Original Research

SAGE Open
October-December 2025: 1–18

DOI: 10.1177/21582440251380135
journals.sagepub.com/home/sgo
Sage

Trust Under Bounded Rationality: Exploring Human-AI Interaction in Decision-Making Through Large Language Models

Waleed Almutairi[1] and Ibrahim Almatrodi[1]

Abstract

This study investigates the role of artificial intelligence (AI) and large language models (LLMs) within Simon's bounded rationality framework, focusing on factors such as preferences, competence, learning, and persuasion that influence decision-makers' trust in AI outcomes. Data were collected using mixed methods, including surveys and interviews, followed by descriptive and thematic analyses to explore the trust dynamics in human-AI interactions under bounded rationality. Participants highlighted the effectiveness of AI systems in decision-making constrained by bounded rationality and discussed how AI systems might mitigate these limitations. The findings emphasize the critical role of trust in facilitating effective human-AI interactions, indicating that AI-provided explanations not only support decision-making but also enhance users' trust in these systems. This study identifies trust as a multifaceted and dynamic aspect of human-AI interactions, suggesting that AI developers can improve trustworthiness through transparency, demonstrated competence, and continuous learning. Enhancing these factors is expected to drive widespread adoption and improve the overall user experience with AI systems.

Keywords

artificial intelligence, large language models, decision-making, bounded rationality, trust, human-computer interaction

Introduction

Recent advancements in large language models (LLMs), a sub-discipline of artificial intelligence (AI), have transformed various sectors by enabling sophisticated language processing, necessitating an evaluation of their accuracy (Makridakis et al., 2023). Models such as GPT-3 have gained prominence for their exceptional ability to generate and interpret human-like text, revolutionizing applications across sectors (Takemoto, 2024; see Figure 1).

Omiye et al. (2024) argued that further training involving industry-specific queries can contribute toward developing intelligence applications for LLMs, enhancing their conceptual comprehension and reasoning capabilities. For instance, Saudi Arabia's Vision 2030 integrates AI to advance strategic decision-making across sectors (Zhavoronkov, 2022), with organizations adopting LLMs to enhance customer service, comprehend customer needs, and offer immediate solutions.

Despite its widespread adoption, challenges persist, including concerns related to trust, interpretability, transparency, and human cognitive limitations in decision-making scenarios. Issues around data governance, biases, privacy, security, and transparency have intensified with the increasing integration of LLMs into daily life (Piñeiro-Martín et al., 2023).

This study holds practical significance for AI developers, organizational decision-makers, researchers, and policymakers. By investigating the trust-enhancing features, this study aims to promote responsible AI integration across various sectors. Ultimately, it seeks to foster a deeper understanding among stakeholders involved in

[1]King Saud University, Riyadh, Saudi Arabia

Corresponding Author:
Waleed Almutairi, King Saud University, Prince Turki Bin Abdulaziz Al Awal Road, Riyadh 11451, Saudi Arabia.
Email: walmut44@gmail.com

Data Availability Statement included at the end of the article

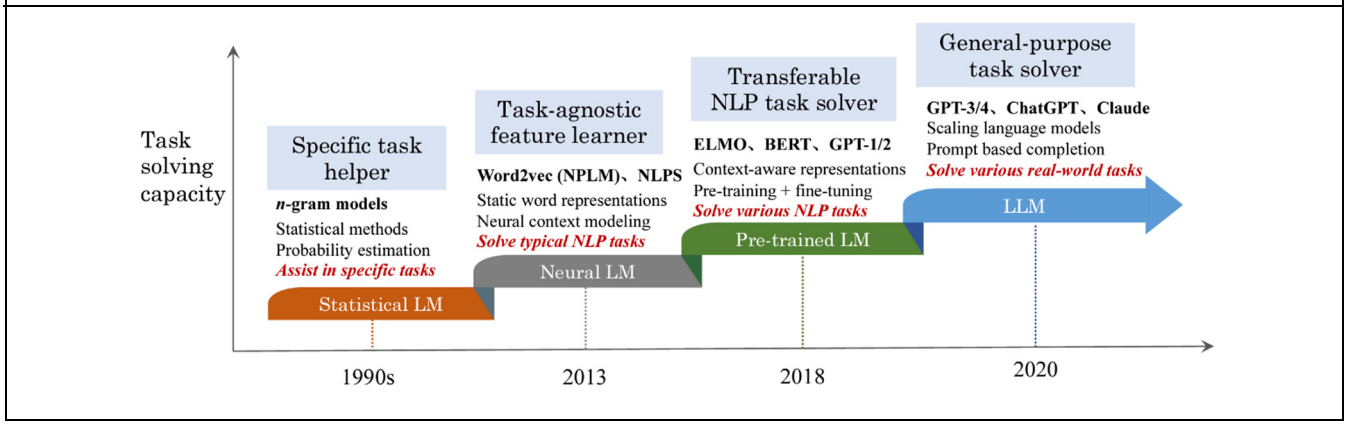

Figure 1. Evolution process of the four generations of language models (Zhao et al., 2023).

the design, implementation, and governance of AI technologies, thereby promoting trust and broader adoption.

Research Gap

The widespread adoption of LLMs in decision-making has amplified concerns regarding biases and trust, necessitating further investigation. This study addresses the gap in understanding how trust in AI systems, particularly LLMs, operates within Simon's bounded rationality framework. Specifically, it examines how factors such as AI competence, adaptability, and alignment with user values influence decision-makers' trust, especially when they rely on LLM outputs. These factors are critical, as trust mitigates cognitive and informational constraints inherent in bounded rationality. To address these challenges, applications such as explainable AI offers potential solutions by clarifying how model features influence outcomes, thereby bridging the gap between complex AI models and human cognitive constraints (Wang et al., 2019). By examining these dynamics, this study seeks to inform the development of trustworthy AI systems that support effective decision-making under cognitive constraints.

Literature Review

Decision-making involves logical evaluation and estimation of probable outcomes under conditions of uncertainty (Knight, 1921). Within the organizational context, Schoemaker and Russo (2016) examined decision-making models, highlighting rational, organizational, and political dimensions that shape complex choices. They noted that as firms grow, decision-making becomes increasingly complex and biased, influenced significantly by internal politics and constrained heuristics. Similarly, Barnard (as cited in Novicevic et al., 2011) distinguished organizational from individual decisions, defining decision-making as a conscious choice among alternatives to achieve desired goals. Novicevic et al. (2011) noted that this process can be both rational and intuitive.

Moreover, Arrow's (2012)*impossibility theorem* highlighted challenges in aligning individual preferences with collective goals in political science and welfare economics, suggesting that rational pursuit of self-interest often dominates. Tversky and Kahneman (1974) further challenged perfect rationality, demonstrating cognitive biases such as loss aversion, where individuals prioritize avoiding losses over achieving gains. Similarly, Janis (1972, as cited in Hart, 1991) described groupthink, where consensus-driven groups suppress critical thinking, potentially resulting in suboptimal decisions.

In social sciences, Edwards (1954) viewed decision-making as a process of choosing between different states, assuming complete information and rationality to maximize utility. Schilirò (2012) supported this, positing that individuals select options to maximize utility under uncertainty. Historically, Bentham equated utility maximization with achieving the greatest happiness, laying the foundation for economic psychology. Kahneman et al. (1997) revisited this, proposing experienced utility to redefine rationality (Read, 2020). Additionally, Neumann and Morgenstern introduced the expected utility hypothesis, addressing decision-making under risk and uncertainty (Prokop, 2023).

Weirich (1983) distinguished decisions from actions, noting that certainty in decision-making requires understanding consequences, whereas uncertainty involves unknown outcomes. Knight (1921) differentiated risk (measurable) from uncertainty (non-measurable) using the term uncertainty for non-quantifiable risk scenarios (Knight, 1921).

Bounded Rationality

Simon (1955, 1997) introduced bounded rationality, departing from the classical notions of rationality by acknowledging cognitive and environmental constraints.

He argued that decision-makers, unable to access all alternatives or predict all outcomes, seek satisfactory rather than optimal solutions, a concept termed satisficing. Simon's rational theory emphasized in-depth observations of real-world human behavior to understand decision-making under these constraints (Simon, 1955, 1997). Specifically, he identified limitations such as incomplete knowledge of alternatives, uncertainty about external events, and computational challenges in predicting outcomes (Simon, 1997). These constraints shifted the understanding from complete rationality, which assumes full knowledge and evaluation of all consequences, to bounded rationality (Simon, 1979). Consequently, Simon rejected the concept of "economic man," who maximizes benefits, in favor of the "administrative man," who prioritizes satisficing (Barros, 2010).

