# Impact of Team Models in Hierarchical Human-Agent Decision-Making Teams

Ranjani Narayanan[a] and Karen M. Feigh[b]
*School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, U.S.A.*
{*rnarayanan35, karen.feigh*}*@gatech.edu*

Keywords: Mental Models, Decision Support, Human-AI Teaming, Decision-Making.

Abstract: With increasing opportunities for the integration of AI-based decision support tools, human interactions with AI systems must be studied under different contexts. The quality of interactions is said to improve when humans have a mental model (MM) of their AI teammates. This study tests this hypothesis for a hierarchical human-AI triad in a decision-making setting. The impact of humans' MM of AI on various performance metrics is studied in the presence and absence of mutually compatible MMs between the two agents. Mutually incompatible models lead to dissonance between the agents, causing an overall degradation in their joint activity. Results showed that operators exercised increased vigilance when they did not have a MM of their AI teammates, but having an accurate MM improved decision utility over time without reducing speed or increasing participants' task load. It also led to greater acceptance of the AI's suggestions without inducing biases towards the AI. Additionally, operators reported lesser effort and mental demand and had more accurate judgments of the relative competence of the two agents during dissonance. These findings motivate further research on understanding how different levels of MMs between humans and their AI teammates leads to different outcomes in complex collaborative settings.

## 1 INTRODUCTION

Literature suggests that the inclusion of AI as a teammate, as opposed to a mere tool, would enhance performance, the quality of interactions, and humans' satisfaction of working with the AI, as a teammate would be able to adapt to their human counterparts (Andrews et al., 2023; O'Neill et al., 2022). *Team* members must have "clearly defined differentiated roles and responsibilities, hold task-relevant knowledge, and be interdependent (i.e., must rely on one another to accomplish goals). Teams can be distinguished from groups, in which members have homogeneous expertise, roles, and responsibilities" (Converse et al., 1993; Orasanu and Salas, 1993). Per this definition, the portrayal of humans and AI collectives as teams is dubious (Groom and Nass, 2007) as their interactions have largely been studied as groups.

A team decision process "involves gathering, processing, integrating, and communicating information in support of arriving at a task-relevant decision" (Converse et al., 1993). They state that this activity does not always require consensus between individuals, nor are all individuals involved in all aspects of the process. The role of members is to filter out irrelevant information, apply their knowledge and skills to solve role-specific problems, communicate the relevant information with other team members, and make recommendations. With this definition, a simple hierarchy would entail assigning authority of the final decision to a single individual, while the team supports the decision-maker with assessments and information as per the given situation. Due to a host of legal, ethical, moral, safety, and trust-related issues surrounding entrusting authority to automation (Awad et al., 2018; Kingston, 2018; Bartneck et al., 2021), we posit that humans will likely continue to be placed at the apex of the decision-making ladder for safety-critical situations. Human-AI or AI-only teams would function towards supporting the decision-maker.

Mental Models have been used as a basis for understanding human-human and human-automation interactions. Johnson-Laird originally described a MM as the mechanism by which humans understand the world, creating working models in their minds. These are abstract, long-term knowledge structures used to describe, explain, and predict the world

[a] https://orcid.org/0000-0003-4351-6501
[b] https://orcid.org/0000-0002-0281-7634



Narayanan, R. and Feigh, K. M.
Impact of Team Models in Hierarchical Human-Agent Decision-Making Teams.
DOI: 10.5220/0013097400003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 1: GRAPP, HUCAPP and IVAPP*, pages 452-463
ISBN: 978-989-758-728-3; ISSN: 2184-4321

around (Johnson-Laird, 1983). Shared Mental Models (SMMs) extend the MM concept to teams. The SMM theory hypothesizes that if individuals in a team possess similar MMs of the shared task and of each other they can effectively establish mutual awareness that helps in inter-agent information sharing and expectation setting. It also fosters trust and accountability in teammates and facilitates anticipatory behavior during difficult situations without the need for explicit communication (Kleinman and Serfaty, 1989). These capabilities are crucial for effective collaboration between teammates, especially when they are fundamentally different such as humans and AI agents.

(Converse et al., 1993) laid out four components that constitute a MM: 1) Equipment Model, 2) Task Model, 3) Team Interaction Model, and 4) Team Member Model. This taxonomy has been reduced based on the two content-specific domains: 1) the Task model and 2) the Team Model (Scheutz et al., 2017). The former consists of task procedures, possible outcomes and how to handle them, technical systems involved in performing the task, and how to perform them. On the other hand, the Team Model is comprised of tendencies, beliefs, and personalities of team members, how the team is structured, its roles, modes, and frequency of communications. SMMs benefit the team by enabling members to form accurate explanations and expectations of the task, coordinate their actions, and adapt their behavior to accommodate the demands of the task and other team members (Mathieu et al., 2000). However, the complexity and stability of these models are not the same. Team Models are less stable because they are dynamic and not just dependent on the situation but also on the team members involved. Individuals have high variability in their prior MMs, past experiences, education, socio-economic backgrounds, and individual personalities, all of which influence their MM formation (Converse et al., 1993). Given the necessity to establish a SMM between humans and AI agents for successful teamwork, their Team Models make an interesting object of study and warrant attention.

