JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

UNDERSTANDING TRUST IN AI: HUMAN PERCEPTIONS OF COLLABORATIVE DECISION-MAKING SYSTEMS

Mr. Tejas Vinod Agarwal[1] Miss. Vaishnavi Dattatray Patil[2]
Prof. Rajat Hedav[3] Prof. Prajali Patil[4]
12 Students, Dr. D. Y. Patil Centre for Management & Research, Chikhali, Pune, India
34 Assistant Professor,Department of Master of Computer Applications, Dr. D. Y. Patil Centre for Management & Research, Chikhali, Pune, India

***Abstract* :Artificial Intelligence, in the modern day, is an intrinsic part of contemporary decision support systems, revolutionizing sectors like healthcare, finance, and customer service with data-driven accuracy and automation. Not with standing their technical capabilities, the success of AIs in supporting collaborative decision-making also hinges on the level of human trust generated. This study, entitled "Understanding Trust in AI: Human Perceptions of Collaborative Decision-Making Systems," examines how humans perceive, establish, and sustain trust in AI systems working together with human decision-makers. The research adopts a mixed-methods design, where quantitative surveys along with statistical analysis are combined with qualitative interviews and field observation. It considers how transparency, interpretability, system reliability, and user control shape trust, along with demographic, cultural, and contextual influences. The study focuses on three broad domains-healthcare, finance, and customer service-that represent different degrees of decision-criticality and user interaction. It looks at how trust generation and calibration vary across these domains using comparative analysis. The results show that explanation modes-feature-based, counterfactual, and example-based-and uncertainty report formats have a significant effect on user trust and overreliance behaviors. Adaptive transparency-or AI explanations self-modulating with user knowledge and context of situation-was found to elicit more balanced trust levels by reducing overreliance and distrust. The study also points out that the repair functions of trust-open error reporting and ethical accountability protocols-rebuild user trust in case of system breakdowns. In all, the research provides an integrated framework for explaining and engineering trust in human-AI collaboration.**

***Keywords* - Artificial Intelligence (AI), Human–AI Collaboration, Trust in AI, Explainable AI (XAI), Transparency, Interpretability, System Reliability, User Control, Adaptive Transparency, Uncertainty Communication, Overreliance and Underreliance, Trust Calibration, Trust Repair, Ethical AI.**

I.INTRODUCTION

Artificial intelligence (AI) is more and more a part of our daily experience, revolutionizing many industries and operations. While its ability to analyse has been praised, moving from AI being a utility to a co-partner, particularly in high-stakes decision-making situations, raises new issues. This change poses ultimate questions related to trust and perceptions. As AI devices become more autonomous and the output less predictable, it is important to know how and why humans observe, interpret, and eventually trust these devices. This research investigates the determinants of trust in AI, especially in collaborative decision-making models. AI development in decision-making can be broadly divided into major development stages, each with its own unique technological innovations, application processes, and degrees of human-AI collaboration. The artificial intelligence (AI) influence on decision-making has increased significantly, impacting how decisions are made in sectors. From Carly deterministic systems to human-AI collaborative frameworks of today, Al's evolution is defined by significant technological developments, increasingly complex algorithms, and a changing dynamic between AI systems and human users. This is broken down into four significant stages:

- Rule-Based Systems (1950s-1980s)
- Machine Learning (1980s-2000)
- Deep Learning and Neural Networks (2000s-2010s)
- Human-AI Collaboration (2010s Present)

1. Rule-Based Systems (1950s-1980s): The Building Blocks of AI Decision-Making: The earliest decision-support systems (DSS) were rule-based, based on "if-then" rules expressed by human experts. These systems, like MYCIN in medical diagnosis, were narrow in scope and could only perform well-defined, narrow tasks. Rule-based systems could not learn; after they were programmed, their rules remained fixed, and they could not get better with new information (Shurtliff, 1976).

- **Strengths:** Logical clarity, deterministic outcomes, and ease of interpretation.
- **Weaknesses:** Rigidity, scalability problems, and inability to learn from new situations (Nilsson, 1998).