Simon's (1955, 1997) concept of bounded rationality and satisficing elucidated how human judgment frequently deviates from rationality due to incomplete information, enabling recognition of decision-making constraints. However, his framework did not address specific biases or their impacts on judgment. By contrast, Tversky and Kahneman (1974) identified systematic biases influencing human judgment, proposing heuristics as simplified decision-making techniques under such constraints.

Decision-Making Technologies

Decision support systems, which emerged in the early 1970s, enhanced decision-making efficiency (Pearson & Shim, 1995; Shim et al., 2002). Suri (2024) noted that decision-making is critical to the function of organizations or societies, with far-reaching consequences for individuals. Decision-making models have progressed significantly over time, with traditional methods replaced by current practices that depend on data, technology, and scientific frameworks. Teng (2023) noted that these practices leverage the power of data, enabling information collection and processing required for personal and organizational decisions, thereby opening up new opportunities for enlightened choices (Yousuf & Zainal, 2020).

The integration of AI and machine learning in decision-making processes has contributed to several significant improvements, particularly given that AI algorithms are capable of processing large amounts of data at a far greater speed than is possible for humans (Bao et al., 2023). Tu et al. (2023) highlighted the central role of technology in data collection and processing.

Large Language Models

LLMs, a subset of AI, leverage natural language generation (NLG) and natural language understanding (NLU) to facilitate sophisticated human-machine communication. NLG enables machines to produce meaningful text (Sheng et al., 2021), while NLU allows analysis of human language by extracting aspects such as concepts and emotions (Khurana, 2023).

LLMs rely on *transformer* architecture, a largely ignored mechanism that allows LLMs to understand context, generate coherent sentences, and address diverse issues across multiple domains (Zhao et al., 2023). Additionally, the integration of advanced attention mechanisms with the ability to transfer learning improves LLMs' responsiveness and accuracy (Vaswani et al., 2017). Figure 1 illustrates the evolution of language models in terms of their task-solving capacity (Zhao et al., 2023). Moreover, larger models demonstrate improved performance, capable of predicting complex text outputs, such as paragraphs or chapters (Hadi et al., 2023).

Explainable Artificial Intelligence

Explainable artificial intelligence (XAI) delivers transparent explanations for AI actions. Research over decades emphasized the need for addressing transparency, human interaction, and trust (Fox, 2017). In addition, opaque AI inferences challenge AI acceptance, emphasizing the need for interpretability (Adadi, & Berrada, 2018).

Various factors drive the demand for explainability in AI, particularly in predictive *black-box* or *complex models*. Goodman and Flaxman (2017) noted the necessity for understanding and interrogating machine learning systems. Furthermore, researchers developed interpretable algorithms to clarify complex mechanisms for decision makers, forming the foundation of XAI (Lakkaraju et al., 2019). However, data complexity often limits model explainability (Guidotti et al., 2019). Additionally, Doshi-Velez and Kim (2017a) identified decision impacts, error correction, and societal alignment as critical social aspects influencing explanations.

Bias and Trust

Information technology has transformed the decision-making processes in various areas by enabling improved data processing and informed choices (Yousuf & Zainal, 2020). However, real-world data often suffer from biases, small sample sizes, or inaccuracies, posing significant challenges (Whang et al., 2023). Therefore, acknowledging the vital role played by technology in data processing is important (Tu et al., 2023). Ethical concerns related to biases in LLMs threaten fairness and equity, as evidenced by preference or prejudice toward one set of people or ideas (Ferrara, 2023). Examples include perpetuating harmful stereotypes (Sheng et al., 2021), cultural bias leading to miscommunication (Cao et al.,

2023), and political bias favoring specific ideologies (Hartmann et al., 2023).

The tendency of LLMs to generate fabricated information, known as hallucinations, erodes trust, particularly in sensitive contexts, necessitating rigorous validation (F. Liu, Lin, et al., 2023; Valmeekam et al., 2023; Zhang et al., 2023). Trust, defined as accepting vulnerability to an entity, particularly when potential risks outweigh benefits, relies on ability, benevolence, and integrity (Zand, 1972, as cited in Cugueró-Escofet & Rosanas-Martí, 2019; Mayer et al., 1995, as cited in Cugueró-Escofet & Rosanas-Martí, 2019). Challenges such as algorithm aversion, where humans avoid AI after observing errors, and transparency issues in black-box models further underscore the need for trust-enhancing strategies (Cheng et al., 2019; Dietvorst et al., 2015; Doshi-Velez et al., 2017b; Lee, 2004). However, overloading users with information can adversely impact decision-making (Cummings, 2017).

Theoretical Framework

This study explores trust's pivotal role in decision-making within the bounded rationality context, focusing on LLMs and their application in AI-driven processes. Bounded rationality highlights cognitive and informational constraints that limit individuals' problem-solving capabilities. Trust, influenced by the transparency and understandability of AI systems, shapes effective decision-making in organizations. AI applications are designed to enhance model outcomes by improving interpretability and justifiability (Lakkaraju et al., 2019). Moreover, AI's ability to mimic human reasoning underscores parallels between computational mechanisms and cognitive processes (Simon, 1996). Cugueró-Escofet and Rosanas-Martí (2019) explored trust within the bounded rationality framework, emphasizing cognitive limitations' impact on human-human interactions, which this study extends by examining trust factors in AI systems.

Factors of Trust in AI Systems

Preferences. Bounded rationality underscores two key trust-related concepts: challenges in assessing benefits under uncertainty due to unpredictable probabilities and the influence of future preferences on decisions (Cugueró-Escofet & Rosanas-Martí, 2019). Thomas et al. (2024) demonstrated that user feedback improves LLM prompt accuracy, as witnessed in Bing's search ranker training, aligning AI outputs with user needs. Similarly, representation alignment through human feedback tailors LLMs to reflect values such as helpfulness and truthfulness, enhancing performance (W. Liu, Wang, et al., 2023). By contrast, Simon (1997) linked decision-making to a hierarchy of goal-driven choices within a means-ends framework, suggesting that these frequently serve as stepping stones toward higher objectives (Cugueró-Escofet, & Rosanas-Martí, 2019). Bounded rationality can limit an individual's ability to make decisions that align with core (or basic) values. However, familiarity with particular situations enhances the capacity to make decisions capable of optimizing higher-end values, leading to clear preferences regarding more immediate goals (Fischhoff et al., 1988, as cited in Cugueró-Escofet & Rosanas-Martí, 2019).

Competence. Bidault and Jarillo (1997, as cited in Cugueró-Escofet & Rosanas-Martí, 2019) noted that human competence aligns with behavioral literature's concept of ability, a fundamental aspect of trust, particularly in professions requiring technical expertise. When evaluating LLMs, competence extends to an algorithm's capacity to accurately interpret data, as well as to learn and make decisions aligned with predefined objectives (or values) set by users. LLMs exhibit varying degrees of confidence and performance across different tasks, mirroring human cognitive biases, indicating the complexity of AI competence and the need for further exploration (Singh et al., 2023). Moreover, integrating separate AI systems (i.e., agents) to work collectively offers a pathway for mastering complex domains, thereby addressing essential needs and advancing toward artificial general intelligence. This approach leverages the specialized competencies of individual agents to facilitate comprehensive mastery over multifaceted challenges and enhance AI systems' overall effectiveness (Seals et al., 2023).

Learning. Trustees' regret over initial choices can influence future decisions, shaped by limited information and external constraints such as time. Within the context of AI, LLMs can learn from their outcomes, and with more information and data, improve over time. AI uses multiple training mechanisms consisting of (1) pre-training, (2) fine-tuning, (3) reinforcement learning from human feedback, and (4) adapters (Hu et al., 2023; Lester et al., 2021; Ouyang et al., 2022). These processes resemble human ability to learn from mistakes. Thus, additional information and adjustments to AI model training enhance future accuracy and reduce errors. This iterative enhancement reflects the dynamic nature of the AI system, which continuously evolves to produce increasingly reliable and accurate results based on the feedback obtained from previous performances and refined computational mechanisms.