Studies in human-AI teams primarily focus on three factors: 1) the AI's MM of the task and teammate (human), 2) the human's MM of the task and teammate (AI), and 3) the joint (shared) MMs. Within this framework, we aim to study the utility of accurate Team Models among decision-makers, especially during inter-agent dissonance. Dissonance between agents may occur due to inadequate calibration, faulty information sources, different policies for training (preferences), improper distribution of information, etc. In real-life settings, with growing team sizes and complicated decision workflows, where only key information may be shared, lack of communication and effective information exchange may lead to mutually incompatible MMs between members of a team. This creates 'dissonance' between them. The role of the operator must be to identify and overcome the undesirable effects of such dissonance. Team Models in human-AI dyads have been beneficial in reducing operator bias, improving performance, workload, and trust, as the operator understands the circumstances under which the AI is reliable (Kulesza et al., 2012; Walsh et al., 2024; Yang and Dorneich, 2018). This enables complementary behavior and augments performance that neither humans nor AI can achieve alone (Kamar, 2016; Kamar et al., 2012).

A recent study (Walsh et al., 2024) provided empirical evidence that mutual understanding between humans and AI teammates positively influences performance and workload for decision-making when humans' Task Models are imperfect. The work also identified the benefits and trade-offs associated with the human or the AI possessing an accurate Team Model of the other. However, as with most of these studies, this work focused on human-AI dyads with members acting as 'groupmates' thereby simulating relatively simple aspects of collaboration. The distribution of information within these dyads has been homogeneous as members have shared common roles and responsibilities. There is little inter-dependency and greater overlap, leaving little opportunity to understand the true aspects of teaming. Therefore, empirical research in human-AI teams must be conducted for multi-agent complex teams to verify if prior findings hold across different settings.

In this study, we expand the definition of mixed-agent teams by studying a triad of a human decision-maker equipped with two AI agents for a geospatial decision-making task. The agents are given information specific to their assignments, while the human has high-level task information to monitor the agents. Interdependent activities between the two agents involve filtering out irrelevant information and providing the human decision-maker with relevant information. This can only be achieved when they operate with mutually compatible 'MMs'. The two low-level agents must also generate decision recommendations, for which they will have a finer-grained task understanding than their human manager.

This study aims to empirically verify the utility of Team Models in hierarchical human-agent (triad) teams by answering the following research questions:

- **RQ1.** How does providing a (human) decision-maker/manager with an accurate Team Model affect performance in mixed hierarchical teams?
- **RQ2.** How do Team Models affect the decision-

maker's workload and perception of the AI agents within a hierarchical team?

## 2 RELATED WORKS

The focus of design of socio-technical systems has started shifting towards not just making agents smarter, i.e. improving their task-related capabilities, but taking a human-centered approach by trying to understand humans' cognitive needs while coordinating and cooperating with AI agents at large. In this section, we briefly visit the concept of mental models to highlight the role of humans' understanding of their AI teammates and how explainable AI has been used to bridge the gap between humans' understanding and the true nature of the AI systems they interact with.

### 2.1 Mental Models

Humans create mental models of any system they interact with (Norman, 2013), including AI agents (Kulesza et al., 2012; Tullio et al., 2007). These models are not static knowledge structures. They change on continuous interaction with the system. However, they are sparse and parsimonious, leading the user to form useful approximations that they can rely on to understand the target system (Norman, 2013). Humans' mental models can be broadly classified into Task or Team Models, depending on the major components constituting them. While Task models are relatively more stable, Team Models are more susceptible to change (Converse et al., 1993). This is because Team Models constitute team-specific characteristics such as the roles, responsibilities, and interaction patterns between team members. It also depends on the situation and the particular team members involved. However, as the role of human operators transitions from handling the task at hand to supervising and coordinating with multiple agents, humans' mental models of the AI i.e., their Team Models will be a major determinant of team success. They will help humans discern the conditions under which AI systems may be relied upon, and help them identify the circumstances within which the operator has to exercise increased caution and monitor the agents more closely. (Kulesza et al., 2012) demonstrated that participants who were able to create sound mental models of a recommender system's reasoning and those who most improved their mental models made the system operate to their satisfaction. A lack of understanding of the AI agents has shown to induce algorithmic aversion (Dietvorst et al., 2014), leading the human to completely avoid using the agents' suggestions. Alternatively, humans may excessively rely on AI systems and base their final decisions on the AI's flawed recommendations, which is termed as "automation bias" (Mosier and Skitka, 1999; Cummings, 2004; Wagner et al., 2018; Robinette et al., 2017).

Norman emphasized the role of the designer in bridging the gap between the conceptual model of the target system and the actual mental model that users develop as a result of their interactions (Norman, 2013). To improve the user's mental model of the AI, the designer may employ better instruction, training, improvements in design workflows, or provide the user with different kinds of explanations. A humans' mental model of the AI may constitute an understanding of the underlying algorithm (Kulesza et al., 2012; Kulesza et al., 2013), its predictions (Ribeiro et al., 2016), or error boundaries (Bansal et al., 2019). Crucially, the goal of designers must be to understand how to reinforce the development of some or most of these aspects of MMs. In the following section, we will address the role of explanations in improving humans' MMs of their AI teammates.