2. Machine Learning (1980s-2000s): The Advent of Adaptive Systems: Machine learning (ML) brought about adaptive decision-making, where systems could learn from experience. Decision trees, neural networks, and Bayesian models permitted probabilistic decision-making through detecting patterns and trends in historical data

- **Strengths:** Can adapt to learn, can manage uncertain inputs, can generalize from examples.
- **Weaknesses:** Needs large datasets, low interpretability for high-complexity models (Goodfellow et al., 2016).

3. Deep Learning and Neural Networks (2000s-2010s): High-Performance Decision-Making: Advances in deep learning, a type of ML, allowed AI systems to analyse large sets of data and independently extract high-level patterns. Using multi-layered neural networks, deep learning models were capable of very high accuracy in image and speech recognition, natural language processing, and disease diagnosis.

- **Strengths:** Precise, feature extraction automated, for complex, unstructured data.

4. Human-AI Collaboration (2010s-Present): Towards Transparent and Ethical Decision Systems: Now that AI has progressed, more stress is being laid on the development of collaborative systems in which AI helps humans make decisions.

- **Strengths:** Enhances human capabilities, enhanced interpretability, and ethical conformance.

II. Literature Review

- **Lee & See, 2004** Human and AI system collaboration is transforming decision-making in numerous sectors, including medicine, finance, and transportation. With AI systems becoming more embedded in decision-making positions, it is vital to understand human perception and trust of these systems to ensure that collaboration is effective. Trust is one of the most key elements of this equation, as it determines whether or not users embrace and trust the decisions or recommendations that AI offers. In the absence of trust, even the most sophisticated AI systems can be underutilized or misused, shortening their capacity to make a positive impact
- **Doshi-Velez & Kim, 2017** Al's trust is not a one-dimensional construct but a multi-dimensional one encompassing multiple important components, such as reliability, transparency, and explainability. For users to be able to trust Al, they need to comprehend the processes and rationale behind Al's decision-making functions. This is especially true for machine learning algorithms, which tend to act like "black boxes," making choices based on large datasets but not giving users explicit reasons they can understand
- **Amann et al., 2020** Transparency is one of the critical factors that contribute to trust. Users must be informed about how the AI arrives at a specific decision, particularly in serious domains like medicine or law services where mistakes can carry major consequences. For example, in medicine, AI systems used to aid in making diagnoses should not only provide precise results but also specify explanations upon which professionals are able to interpret and confirm
- **Siau & Wang, 2018** Trust may be dynamic, nonetheless: while it may grow as the AI system demonstrates its correctness in the long run, it may also collapse quickly if the system crashes unexpectedly or makes decisions that are deemed wrong or unjustified by users
- **Dzindolet et al., 2003** Besides, AI trust is not fixed and can change depending on interaction over time. As people use AI systems, they learn about the performance of these systems in different contexts. Positive interactions lead to building trust, but negative interactions, particularly those without proper explanations, can erode it. This phenomenon is referred to as trust calibration, whereby users tend to adjust their trust levels according to its real performance versus expectations

I. RESEARCH METHODOLOGY

The current research, "Understanding Trust in AI: Human Perceptions of Collaborative Decision-Making Systems," utilizes a mixed-methods design integrating quantitative and qualitative methods for exhaustive exploration of the determinants of user trust in AI. The study takes a sequential explanatory strategy with quantitative surveys preceded by qualitative interviews to obtain convergent findings. Information was gathered from the respondents over three significant fields—healthcare, finance, and customer service—chosen based on their differing criticalities of decision-making and levels of AI implementation. A stratified random sampling technique was employed to capture urban and semi-urban areas, presenting a sample size of around 370 people. Primary data were collected using guided questionnaires derived from the Trust in Automation Scale (Jian et al., 2000) and Technology Acceptance Model (Davis, 1989), whereas secondary data were collected from research papers, industry reports, and AI case studies. Statistical packages like SPSS and Excel were applied for descriptive analysis, ANOVA, Chi-square, and regression tests to find out important relationships between variables like transparency, reliability, and frequency of user interaction. Semi-structured interview qualitative data were analysed thematically to reflect insights of a more profound nature into user beliefs, ethicality, and trust calibration processes. Cronbach's Alpha (>0.8) was used to ensure reliability, while expert review and pilot testing confirmed content validity. Ethicality was ensured by informed consent, anonymity, and compliance with the research ethics policy of the university. This interdisciplinary approach guarantees a comprehensive understanding of the cognitive, emotional, and contextual facets of trust in human–AI interaction and offers a solid framework for the creation of transparent, explainable, and ethically sound AI systems.

