Confidence-Based Trust Calibration in Human-AI Teams

Michael Ibrahim
Computer Engineering Department-Faculty of Engineering, Cairo University, Giza, Egypt 12613

***Abstract*—Effective human-AI collaboration is contingent upon calibrated trust, wherein users depend on AI systems when accuracy is probable and rely on human judgment when errors are likely. In this study, a confidence-based mechanism for trust calibration within human-AI teams is examined. A decision-making strategy is proposed in which task delegation is governed by the AI's confidence: when the confidence surpasses a specified threshold, the AI's recommendation is adopted; otherwise, the decision is deferred to the human. Through simulation experiments on a binary classification task, performance outcomes are compared. The AI system achieves an accuracy of 77.7%, whereas the human decision-maker, modeled with a confidence-sensitive accuracy function $p_h(c) = 0.95 - 0.3c$, attains an overall accuracy of 71.9%. Team performance is evaluated across a range of AI confidence thresholds (0.50 to 0.99), revealing that an intermediate threshold yields optimal team accuracy of 84.14%, substantially exceeding the performance of either agent individually. The findings provide a detailed analysis of confidence-based delegation, align with existing research on trust calibration, and underscore critical design implications for the development of human-centric AI systems.**

***Keywords*—Human-AI collaboration; trust calibration; confidence-based delegation; decision-making strategies**

I. Introduction

Artificial Intelligence has been increasingly deployed to support human decision-making across domains such as healthcare, finance, and public administration. From the assistance of clinicians in diagnostics to the facilitation of faster financial fraud detection, AI systems are now influencing decisions that carry significant consequences for individuals and society [1], [2], [3].

In many of these applications, AI systems are designed to act as decision aids, producing predictions or recommendations while the final judgments are left to human operators. The effectiveness of such human-AI teams is determined not only by the technical accuracy of the AI but also by the quality of human-AI collaboration, particularly the calibration of human trust in the system [4], [5]. Misaligned trust, manifested as either over-reliance on flawed outputs or unwarranted skepticism toward reliable ones, has been shown to significantly impair decision quality [6].

Trust calibration is defined as the alignment between a user's trust in an AI system and the system's actual performance or reliability. When trust is properly calibrated, users tend to rely on the AI when it is likely to be correct and revert to their own judgment when the AI is likely to err [7]. This balance is particularly critical in high-stakes environments, where both AI errors and human biases may lead to severe consequences [8].

To foster calibrated trust, many AI systems have been equipped with confidence scores or uncertainty estimates that accompany their predictions. These indicators are intended to help users discern when AI recommendations should be accepted or questioned [4]. However, the provision of confidence information alone has not been found sufficient. Empirical studies indicate that users' own confidence levels and cognitive biases play a significant role in determining whether reliance on AI systems is appropriately adjusted [5], [7]. For instance, overconfident users may disregard valuable AI input, whereas underconfident users may accept flawed suggestions without critical evaluation.

Consequently, trust calibration emerges as a dual challenge, requiring both transparent AI systems and mechanisms that account for human self-assessment and behavioral tendencies [6]. In response to this challenge, a confidence-based delegation strategy is examined, wherein AI predictions are adopted only when their confidence surpasses a predefined threshold; otherwise, the decision is deferred to the human operator.

This strategy is intended to align decision authority with the most reliable agent in each situation. By tuning the threshold, the balance between human and AI decision-making can be dynamically adjusted, ranging from heavy reliance on AI to near-complete human control. The proposed approach is evaluated within a simulated binary classification task, in which an AI system with fixed accuracy and a human agent with confidence-sensitive performance are modeled. The results indicate that an intermediate confidence threshold yields the highest overall team performance, surpassing that of either the human or AI operating independently.

These findings provide new insights into trust calibration strategies for human-AI collaboration and contribute to the design of systems that optimize joint decision-making outcomes in practical applications.

