# Using Mental Models to Understand the Effect of AI Transparency on Trust

**Teerthaa Parakh , Ryan Bowers , Karen Feigh**
Georgia Institute of Technology
{teerthaa.parakh, rbowers32, karen.feigh}@gatech.edu

## Abstract

Transparency in AI systems has shown mixed effects on human decision-making, sometimes leading to under-reliance, other times to over-reliance. We investigate this inconsistency through the lens of users' mental models: internal representations people form about how AI systems behave. Focusing on AI confidence scores as a form of transparency, we examine how users' trust and reliance is affected when they recognize that the AI is aware of its own capabilities. We propose a game-based experimental framework inspired by real-world Command-and-Control scenarios, requiring collaborative decision-making between a human and multiple AI agents. This setup allows us to study how confidence scores shape mental model formation and influence both per- decision step reliance (i.e., how much users depend on AI agents at each decision) and their overall trust in the AI agents and as well as human-AI team performance.

## 1 Introduction

Theory of Mind (ToM) is the ability of an agent (either human or artificial) to attribute mental states including their beliefs, intentions, and desires to themselves and others. Similar to Theory of Mind is the concept of mental models, which are simplified representations that humans use to describe, explain, and predict systems that are too complex to fully understand [Johnson-Laird, 1980; Johnson-Laird, 1986; Van den Bossche *et al.*, 2011; Cannon-Bowers *et al.*, 1993; Andrews *et al.*, 2023]. In this sense, Theory of Mind systems use mental models formed of another agent to infer cognitive states.

Understanding humans' mental models is crucial in human-AI teams, as these models help explain observed human behavior and, in turn, inform improvements to AI systems. Conceptually, we can think of the environmental state as the input to the human and the observed behavior as the output. By conceptually dissecting the human brain, i.e. eliciting the mental model, we can uncover the cause–effect relationships driving their decisions. Moreover, these mental models can serve as strong priors for developing AI agents with Theory of Mind capabilities, enabling them to better understand, predict, and adapt to human behavior.

Mental models are particularly valuable to examine when cause–and–effect relationships are complex and indirect. One such case is AI agent transparency and its impact on reliance in AI systems, and consequently, on human–AI team performance. It has been frequently studied as a means to calibrate reliance in human-AI teams [Mehrotra *et al.*, 2024], particularly in the form of confidence scores (an agent's degree of certainty in its actions) and explanations. Several studies [Cai *et al.*, 2019; Lundberg *et al.*, 2018; Schmidt and Biessmann, 2019] have demonstrated the effectiveness of transparency in improving reliance calibration. However, other studies highlight significant limitations and mixed outcomes of transparency approaches [Bansal *et al.*, 2021; Miller, 2023; Alfrink *et al.*, 2023; Zhang *et al.*, 2020]. Additionally, implementing transparency faces challenges in determining appropriate information content, timing, and presentation without increasing cognitive workload, leading to inconsistent results in reliance calibration across different contexts [Zerilli *et al.*, 2022].

Due to these mixed findings, a mental model approach becomes especially valuable. Understanding how humans form mental models of AI agents can provide deeper and more generalizable insights into when and why transparency succeeds or fails in calibrating reliance. However, *only a few studies* [Bansal *et al.*, 2019; Schraagen *et al.*, 2020; Nourani *et al.*, 2021] have explicitly examined this through direct mental model elicitation. Even then, these studies have focused on much simpler systems involving one-step decision-making and a human-AI dyad.

Real-world systems are complex and frequently involve humans interacting with multiple AI agents. This makes it particularly important to study team compositions beyond the dyad. Examining the effect of transparency on reliance in such complex structures is even more challenging, thus realizing the need to explore this cause–effect relationship through the lens of mental models.

To study the effect of transparency in complex scenarios, we developed a strategic Command and Control game using the GameTeq software suite [MetaTeq, 2025]. The game is inspired by real-world hierarchical team structures in which humans make higher-level decisions supported by multiple AI agents in scenarios that involve cascaded decision-making

and delayed outcomes. These scenarios are complex because each decision step's available choices and outcomes depend on the choices made in prior decision steps, which in turn depends on the team's trust dynamics. *Such cascaded decision-making scenarios are understudied in human-in-the-loop contexts*, despite occurring frequently in real-world environments.