### 2.2 Improving Mental Models

ML algorithms, particularly Deep Neural Networks (DNNs), work as black-box models that do not provide much traceability of the predictions or outcomes. Unlike simplistic rule-based approaches that allow operators to trace back the steps leading to a decision, DNNs are complex and opaque. This lack of interpretability, a key requirement for human reasoning, precludes the development of any reliable mental model of the AI system. Providing explanations to the user has been a popular approach taken by most AI researchers and developers. (Mueller et al., 2019) state that MMs help users reason with the system they interact with, leading to a generation of explanations. Explanations (provided during system interaction), in turn, help facilitate the development of MMs by proposing causal relationships or directing user attention. However, there is no guarantee of the effectiveness of an explanation if it does not consider the user by understanding what they want and what they already know (Andrews et al., 2023).

For example, one of the factors that influence the calibration of humans' expectations of an AI is providing explanations for failures, either through natural language methods (Das et al., 2021), confidence scores (Zhang et al., 2020), or both. (Yin et al., 2019) found that a model's stated accuracy has a significant effect on people's trust even after observing a high accuracy in practice. However, if a model's ob-

served accuracy is low, then after observing this accuracy, the stated accuracy has at most a very small effect on people's trust in the model. But there is a catch. The stated accuracy may not always correctly represent an algorithm's performance. Humans are less likely to follow the AI recommendation when the stated score is an incorrect estimate of risk. Additionally, humans have trouble processing probabilistic information that signifies confidence calibrations of some AI models (Handmer and Proudley, 2007). While explanations have a functional role in improving transparency and understanding, it is not yet clear under which conditions they improve collaboration and human productivity (Doshi-Velez and Kim, 2017; Poursabzi-Sangdeh et al., 2021).

As we transition to multi-agent settings, there is a need to identify and understand what aspects of the MM need to be strengthened to enable operators to function effectively within the team. For example, a domain expert may be well-equipped to form reliable Task Models, but a lack of familiarity with AI-based DSS will lead to misalignment between their expectations and what the AI is truly capable of. In a decision hierarchy, the manager who is responsible for coordinating with the other agents does not have access to all the relevant information to inform their decisions. They may also not have access to the raw information, or the exact values that the individual agents assign to key information. Explanations, in the form of salient features from the raw information is sometimes used to inform the operators for better situational assessments (Andrews et al., 2023). Further, in time-sensitive situations, they may be unable to decompose the activities of the individual agents to identify the root cause of any failures. Therefore, managers must be able to learn and adapt through external observation, feedback from the environment, and continuous interaction with the task and teammates.

In this study, we hypothesize that providing human managers with accurate prior MMs of their AI teammates will equip them to better judge the agents' capabilities. We believe that they will enable the operator to reconcile any differences between two agents, due to dissonance in their alignment/calibration, and achieve task-related objectives more effectively.

# 3 METHODOLOGY

This section describes the experimental task, interface design, and the metrics used to understand the impact of Team Models on decision-making activity.

## 3.1 Experimental Task

Participants in our experiment are tasked as decision-makers (team managers) in a disaster-relief team responsible for delivering resources to afflicted regions in a storm-struck city. The goal is to place two key resources, i.e. Food and Generator, based on information about the city's Population distribution and Power Outage conditions, respectively. Two additional sources of information, i.e. No-go and Flooding conditions, must be used to inform the placement of *both* resources. Two AI agents assist the participants, each providing suggestions for the placement of Food and the Generator based on significantly detailed data from these information sources. To avoid information overload, the agents provide a condensed map of the common information sources i.e., No-go and Flooding conditions to their manager. This condensed map is a composite of the two raw information sources and is subject to variation between the two agents based on whether they have mutually compatible Team Models. In the presence of Team Models, the raw inputs are processed similarly by the two agents and the composite map presented by both agents to the participant is consistent. This is an underlying feature of Team Models because mutual awareness of how a partner may perceive certain information drives consistent and compatible behaviors in teams. During dissonance, each agent provides its own version of the composite map, resulting in conflicting versions of the same raw information, thereby simulating an adverse effect of dissonance that may lead to confounding the end user. The participant may accept or override either agent's suggestion and place the resource at their discretion for each task.

Information sources presented to the participant are in the form of 8x8 grid maps across which resources, that are 2x2 grids, must be placed. Each grid on the map corresponds to a specific color that indicates the utility of that location. The four colors (Green, Yellow, Orange, and Black) are ordered based on their relative utility assigned to the grids. Green and Black grids have the highest and lowest utilities respectively. Participants must place the grids in such a way that the maximum 2x2 utility area for the relevant maps is serviced. They are scored on the combined utility for placement of *both* resources.

The experiment was conducted to study the influence of the human manager's Team Model under two main conditions: (1) when there was no inter-agent dissonance, and (2) during inter-agent dissonance. Both versions of the AI agents are designed to have a 50% success rate, to prevent any systemic bias in the participant's interpretation of their compe-

tencies. The sequence in which the AI agents err is randomized but kept consistent across all the experimental groups. The second independent variable is the presence (or absence) of the participants' Team Model of the agents. Team Models in participants are instilled through instructions. Participants with a Team Model are told about the mutual inconsistency between the agents and their effect on the composite maps presented. The participants without a Team Model are given no such instruction. Team Models give them the required understanding to identify and diagnose the source of the agents' errors and evaluate their capabilities in providing assistance. The utility of the participant's Team Model is evaluated during both conditions: with and without inter-agent team dissonance. Thus, the design of our experiment is a 2x2 fully crossed between-subject experiment. This setup is intended to simulate situations under which team dissonance leads to improper coordination among AI team members and their managers.