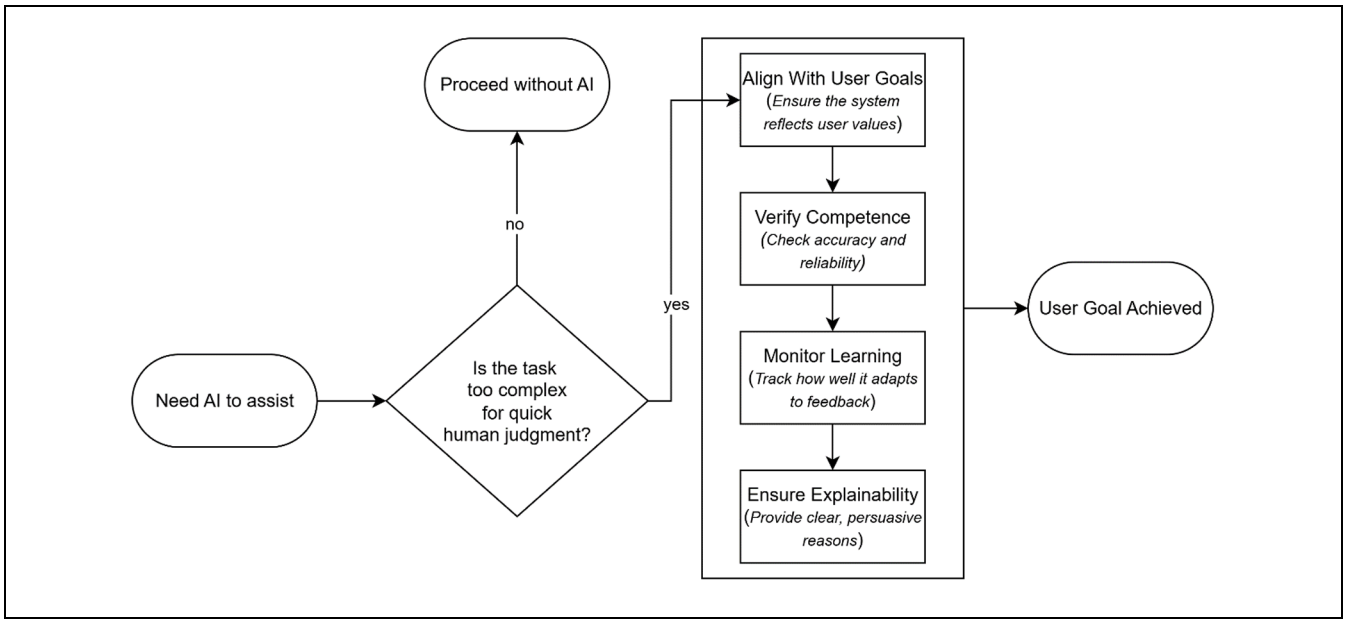

Figure 2. Trust building framework in human-AI decision-making under bounded rationality.

Persuasion. Persuasion, alongside incentives, aligns individual actions with organizational goals, particularly when goals lack immediate appeal (Barnard, 1938, as cited in Cugueró-Escofet & Rosanas-Martí, 2019). Contemporary business education and experience underscore the enduring value of effective communication and persuasion in achieving these goals. Uneven knowledge distribution between parties can undermine trust; for instance, if one party possesses greater expertise, the less-informed party may anticipate suboptimal decisions, reducing willingness to trust (Cugueró-Escofet & Rosanas-Martí, 2019). Persuasion mitigates this by influencing decisions and fostering trust. In knowledge-based and strategic contexts, persuasive communication, driven by informativeness and precision, shapes decision-making (Penczynski, 2016). In AI, black-box models' lack of transparency hinders user understanding and trust (Carli et al., 2022). XAI addresses this gap by delivering clear explanations, enhancing persuasiveness, and enabling informed decisions through transparent outputs.

Development of Theoretical Framework

This study employs Simon's bounded rationality to examine trust's impact on decision-making outcomes in AI systems, specifically LLMs, by emphasizing transparency and interpretability, fostering greater trust in AI decisions. Additionally, trust under bounded rationality encompasses multiple factors—preferences, competence, learning, and persuasion—that influence confidence in LLMs (see Figure 2). Insights from human-to-human interactions, as explored by Cugueró-Escofet and Rosanas-Martí (2019), highlight distinctions in human-to-machine interactions. These four factors (i.e., *preferences*, *competences*, *learning*, and *persuasion* form the core of this theoretical framework, guiding the analysis of trust in AI-assisted decision-making.

Methodology

Study Design

This study adopted a multidisciplinary mixed-methods approach, integrating behavioral economics, cognitive science, management information systems, and AI to examine trust in AI systems—particularly LLMs—under bounded rationality. The complex interplay of human cognitive limitations and machine-learning capabilities necessitated a combined qualitative and quantitative design to achieve both depth and empirical validation.

Specifically, the qualitative phase was prioritized to generate deep theoretical insights from managers' experiences and perceptions. This phase identified the context-specific factors influencing human trust in AI-driven decisions through in-depth interviews. Subsequently, a quantitative phase, utilizing structured questionnaires, was conducted to statistically measure and validate these factors across a broader managerial sample.

Qualitative exploration before quantitative validation allowed initial qualitative findings to inform the quantitative instrument design, enhancing both the relevance and robustness of the results. Thus, integrating these complementary methods provides both the depth and

Table 1. Theoretical Framework.

Theory	Focus	Trust under bounded rationality for AI decision-making
Bounded Rationality	1. Preferences 2. Competence 3. Learning 4. Persuasion	The contribution of this research paper

Table 2. Interview Participants and Assigned Codes.

Participants' number	Participants' role	Departments	Code Name
P1	CEO	Applications Sector	CEO1
P2	VP	Applications Sector	VP1
P3	Director	Data Integration	DI1
P4	Director	Software Development	DI2
P5	ML Engineer	Artificial Intelligence	ML1
P6	ML Engineer	Artificial Intelligence	ML2
P7	General Manager	Business Continuity	GM1
P8	General Manager	Security	GM2
P9	General Manager	Innovation	GM3
P10	General Manager	Data Science and Analysis	GM4
P11	Director	Data Quality	DI3
P12	Data Scientist	Data Analysis	DS1
P13	Data Analyst	Data Analysis	DA1

breadth essential to comprehensively address the complex interplay of human-machine trust in organizational decision-making contexts (Almalki, 2016).

We designed the study as a minimal-risk behavioral experiment with adult volunteers. Participants completed research questionnaires; we did not require real-world or clinical decisions, and we collected no sensitive personal information. We did not log participant identifiers, and we stored no names, email addresses, IP addresses, or device identifiers. The participants could withdraw at any time without penalty or loss of compensation. The study's potential benefits are substantial: it advances understanding of how people calibrate trust in AI recommendations under bounded rationality, which can inform safer, more transparent decision-support systems in domains such as education, customer support, and operations.

Development of Qualitative Phase

The qualitative analysis utilized semi-structured interviews to explore trust nuances in AI-assisted decision-making within organizational contexts. Interview questions, developed from key literature themes—preferences, competence, learning, and persuasion within bounded rationality—ensured a systematic investigation (see Table 1). Thirteen purposively selected participants, including one chief executive officer, one vice president, four general managers, three directors, and four employees from various departments, provided diverse perspectives (see Table 2). Each interview lasted approximately 1 hr, capturing varying levels of AI engagement. Thematic analysis, following Braun and Clarke (2006), identified and analyzed patterns in the qualitative data, directly informing the quantitative questionnaire's development for validation across a larger sample.

Development of Quantitative Phase

The quantitative phase utilized a structured questionnaire distributed online to a purposively selected sample of employees within an organization responsible for supporting and advancing the national agenda for data and artificial intelligence. The sample encompassed approximately 300 employees who were invited based on their active engagement or interest in AI-related activities within the organization. Using Cochran's standard statistical formula, a sample size of 169 respondents was required at a 95% confidence level and 5% margin of error. A total of 170 valid responses were received, marginally exceeding the minimum requirement. Internal communication channels facilitated efficient dissemination, ensuring accessibility across departments.

Data Collection Tools

Online questionnaires, distributed via internal communication channels, enabled convenient participation across departments. Once significant data were gathered, surveys and interviews were transcribed, and analysis techniques such as Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), Structural Equation Modeling (SEM), and t-tests were applied to extract relevant insights from both quantitative and qualitative data.

Results

Qualitative Results

Extensive attention to AI across various platforms, from press coverage to academic research, underscores its transformative potential for organizations (Bughin et al., 2018). Both organizations and AI users emphasize the need for AI systems to be transparent, explainable, ethical, and unbiased, highlighting the importance of responsible development and implementation of AI to achieve transformative outcomes (Zhao & Gómez Fariñas, 2023). Trust in human-machine interactions, particularly with LLMs under bounded rationality, presents complex challenges (Glikson & Woolley, 2020). This study investigates trust's role in AI-supported decision-making, identifying key factors that foster trust in environments where AI augments or directs human decisions. Through thematic analysis, qualitative findings reveal how these factors shape trust, as detailed below.