We use a conceptual model of AI [Gero *et al.*, 2020] in our cooperative game setting. Conceptual models serve as more practical representations of the target system, especially since an AI agent's internal design is not always apparent from its observed behavior. Following [Gero *et al.*, 2020], our model includes three key components: (A) **Local behavior**: how individual decisions or actions are made by the AI agent, (B) **Global behavior**: broader aspects, such as whether the AI agent performs well at its task, operates effectively within its role in the system or game, is aware of its own capabilities, and can recognize when it lacks knowledge, and (C) **Knowledge distribution**: conceptions of what the AI agent knows, for example, whether it is aware of specific people, events, or attributes.

Our experimental setup is designed to investigate how confidence score influences reliance through users' mental models in complex decision-making environments, specifically addressing:

1. Which components of users' mental models are influenced by AI confidence scores?
2. Can AI confidence scores help calibrate users' per-decision step reliance in cascaded decision-making scenarios?
3. Can AI confidence scores increase the overall trust on AI?

# 2 Related Work

## 2.1 Shared Mental Models

A rich body of research on mental models and team shared mental models (SMMs) [Cannon-Bowers *et al.*, 1993; Johnson *et al.*, 2008] has shown that SMMs are strong drivers of team performance and fluency [Klimoski and Mohammed, 1994]. More recently, SMMs have also been applied to human-AI teams [Scheutz *et al.*, 2017], [Kaur *et al.*, 2019; Kelly *et al.*, 2023].

[Bansal *et al.*, 2019] found that humans formed an accurate mental model of an AI's error boundary when it was parsimonious, non-stochastic and low-dimensional, which led to improved team performance.[Gero *et al.*, 2020] used think-aloud protocols to elicit mental models of an AI agent in a word-guessing game and argued that the accuracy of these mental models should be assessed against the AI's conceptual model, since a system's architecture and training do not always reflect its actual behavior. [Nourani *et al.*, 2021] studied how cognitive biases influence mental model formation when humans are presented with AI explanations.

While there is research on human-AI shared mental models, most studies focus on single-decision tasks, where humans simply accept or reject AI suggestions and receive feedback on the outcome immediately after making their decision. To date, few works ([Gupta *et al.*, 2024] being one exception) have investigated these scenarios.

Furthermore, work by [Siu *et al.*, 2021] on human-AI teaming in the Hanabi game has demonstrated that even when objective performance metrics, such as game scores, remain the same, subjective human perceptions of AI teammates (including trust, interpretability, and teamwork quality) can vary significantly. This divergence in subjective measures highlights the importance of studying mental models, as these perceptions reflect the mental models that users form about AI.

## 2.2 Transparency and Uncertainty

Inter-teammate transparency - the degree to which team members have knowledge of each other's roles, capabilities, and decision-making processes - has long been recognized as critical to effective team behavior.Transparency has been shown to increase trust, team performance, and the rate of inter teammate habituation [Bhatt *et al.*, 2021; Zerilli *et al.*, 2022; Tomsett *et al.*, 2020].

It can take many forms, such as explanations and confidence scores. For example, [Bansal *et al.*, 2021] showed that displaying explanations only for high confidence predictions, but not for low confidence ones, can reduce human overtrust. Similarly, [Zhang *et al.*, 2020] found confidence scores to be effective for calibrating trust. [Khastgir *et al.*, 2018] combined explanations and confidence scores in the context of autonomous vehicles to support trust alignment.

Many team transparency frameworks include the agent's degree of uncertainty as a critical component. For example, the SA based Agent Transparency framework [Chen *et al.*, 2018] places uncertainty within Situation Awareness Level 3, alongside the projection of future outcomes and current limitations.

Prior work has investigated how this communicated uncertainty influences human trust and performance in human AI teams. These studies have yielded contradictory results—some found that communicating uncertainty increases trust [Reyes *et al.*, 2025; Zhang *et al.*, 2020] and performance [Marusich *et al.*, 2024; Vodrahalli *et al.*, 2022], while others reported mixed or inconclusive outcomes [Greis *et al.*, 2024; Cao *et al.*, 2023].