RESEARCH DESIGN

The present research on Human-AI Collaboration in Decision Making and User Perceptions and Trust in AI Systems has utilized a mixed-methods research design. integrating quantitative and qualitative approaches. The quantitative aspect has used surveys and experiments to gather quantifiable data regarding trust behaviors, and on the other hand the qualitative aspect have used particular computational model to examine the subtleties of user perceptions and experiences. This research design aim to give an in-depth look at how trust is established, sustained, and restored in human-AI collaboration in various areas (Teddlic & Tashakkori, 2009).

This structure lends itself to an empirical method of quantifying variables and establishing quantitative correlations, and qualitative investigation into determining deeper reasons, meanings, and intricate phenomena surrounding trust and perceptions in AI systems (Creswell & Creswell, 2018). The synthesis of the methods seeks to strike a balance between depth and generalizability and foster an overall understanding of human-AI collaboration in decision-making scenarios. The research have also investigated various contexts to study context-dependent trust mechanisms and the effects of varying decision-making environments on user attitudes (Plano Clark & Ivankova, 2016)

SELECTION OF DECISION-MAKING CONTEXTS

The choice of healthcare, finance, and customer service as primary decision-making contexts for exploring Human-AI Collaboration and user trust is intentional, exposing the varied levels of risk, decision criticality, and types of user engagement within each discipline. Each of these domains offers distinct challenges and their respective impact on trust, influenced by the unique requirements, ethical aspects, and user expectations that define each. Succinct overview of decision-making domains is depicted hereunder:

1. Healthcare: Healthcare decisions have very serious, life-critical options where accuracy, reliability, and trust are highly important. AI technologies are being used more and more for purposes such as diagnostics assistance, predictive modeling, and treatment proposals. Since, healthcare decisions have a direct bearing on patient health dimensions, trust in AI systems in this field is highly reliant on transparency, precision, and the capacity to convey limitations and maintain ethical norms. Any, misstep will create severe health consequences, which make healthcare a worthwhile field for examining how system openness, regulatory compliance, and user trust correlate in consequential, high-risk choices.

2. Finance: Financial choice has intricate risk analyses, market forecasts, and investment approaches with possible substantial financial repercussions. AI systems within the finance sector are applied extensively for fraud detection, loan granting, and investment guidance, where precision and regulatory compliance are critical. Stakes are high, as any slight inaccuracies could have significant consequences or regulatory issues. Trust in financial AI systems is influenced by data protection, compliance, and unambiguous accountability framework. This environment offers a potential to examine how trust relies on users' belief in system's reliability, justice, and adherence to industry norms, especially in high-stakes situations among various users, ranging from individual investors to big companies.

3. Customer Service: Customer service offers a contrast, with reduced-risk interactions but frequent and high-volume encounters. Al-driven chatbots and automated customer support systems are often managing inquiries, problem-solving, and providing personalized suggestions. In these, trust is established based on responsiveness, user privacy, and system's domain understanding ability, with an expectation of resolving complex problems smoothly being escalated to human agents if necessary.

Sector	Hypothesized Population (N)	Corrected Sample Size (n)	Urban (70%)	Semi-Urban (30%)
Healthcare	10,000	371	260	111
Finance	8,000	368	258	110
Customer Service	12,000	373	261	112

Table 1 : Hypothesized Population

Desired Confidence Level and Margin of Error:

- **Confidence Level:** 95% (Z-score of 1.96).
- **Margin of Error:** 5% (0.05).

For the initial determination of sample size, Cochran's formula was applied:

$$n_0 = \frac{z^2 . p . (1-p)}{e^2}$$

With the p=0.05, for maximum visibility ensuring the initial sample size would be:

$$n_0 = \frac{(1.96)^2 . 0.05 . (1-0.05)}{(0.05)^2} = \frac{(3.8416 * 0.25)}{0.0025} = 384.16 \cong 385$$

Accordingly, the initial sample size to provide representativeness is 385 persons. Because the populations in each sector are limited, Cochran's finite population correction was used for the purpose of correction to refine the sample size for each sector.

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}}$$

For Healthcare (N = 10,000): $n_{healthcare} = \frac{385}{1+\frac{385-1}{10000}} = \frac{385}{1+0.0384} = 371.05 \cong 371$

For Finance (N = 8,000): $n_{finance} = \frac{385}{1+\frac{385-1}{8000}} = \frac{385}{1+0.048} = 367.89 \cong 368$

For Customer Service (N=12,000): $n_{customerservice} = \frac{385}{1+\frac{385-1}{12000}} = \frac{385}{1+0.0319} =$

$373.29 \cong 373$

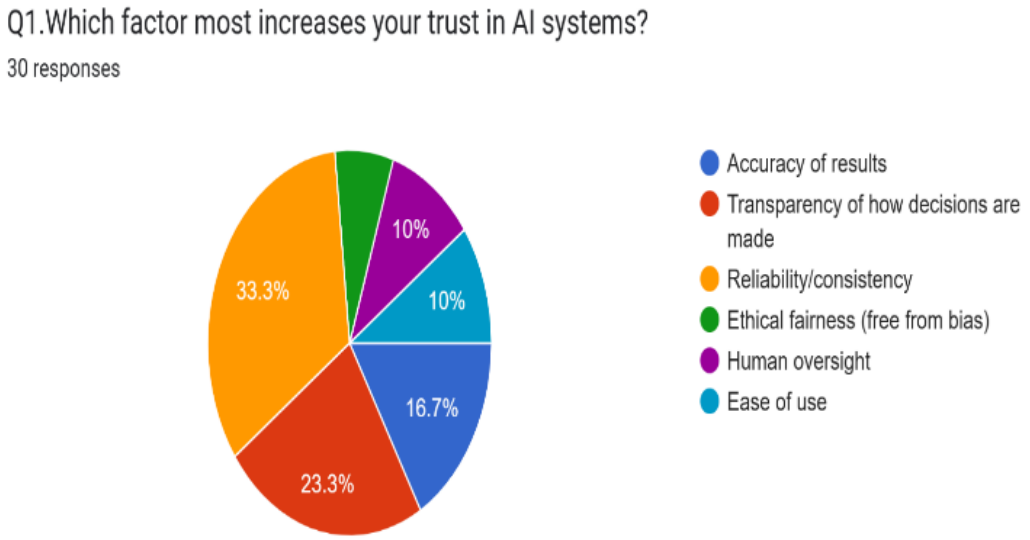
Additionally, these samples within cach sector were apportioned according to the urban (70%) and semi-urban (30%) proportions.

For Healthcare: Urban (70%): 0.7 × 371 = 259.7 ≈ 2t and Semi-Urban (30%): 0.3 × 371 = 111.3 ≈ 111
For Finance: Urban (70%): 1.7 × 368 = 257.6 = 258 and Semi-Urban (30%); 0.3 × 368 = 110.4 ≈ 110
For Customer Service: Urban (70%): 0.7 × 373 = 261 1 ≈ 261 and Semi-Urban (30%): 0.3×373-111.9-112.