Theme 1: Criticality of Incorporating Preferences in AI Decision-Making. Participants emphasized that personal preferences play a crucial role in interactions with AI systems, and embedding these preferences within AI algorithms is essential for creating effective and user-centric systems. The application of these algorithms enables AI models to understand, predict, and adapt to user needs, ensuring that AI-driven solutions are not only technically proficient but also aligned with individual expectations. AI systems that consider these preferences can provide more relevant and practical recommendations, ultimately increasing their trust. For example, Data Integration Director (DI1) noted:

> *"Preferences are important factors that people consider. These preferences should be compatible with our values. For example, if AI suggests a correct answer that does not match our preferences and values, taking a direction that conflicts with my beliefs, this would certainly affect my trust in accepting answers from this system."* (DI1)

Effective interaction with AI models is important for establishing their technical capabilities and the user's ability to intelligently engage with these systems. This has led to the emergence of the role known as the Prompt Engineer, a specialist skilled in querying and instruction, to improve AI output. Data Quality Director (DI3) observed:

> *"I rarely notice a conflict between AI's answers and my preferences. This is because of how the user engages with these models. If the user gradually interacts with the model, the output quality increases significantly."* (DI3)

These findings align with existing theoretical frameworks, specifically, Simon's bounded rationality (1955) and satisficing concepts. Respondents emphasized that understanding and integrating user preferences enable AI systems to generate outcomes closely aligned with the acceptable satisfaction thresholds established by users, thereby bridging the cognitive limitations inherent in human decision-making. Participants highlighted applications of AI systems that leverage user preferences, particularly in contexts such as e-commerce platforms. As the General Manager of Innovation (GM3) noted:

> *"We find many e-commerce websites recommend and suggest a specific product when algorithms notice customers' behavior. This generally reflects the importance of preferences and their role in gaining trust."* (GM3)

Preferences must align with personal and organizational values, which is consistent with Simon's (1997) argument that although goal selection can be rational (i.e., in terms of efficacy and logic), the selection of each goal remains subjective (Cugueró-Escofet & Rosanas-Martí, 2019). Thus, trust naturally increases when AI systems generate recommendations or decisions that are compatible with user values. Furthermore, users felt that the system understood their needs and respected their perspectives.

Theme 2: Competence and Ability of AI Systems: Effectiveness Through Technical Proficiency. Alongside preferences, the competence and ability aspects demonstrate technical capabilities, including the power of such a system to process data and offer solutions that align with both human and organizational objectives. As Vice President (VP1) stated:

> *"Yes, ability is one of the essential features that fosters trust. This also applies to AI systems, where 'ability' implies the capacity of systems to process complex data, obtain insights, and understand the direction the topic is heading."* (VP1)

Data Quality Director (DI3) added:

> *"Correct answers cannot determine the ability of a particular model. For me, ability is how AI models process information as if thinking like humans, effective for deduction, reasoning, and formulating results closer to human thought."* (DI3).

This competence addresses bounded rationality constraints—uncertainty, incomplete information, and computational complexity—by processing data beyond human cognitive capacity. As Simon (1972) noted:

> *"Uncertainty about the consequences that would follow from each alternative, incomplete information about the set of alternatives, and complexity preventing the necessary computations from being carried out."* (p. 169).

AI systems can mitigate these constraints by providing enhanced data processing capabilities and removing (or reducing) biases inherent in human analysis. Furthermore, it can translate complex datasets into understandable and usable information, thereby supporting effective decision-making.

Theme 3: Role of Learning Mechanisms in AI Systems to Increase Trust in Outcomes. Learning mechanisms, including human feedback, enable AI systems to refine outcomes and foster trust. CEO1 and DS1 shared the same thoughts, demonstrating the critical role of ongoing training and human feedback in refining AI models. They noted:

> *"One of the key successes is humans intervention to correct these models using feedback. When Google launched its translator, it relied on people correcting the translation."* (CEO1).
> *"The role of human feedback is important in AI systems. Methods like reinforcement learning from human feedback align machine learning models with human goals."* (DS1).

Thus, by aligning AI learning processes with human goals, these mechanisms enhance decision-making, thereby expanding the scope of rationality (Cugueró-Escofet & Rosanas-Martí, 2019). Fischhoff et al. (1988) emphasized trial-and-error learning, where individuals derive decision-making frameworks from their experiences for making decisions in similar situations in the future (as cited in Cugueró-Escofet & Rosanas-Martí, 2019). By contrast, Simon (1987) highlighted intuition's role in familiar contexts, where past experiences and their emotional and cognitive impressions simplify the decision-making process by bypassing extensive analytical reasoning (as cited in Cugueró-Escofet & Rosanas-Martí, 2019).

Theme 4: Persuasion to Influence Decision-Making Through Intelligent Interaction. AI systems can effectively present information, recommendations, and insights in ways that resonate with human cognitive biases and preferences. Machine Learning Engineer (ML1) noted:

> *"Sometimes, subtle persuasion may occur in context, but the evidence and reasons provided by AI systems are critical, whether interacting with a human or a machine, significantly fostering trust."*

Various studies support this view. For instance, Simon (1997) highlighted that in the bounded rationality context, given cognitive limitations and imperfect information, not all variables are equally understood or valued. This underscores the benefits of using communication methods such as persuasion (Simon, 1997). Persuasion is not only employed to push a particular agenda or decision but also to facilitate information flow and frame it in a manner that is accessible and influential. Data Quality Director (DI3) recognized persuasion as an approach generally used across the Internet, noting that:

> *"This behavior, common on the Internet and not just in AI models, involves providing some results while ignoring others, resembling isolation rather than persuasion. AI models remain limited by their training data."*

Transparency, integral to persuasion, influences how trust and credibility are perceived in AI systems. It serves as a foundational element capable of building trust in marketing, public relations, and AI-driven decision-making, as it involves how information is collected, disseminated, and employed, along with how decisions are made, particularly when they directly impact users.

Quantitative Results

Building on the qualitative findings and themes of preferences, competence, learning, and persuasion, this quantitative analysis tested four hypotheses to validate their impact on user trust in AI decision-making:

- H1 (Preferences): Users trust AI decisions more when aligned with individual preferences.
- H2 (Competence): Greater perceived AI competence significantly enhances user trust in AI decision-making.
- H3 (Learning): AI systems that continuously adapt and learn from previous outcomes are perceived as more reliable, enhancing user trust.

Table 3. Descriptive Statistics for Survey Items Q3–Q10.

	Q3	Q4	Q5	Q6	Q8	Q10
Valid	170	170	170	170	170	170
Missing	0	0	0	0	0	0
Mean	3.224	3.265	3.106	3.947	4.118	3.765
Std. Deviation	0.984	0.982	1.167	1.095	0.916	1.237
Skewness	−0.651	−0.440	−0.615	−0.853	−0.704	−0.510
Std. Error of Skewness	0.186	0.186	0.186	0.186	0.186	0.186
Kurtosis	−0.466	−0.530	−0.839	0.140	−0.482	−0.870
Std. Error of Kurtosis	0.370	0.370	0.370	0.370	0.370	0.370
Minimum	1.000	1.000	1.000	1.000	2.000	1.000
Maximum	5.000	5.000	5.000	5.000	5.000	5.000

Table 4. Scale Reliability Estimates for Items Q3–Q10: McDonald's ω and Cronbach's α with 95% Confidence Intervals.

			95% CI	
Coefficient	Estimate	Std. Error	Lower	Upper
Coefficient ω	0.735	0.031	0.676	0.793
Coefficient α	0.731	0.040	0.653	0.808

- H4 (Persuasion): Persuasive AI communication, particularly when aligned with user values, positively influences trust in AI-driven decisions.

These hypotheses are tested using descriptive statistics, EFA, CFA, SEM, independent sample t-tests, and analysis of variance (ANOVA) to rigorously assess relationships.

As Table 3 illustrates, respondents exhibited moderate confidence in AI's ability to respect personal preferences (Q3: $M = 3.22$, $SD = 0.98$), reflecting cautious trust in AI personalization. Greater trust emerged in AI's competence for handling complex information (Q6: $M = 3.95$, $SD = 1.10$), suggesting strong perceived capability in analytical tasks. Participants valued AI's continuous learning capability as crucial for trust formation (Q10: $M = 3.77$, $SD = 1.24$), emphasizing adaptive and evolving systems. Similarly, AI's persuasive communication aligning with user values and preferences was highly regarded (Q8: $M = 4.12$, $SD = 0.92$), underscoring transparency and ethical alignment.

These findings highlight transparency, competence, and adaptive learning capabilities as key drivers of trust, aligning with Simon's (1972) bounded rationality theory, where AI bridges cognitive limitations through clear, reliable, and adaptive decision-making support.

The scale (Table 4) exhibited strong internal consistency, as indicated by McDonald's Omega ($\omega = 0.74$, 95% CI [0.68, 0.79]) and Cronbach's Alpha ($\alpha = 0.73$, 95% CI [0.65, 0.81]), confirming its reliability for further statistical analyses.

Table 5. Factor Loadings for the Two-Factor Solution of Items Q3–Q10.

	Factor 1	Factor 2	Uniqueness
Q6	0.834		0.294
Q8	0.795		0.466
Q10	0.369		0.763
Q5		0.722	0.523
Q3		0.716	0.523
Q4		0.617	0.549

EFA (Table 5) identified two distinct factors explaining 48% of the cumulative variance:

- **Factor 1 (Competence):** Items Q6 (*0.83*) and Q8 (*0.80*) exhibited strong loadings, indicating AI's robust technical capability and reliability, labelled **"Competence."**
- **Factor 2 (Relation Trust):** Items Q3 (*0.72*), Q4 (*0.62*), and Q5 (*0.72*) loaded significantly, representing alignment of values and preferences, labelled "Relational Trust."

A parallel analysis further validates retaining these two factors, aligning with theoretical frameworks that

Table 6. Fit Indices for the Two-Factor CFA Model.

Index	Value
Comparative Fit Index (CFI)	0.998
Tucker–Lewis Index (TLI)	0.997
Bentler–Bonett Non-Normed Fit Index (NNFI)	0.997
Bentler–Bonett Normed Fit Index (NFI)	0.976
Parsimony Normed Fit Index (PNFI)	0.527
Bollen's Relative Fit Index (RFI)	0.955
Bollen's Incremental Fit Index (IFI)	0.998
Relative Non-Centrality Index (RNI)	0.998

Table 7. Fit Indices for the Two-Factor SEM Model.

Index	Value
Comparative Fit Index (CFI)	0.999
Tucker–Lewis Index (TLI)	0.998
Bentler–Bonett Non-Normed Fit Index (NNFI)	0.998
Bentler–Bonett Normed Fit Index (NFI)	0.979
Bollen's Relative Fit Index (RFI)	0.956
Bollen's Incremental Fit Index (IFI)	0.999
Relative Non-Centrality Index (RNI)	0.999
Root Mean Square Error of Approximation (RMSEA)	0.018
RMSEA 90% CI lower bound	0.000
RMSEA 90% CI upper bound	0.097
RMSEA p-value	0.651
Standardized Root Mean Square Residual (SRMR)	0.037
Hoelter's critical N ($\alpha = .05$)	505.858
Hoelter's critical N ($\alpha = .01$)	664.064
Goodness of Fit Index (GFI)	0.997
McDonald Fit Index (MFI)	1.007

differentiate technical competence from relational trust in AI systems. Practically, the distinction between these two dimensions emphasizes both the competence (technical ability and practical usefulness) and relational (user-AI relationship) elements essential for developing comprehensive user trust in AI.

CFA (Table 6 and 7) confirmed an excellent model fit for the proposed factors. Specifically, the Comparative Fit Index (CFI = 0.998), Tucker–Lewis Index (TLI = 0.997), and Incremental Fit Index (IFI = 0.998) exceeded the recommended threshold of 0.95, indicating strong evidence of construct validity. Similarly, the Bentler–Bonett Normed Fit index (NFI = 0.976) and the Relative Fit Index (RFI = 0.955) were above the commonly accepted benchmark, further supporting the model fit. However, the Parsimony-Normed fit index (PNFI = 0.527) was relatively low, suggesting slight model complexity. Overall, these fit indices provide robust statistical evidence validating the factor structure and underlying theoretical framework of trust in AI decision-making.

The regression analysis (Table 8) showed significant model improvement $(\Delta x^2 = 18.16,\ p < .001)$, with explanatory transparency enhancing trust $(McFadden\ R^2 = .209)$ and $(Nagelkerke\ R^2 = .253)$, indicating a moderately strong effect. This highlights the critical importance of explanatory transparency in AI systems for enhancing user trust.

Similarly, results in Table 9 demonstrate a statistically significant improvement in model fit $(\Delta x^2 = 9.19,\ p = .010)$. This suggests that respondents strongly believed that an AI system's ability to learn and adapt to past decisions substantially increases its reliability. While the explained variance was modest (McFadden $R^2 = .100$, Nagelkerke $R^2 = .126$), this outcome underscores the importance of continuous learning capabilities in enhancing AI trustworthiness.

Table 10 shows a marginally significant improvement in model fit ($\Delta x^2 = 5.18$, p = .075). Although the improvement is smaller compared to questions 7 and 9, it suggests that persuasive communication aligned with user values has a positive, though less pronounced, effect on trust. The smaller effect sizes (McFadden $R^2 = .037$, Nagelkerke $R^2 = .053$) imply that transparency and adaptability outweigh persuasion in fostering trust.

Table 8. Regression Model Summary of Item Q7.

Model	Deviance	AIC	BIC	df	ΔX^2	p	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox and Snell R^2
M_0	86.754	88.754	91.889	169			0.000		0.000	
M_1	68.593	74.593	84.000	167	18.161	< .001	0.209	0.253	0.159	0.101

Table 9. Regression Model Summary of Item Q9.

Model	Deviance	AIC	BIC	df	ΔX^2	*p*	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox and Snell R^2
M_0	91.822	93.822	96.957	169			0.000		0.000	
M_1	82.632	88.632	98.040	167	9.189	.010	0.100	0.126	0.088	0.053

Table 10. Regression Model Summary of Item Q11.

Model	Deviance	AIC	BIC	df	ΔX^2	*p*	McFadden R^2	Nagelkerke R^2	Tjur R^2	Cox and Snell R^2
M_0	141.975	143.975	147.111	169			0.000		0.000	
M_1	136.791	142.791	152.198	167	5.184	.075	0.037	0.053	0.036	0.030

Table 11. Independent Samples T-Test of Item Q7.

Variable	Test	Statistic	df	*p*-value	Cohen's d	SE Cohen's *d*	95% CI Lower	95% CI Upper
Competence	Student	−4.883	168.00	<.001	−1.462	0.423	−2.067	−0.853
	Welch	−4.138	12.17	.001	−1.336	0.405	−2.109	−0.534
Relational Trust	Student	−2.800	168.00	.006	−0.838	0.345	−1.431	−0.244
	Welch	−2.508	12.33	.027	−0.791	0.340	−1.442	−0.115

Table 12. Independent Samples T-Test of Item Q9.

Variable	Test	Statistic	df	*p*-value	Cohen's *d*	SE Cohen's *d*	95% CI Lower	95% CI Upper
Competence	Student	−3.340	168.000	.001	−0.964	0.345	−1.537	−0.388
	Welch	−2.277	12.809	.041	−0.767	0.325	−1.393	−0.118
Relational Trust	Student	−2.057	168.000	.041	−0.594	0.311	−1.162	−0.023
	Welch	−1.803	13.500	.094	−0.553	0.308	−1.146	0.058

Table 13. Independent Samples T-Test of Item Q11.

Variable	Test	Statistic	df	*p*-value	Cohen's *d*	SE Cohen's *d*	95% CI Lower	95% CI Upper
Competence	Student	−1.724	168.000	.087	−0.373	0.223	−0.799	0.053
	Welch	−1.604	31.121	.119	−0.360	0.222	−0.790	0.077
Relational Trust	Student	−2.378	168.000	.019	−0.515	0.228	−0.942	−0.086
	Welch	−1.932	28.751	.063	−0.459	0.226	−0.896	−0.015

Independent samples t-test (Table 11) revealed significant differences in trust related to AI competence when explanations of the decision-making process were provided. Specifically, respondents who valued competence-based explanations exhibited a significantly higher trust level (Student's t (168) = −4.883, $p < .001$, Cohen's $d = -1.462$), indicating a large practical effect. This effect remained strong under the Welch test (t (12.17) = −4.138, $p = .001$, Cohen's $d = -1.336$). For relational trust, results were significant but moderate (Student's $t(168) = -2.800$, $p = .006$, Cohen's $d = -0.838$), confirmed by Welch's test ($t(12.33) = -2.508$, $p = .027$, Cohen's $d = -0.791$). These results underline the importance of clear explanations for both competence and relational trust.