This discrepancy is likely because human-AI trust does not depend *directly* on the AI's level of certainty/uncertainty. Rather, uncertainty is an aspect of the team's SMM that calibrates trust and enables more fluid team performance [Tomsett *et al.*, 2020]. If the AI states that it has low certainty in a given action/suggestion, the human's reliance in that *specific* action should decrease, but their trust in the agent as a whole may increase (similar to the phenomenon where a person is perceived as more trustworthy and knowledgeable when they admit their mistakes or refrain from being overly confident in their decisions). In this work, we approach the issue of human-AI team uncertainty through this more nuanced perspective, and consider confidence scores as a method of communicating uncertainty.


Figure 1: The game environment consists of two regions: the blue ally region and red enemy region. Each region contains a carrier at its center that forces must protect while attempting to destroy the opponent's carrier by launching aircraft. Enemies can enter the ally region at any time, requiring strategic resource allocation for both offensive and defensive operations. The ally side has a total of 5 aircraft available and must decide how many to deploy in each region across 7 time steps. The interface includes a clock and chat box in the bottom right where AI suggestions appear and participants enter their decisions, while scores for both forces are displayed in the top right corner.

## 3 System Design

In our experimental framework, participants engage in a resource allocation game involving limited resources and strategic decision-making under time pressure. The game consists of two carrier strike groups – an ally and an enemy group – each comprising an aircraft carrier and a few aircraft deployed at sea, shown in Figure 1.

### 3.1 Game Design

The game simulates an offensive-defensive scenario where resources must be allocated across multiple steps as enemies progressively enter different regions, which requires participants to think strategically about resource distribution to achieve victory. Our experimental setup is distinguished by its cascaded decision-making structure, which differs from traditional one-step decision paradigms commonly used in mental model research. At each decision step, participants receive recommendations from two AI agents via chat messages (Figure 2), with each agent suggesting the number of aircraft needed for their respective regions. This task design constrains the participant's responsibility to high-level resource allocation, while low-level decisions (such as aircraft movement and enemy targeting) are either hard-coded or handled by AI agents. Participants then assess enemy positions and health status, evaluate their own aircraft positions and health, consider AI agent recommendations, and input their final decision regarding aircraft deployment through the chat interface. Participants are required to make each choice within a relatively short time period (30 seconds), adding additional time pressure.

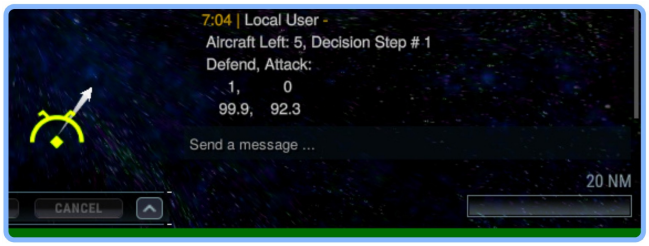


Figure 2: Chat interface displaying AI agent suggestions, remaining aircraft count, and current decision step in the game.

### 3.2 Decision Agents and Confidence Scores

The two decision agents are implemented as actor-critic networks trained using Proximal Policy Optimization in a Centralized Training, Decentralized Execution approach [Amato, 2025] [Yu *et al.*, 2022]. Each agent's actor network has partial observability of the game state, only observing their respective region (either offense or defense). The agents' critic networks observe the full game state. In response to observations at each major decision point, the agents output a multi-discrete action that represents how many aircraft to send to each region and which opponents each aircraft should target. The agents' confidence scores are calculated as the softmax probabilities generated by the model [Hendrycks and Gimpel, 2017]. For testing, we selected cases from different training stages where the AI's confidence scores were well-calibrated (i.e. high confidence scores are paired with effective actions).

## 4 Methodology

We investigate how AI transparency affects human reliance on AI agents and human-AI team performance through the

lens of mental model formation. Using a between-subjects user study, we examine how confidence scores influence participants' decision-making, trust calibration, and mental model development in a collaborative resource allocation task. Our study protocol has been approved by our institution's IRB board.