DATA ANALYSIS AND INTERPRETATION

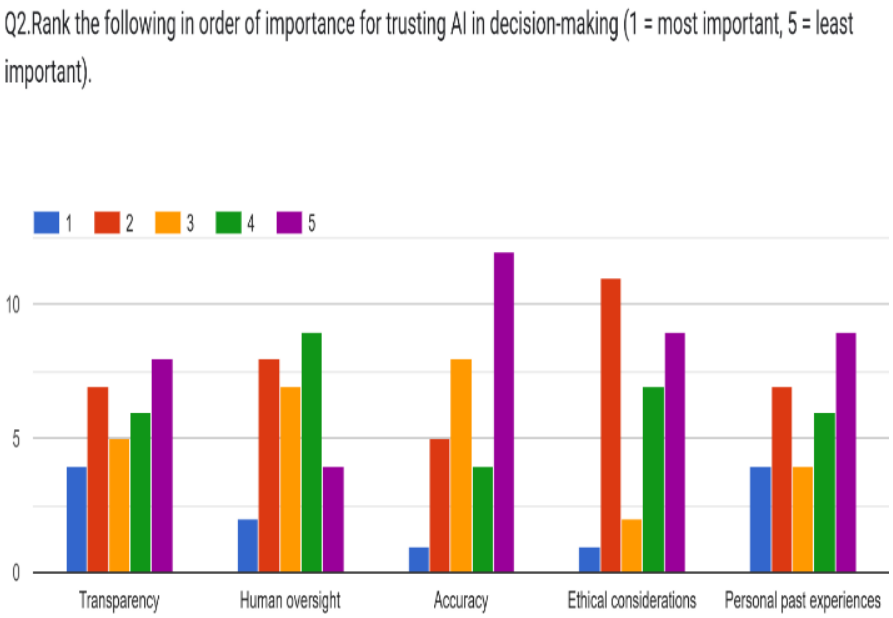
The primary data for the research were gathered by a comprehensive questionnaire framed to address essential areas that affect trust in AI systems. The questionnaire starts with Demographics and Context-Specific Factors, collecting baseline data like age, gender, education level, and exposure to AI, along with particular contexts of AI use (e.g., health, finance, customer support). This segment presents a baseline level of understanding of respondents' backgrounds and usage behaviors

The second part, Trust-Building and Repeated Interactions, describes how users' trust grows as a result of repeated interactions with AI systems and seeks to discover how frequency influences confidence in and satisfaction with Al-driven decision-making. Models of Human-AI Interaction comes next, discussing users' interaction model preferences for different models like fully automated, assistive, or semi-automated models and how these influence trust and comfort in decision processes. To capture differences informed by social and cultural backgrounds, the Cultural, Social, and Demographic Differences in Trust-Building section explores how cultural norms, age, gender, and educational background might influence users' views and comfort with AI systems.



Graph 1 : Factor in AI System

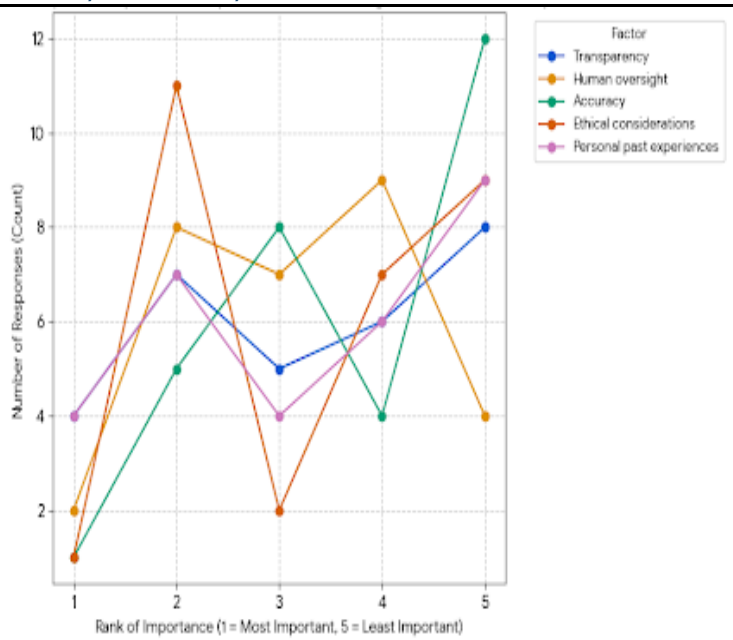
Reliability/consistency is the strongest contributor to trust: 33.3%. This suggests that people value an AI system that acts predictably and gives stable performance.The second most important factor is transparency in how decisions are made, at 23.3%. People want to understand why AI makes certain decisions.The accuracy of results accounts for 16.7%, showing that performance still matters but is not the dominant factor.Ethical fairness, ease of use, and human oversight received 10% each, which is perceived as valued but less influential in comparison to reliability and transparency.