## 3.2 Experimental Interface

The experimental interface consists of three main areas shown in Figure 1. The left-hand side consists of the two information attributes along with the composite maps generated by each AI agent, contained within the 'Data Sources' section. The middle section shows the grid map of the affected regions in the city, and the right-hand side includes controls to submit the resource placement. Panels in 'Data Sources' allow the user to view the heat-map overlay and the associated utility of each grid location on the map. Two (dotted) icons in blue and red are overlaid on the heatmap to indicate the AI's suggestions for the placement of Food and Generator. Two (solid line) markers are provided to the participant for their final placement of the resources. A color scale indicating the relative utilities of each grid is displayed adjacent to the map.

The composite map is a tool that helps the participants visualize all or different combinations of information attributes at once. In this case, each AI agent synthesizes information from No-go and Flooding data sources to provide the user with a condensed representation of all the information. [1] The first instance of this was described by (Illingworth and Feigh, 2021). Information from the composite map along with Population or Power Outage data sources must be combined (using linear superposition) to determine the appropriate location to place Food and Generator respectively. Without Team Models, the two agents would alternatively provide the decision-maker with an incorrect composite map for 6 out of 12 tasks in total. With mutually compatible Team Models, the two composite maps are identical but are incorrect for 6 out of 12 tasks, depriving the user of the correct information from the No-Go and Flooding maps.

Once participants submit their final decision, feedback in the form of the individual utility (for each resource) and joint utility is provided to the user. They then proceed to the next task. Information sources get updated for every task.

## 3.3 Procedure

Participants start by signing an online consent form and proceed to read a set of instructions. Once the instructions are completed, participants are tested for their task understanding using a questionnaire. Only on clearing this test are they allowed to proceed. They are provided with five training tasks that mimic the actual tasks they will be faced with. On completion of training, participants in the treatment group with Team Models are provided instructions about the agents. They then proceed to complete the main experimental tasks. On completion, participants rate the AI along several dimensions described in section 3.4. They are also asked to rate their workload using the NASA-TLX (Hart and Staveland, 1988).

## 3.4 Measures

This section describes the subjective and objective metrics used to answer our research questions.

*Team performance* evaluates the accuracy of the users' decisions based on the utility of the selected location. The utility for placement of each resource is the linear superposition of the utilities from the relevant maps for that location. Task scoring is the arithmetic average for the placement of both resources. The scores are normalized (min-max) to the range of 0-100%. Along with decision utility, task load is objectively measured using instances of users' information access. Information access for each attribute is represented by the number of clicks on that information source. The sum of all such clicks acts as an objective, proxy measure for the effort expended by the user in evaluating the agents' suggestions toward making the final decision. In the same vein, the time (to the nearest second) to make each decision is measured from the time the information is displayed on

[1] The use of the term GenAI in the toggle panel on the interface indicates that the composite map presented to the user is generated by the AI agent responsible for the placement of the Generator. This term is not to be confused with the popular term 'GenAI' for 'Generative AI'. The distinction has been made clear to the user through instructions.

Table 1: Experiment Design: Treatment Groups.

| Participant Team Model \ Inter-agent Team Dissonance | No Dissonance | Dissonance |
|---|---|---|
| **Team Model** | Participants have MM of the agents; agents have MM alignment | Participants have MMs of the agents; agent MMs are not mutually aligned |
| **No Team Model** | Participant does not have MMs of agents; agents have MM alignment | Neither participants have MMs of agents, nor do agents have MM alignment |

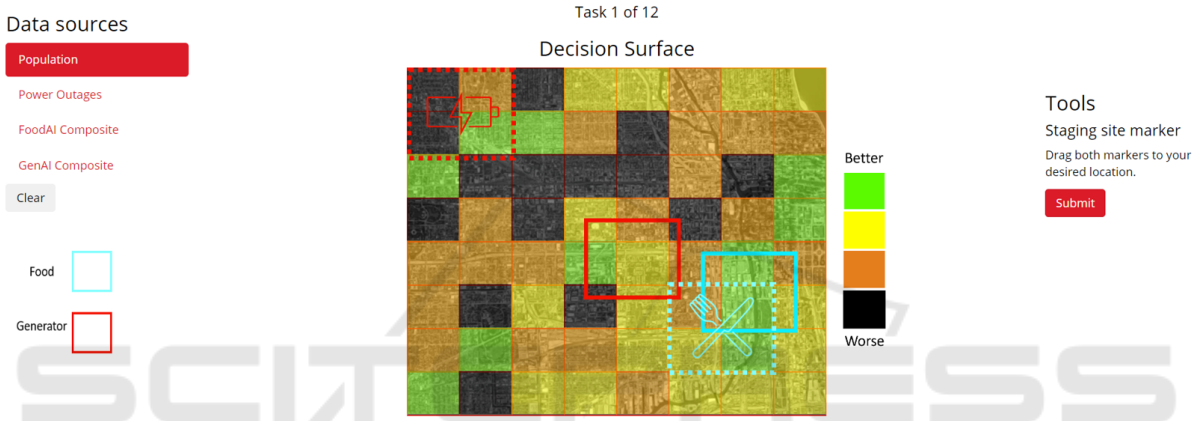



Figure 1: Gorilla Experimental Interface: (1) The toggle panel to the left indicates information sources. (2) The map at the center shows the affected areas in the city. AI Suggestions in the form of red and blue dotted icons are placed at different locations on the map. Blue and Red solid markers indicate resources for placement. Color scaling to the right indicates the relative utility associated with each color. (3) Tools on the right consist of a 'Submit' button for submitting responses.

the screen until the user hits the 'submit' button to transition to the next task.