AI's learning and adaptability features significantly influenced trust in competence (Table 12). Welch's adjusted test supported this finding with a moderate effect (t (12.809) = −2.277, $p = .041$, Cohen's $d = -0.767$). Relational trust displayed significance (Student's t (168) = −2.057, $p = .041$, Cohen's $d = -0.594$), although Welch's test indicated marginal significance (t (13.500) = −1.803, $p = .094$, Cohen's $d = -0.553$). These findings highlight that AI's learning capability significantly influences users' perceptions of its competence and modestly impacts relational trust.

For items concerning persuasive AI language aligned with user values (Table 13), the effect on perceptions of AI competence was not statistically significant (Student's $t(168) = -1.724$, $p = .087$; Welch's $t(31.121) = -1.604$, $p = .119$), indicating only a small practical impact. However, relational trust showed a significant result in the Student's t-test ($t(168) = -2.378$, $p = .019$, Cohen's $d = -0.515$) but was marginally non-significant under

Table 14. ANOVA Results Examining Differences in Competence Across Groups of Q1.

Homogeneity Correction	Cases	Sum of Squares	df	Mean Square	F	p
Welch	Q1	4.998	3	1.666	0.926	.454
	Residuals	138.669	13.782	10.061		

Table 15. ANOVA Results Examining Differences in Relational Trust Across Groups of Q1.

Homogeneity Correction	Cases	Sum of Squares	df	Mean Square	F	p
Welch	Q1	5.079	3	1.693	0.915	.459
	Residuals	129.245	13.853	9.330		

Welch's adjustment ($t(28.751) = -1.932$, $p = .063$, Cohen's $d = -0.459$). These findings suggest that persuasive language moderately affects relational trust, whereas its influence on the perception of AI competence is less definitive.

ANOVA (Table 14) examined how experience levels influence perceptions of AI competence and relational trust. The results for competence indicated no significant differences across various experience levels (Welch's $F(3, 10.06) = 0.926$, $p = .454$). Group means suggested subtle variations; however, these do not reach statistical significance, suggesting that individual experience may not distinctly shape perceived technical competence toward AI.

Similarly, the analysis of relational trust (Table 15), corrected for homogeneity with Welch's test, also revealed no statistically significant differences across experience levels (Welch's $F(3, 13.853) = 0.915$, $p = .459$). The mean scores indicated minimal differences across groups, reflecting consistency in relational trust perceptions regardless of users' AI familiarity.

Discussion

This study explored trust formation in human-AI interactions within the context of bounded rationality. The integration of qualitative insights from organizational experts and quantitative data from a structured questionnaire provided a comprehensive overview of the critical factors influencing trust. The analysis identified preferences, competence, learning, and persuasion as critical factors shaping trust in AI decision-making.

Quantitative findings revealed moderate-to-high levels of trust in AI competence ($M = 3.95$, $SD = 1.10$) and relational trust ($M = 3.77$, $SD = 1.24$), aligning closely with the theoretical framework of bounded rationality (Simon, 1955, 1972). These results underscore AI's technical proficiency and alignment with user preferences as pivotal for trust, enabling systems to address cognitive limitations through reliable and user-centric outputs.

Reliability assessments yielded strong internal consistency (Omega $\omega = .74$; Cronbach's $\alpha = .73$), indicating the robustness of the employed scales. EFA further distinguished competence and relational trust as separate constructs with clear factor loadings, highlighting the dual nature of trust in AI—both technical reliability (competence) and user interaction (relational)—aligned with prior research (Cugueró-Escofet, & Rosanas-Martí, 2019).

SEM analysis supported a good model fit across all indices (e.g., CFI = 0.999, RMSEA = 0.018, SRMR = 0.037), reinforcing model validity. This robust validation aligns with Simon's (1972) bounded rationality theory, indicating that well-designed AI systems can effectively address cognitive limitations by enhancing decision-making capacity through transparent and reliable outputs.

Logistic regression analyses revealed significant insights into the factors that shape trust in AI systems. Specifically, results indicated that the AI system's ability to learn and adapt from past decisions significantly enhanced perceived reliability, as evidenced by a statistically significant improvement in model fit ($\Delta\chi^2 = 9.19$, $p = .010$), with moderate variance explained (McFadden $R^2 = .100$, Nagelkerke $R^2 = .126$). Additionally, persuasive language from AI systems aligned with user values yielded marginally significant results ($\Delta\chi^2 = 5.18$, $p = .075$), suggesting that while persuasive communication contributes positively to trust, its impact is comparatively weaker than transparency and adaptability.

T-tests revealed significant differences in trust perceptions based on AI's competence and relational factors across specific items (e.g., competence differences were significant in Q7 and Q9, with strong Cohen's d values, but less pronounced in Q11). This suggests that although both competence and relational trust are important, the

significance of AI's perceived technical competence may be more pronounced in trust formation, especially when explanations and transparency are effectively implemented.

Collectively, the ANOVA findings highlight a relatively stable relationship between individual experiences and trust dimensions. Both competence and relational trust appeared to be consistent across different experience levels, indicating that trust in AI systems may rely on factors beyond individual experiences.

Thematic analysis reinforced these quantitative insights. The participants consistently emphasized the importance of AI's ability to incorporate personal preferences, demonstrate technical proficiency, and employ robust learning mechanisms. Expert interviews indicated that integrating user preferences into AI decision-making processes not only enhances system effectiveness but also significantly boosts user trust, consistent with previous findings (Simon, 1997, as cited in Cugueró-Escofet & Rosanas-Martí, 2019).

Therefore, participants emphasized the importance of technical capabilities and the power of AI systems in processing complex data. This reflects the key contribution of AI in managing cognitive limitations inherent in human decision-making, highlighting its practical value in organizational contexts.

The importance of learning mechanisms in building trust was further validated qualitatively, aligning with the literature, which indicates that feedback loops and continuous learning are vital for enhancing AI systems. Simon's opinion of intuition in decision-making (1987, as cited in Cugueró-Escofet & Rosanas-Martí, 2019) is well aligned with AI learning mechanisms. Simon noted that intuition operates effectively in familiar contexts, where past experiences and the resulting emotional and cognitive impressions can simplify decision-making processes and reduce the need for extensive analytical reasoning. Similarly, as AI systems learn from previous data and interactions, they develop a form of 'digital intuition,' which allows them to form faster and more effective decisions based on learned patterns, without always resorting to deep analytical processes. This capability mirrors human intuition but is derived from huge data that AI systems can process, learn from, and recall.

Quantitative analysis revealed persuasion as relevant, but notably less influential than technical competence and capability signals. The empirical results demonstrated that persuasive interactions, although valuable, constitute a sub-factor under the broader umbrella of competence, indicating that users primarily interpret persuasive cues as supplementary evidence of competence rather than as independent drivers of trust.

Moreover, while acknowledging its role, the overall findings highlighted that persuasion's influence on trust formation, especially under bounded rationality conditions, remains secondary to the direct indicators of competence, learning capability, and technical reliability. This suggests that persuasive framing effectively supports trust formation but primarily operates through reinforcing perceptions of competence rather than as an independent trust-building factor.

Linking the Findings with Literature

Prior research has underexplored bounded rationality's impact on trust in AI interactions. This study addressed this gap by examining how AI mitigates cognitive limitations in organizational decision-making. Developing a clear understanding of how advanced AI tools influence decision-making processes can enable users and developers to establish more informed strategies, particularly during the development and usage phases of a system. This informed approach is crucial for maximizing the benefits of AI while simultaneously minimizing any potential risks, particularly under bounded rationality, where decision-making is constrained by cognitive limitations. This study emphasized that such an understanding is vital for crafting AI systems that are not only technically proficient but also trustworthy and aligned with human values and organizational goals.

Additionally, this study established the importance of considering personal preferences when designing AI systems. Furthermore, the integration of user preferences into AI algorithms is fundamental for creating effective and user-centric systems. As demonstrated in this study, this approach extends beyond technical proficiency, ensuring that AI-driven solutions align with individual preferences, thereby enhancing user satisfaction and trust.

This study also illustrated that incorporating user preferences into AI algorithms fosters satisficing, where satisfactory rather than optimal solutions align with cognitive constraints (Simon, 1997, as cited in Barros, 2010). Moreover, the competence of AI systems is defined by their technical capabilities, including the power to process complex data efficiently, critical for enhancing human decision-making capacity. This aligns with Simon's (1972) bounded rationality theory, which acknowledges the inherent limitations in human cognition, uncertainty concerning potential outcomes, incomplete information on available alternatives, and computational complexity inherent in decision-making. AI systems can significantly mitigate these constraints by offering advanced data-processing capabilities that transcend human abilities. Furthermore, as AI systems learn and evolve, they will become better equipped to handle real-world applications, thus reinforcing user trust in technology.