Graph 2 : Importance of Trusting AI

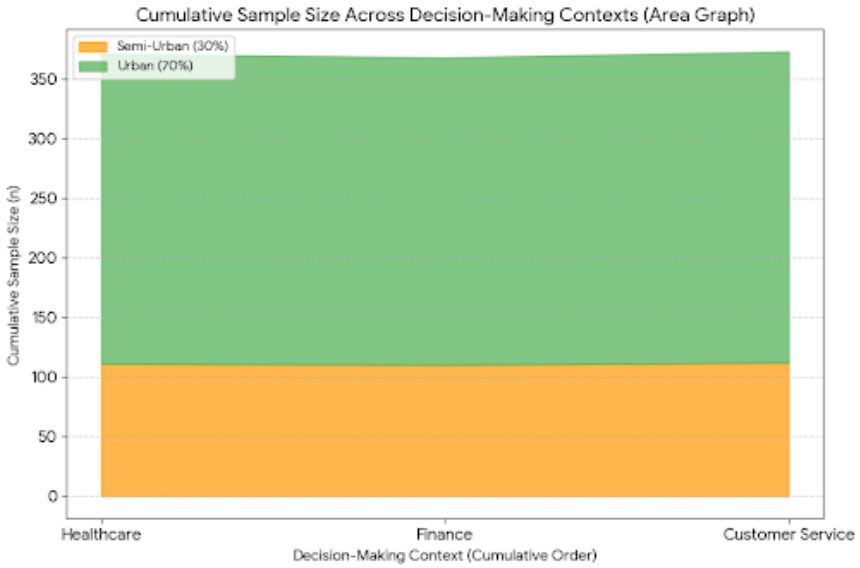
Ethical considerations received a high number of "1 – most important" rankings, meaning many respondents see ethics as a primary requirement for trust.Human oversight and transparency show mixed importance: they have responses across all ranks, indicating differing user beliefs.Accuracy has relatively few "1" rankings and many mid-level rankings; in particular, 3 and 4, suggesting it is important but not the top priority for most people.Personal past experiences received more "4"s and "5"s, meaning it is considered less important to trust AI compared to other factors.

Statistical packages like SPSS and Excel were applied for descriptive analysis, ANOVA, Chi-square, and regression tests to find out important relationships between variables like transparency, reliability, and frequency of user interaction. **Semi-structured interview qualitative data** were analysed thematically. Reliability was ensured using **Cronbach's Alpha (>0.8)**, and content validity was confirmed by expert review and pilot testing.



Graph 3 : Tradeline Graph for Field of importance in AI

Ethical considerations peaks at Rank 1 sharply, reinforcing that ethics is viewed as the most important element for trust. Human oversight peaks at Rank 4, indicating that many perceive it as important but not the single most critical element. Transparency has moderate counts across the ranks, ranging from a cluster between Ranks 3–5, showing varied opinions on its importance.Accuracy is spread out relatively evenly, showing interest but not a dominant level of importance.The personal past experiences trend toward higher rank numbers, that is, less important, which echoes the findings in the bar chart.



Graph 4 : Area Graph for Different Sector

The overall sample is equal for contexts, and urban respondents compose the majority, as reflected by the large green area. Semi-Urban respondents consistently represent about 30% of the sample, shown by the orange area. The stacked format reveals that sample distribution is uniform across Healthcare, Finance, and Customer Service — the proportion of urban versus semi-urban participants does not change.

Results and Discussion

The following section provides the expected results of the envisaged research and elaborates on their significance for comprehension and engineering trust in collaborative decision-making systems based on AI. Although empirical verification through the envisaged methodology will be pursued, the subsequent section details the envisaged outcomes based on theoretical underpinnings, existing literature, and logical estimates.

Expected Results of Controlled Experiments

1.1 Explanation Styles:
Explanation based on features should be more effective for domain experts who already have the required background knowledge in order to understand technical hints. Experts should report greater calibrated trust and more accurate judgments when explanations specify which features guide AI decisions.
Counterfactual explanations can be particularly useful for non-technical users, who appreciate contrastive reasoning (e.g., "If variable X were other, the choice would have been different"). These explanations assist non-experts in placing AI reasoning in context.
Example-based explanations are expected to fill the gap by providing intuitive landmarks. Experts and non-technical users might both use these as helpful tools to construct mental models of AI decision-making processes.

1.2 Uncertainty Communication
Numeric uncertainty (probability, confidence intervals) will most probably increase numerate users and domain experts' trust calibration, leading to better error detection and lower overconfidence.
Qualitative uncertainty descriptors (e.g., "high confidence" or "low confidence") are likely to be more usable for lay users, avoiding cognitive overload but possibly at the price of vagueness. Trust calibration will be enhanced when such descriptors are standardized and regular.