Along with these performance metrics, *users' agreement levels* with the AI agent are measured as the proportion of instances (from the 12 tasks) for which the user agreed with either AI's suggestions. This represents the participants' tendency to utilize the AI. The appropriateness of these agreement levels was measured by the proportion of instances for which the user agreed with an AI's suggestion when it was correct and disagreed with the AI's suggestion when it was incorrect. The overall agreement levels are a measure of calibrated reliance on AI.

Users' experience of *workload* during their interaction with the AI team was measured using the NASA TLX. We used a scale of 1-21 for recording responses for each sub-scale. To identify users' perceptions of each teammate, participants were asked to rate their experience of working with each agent across the following dimensions: *Intuitiveness* (not intuitive - intuitive), *Confidence* (not confident - confident), *Competency* (incompetent - competent), *Benefit in teaming* (useless - beneficial), *(Frequency of) Trouble in Decision Making* (never - always), *Task Understanding* (mis-understood - understood), *Willingness to Work (together) again* (never - definitely). Responses to these questions were recorded on a Likert Scale of 0-5 with the lowest and highest ratings for each metric indicated within the parentheses.

## 3.5 Participants

Data were collected from 80 participants with 20 participants in each experimental group. The male-to-female ratio in our participant pool was 50:29 and 1 participant was undesignated. Ages ranged between 20 to 69 with a median age of 36 years. All participants resided in the United States and reported fluency in English. There were no reports of color blindness. We used an online experiment-building platform called *Gorilla* and conducted recruitment using the online crowd-sourcing platform, *Prolific*. A

simpler version of this setup was first presented by (Walsh and Feigh, 2021). Additional compensation of 25% was provided to the top 10% performers to encourage high-effort participation. The study was IRB-approved at Georgia Institute of Technology.

# 4 RESULTS

We explored the impact of the human decision-maker's Team Model, during both conditions of inter-agent dissonance. The Kruskal-Wallis test was used to determine statistical differences between groups.

A comparison of team performance between groups (see Fig. 2a) without inter-agent dissonance indicated little effect of providing participants with a Team Model ($\chi^2(1) = 0.042, p = 0.8378$). Curiously, the lack of Team Models between agents led participants without Team Models to make better decisions on average ($\chi^2(1) = 9.28, p < 0.01$). However, improvements over the 12 tasks (slope = $1.029 \pm 0.4, p < 0.05$) were significant among participants with Team Models. These improvements were more gradual and insignificant among participants without Team Models (slope = $0.567 \pm 0.4, p = 0.1390$). Figure 3a indicates the trends for performance over the 12 tasks. Here, we see little differences in the initial performance levels among participants with Team Models during dissonance and the participants paired with AI agents without dissonance. This lack of difference in the initial stages can be attributed to an 'illusion of understanding'. Despite access to the correct version of the composite maps, this illusion may create initial degradation in performance which was quickly overcome with increased interactions. Surprisingly, participants without Team Models had improved scores. This may be due to increased vigilance on their behalf, which led to active monitoring (as we will see from information access and time to make decisions) as they perceived inconsistency in the information provided by the two agents. Testing for an extended number of trials is necessary to determine if these effects persist. Conversely, Team Models only had a marginal effect on performance for human-agent teams without dissonance because when agents were incorrect, the decision-maker did not have access to the correct version of the composite maps. They would have had to rely on underlying task heuristics, which may be sparse leading to sub-optimal outcomes.

The effect of possessing Team Models in participants had significant effects on their task load (information access) (during dissonance: $\chi^2(1) = 8.28, p < 0.01$, without dissonance: $\chi^2(1) = 31.03, p < 0.001$).

Table 2: Summary of results from Kruskal Wallis test for comparison of performance metrics in decision-makers with and without Team Model.

| Metric | Condition | $\chi^2(1)$ | $p$ |
|---|---|---|---|
| Team Performance | No Dissonance | 0.042 | 0.8378 |
| | Dissonance | 9.28 | $<$ **0.01** |
| Information Access | No Dissonance | 31.03 | $<$ **0.001** |
| | Dissonance | 8.28 | $<$ **0.01** |
| Time Taken | No Dissonance | 50.25 | $<$ **0.001** |
| | Dissonance | 31.86 | $<$ **0.001** |
| Agreement with AI | No Dissonance | 15.43 | $<$ **0.001** |
| | Dissonance | 6.40 | $<$ **0.05** |

The time required by the participants to arrive at the final decision was also significantly reduced with Team Models (during dissonance: $\chi^2(1) = 31.86, p < 0.001$, without dissonance: $\chi^2(1) = 50.25, p < 0.001$). However, the amount of time taken per decision also declined over the 12 tasks among participants who did not have a Team Model but were equipped with non-dissonant agents that had mutually compatible MMs, (slope = $-1.74 \pm 0.5, p < 0.001$). The ease for participants to recognize the behavior due to the coordinated activity of the two agents may explain this trend. Participants spent less time making decisions when they possessed Team Models and also benefited (with increased task and team familiarity) as long as Team Models existed between the two AI agents causing them to be non-dissonant.

A comparison of the reliance of the manager on AI agents revealed a significant rise in the agreement levels when participants had an accurate Team Model (Figure 4). To measure the level of reliance on the AI agents, we measure the agreement levels between participants' final decisions and either AI's suggestions. During dissonance, this difference is of the order $\chi^2(1) = 6.40, p < 0.05$, whereas without dissonance, the difference is of the order $\chi^2(1) = 15.43, p < 0.001$. We also measured how appropriate the reliance was. Notably, there was no difference when the participant was provided with a Team Model of the agents. They were more accepting of the AI's suggestions when they had a Team Model while being wary of its suggestions when they did not possess a Team Model. Providing a Team Model did not lead to blind compliance as participants were better equipped to identify when to follow either AI's suggestions.

Lastly, we measured participants' subjective workload using the NASA-TLX. Team Models influenced operators' workload when faced with conflicting information as the two agents were uncoordi-

nated. Participants with Team Models reported significantly less Mental Demand ($\chi^2(1) = 5.38, p < 0.05$) and Effort ($\chi^2(1) = 5.43, p < 0.05$).

Along with workload, participants were asked to rate the Agents across several dimensions to assess the perception of their Teammates Team Models did not affect the ratings given to either agent along any of the specified dimensions. However, participants without a Team Model sensed greater frequency in the occurrence of errors for the AI that made suggestions for the placement of Food ($\chi^2(1) = 4.36, p < 0.05$) during inter-agent team dissonance. The lack of Team Models may lead to a tendency among users to assess two equally competent agents differently, especially when there is dissonance among members of the team. This is not surprising because, within our experiment, dissonance was designed to simulate a condition where there was disagreement between the agents on performing an interdependent task of synthesizing common pieces of information. Without a Team Model, operators would need more task training and experience with the two agents to recognize that the two agents are similar in competency.

## 5 DISCUSSION

With the rising development and integration of AI-based agents in a variety of task settings, it is necessary to understand how their interactions with humans will evolve. While traditional human-AI teaming continues to be studied under simplistic decision workflows, multi-operator settings would evolve into a complex reorganization of activities that constitute a hierarchy. Individuals operating in such teams would have varied roles, and responsibilities, where establishing SMMs in all aspects of the collaboration will prove increasingly challenging. Especially, as AI systems remain largely opaque, a detailed understanding of the AI teammate or lack thereof may have severe consequences in collaborative output. This study aimed to understand whether accurate Team Models of the AI serve the purpose of aiding the decision-maker within a hierarchical setting, especially during dissonance between the members of the team.

We recognized that dissonance in teams led to surprising outcomes in terms of the decision utility. Inconsistency in the mutual perception and integration of information between the two agents led to increased vigilance among decision-makers without accurate Team Models. Providing a mental model of their teammates led to an initial 'illusion of understanding', which led to some initial reductions in decision utility. However, the long-term benefits of Team Models are that they help decision-makers quickly identify and correct inter-agent inconsistencies to improve decision outcomes and close any performance gaps. We urge researchers to verify these results in other operational environments to empirically identify the utility of Team Models in multi-agent settings.

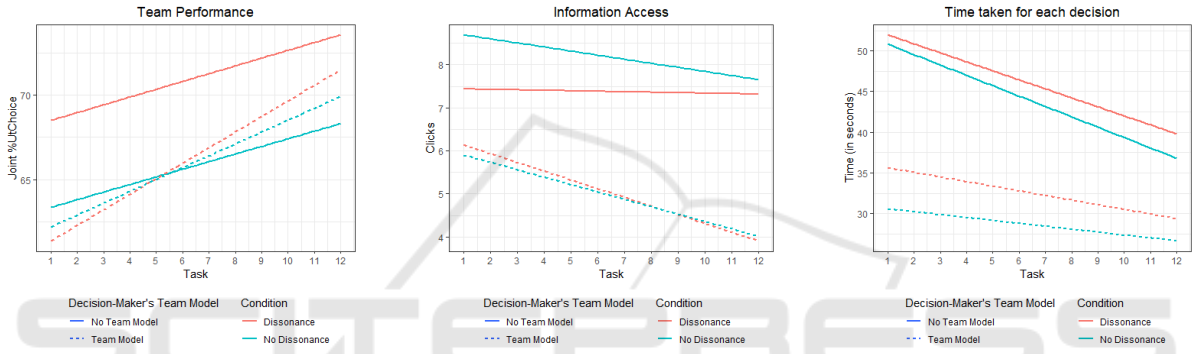
In situations where the risks associated with decision outcomes are severe, increased training must be exercised between members for MMs between them to form and converge. Convergence does not mean an overlap of information within MMs. Especially in multi-agent settings, MM convergence should aim towards arriving at shared expectations between individuals of a team for any given situation (Cannon-Bowers and Salas, 1990). Such training may also include ways for the decision-makers to verify the veracity of information the agents operate on, to improve transparency. As we transition to larger teams, dissonance is more likely to occur especially in scenarios where communication is restrictive. Alternative ways for the decision-maker to tackle such inconsistencies may be useful in achieving better outcomes.

For tasks that prioritize speed over decision utilities, mutual predictability in agentic behavior may be a driver of high performance. This is because individuals without accurate Team Models also tend to improve their decision-making speed with increased familiarity with the task and teammates, without necessarily compromising on the quality of the decisions. Although providing Team Models benefits the operators by reducing their overall task loads, in situations where achieving an accurate, thorough, and detailed understanding of the teammates is challenging, improving the predictability of AI teammates may yield sufficiently desirable outcomes.

Overall, the benefits of providing Team Models among decision-makers are significant in terms of the quality of ensuing interactions. The use of AI-based tools can be maximized only when users are more likely to use them rather than indiscriminately avoiding them. Providing a Team Model helps improve the acceptance of the agents' suggestions without creating over-reliance or complacency among individuals. Decision-makers can perceive AI errors more thoroughly which would eliminate any potential automation biases (Mosier and Skitka, 1999). In reality, as 'MMs' between agents diverge, communication breaks down, and it becomes difficult for agents to describe their state and their actions in a way that the decision-maker will understand, thereby introducing a breakdown of transparency (Scali and Macredie, 2019). In addition to reducing mental demand and effort perceived by decision-makers, during dissonance, having reliable Team Models lends more accurate and


(a) (b) (c)

Figure 2: Comparison of performance metrics measured across both conditions of dissonance (a) Team performance in terms of decision utility; (b) Information access (mouse clicks); (c) Time taken per decision.




(a) (b) (c)

Figure 3: Comparison of trends in performance metrics measured across both conditions of dissonance (a) Team performance in terms of decision utility; (b) Information access (mouse clicks); (c) Time taken per decision.

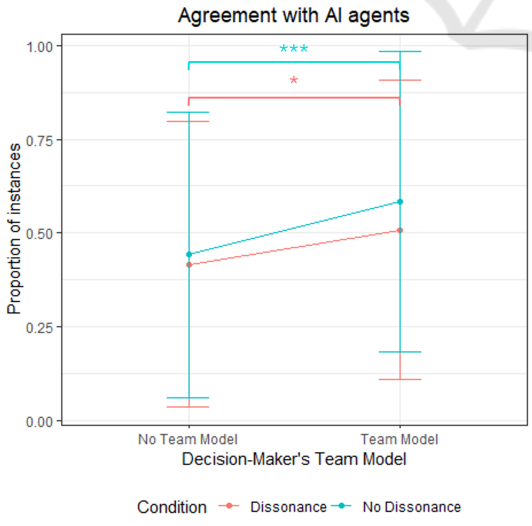



Figure 4: Overall agreement levels with either AI agent.

equitable perceptions by humans of their AI teammates. Researchers in the field of Explainable AI should aim to identify what aspects of the humans'

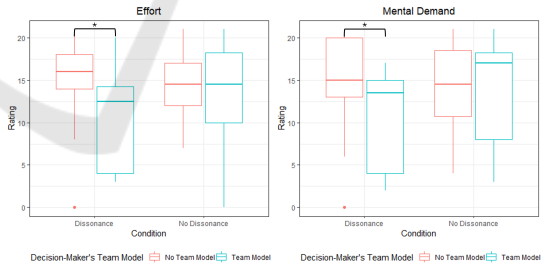



Figure 5: Comparison of ratings for workload dimensions.

mental models of the AI need reinforcement, such that the explanations deliver the maximum positive impact toward the interaction between the human and AI agents, without overwhelming the user with additional information.

## 6 LIMITATIONS

While our study provides insight into the interactions between humans and their automated teammates in a

hierarchical decision setting, we wish to shed light on some limitations in the design choices for our study.

This study was designed for crowd workers to understand decision-making in a hierarchical triad of human-agent teams. The task was simplified and training was provided to equip participants with sufficient understanding to complete the task. We also acknowledge that our decision environment may not have *accurately* simulated the varied levels of risks associated with real-life decision environments. This restricts our ability to fully capture the true attitudes and preferences that experts may demonstrate during the interaction. To understand the naturalistic decision processes among domain experts and the role of accurate Team Models in facilitating the interaction process, we recommend that researchers and design practitioners study the validity of our findings in higher-fidelity testing environments, where they may replicate the risks associated with the task. Our findings serve an instructional value, in that they prove the existence of some of the underlying trends that we may observe in real-life situations.

Participants in our study also had access to high-level relevant information, such that they may use to improve the decisions of their AI teammates. This design choice was made to identify whether learning occurs within the duration of our study. In cases where there are intangibles, due to a non-uniform distribution of key information, where the user cannot possibly make up for performance, the decision utility levels as seen in the No Team Model and Dissonance condition of our study may not hold valid. In such situations, the user has to have a better understanding of the AI teammate or must be provided with other means of verifying the fidelity of AI recommendations, to prevent any degradation in team performance. In disaster-relief situations, where the decision utility may map to the number of lives saved, even small increments in the quality of decisions may have major consequences. Thus, the success of collaborative decision-making should be studied in the context of the environment.

We also believe that extending the number of trials will help us develop better insights into how some of our performance metrics may evolve over longer periods. Our study was designed to be completed within an hour (maximum time taken by users within a pilot study, from instruction to post-study questionnaires), to ensure participant engagement and retention. The design of the study, through instructions and training, ensured that participants could identify, learn, and adapt through increased engagement with the task and agents. However, we hesitate to say if the differences between groups for any of the performance metrics will be sustained or if they will level out with greater experience in performing the task. Future work must study how the observed trends evolve over extended periods to determine the existence of long-term gains in the interaction of these MM levels between humans and AI.

Finally, we acknowledge that the definition of Team Models within our study was limited to the user's understanding of the agents' proclivity to making errors and how mutually incompatible agents processed information differently. Team Models are multi-dimensional and could be represented/modeled in different ways. Studies incorporating the other dimensions, such as teammate skills, communication patterns, preferences, etc. will help researchers understand what aspects of Team Models best improve upon the performance in hybrid teams. It is also worth investigating which aspects of the Team Model are prone to degradation and how design workflows, explanations, and other mechanisms may improve the humans' MMs of AI.

# 7 CONCLUSION

This work is the first in a series of studies that will study the impact of teammate MMs in hierarchical human-AI teams for decision-making. We studied the impact of providing the decision-maker with an accurate Team Model, when there was inter-agent dissonance and when there was not. With partial and abstracted Task Models at the apex of the decision hierarchy, the lack of teammate understanding leads to greater vigilance, task load, and reduced speed on behalf of the decision-maker during dissonance. Team Models help decision-makers close performance gaps without causing greater stress or reducing their speed of decision-making. They help improve humans' perception of the agents and reduce the subjective workload, especially during dissonance. These findings drive the importance of designing AI-based decision-support tools that not only support task-based needs but also drive transparency in operations leading to better MM formation among its users. As AI-powered tools proliferate, studies should focus on the interaction effects between decision-makers' teammate understanding in complex settings and various task environments to drive better design choices in creating intelligent decision-support tools.

## ACKNOWLEDGEMENTS

This work was supported by the Office of Naval Research Command Decision Making Program under Contract N00014-24-1-2135. The results do not reflect the official position of this agency.

## REFERENCES

Andrews, R. W., Lilly, J. M., Srivastava, D., and Feigh, K. M. (2023). The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018). The moral machine experiment. *Nature*, 563:59–64.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-ai team performance. In *AAAI Conference on Human Computation & Crowdsourcing*.

Bartneck, C., Lütge, C., Wagner, A., and Welsh, S. (2021). *An Introduction to Ethics in Robotics and AI*.

Cannon-Bowers, J. and Salas, E. (1990). Cognitive psychology and team training: Training shared mental models and complex systems. *Human Factors Society Bulletin*, pages 1–4.

Converse, S., Cannon-Bowers, J. A., and Salas, E. (1993). Shared mental models in expert team decision making. *Individual and group decision making: Current issues*, 221:221–46.

Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. *Collection of Technical Papers - AIAA 1st Intelligent Systems Technical Conference*, 2.

Das, D., Banerjee, S., and Chernova, S. (2021). Explainable ai for robot failures. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM.

Dietvorst, B., Simmons, J., and Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of experimental psychology. General*, 144.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Groom, V. and Nass, C. (2007). Can robots be teammates? benchmarks in human-robot teams. *Interaction Studies*, 8:483–500.

Handmer, J. and Proudley, B. (2007). Communicating uncertainty via probabilities: The case of weather forecasts. *Environmental Hazards*, 7:79–87.

Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload*, 1(3):139–183.

Illingworth, D. A. and Feigh, K. M. (2021). Impact mapping for geospatial reasoning and decision making. *Human Factors*, page 0018720821999021.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Number 6. Harvard University Press.

Kamar, E. (2016). Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 4070–4073. AAAI Press.

Kamar, E., Hacker, S., and Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '12, page 467–474, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Kingston, J. (2018). Artificial intelligence and legal liability.

Kleinman, D. L. and Serfaty, D. (1989). Team performance assessment in distributed decision making. In *Proceedings of the symposium on interactive networked simulation for training*, pages 22–27. University of Central Florida Orlando, FL.

Kulesza, T., Stumpf, S., Burnett, M., and Kwan, I. (2012). Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1–10, New York, NY, USA. Association for Computing Machinery.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. (2013). Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., and Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of applied psychology*, 85(2):273.

Mosier, K. L. and Skitka, L. J. (1999). Automation use and automation bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 43(3):344–348.

Mueller, S. T., Hoffman, R. R., Clancey, W. J., Emrey, A., and Klein, G. (2019). Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. *ArXiv*, abs/1902.01876.

Norman, D. A. (2013). *The Design of Everyday Things*. MIT Press.

Orasanu, J. M. and Salas, E. (1993). Team decision making in complex environments. page 327–345. Ablex Publishing.

O'Neill, T., McNeese, N., Barron, A., and Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, 64(5):904–938. PMID: 33092417.

Poursabzi-Sangdeh, Forough, Goldstein, G, D., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI

'21, New York, NY, USA. Association for Computing Machinery.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Robinette, P., Howard, A., and Wagner, A. (2017). *Conceptualizing overtrust in robots: Why do people trust a robot that previously failed?*, pages 129–155. Springer International Publishing.

Scali, G. and Macredie, R. (2019). Shared mental models as a way of managing transparency in complex human-autonomy teaming.

Scheutz, M., DeLoach, S. A., and Adams, J. A. (2017). A framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making*, 11(3):203–224.

Tullio, J., Dey, A. K., Chalecki, J., and Fogarty, J. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 31–40, New York, NY, USA. Association for Computing Machinery.

Wagner, A. R., Borenstein, J., and Howard, A. (2018). Overtrust in the robotic age. *Communications of the ACM*, 61(9):22–24.

Walsh, S., Narayanan, R., and Feigh, K. M. (2024). The role of shared mental models in ai-advised decision support. (under review).

Walsh, S. E. and Feigh, K. M. (2021). Differentiating 'human in the loop' decision process. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3129–3133.

Yang, E. and Dorneich, M. C. (2018). Affect-aware adaptive tutoring based on human–automation etiquette strategies. *Human factors*, 60(4):510–526.

Yin, M., Vaughan, J., and Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. pages 1–12.

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