Table 16. Theoretical Framework.

Theory	Focus	Description and Justification
Bounded Rationality	1. Preferences 2. Competence 3. Learning 4. Persuasion	1. Integrating user preferences into AI systems is critical for aligning AI-driven decisions with individual needs and values, thereby enhancing acceptance, trust, and mitigating constraints imposed by bounded rationality. *Preference alignment significantly influenced the Relational Trust factor, as reflected in value alignment.* 2. Demonstrating the technical competence of AI systems is essential for processing complex data and providing actionable insights, helping to reduce the cognitive load on decision makers. *Competence emerged as a dominant factor in the empirical data, significantly influencing perceptions of trust.* 3. Continuous learning mechanisms, including iterative feedback loops, enhance AI's adaptability and effectiveness in dynamic environments, improve AI decision-making reliability, and thus user trust. *Learning was notably loaded in empirical data on competence, significantly predicting perceived reliability, thus confirming its role as a subcomponent of competence.* 4. Utilizing persuasive communication techniques enables AI to present information effectively and minimize human cognitive biases, thereby enhancing decision-making processes. However, persuasion primarily enhances perceptions of AI competence rather than building trust independently. *Persuasion was moderately correlated with competence and was empirically shown to be less impactful compared to direct competence indicators, reinforcing competence rather than independently driving trust decisions.*

Persuasion plays a critical role in environments characterized by bounded rationality, particularly when not all variables are equally understood or valued. Persuasion has emerged as a significant element in the learning process, helping overcome decision-makers' cognitive constraints and bridge gaps in understanding and valuation.

Simon's (1997) insights further reinforce this perspective, acknowledging that, in the bounded rationality context (marked by cognitive limitations and imperfect information), communication strategies (i.e., persuasion) become essential. These limitations prevent uniform understanding or valuing of all information, demonstrating that persuasive communication can play a transformative role in influencing decision-making processes (Simon, 1997).

A case study in a Saudi Arabian organization focused on data and AI (see Table 16) applied a bounded rationality framework to examine decision-making processes, revealing three key insights:. (a) the framework's relevance in data- and AI-driven environments, where information complexity frequently exceed human cognitive capacity, (b) AI and data analytics tools' capacity to mitigate or exacerbate these biases and cognitive limitations, and (c) AI's effectiveness in supporting decision-makers to overcome these constraints (Barros, 2010). These findings inform strategies for designing trustworthy, user-centric AI systems aligned with human values and organizational goals, balancing benefits and risks.

Thus, the use of this theoretical framework provided crucial insights into the dynamics of human-AI interaction and decision-making, contributing to the understanding of practical applications and implications of bounded rationality in modern organizational settings.

Limitations and Future Research

This study's sample, drawn from employees in a single AI-focused organization, limits the generalizability of the findings to other industries. The cross-sectional design captured trust perceptions only at a specific point in time, missing evolving dynamics over time. While preferences, competence, learning, and persuasion emerged as key trust factors in AI decision-making within the bounded rationality framework, other potential influences on trust may remain unexplored, limiting the findings' comprehensiveness. Additionally, the qualitative thematic analysis involved the researcher's interpretation, potentially introducing bias.

Future research should address these limitations by including diverse organizational contexts to enhance generalizability, adopting longitudinal designs to track trust dynamics, and exploring additional trust factors to enrich understanding of human-AI interactions. Incorporating multiple data sources can further reduce bias in qualitative analyses, strengthening the robustness of findings.

Conclusion

This study addressed a critical gap in the literature on trust in decision-making processes involving AI systems, particularly through the lens of bounded rationality. By investigating the roles of preferences, competence,

continuous learning, and persuasion in human-AI interactions, the study offered significant insights into building and maintaining trust. The findings demonstrated that aligning AI-driven decisions with individual user preferences significantly enhances acceptance and trust, supporting Simon's theoretical notions of bounded rationality and satisficing. Moreover, this study highlighted the importance of AI system competence, noting that users were more likely to trust and prefer AI systems that consistently exhibit technical proficiency and accurate decision-making. The relationship between AI competence and user trust was identified as a key factor influencing the adoption of AI models across various sectors, including medical diagnostics, financial risk assessment, and autonomous vehicles. The study further emphasized the value of continuous learning capabilities in AI, underscoring how systems that learn and adapt from performance errors enhance their reliability and user satisfaction over time. While persuasion was found to be beneficial, its role was supportive rather than independently critical, primarily reinforcing the perceptions of AI competence. Overall, this study provided a comprehensive theoretical framework that clarified how AI can address cognitive constraints inherent in decision-making. The empirical findings not only contribute to theoretical advancements but also offer valuable guidance for AI developers and organizational decision-makers seeking to enhance the trustworthiness and effectiveness of AI technologies in real-world settings. Future research should explore additional trust factors, such as how trust evolves in real-time AI interactions to broaden these insights.

Acknowledgments

The paper is based on the first author's unpublished master's thesis.

ORCID iDs

Waleed Almutairi https://orcid.org/0009-0006-0854-3036
Ibrahim Almatrodi https://orcid.org/0000-0003-2363-7357

Ethical Considerations

Conducted as part of a master's thesis in the E-Business Program, Department of Management Information Systems, King Saud University, the department approved this study, and no ethical issues were involved in the study. All respondents of the research data were informed and consented.

Consent to Participate

The authors ensured that the participants' consent was obtained before each interview and each questionnaire. The author(s) asked each employee at the start of the interview about consent to process and use their data for scientific publications. Also, reminded them that they could refuse to answer any question or choose to leave the interview at any time.

Author Contributions

conceptualization, I.A., W.A.; methodology, I.A., W.A.; validation, W.A.; formal analysis, W.A; data curation, W.A.; writing—original draft preparation, W.A., I.A; writing—review and editing, W.A; visualization, W.A.; supervision, I.A. All authors have read and agreed to the published version of the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data Availability Statement

The data supporting this study's findings are available from the corresponding author upon reasonable request.

References

Almalki, S. (2016). Integrating quantitative and qualitative data in mixed methods research–Challenges and benefits. *Journal of Education and Learning*, *5*(3), 288–296. https://doi.org/10.5539/jel.v5n3p288

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Arrow, K. J. (2012). *Social choice and individual values*, (Vol. 12). Yale University Press.

Bao, Y., Gong, W., & Yang, K. (2023). A literature review of human–AI synergy in decision making: From the perspective of affordance actualization theory. *Systems*, *11*(9), 442. https://doi.org/10.3390/systems11090442

Barros, G. (2010). Herbert A. Simon and the concept of rationality: Boundaries and procedures. *Brazilian Journal of Political Economy*, *30*(3), 455–472. https://doi.org/10.1590/S0101-31572010000300006

Bidault, F., & Jarillo, J. C. (1997). Trust in economic transactions. In P. Y. G. F. Bidault & a. G. Marion (Eds.), *Trust, firm and society* (pp. 81–94). London: Macmillan Business.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. *McKinsey Global Institute* p. 4. https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-modeling-the-impact-of-ai-on-the-world-economy

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). *Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study*. https://doi.org/10.48550/arXiv.2303.17466

Carli, R., Najjar, A., & Calvaresi, D. (2022). Risk and exposure of XAI in persuasion and argumentation: The case of manipulation. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems* (pp. 204–220). Springer International Publishing. https://doi.org/10.1007/978-3-031-15565-9_13

Cheng, H. F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems* (pp. 1–12). ACM. https://doi.org/10.1145/3290605.3300789

Cugueró-Escofet, N., & Rosanas-Martí, J. (2019). Trust under bounded rationality: Competence, value systems, unselfishness and the development of virtue. *Intangible Capital*, *15*, 1–21. http://dx.doi.org/10.3926/ic.1407

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In D. Harris & W. C. Li (Eds.), *Decision making in aviation* (pp. 289–294). Routledge. https://doi.org/10.4324/9781315095080-17

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology*. *General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Doshi-Velez, F., & Kim, B. (2017a). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://doi.org/10.48550/arXiv.1702.08608

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., & Wood, A. (2017b). Accountability of AI under the law: The role of explanation. https://doi.org/10.48550/arXiv.1711.01134

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*(4), 380–417. https://doi.org/10.1037/h0053870

Ferrara, E. (2023). Should ChatGPT be biased? *Challenges and risks of bias in large language models*. https://doi.org/10.48550/arXiv.2304.03738

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1988). Knowing what you want: Measuring labile values. *Decision Making: Descriptive, Normative and Prescriptive Interactions*. *Cambridge University Press* (pp. 398–421). https://psycnet.apa.org/doi/10.1017/CBO9780511598951.020

Fox, M., Long, D., & Magazzeni, D. (2017). Explainable planning. *arXiv preprint arXiv:1709.10256*. https://doi.org/10.48550/arXiv.1709.10256

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., & Mirjalili, S. (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea preprints*. https://doi.org/10.36227/techrxiv.23589741.v1

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4316084

Hu, Z. a.-P.-W. (2023). LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2304.01933*. https://doi.org/10.48550/arXiv.2304.01933

Janis, I. L. (1972). *Victims of Groupthink: A Psychological Study of Foreign-Policy Decisions and Fiascoes*. Houghton Mifflin.

Kahneman, D., Wakker, P. P., & Sarin, R. (1997). Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, *112*(2), 375–406. https://doi.org/10.1162/003355397555235

Khurana, D. A., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Knight, F. H. (1921). *Risk, uncertainty and profit*, (Vol. 31). Houghton Mifflin.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131–138). ACM. https://doi.org/10.1145/3306618.3314229

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*. https://doi.org/10.48550/arXiv.2104.08691

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J. R. (2023). Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*. https://doi.org/10.48550/arXiv.2305.10355

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., & Wang, L. (2023). Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*. https://doi.org/10.48550/arXiv.2306.14565

Liu, W., Wang, X., Wu, M., Li, T., Lv, C., Ling, Z., … Huang, X. (2023). Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*. https://doi.org/10.48550/arXiv.2312.15997

Makridakis, S., Petropoulos, F., & Kang, Y. (2023). Large language models: Their success and impact. *Forecasting*, *5*(3), 536–549. https://doi.org/10.3390/forecast5030030

Mayer, R., Davis, J., & Schoorman, F. (1995). An Integrative Model of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Novicevic, M. M., Clayton, R. W., & Williams, W. A. (2011). Barnard's model of decision making: A historical

predecessor of image theory. *Journal of Management History*, *17*(4), 420–434. https://doi.org/10.1108/17511341111164427

Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J., & Daneshjou, R. (2024). Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of internal medicine*, *177*(2), 210–220. https://doi.org/10.48550/arXiv.2309.00087

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744. https://doi.org/10.48550/arXiv.2203.02155

Hart, P. (1991). Irving L. Janis' victims of groupthink. *Political Psychology*, *12*(2), 247–278. https://doi.org/10.2307/3791464

Pearson, J. M., & Shim, J. P. (1995). An empirical investigation into DSS structures and environments. *Decision Support Systems*, *13*(2), 141–158. https://doi.org/10.1016/0167-9236(93)E0042-C

Penczynski, S. P. (2016). Persuasion: An experimental study of team decision making. *Journal of Economic Psychology*, *56*, 244–261. https://doi.org/10.1016/j.joep.2016.07.004

Piñeiro-Martín, A., García-Mateo, C., Docío-Fernández, L., & López-Pérez, M. C. (2023). Ethical challenges in the development of virtual assistants powered by large language models. *Electronics*, *12*(14), 3170. https://doi.org/10.3390/electronics12143170

Prokop, D. (2023). Von Neumann–Morgenstern utility function. *Encyclopedia Britannica*. https://www.britannica.com/topic/von-Neumann-Morgenstern-utility-function

Read, D. (2020). Experienced utility: utility theory from Jeremy Bentham to Daniel Kahneman. In *Judgement and choice: Perspectives on the work of Daniel Kahneman* (pp. 45–61). Psychology Press. https://doi.org/10.1080/13546780600872627

Schilirò, D. (2012). Bounded rationality and perfect rationality: Psychology into economics. *Theoretical and Practical Research in Economic Fields, III*(2), 99–108. https://doi.org/10.2478/v10261-012-0007-0

Schoemaker, P. J. H., & Russo, J. E. (2016). Decision-making. In M. Augier & D. J. Teece (Eds.), *The Palgrave encyclopedia of strategic management* (pp. 1–5). Palgrave MacMillan. https://doi.org/10.1057/978-1-349-94848-2_341-1

Seals, S. M., & Shalin, V. L. (2023). Evaluating the deductive competence of large language models. *arXiv preprint arXiv:2309.05452*. https://doi.org/10.48550/arXiv.2309.05452

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*. https://doi.org/10.48550/arXiv.2105.04054

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future decision support technologies. *Decision Support Systems*, *33*(2), 111–126. https://doi.org/10.1016/S0167-9236(01)00139-7

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Simon, H. A. (1972). Theories of bounded rationality. In C. B. McGuire & R. Radner (Eds.), *Decision and organization* (pp. 161–176). Elsevier.

Simon, H. A. (1979). Rational decision making in business organizations. *American Economic Review*, *69*(4), 493–513. https://www.jstor.org/stable/1808698

Simon, H. A. (1987). Making management decisions: The role of intuition and emotion. *Academy of Management Perspectives*, *1*(1), 57–64. https://doi.org/10.5465/ame.1987.4275905

Simon, H. A. (1996). The sciences of the artificial. *The MIT Press*. https://mitpress.mit.edu/9780262691918/the-sciences-of-the-artificial/

Simon, H. A. (1997). *Administrative behavior* (4th ed.). Free Press. (Original work published 1947).

Singh, A. K., Devkota, S., Lamichhane, B., Dhakal, U., & Dhakal, C. (2023). The confidence-competence gap in large language models: A cognitive study. *arXiv preprint arXiv:2309.16145*. https://doi.org/10.48550/arXiv.2309.16145

Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology. General*, *153*(4), 1066–1075. https://doi.org/10.1037/xge0001547

Takemoto, K. (2024). The moral machine experiment on large language models. *Royal Society Open Science*, *11*(2), 231393. https://doi.org/10.1098/rsos.231393

Teng, Y. A., Zhang, J., & Sun, T. (2023). RETRACTED: Data-driven decision-making model based on artificial intelligence in higher education system of colleges and universities. *Expert Systems*, *40*(4), e12820. https://doi.org/10.1111/exsy.12820

Thomas, P., Spielman, S., Craswell, N., & Mitra, B. (2024, July). Large language models can accurately predict searcher preferences. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval* (pp. 1930–1940). https://doi.org/10.48550/arXiv.2309.10621

Tu, X., Zou, J., Su, W. J., & Zhang, L. (2023). What should data science education do with large language models. *arXiv preprint arXiv:2307.02792*, *3*. https://doi.org/10.48550/arXiv.2307.02792

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Valmeekam, K., Marquez, M., Olmo, A., Sreedharan, S., & Kambhampati, S. (2023). Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, *36*, 38975–38987. https://doi.org/10.48550/arXiv.2206.10498

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://doi.org/10.48550/arXiv.1706.03762

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019).Designing theory-driven user-centric explainable AI. In *Proceedings of*

the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–15). ACM. https://doi.org/10.1145/3290605.3300831

Weirich, P. (1983). A decision maker's options. *Philosophical Studies*, *44*(2), 175–186. https://doi.org/10.1007/BF00354098

Whang, S. E., Roh, Y., Song, H., & Lee, J. G. (2023). Data collection and quality challenges in deep learning: A data-centric AI perspective. *VLDB Journal*, *32*(4), 791–813. https://doi.org/10.1007/s00778-022-00775-9

Yousuf, H., & Zainal, A. Y. (2020). Quantitative approach in enhancing decision making through big data as an advanced technology. *Advances in Science, Technology and Engineering Systems Journal*, *5*(5), 109–116. https://doi.org/10.25046/aj050515

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., … Shi, S. (2023). Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*. https://doi.org/10.48550/arXiv.2309.01219

Zhao, J., & Gómez Fariñas, B. (2023). Artificial intelligence and sustainable decisions. *European Business Organization Law Review*, *24*(1), 1–39. https://doi.org/10.1007/s40804-022-00262-2

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223, *1*(2). https://doi.org/10.48550/arXiv.2303.18223

Zhavoronkov, A. (2022). Vision 2030 and AI. Retrieved from Forbes. *The New Press*. https://www.forbes.com/sites/alexzhavoronkov/2022/07/14/the-new-saudi-arabiavision-2030-and-ai/?sh = 45a2fc481805

Zand, D. E. (1972). Trust and managerial problem solving. *Administrative Science Quarterly*, *17*, 229–239. https://doi.org/10.2307/2393957