C.32: Summary of Censored and Uncensored Observations by Experimental Condition

Condition	Total	Recovered (Event)	Censored
No Configuration	974	377	61.29%
Configuration	972	404	58.44%

Table C.33: Cox proportional hazards model of confidence recovery

Predictor	Coef.	SE	HR (95% CI)	p
Configuration	0.15	0.07	1.16 [1.01, 1.33]	.041

$n = 1,946$, Events = 781, Concordance $C = 0.52$