Anticipated Outcomes from Longitudinal Field Study

2.1 Trust Trajectories:
Trust will rise consistently in the early exposure phase, levelling off after multiple accurate interactions.
Recurring system mistakes will result in temporary dips in trust, but recuperation will be contingent upon the existence of transparent explanations and accounting systems.

2.2 Organizational Context:
Within high-stakes settings (e.g., medicine), organizational tradition and peer monitoring will have a strong bearing on trust. Clinicians, for example, may double-check AI suggestions prior to acceptance, influencing mutual dependence.
Mechanisms for accountability, e.g., audit logs and decision review committees, should minimize overdependence while maintaining trust in AI support.

Conclusion

The following section provides the expected results of the envisaged research and elaborates on their significance for comprehension and engineering trust in collaborative decision-making systems based on AI. Although empirical verification through the envisaged methodology will be pursued, the subsequent section details the envisaged outcomes based on theoretical underpinnings, existing literature, and logical estimates.

1. Expected Results of Controlled Experiments

Explanation based on features should be more effective for domain experts who already have the required background knowledge in order to understand technical hints. Experts should report greater calibrated trust and more accurate judgments when explanations specify which features guide AI decisions.
Counterfactual explanations can be particularly useful for non-technical users, who appreciate contrastive reasoning (e.g., "If variable X were other, the choice would have been different"). These explanations assist non-experts in placing AI reasoning in context.
Example-based explanations are expected to fill the gap by providing intuitive landmarks. Experts and non-technical users might both use these as helpful tools to construct mental models of AI decision-making processes.

Uncertainty Communication
Numeric uncertainty (probability, confidence intervals) will most probably increase numerate users and domain experts' trust calibration, leading to better error detection and lower overconfidence.
Qualitative uncertainty descriptors (e.g., "high confidence" or "low confidence") are likely to be more usable for lay users, avoiding cognitive overload but possibly at the price of vagueness. Trust calibration will be enhanced when such descriptors are standardized and regular.

Levels of Autonomy
Recommendation-only systems should promote active human involvement, increased monitoring, and less automation bias but with the penalty of higher cognitive workload.
Action-automation systems should enhance efficiency and task performance speed but potentially pose risks of overdependence, particularly when operating under time pressure. Users can submit to automation even when mistakes are apparent.

Future Scope
Research on trust in AI-based collaborative decision-making systems creates a number of promising directions for continued future work. As AI technologies continue to advance and become intertwined in various human-centered applications, continued research will be needed to respond to new challenges, broaden existing findings, and streamline design practice. The following are suggested directions for future work:

1. Multi-Agent Human–AI Collaboration

Future studies need to examine environments where humans interact with multiple AI agents at once, each having different capabilities or expertise levels. This will address how users address trust distribution across many AI systems and negotiate conflicting recommendations.

2. Cross-Cultural and Contextual Studies

Perceptions of trust are heavily shaped by cultural values, norms, and societal expectations. Cross-cultural and cross-organizational comparative studies can uncover variation in trust establishment around the world and guide culturally sensitive AI design strategies.

3. Long-term Trust Dynamics

Although this research has a longitudinal aspect, trust should be studied over the long term (e.g., years) in subsequent studies to understand long-term adoption, adaptation, and trust repair processes following persistent system crashes or updates.

4. Mechanisms of Trust Repair

Disasters are bound to happen in AI systems. Subsequent studies should analyze how communication plans (e.g., apology, error admission, corrective patches) and system redesigns impact the mechanism of trust repair once broken.

- **References**
- 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

- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human–AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1), 2–11.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *aria preprint arXiv:1702.08608*.

- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660.
- 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
- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors, 48*(2), 241–256.

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

- 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

- 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

- Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors, 48*(2), 241–256. https://doi.org/10.1518/001872006777724408

- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction, 36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

- Vincent, J. (2021). Why we should be wary of AI that pretends to be human. *The Verge*. https://www.theverge.com/2021/3/2/22308088/ai-human-trust-deception-risks

- Zhou, Y., & Chen, L. (2021). A survey on trust in human–AI interaction. *ACM Computing Surveys, 54*(6), 1–36. https://doi.org/10.1145/3452233