II. Related Work

A. Trust Calibration and Miscalibration in Human-AI Interaction

Trust calibration lies at the core of effective human-AI interaction, serving as the process by which human users adjust their trust in an AI system to align with the system's actual competence and reliability. Properly calibrated trust allows users to leverage AI strengths while maintaining critical oversight, ensuring that reliance on automation remains proportional to its demonstrated performance. Conversely, miscalibrated trust, manifesting as either overtrust or undertrust, can undermine system efficacy, situational awareness, and

ultimately safety. This is particularly consequential in high-stakes domains such as healthcare, autonomous vehicles, and maritime navigation, where erroneous trust judgments can have severe operational and ethical implications.

Kneer, Loi, and Christen [9] explore these dynamics in the context of human–AI interaction, emphasizing how automation bias and algorithm aversion emerge when users misperceive AI capabilities or misinterpret its outputs. Such errors may arise from opacity, inconsistent performance, or limited user understanding. As a result, trust calibration emerges as both a technical and psychological challenge.

Extending this perspective, Kulal [10] argue that even when AI systems provide statistically calibrated outputs, users may misalign their trust due to cognitive biases, such as anchoring on superficial cues or overgeneralizing prior experiences. This highlights that achieving trust calibration is not solely a matter of improving algorithmic confidence, it also requires user-centered design informed by cognitive science and HCI. Thus, future systems must integrate these insights to foster more resilient, interpretable, and human-aligned decision-making partnerships.

B. Confidence and Uncertainty Displays

A prominent approach to improving trust calibration in human-AI interaction involves communicating the AI system's confidence or uncertainty levels. These displays aim to make the system's internal reasoning and limitations more transparent, allowing users to adjust their reliance on AI outputs according to perceived reliability. When well-designed, such indicators can foster calibrated trust, reduce automation bias, and help users recognize when human oversight is necessary. However, the effectiveness of confidence and uncertainty displays depends critically on users' ability to correctly interpret and integrate these signals into their decision-making processes.

Prior work has shown that presenting confidence alone is insufficient if users lack the contextual or cognitive tools to make sense of probabilistic information. For instance, Zhang and Xu [11] demonstrate in the maritime domain that multimodal indicators, such as uncertainty bars, linguistic cues, and system status indicators, are more effective than confidence scores alone in promoting trust calibration and preventing overreliance or disuse. These findings highlight the importance of designing interfaces that help users develop accurate mental models of AI decision boundaries.

Similarly, byasa and Rahmania [12] emphasize the value of explanation-assisted confidence displays in medical imaging. Their approach integrates visual saliency explanations with model confidence in 3D brain tumor segmentation tasks. This combination improved clinician understanding of when to trust or question AI outputs, suggesting that confidence information is most effective when paired with interpretable explanations.

Taken together, these insights emphasize that uncertainty communication is as much a matter of human-centered design as it is of algorithmic transparency. Effective confidence displays must balance informativeness and interpretability, tailoring the presentation of uncertainty to users' cognitive and contextual constraints. As a result, ongoing work in this area increasingly integrates research from explainable AI, human–computer interaction, and cognitive psychology to develop uncertainty representations that genuinely support trust calibration.

C. Confidence Alignment in Human–AI Collaboration

Effective trust calibration in human-AI collaboration requires a bidirectional approach: AI systems must communicate confidence that accurately reflects their underlying uncertainty, while human users must be supported in correctly interpreting and responding to these cues. Achieving alignment between human and AI confidence is not merely a technical issue of probabilistic calibration, but also a human-centered challenge involving perception, communication, and adaptive interface design. When either side of this alignment fails, when the AI misrepresents its certainty or when humans misread it, the partnership risks degradation through underutilization, overreliance, or unsafe decision-making.

In dynamic settings such as maritime navigation, Zhang and Xu [11] emphasize the importance of adaptive user interfaces that integrate explainable AI (XAI) components. They find that calibrated confidence and compliance indicators support user understanding of automation, fostering more consistent trust judgments under uncertainty. Their work highlights how interface-level interventions can prevent automation bias and encourage proper human oversight.

Reinforcing this view, von Reeken, Salous, and Abdenebaoui [13] demonstrate that subtle interface variations in conversational AI systems can significantly affect how users align their trust with system recommendations. Their study shows that users benefit from interface cues that signal confidence calibration quality, especially in retrieval-augmented generation systems. When these cues are absent or misaligned, trust can diverge from actual system reliability, leading to overtrust or disuse.

Finally, Marusich et al. [14] underscore the necessity of designing for trust calibration in time-sensitive, high-stakes applications. They advocate for adaptive confidence visualizations and user training strategies that evolve with changing task difficulty and user experience. Their findings suggest that integrating behavioral modeling into the interface design, such as adjusting explanation granularity based on user trust metrics, can prevent cascading decision errors and improve joint performance.

Taken together, these studies affirm that trust calibration is best achieved through a combination of calibrated AI outputs and adaptive, user-sensitive interface design. Future systems should be designed not just to communicate confidence but to help users meaningfully act on it.

D. Confidence-Based Delegation Policies

Confidence-based delegation represents a structured mechanism for managing trust and control within human-AI collaboration. In such frameworks, decision authority dynamically shifts between the human and the AI system according to confidence thresholds, when the AI's confidence surpasses a predefined level, the system autonomously executes a decision; when confidence falls below that level, control is transferred to

the human operator. This approach formalizes trust calibration by operationalizing it into quantifiable and adaptive decision policies, thereby aligning task allocation with system reliability and uncertainty.

This paradigm offers several benefits. It allows for efficient division of labor, ensuring that AI systems handle routine or high-certainty cases while humans intervene in ambiguous or high-risk situations. By doing so, it mitigates both over-reliance and under-reliance on automation, promoting a balanced collaborative dynamic. However, the success of confidence-based delegation hinges on how well the confidence thresholds are defined, validated, and communicated. Misaligned thresholds can either overwhelm human operators with unnecessary interventions or, conversely, lead to excessive automation that undermines human oversight and accountability.

Recent work by Le Denmat, Desender, and Verguts [15] underscores the importance of aligning system confidence with real-time human decision dynamics. Their study shows that agents can learn optimal confidence thresholds through experience, adjusting delegation based on trial-level feedback about decision accuracy. This learning-based approach allows human-AI systems to optimize collaboration patterns dynamically and avoid rigid reliance on static thresholds. This work builds on this by examining how performance varies across different threshold settings and demonstrating that fine-tuned delegation improves joint outcomes significantly. Such insights are crucial for deploying human-AI systems in dynamic, high-stakes environments where uncertainty varies widely across tasks.

III. Methodology

A trust calibration strategy was investigated for a human-AI team performing repeated binary classification tasks, such as diagnosing medical images or identifying fraudulent financial transactions. The team consisted of two agents: an AI model and a simulated human decision-maker. Both agents produced a class prediction and an associated confidence score for each decision instance. A simple confidence-based delegation policy was introduced, wherein the AI's confidence determined which agent's decision was adopted.

A. Confidence-Based Delegation Policy

Formally, the delegation policy was defined by a threshold parameter $\tau \in [0.5, 1.0]$. For each instance, the AI produced a class prediction and a confidence score $C_{\text{AI}} \in [0.5, 1.0]$. The confidence was restricted to be at least 0.5 since probabilities below this value would imply a different predicted class. The policy operated as follows:

- If $C_{\text{AI}} \geq \tau$, the AI's prediction was selected.
- If $C_{\text{AI}} < \tau$, the human's prediction was selected.

This threshold-based policy enabled dynamic delegation: the AI was trusted when sufficiently confident, while decisions were deferred to the human otherwise. Lower thresholds led to increased AI usage, whereas higher thresholds resulted in greater reliance on the human agent. It was hypothesized that an intermediate threshold could yield optimal overall accuracy by balancing the complementary strengths of both agents.

B. Simulation Framework

The decision-making process of the human-AI team was simulated to evaluate the proposed strategy.

a) AI Model: The AI was modeled as a binary classifier with fixed accuracy and calibrated confidence scores. Specifically, an AI accuracy of 77.7% across decision instances was simulated. The AI's confidence C_{AI} was sampled uniformly from the interval [0.5, 1.0]. The correctness of each prediction was determined probabilistically to ensure the specified overall accuracy.

b) Human Model: The human agent was modeled with imperfect confidence calibration. For each decision, a confidence value $c \in [0, 1]$ was reported. The probability of correctness given the confidence was defined as:

$$p_h(c) = 0.95 - 0.3c$$

This formulation captured typical miscalibration patterns, such as overconfidence at high confidence levels and underconfidence at low ones. For instance, when $c = 1.0$, the actual accuracy was 65%; when $c = 0.2$, accuracy increased to 89%. Confidence values were drawn from a distribution yielding an overall accuracy of 71.9%.

c) Independence Assumption: For analytical tractability, statistical independence between the AI and human confidences and correctness outcomes was assumed. Although correlated errors may occur in practice, this assumption allowed isolation of the effects of the delegation policy.

C. Decision and Evaluation Procedure

For each simulated decision instance, the following procedure was executed:

1) An AI confidence score C_{AI} was sampled, and correctness of the AI prediction was determined.
2) A human confidence score c was sampled, and correctness was determined based on $p_h(c)$.
3) The threshold policy was applied to select either the AI or human prediction.
4) The correctness of the chosen prediction was recorded.

Over many instances, two key metrics were computed:

- Team Accuracy: The proportion of correct final decisions.
- AI Usage Rate: The proportion of instances where $C_{\text{AI}} \geq \tau$.

By varying τ from 0.50 to 0.99, the effects on performance were examined. At the extremes, $\tau = 0.50$ corresponded to complete reliance on the AI (team accuracy = 77.7%), while $\tau = 0.99$ corresponded to near-exclusive human decision-making (team accuracy = 71.9%). Optimal team performance was expected at intermediate thresholds, where complementarities between AI and human decision-making were maximized.

D. Experimental Setup

Simulations were conducted at thresholds $\tau \in \{0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.99\}$, with tens of thousands of instances sampled per setting to ensure statistical robustness. For each threshold:

- AI confidence C_{AI} was sampled uniformly from [0.5, 1.0].
- AI correctness was probabilistically determined to ensure a 77.7% overall accuracy.
- Human confidence c was sampled to yield a 71.9% accuracy based on $p_h(c)$.
- The delegation rule was applied, and the correctness of the selected agent's decision was recorded.

This simulation framework enabled systematic evaluation of how the delegation threshold influenced both decision accuracy and the division of labor between the human and AI agents.

IV. Results

A. Team Performance vs. Threshold

Table I summarizes team accuracy and AI usage rates across different confidence threshold values. As the threshold τ increased from 0.50 to 0.99, team accuracy followed an inverted U-shaped curve. At $\tau = 0.50$, where the AI always makes the decision, team accuracy matched the AI's standalone performance of 77.7%. Increasing the threshold to $\tau = 0.60$ boosted team accuracy to 82.0%, and a further increase to $\tau = 0.70$ yielded the highest observed accuracy of **84.1%**.

At this optimal threshold, the AI handled approximately 75% of the instances, while the human handled the remaining 25%. Beyond $\tau = 0.70$, accuracy began to decline: at $\tau = 0.80$, it dropped to 83.0%; at $\tau = 0.90$, to 79.0%; and by $\tau = 0.99$, to 72.0%, only marginally above the human-alone baseline of 71.9%. AI usage rates decreased accordingly: from 100% at $\tau = 0.50$ down to approximately 3% at $\tau = 0.99$.

Table I demonstrate a clear "sweet spot" at $\tau = 0.70$, where team accuracy is maximized. The gain is substantial: a 6.4 percentage point improvement over the AI alone (77.7%) and a 12.2 point gain over the human alone (71.9%). These results validate the hypothesis that confidence-based delegation can yield performance gains by exploiting complementary strengths.

B. AI Usage vs. Threshold

As shown in the AI usage column of Table I, increasing the confidence threshold gradually reduces the frequency with which the AI's prediction is used. From full usage at $\tau = 0.50$, AI involvement drops to 75% at $\tau = 0.70$, then falls steeply to 50% at $\tau = 0.80$, 25% at $\tau = 0.90$, and near-zero at $\tau = 0.99$. This shift implies increasing reliance on the human as the threshold rises, and team performance decreases accordingly when the human is responsible for too many decisions.

Interestingly, the best accuracy occurs when the AI is used about 75% of the time, aligning with its higher individual accuracy compared to the human. In scenarios where the human were more accurate than the AI, we would expect the optimal threshold to shift accordingly.

C. Interpretation of Results

The results demonstrate that confidence-based delegation is a viable and effective strategy for trust calibration in human-AI teams. The optimal threshold reflects a division of labor that exploits each agent's strengths. At $\tau = 0.70$, the AI handles the decisions where it is relatively confident, while the human steps in where the AI is uncertain. This allocation improves the team's collective performance.

To understand why this threshold works, consider that when the AI is less than 70% confident, it is often wrong on those predictions. Our human model, while less accurate overall, is calibrated to sometimes outperform the AI on low-confidence cases. Conversely, when the AI is highly confident (e.g., $C_{\text{AI}} \geq 0.90$), it tends to be right, and deferring to it yields better results than relying on the human.

Thus, the delegation policy implicitly partitions the decision space: confident, likely-correct decisions go to the AI, and ambiguous, error-prone cases go to the human. Provided the AI's confidence is reasonably well-calibrated, this partitioning enhances team accuracy and supports appropriate reliance.

In summary, this experiment demonstrates that appropriately calibrated confidence thresholds can yield significant performance gains in human–AI collaboration, validating the utility of confidence as a trust calibration mechanism.

V. Discussion

The effectiveness of confidence-based delegation as a mechanism for trust calibration in human-AI collaboration was demonstrated. By using the AI's own confidence scores to determine reliance, decision-making was dynamically allocated to the most appropriate agent, achieving outcomes superior to either agent alone. The implications, necessary conditions, and design considerations for applying this strategy in practice are discussed below.

A. The Value of Calibrated Delegation

The threshold-based policy examined captures a key principle of appropriate reliance: deference to the AI when high confidence is expressed, and reliance on human judgment when uncertainty is greater. This simple policy significantly improved team performance, mitigating both over-reliance on incorrect AI outputs and disuse of valuable AI recommendations. At the optimal threshold $\tau = 0.70$, many of the AI's confident and unconfident mistakes were avoided through intelligent decision allocation.

This outcome aligns with theoretical work suggesting that automation should be applied when AI confidence is high and human input should be sought otherwise. The observed performance gain, from 77.7% to 84.1%, represents a 29% reduction in error. In real-world domains such as medical diagnostics or fraud detection, such improvements could have meaningful impacts.

These benefits, however, depend on human-AI complementarity: each agent must be capable of correcting the other's

TABLE I. TEAM DECISION ACCURACY AND AI USAGE RATE AT DIFFERENT CONFIDENCE THRESHOLDS

Confidence Threshold τ	Team Accuracy (%)	AI Usage Rate (%)
0.50 (Always AI)	77.7	100
0.60	82.0	90
0.70	**84.1**	75
0.80	83.0	50
0.90	79.0	25
0.95	76.0	10
0.99 (Rarely AI)	72.0	3

errors. In the simulation, this complementarity was modeled by allowing the human to outperform the AI in low-confidence cases. When error overlap is substantial, confidence-based delegation may offer limited advantage.

B. Dependence on AI Confidence Calibration

A foundational assumption is that the AI's confidence is reasonably well-calibrated. In practice, confidence miscalibration, manifesting as overconfidence or underconfidence, can severely undermine delegation policies. Overconfident models may mislead users into trusting erroneous outputs, whereas underconfident models may lead to missed opportunities for automation.

Calibration techniques such as Platt scaling, isotonic regression, and temperature scaling can improve the reliability of confidence scores. Alternatively, "learning to defer" approaches explicitly optimize the decision of when to abstain and defer to a human, effectively learning an optimal threshold τ directly from data.

C. Human Factors and Policy Compliance

Strict adherence to the delegation policy was assumed in the simulation. However, real users may deviate due to cognitive biases, trust propensities, or interface design. For example, a clinician might override a confident AI due to recent negative experiences or accept low-confidence AI suggestions due to overtrust.

Compliance can be improved through design interventions. Visual cues indicating whether a confidence threshold has been met could nudge users toward intended behavior. Training and feedback mechanisms may further support trust calibration by displaying statistics on when AI advice was followed or overridden, along with the corresponding outcomes.

Thresholds may also be personalized. Overtrusting users may benefit from higher thresholds that enforce engagement, while distrustful users may begin with lower thresholds that increase over time as trust builds. Adaptive systems that monitor user behavior and adjust τ dynamically may enhance collaboration.

D. Integration with Explanations and Interpretability

Although confidence was the primary trust signal considered, AI systems increasingly provide explanations such as saliency maps and rationales. These can either support or interfere with trust calibration. An explanation aligning with user expectations can reinforce high-confidence decisions, whereas poor explanations may lead to unjustified distrust.

Combining confidence with selective explanations presents a promising direction. Explanations could be shown for low-confidence cases to assist human judgment, and withheld for high-confidence cases to reduce cognitive load. Future empirical research should evaluate such hybrid strategies.

E. Complementarity and Task Structure

The efficacy of threshold-based delegation depends on complementary error patterns between human and AI agents. When failures occur on different instances, confidence-based delegation yields the greatest gains. Task design can enhance complementarity by training AI models on instances difficult for humans or by employing disagreement-based sampling.

Task difficulty distribution also plays a critical role. Confidence-based delegation performs best when there is a broad range of instance difficulties. In tasks that are uniformly easy or uniformly hard, the confidence spectrum may collapse, limiting the utility of threshold-based strategies.

Moreover, error cost asymmetry should inform threshold selection. In high-risk applications, thresholds might be set conservatively to ensure human oversight. While this study optimized for accuracy, practical deployments may need to optimize precision, recall, or cost-sensitive metrics.

F. Implications for Human-Centered AI Design

Confidence thresholds offer a transparent and tunable mechanism for managing human-AI collaboration. They can be adapted to task domain, risk level, and user behavior, making system behavior predictable and auditable to aid accountability.

Effective implementation requires careful attention to interface design, user training, performance monitoring, and explanation integration. The delegation policy should be visible, justifiable, and adaptable.

The findings support a central tenet of human-centered AI: the objective is not to replace humans, but to design systems in which humans and AI complement each other. Confidence-based trust calibration provides a concrete, interpretable, and effective strategy for achieving such synergy.

VI. CONCLUSION

A simulation-based investigation of confidence-based trust calibration in human-AI teams was presented, focusing on a threshold delegation strategy in which the AI's confidence score determined whether its prediction was used or deferred to the human. It was found that selecting an appropriate confidence threshold could significantly enhance team performance: in the examined scenario, team accuracy peaked at 84.1%,

surpassing both the AI-alone baseline (77.7%) and human-alone performance (71.9%). This improvement highlights the value of calibrated reliance, whereby each agent contributes most effectively in contexts that align with their respective strengths.

These findings are consistent with prior research emphasizing the importance of trust calibration in human-AI interaction. Confidence-based delegation policies provide a transparent and operationalizable mechanism for achieving such calibration, assuming that the AI's confidence scores are well-calibrated and that human decision-makers are appropriately guided to interpret and act on those scores.

For designers of human-centric AI systems, the results suggest that implementing confidence-threshold strategies constitutes a practical and beneficial design approach. Rather than positioning human and AI agents as parallel or redundant decision-makers, a coordinated policy can systematically leverage their complementary capabilities. Nevertheless, the realization of these benefits in applied settings is contingent on factors such as user education, interface design, and the reliability of the AI's uncertainty estimates.

Future research is encouraged to investigate adaptive threshold strategies that adjust dynamically based on user behavior, task context, or evolving trust calibration. Empirical user studies are also needed to assess how individuals interact with confidence-based delegation interfaces, the extent to which prescribed policies are followed, and how interventions, such as visual cues or feedback mechanisms, may enhance compliance and comprehension.

Ultimately, confidence-based trust calibration exemplifies the goals of human-centered AI: systems designed not only for accuracy but also to support and align with human decision-making processes. By enabling reliance on AI systems to the extent that their predictions are trustworthy, progress is made toward the development of AI tools that are both effective and accountable. The approach described herein demonstrates that even a simple, interpretable rule anchored in confidence can yield meaningful improvements in team performance. With ongoing advancements in calibration techniques and human-AI interface design, trust calibration strategies are expected to occupy a central role in the next generation of decision-support systems.

REFERENCES

[1] B. Ebenso, E. Namisango, I. Abejirinde, and M. J. Allsop, "The scale-up and sustainability of digital health interventions in low-and middle-income settings," p. 1634223, 2025.

[2] B. P. SAMEERKUMAR, "Ethical considerations in ai design and deployment," *WORLD*, vol. 25, no. 1, pp. 2166–2173, 2025.

[3] C. Happer, "Ai-enabled signal detection in pharmacovigilance," 2025.

[4] J. Kornowicz, "Human integration in ai: Calibrating trust and improving performance in decision support systems."

[5] F. Ghaffar, "Understanding human decision variability and the effects of ai system data modality on trust and acceptance in human–ai collaboration," Ph.D. dissertation, University of Waterloo, 2025.

[6] J. Joseph, "Designing for dignity: Ethics of ai surveillance in elderly care," *Frontiers in Digital Health*, vol. 7, p. 1643238, 2025.

[7] G. Faisal, "Understanding human decision variability and the effects of ai system data modality on trust and acceptance in human–ai collaboration: A human-centred approach to designing ai systems in healthcare contexts," Ph.D. dissertation, 2025.

[8] M. Tyrovolas, "Advanced methods for explainable artificial intelligence based on fuzzy cognitive maps," Ph.D. dissertation, Πανεπιστήμιο Ιωαννίνων. Σχολή Πληροφορικής και Τηλεπικοινωνιών. Τμήμα ..., 2025.

[9] M. Kneer, M. Loi, and M. Christen, "Trust and responsibility in human-ai interaction," *Available at SSRN 5731431*, 2025.

[10] A. Kulal, "Cognitive risks of ai: Literacy, trust, and critical thinking," *Journal of Computer Information Systems*, pp. 1–13, 2025.

[11] Z. Zhang and H. Xu, "Explainable ai for maritime autonomous surface ships (mass): Adaptive interfaces and trustworthy human-ai collaboration," *arXiv preprint arXiv:2509.15959*, 2025.

[12] J. Abyasa and R. Rahmania, "Axons-3: An xai-augmented approach for advancing trust and transparency in 3d brain tumor segmentation," in *2025 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 2025, pp. 64–71.

[13] T. von Reeken, M. Salous, and L. Abdenebaoui, "Un-trusting the chat: Designing for calibrated trust in retrieval-augmented conversations," in *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, 2025, pp. 1–10.

[14] L. R. Marusich, B. T. Files, M. Bancilhon, J. C. Rawal, and A. Raglin, "Trust calibration for joint human/ai decision-making in dynamic and uncertain contexts," in *International Conference on Human-Computer Interaction*. Springer, 2025, pp. 106–120.

[15] P. Le Denmat, K. Desender, and T. Verguts, "Learning to be confident: How agents learn confidence based on prediction errors," 2025.