## 4.1 Experiment Design

Participants are randomly assigned to one of two conditions (Independent Variable), balanced for demographic factors:

- No Confidence Score group: Participants receive recommendations from the Offensive AI and Defensive AI (suggested numbers of aircraft to attack and to defend, respectively) at each decision step. These suggestions are presented alone.
- Confidence Score group: Participants receive the same recommendations alongside a confidence rating (50% to 100%) representing the level of certainty in the AI's suggestion.

We operationalize uncertainty by providing a confidence rating alongside the AI's recommendation at each decision step, and measure participants' reliance on the AI's suggestions and overall team performance depending on whether confidence scores are present.

## 4.2 Experiment Procedures

Each session consists of 2 practice rounds for familiarization followed by 6 experimental rounds. We selected test scenarios where confidence scores range from 50-100% and are well-calibrated (high-confidence recommendations lead to better outcomes than low-confidence ones). During each round, participants observe the game state, receive AI recommendations via chat interface, and must allocate aircraft within 30 seconds.

## 4.3 Dependent Variables

We measure three main dependent variables to identify which components of the participants' mental models are affected by the AI's confidence scores, and how these components influence participants' reliance and trust in the AI at each decision step.

**DV 1: Reliance on AI Suggestions.**
We measure calibrated reliance using the reliance calibration value (RCV), computed for each game round as:

$$\text{RCV} = \frac{\text{True Positive Cases} + \text{True Negative Cases}}{\text{Total number of AI suggestions}}$$

Here, a True Positive case refers to an instance in which the AI's decision has high confidence and the participant accepts the suggestion, while a True Negative case refers to an instance in which the AI's decision has low confidence and the participant rejects the suggestion.

This metric reflects the participant's understanding of each agent's error boundary and capabilities. A reliance calibration value of one indicates fully calibrated reliance, meaning the participant consistently accepts AI suggestions when the AI is likely to be correct and rejects them when it is likely to be incorrect. This metric is computed separately for each AI agent, as their tasks differ.

**DV 2: Team Performance.**
To measure the downstream impact of transparency and trust, we measure human-AI team performance as the difference between the participant's final carrier health and the opponent's carrier health at the end of each game. It is important to note that even if reliance is properly calibrated, such as when the AI has low confidence and the participant correctly recognizes that the AI's suggestion is not appropriate for the situation, the outcome may still result in failure. This can happen if the participant's own decision is also ineffective due to a lack of skill or understanding of the task.

**DV 3: Mental Model Evolution.**
Questionnaires administered after each round track how participants' mental models of the two AIs evolve throughout the experiment. The questionnaires are designed to probe mental models across the three components defined by [Gero *et al.*, 2020]: Global behavior, Knowledge distribution, and Local behavior. Examples of questionnaire items include:

- How participants perceive the AI's competence, including whether its confidence scores appear calibrated and if any suggestions were outright wrong with high confidence (Global behavior).
- Whether participants observe patterns in AI errors, such as suggestions being more likely incorrect when two opponents are present or at the beginning, middle, or end of the game (Knowledge distribution, Local behavior).
- For the Confidence Score group, at what confidence level participants would likely accept the AI's suggestion (Global behavior).

In addition to these three main dependent variables, we also measure participant's overall trust in the AI at the end of each game round using Likert scales, as well as *Decision Speed*, defined as the time taken by the participant to make a decision.

## 4.4 Hypotheses

The conceptual model of AI, consisting of global behavior, knowledge distribution and local behavior, serves as a ground truth for the comparison and evaluation of human mental models of AI. It is designed to provide a practical representation of the target system, as an AI agent's internal design does not always get reflected in its observed behavior [Gero *et al.*, 2020]. The conceptual model of the AI agent is developed by extensively testing the system to identify behavioral patterns.

We make the following hypotheses:

- **H1 (Mental Model Formation):** Providing AI's confidence score facilitates the formation of accurate **global** behavioral mental models [Bansal *et al.*, 2021].
- **H2 (Effect of Priors/Beliefs):** Confidence scores calibrate users' reliance on the AI when participants' prior beliefs about the AI or task domain are weak.

- **H3:** Higher Reliance Calibration Value indicate greater user trust in the AI, as users recognize that the AI is aware of its own capabilities.

# 5 Discussion

## 5.1 Future Work

We plan to conduct comprehensive user studies to test our hypotheses and highlight the importance of understanding mental models in human-AI collaboration. More specifically, through this experimental setup, we aim to show why uncertainty should be studied as a form of transparency, particularly in terms of how it influences users' mental model formation and, in turn, trust, and how it can be leveraged to achieve appropriate user trust.

Based on our results, we aim to better understand how to communicate uncertainty [Bhatt *et al.*, 2021] in ways that support users in developing accurate mental models of the AI agents (that is, sufficient and appropriate representations of its capabilities and limitations). These mental models help foster appropriate trust in the system and improve team performance beyond what either the human or AI could achieve alone.

A further direction for future work involves developing ToM capabilities for AI agents. The mental models elicited through our framework can serve as valuable priors for ToM-enabled AI agents, enabling them to better infer and adapt to human preferences.

## 5.2 Limitations

Our AI agents are trained without human-in-the-loop learning. When human decisions deviate from AI recommendations, the system may encounter out-of-distribution states that were not present during training. As a result, the confidence scores may become miscalibrated and no longer reliably reflect model performance, thereby undermining their intended role in supporting user understanding and trust.

Secondly, we do not comprehensively address the different sources of uncertainty within our current framework. Uncertainty can arise from multiple factors, including partial observability (where agents lack complete environmental information), environmental noise, and model limitations. Our current approach merges different sources of uncertainty into a single concept, which may overlook the ways users respond differently to each type.

# References


[Alfrink *et al.*, 2023] Kars Alfrink, Ianus Keller, Neelke Doorn, et al. Tensions in transparent urban ai: Designing a smart electric vehicle charge point. *AI & Society*, 38:1049–1065, 2023.

[Amato, 2025] Christopher Amato. An initial introduction to cooperative multi-agent reinforcement learning, 2025.

[Andrews *et al.*, 2023] Robert W. Andrews, J. Mason Lilly, Divya Srivastava, and Karen M. Feigh and. The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175, 2023.

[Bansal *et al.*, 2019] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*, HCOMP-19. AAAI, 2019.

[Bansal *et al.*, 2021] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[Bhatt *et al.*, 2021] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. *arXiv preprint arXiv:2011.07586*, 2021.

[Cai *et al.*, 2019] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), November 2019.

[Cannon-Bowers *et al.*, 1993] Janis A. Cannon-Bowers, Eduardo Salas, and Sharolyn A. Converse. Shared mental models in expert team decision making. In Neal J.J̃r. Castellan, editor, *Individual and Group Decision Making: Current Issues*, pages 221–246. 1993.

[Cao *et al.*, 2023] Shiye Cao, Anqi Liu, and Chien-Ming Huang. Designing for appropriate reliance: The roles of ai uncertainty presentation, initial user decision, and user demographics in ai-assisted decision-making. 2023. Johns Hopkins University.

[Chen *et al.*, 2018] Jessie Y. C. Chen, Shan G. Lakhmani, Kimberly Stowers, Anthony R. Selkowitz, Julia L. Wright, and Michael Barnes and. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3):259–282, 2018.

[Gero *et al.*, 2020] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[Greis *et al.*, 2024] Miriam Greis, Passant El.Agroudy, Hendrik Schuff, Tonja Machulla, and Albrecht Schmidt. Decision-making under uncertainty: How the amount of presented uncertainty influences user behavior. 2024. Available at: {firstname.lastname}@vis.uni-stuttgart.de.

[Gupta *et al.*, 2024] Piyush Gupta, Subir Biswas, and Vaibhav Srivastava. Fostering human learning in sequential decision-making: Understanding the role of evaluative feedback. *PLOS ONE*, 19(5):1–23, 05 2024.

[Hendrycks and Gimpel, 2017] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

[Johnson *et al.*, 2008] Tristan E. Johnson, Mohammed K. Khalil, and J. Michael Spector. The role of acquired shared mental models in improving the process of team-based learning. *Educational Technology*, 48(4):18–26, 2008.

[Johnson-Laird, 1980] P.N. Johnson-Laird. Mental models in cognitive science. *Cognitive Science*, 4(1):71–115, 1980.

[Johnson-Laird, 1986] P. N. Johnson-Laird. *Mental models: towards a cognitive science of language, inference, and consciousness*. Harvard University Press, USA, 1986.

[Kaur *et al.*, 2019] Harmanpreet Kaur, Alex C Williams, and Walter S Lasecki. Building shared mental models between humans and ai for effective collaboration. In *CHI'19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, Glasgow, Scotland, 2019. ACM.

[Kelly *et al.*, 2023] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. Capturing humans' mental models of ai: An item response theory approach. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1723–1734, Chicago, IL, USA, 2023. ACM.

[Khastgir *et al.*, 2018] Siddartha Khastgir, Stewart Birrell, Gunwant Dhadyalla, and Paul Jennings. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation Research Part C: Emerging Technologies*, 96:290–303, 2018.

[Klimoski and Mohammed, 1994] Richard Klimoski and Susan Mohammed. Team mental model: Construct or metaphor? *Journal of Management*, 20(2):403–437, 1994.

[Lundberg *et al.*, 2018] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, and Su-In Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749–760, oct 2018.

[Marusich *et al.*, 2024] Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. Using ai uncertainty quantification to improve human decision-making, 2024.

[Mehrotra *et al.*, 2024] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M. Jonker, and Myrthe L. Tielman. A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM J. Responsib. Comput.*, 1(4), November 2024.

[MetaTeq, 2025] MetaTeq. Metateq official website, 2025. Accessed: 2025-06-16.

[Miller, 2023] Tim Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 333–342, New York, NY, USA, 2023. Association for Computing Machinery.

[Nourani *et al.*, 2021] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, IUI '21, page 340–350, New York, NY, USA, 2021. Association for Computing Machinery.

[Reyes *et al.*, 2025] J Reyes, AU Batmaz, and M Kersten-Oertel. Trusting ai: does uncertainty visualization affect decision-making? *Frontiers in Computer Science*, 7:1464348, 2025.

[Scheutz *et al.*, 2017] Matthias Scheutz, Scott DeLoach, and Julie A Adams. A framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making*, 11(3):203–224, 2017.

[Schmidt and Biessmann, 2019] Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. In *Proceedings of the AAAI-19 Workshop on Network Interpretability for Deep Learning*, February 2019. AAAI-19 Workshop on Network Interpretability for Deep Learning.

[Schraagen *et al.*, 2020] Jan Maarten Schraagen, Pia Elsasser, Hanna Fricke, Marleen Hof, and Fabyen Ragalmuto. Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 64, pages 339–343, 2020.

[Siu *et al.*, 2021] Ho Chit Siu, Jaime D. Peña, Yutai Zhou, Edenna Chen, Victor J. Lopez, Kyle Palko, Kimberlee C. Chang, and Ross E. Allen. Evaluation of human-ai teams for learned and rule-based agents in hanabi. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[Tomsett *et al.*, 2020] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. Rapid trust calibration through interpretable and uncertainty-aware ai. *Patterns*, 1(4):100049, 2020.

[Van den Bossche *et al.*, 2011] Piet Van den Bossche, Wim Gijselaers, Mien R.Segers, Geert Woltjer, and Paul Kirschner. Team learning: Building shared mental models. *Instructional Science*, 39:283–301, 05 2011.

[Vodrahalli *et al.*, 2022] Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. Uncalibrated models can improve human-ai collaboration, 2022.

[Yu *et al.*, 2022] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[Zerilli *et al.*, 2022] John Zerilli, Umang Bhatt, and Adrian Weller. How transparency modulates trust in artificial intelligence. *Patterns*, 3(4):100455, 2022.

[Zhang *et al.*, 2020] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery.