Impact of Abstraction Levels of Context Information on AI-Advised Decision Making for an Entry Descent and Landing Task

Divya K. Srivastava* and Jack Kolb † and Karen M. Feigh.‡
Georgia Institute of Technology, Atlanta GA, 30332

AI-advised Decision Making is a form of human-autonomy teaming in which an AI recommender system suggests a solution to a human operator, who is responsible for the final decision. This work seeks to empower and effectively utilize these human decision makers by supporting their cognitive process of judgement. Previous work has found that providing decision makers with relevant information that the AI uses to generate possible courses of action improves crucial team decision-making metrics, and is a viable alternative solution to explaining or interpreting complex AI models. This paper investigates whether this technique and its observed benefits hold when the relevant information provided to the decision maker is displayed at different levels of abstraction. Findings indicate that this technique of supporting the human's judgement process is effective in (1) boosting the human decision maker's situation awareness and task performance, (2) calibrating their trust in AI teammates, and (3) reducing overreliance on an AI partner, irrespective of the abstraction level at which information was displayed.

I. Introduction

AI recommender systems are a form of decision support systems in which an AI suggests a solution to a human operator, who is responsible for the final decision. These systems are typically considered *black boxes*, and solutions are often presented to the human decision maker with little to no context or insight into how that solution came about. As a result, recommender systems can lead to miscalibrated user reliance and reduced situation awareness.

Recent work has shown that an effective way to increase human-AI team decision-making is to improve the human's initial situation awareness by displaying contextual information (information about the decision environment that is relevant to the decision event) upfront [1]. This technique of providing contextual information is effective because it allows the human to establish an initial judgment about the decision environment, i.e. bring the human in earlier in the decision-making process. The decision-making procedure has been widely modeled as the OODA loop, which consists of four cognitive processes[2]:

- Observation: the collection of data through sensory perception.
- Orientation: the analysis and synthesis of data to form one's current mental perspective.
- Decision: the determination of a course of action based on one's current mental perspective.
- Action: the physical playing-out of the decision.

With many AI-recommender systems, the AI is responsible for collecting relevant information (Observe) and using that information to generate possible solutions (Orient) or an assessment of the state of the world (often defined as judgement) and identification of the decision that needs to be made (Decide). The human is tasked with making the final decision (Decide) and implementating the final decision (Act). Because of the division of labor within this framework, the human must make a decision based on a suggestion and minimal information (if any) provided by the recommender system (now often powered by a trained AI algorithm). Srivastava et al. [1] show that providing information that the AI uses to make its recommendations to the human (thus bringing the human in at the *Observation* part of the decision-making process) increases the human decision maker's initial and final judgment, as well as their ability to discern the recommender's error boundary. Additionally, this technique accurately calibrates the human's trust in the AI recommender.

*Graduate Research Assistant, School of Mechanical Engineering
†Graduate Research Assistant, School of Aerospace Engineering
‡Associate Fellow, Professor and Associate Chair for Research, School of Aerospace Engineering

In this paper, we investigate if these findings hold when the human is shown contextual information at different levels of abstraction, or if the findings of [1] are unique to the level of abstraction used in that study.

II. Background

Team members need to be able to receive and transmit information amongst each other in order to converge on a shared interpretation of the environment, task, and each other's responsibilities and capabilities. This requires being able to communicate at multiple levels of abstraction – some things require more detail to understand and interpret, while others can be understood with less detail. Abstraction can be described as generalization – deciding what details need to be highlighted and which details can be ignored in order to retain and make visible only the key relevant information for performing a particular task [3]. High levels of abstraction are less detailed, showcasing the salient, relevant features, while low levels are more detailed and require more effort to interpret. Some studies have shown that higher level of abstraction can be beneficial in situations where technical, detailed know-how is not needed [4]. However, while abstracting to a higher level is beneficial in some cases, doing so requires losing information that can come in handy in unexpected or difficult situations. A common example of this is consulting user manuals when devices do not function as expected. More information is needed to make sense of the present situation and to be able to move forward. For any high-performing team, members need to be able to adapt and communicate at multiple levels of abstraction in order to formulate the best course of action in any given situation.

In human-human teams, it is easy to understand which level of abstraction is needed at any given time because communication can occur through multiple modes. If someone is relaying information to their partner, and their partner is visibly confused or vocalizes a misunderstanding, that person will likely adapt the level of abstraction at which they are describing the information. Either more or less detail is needed to convey the information. For human-autonomy teams, the abstraction level needs of each teammate will be different since humans and autonomous agents process information differently. Humans and AI may receive the same information in different formats and level of abstractions in order to best understand it. Additionally, the communication with and acquisition of information regarding other teammates is largely informed by the interface/communication module used to display, acquire, and pass information. While the means by which this communication occurs is not insignificant, this work focuses on graphical representations of the information presented to the participant. We hypothesize that in low-risk scenarios, higher levels of abstraction will be more beneficial than lower levels, and that the opposite will be true in riskier scenarios or in scenarios where the AI-suggested plan goes against human expectations.

III. Methodology

This work uses the same experiment platform developed by Srivastava et al [1, 5]. This section describes the main components that are necessary to the development and understanding of this specific body of research work.

A. Task Domain

Participants played the role of the commander of a craft in Mars orbit, tasked with finalizing the Entry, Descent, and Landing (EDL) trajectory of a probe to a landing site. Participants were told that they were aided by an AI Mission Computer that made a parallel evaluation of the proposed trajectory and offered agreement or disagreement with the participant's decision. After seeing the AI's recommendation, the participant had the final call on whether to execute or abort the mission. We imposed no time constraints to the task. For control and reproducibility, the system was presented to participants as "intelligent", however its responses in each scenario were predetermined and fixed.

The task is outlined in Fig. 1. First, the two agents in the team (human participant and AI Mission Computer) independently assessed the state of the world, considering three factors: the orbital positions of GPS satellites, the locations of dust storms, and the atmospheric entry angle required by the craft's current orbital state. Using this information, each evaluated whether the present world state permits a suitable landing (*Initial Shared Situation Awareness Check* represented by the yellow diamond #1).

Then, the AI Mission Computer used the world state conditions (contextual information) to generate a possible trajectory for the probe's landing. The team was shown a set of six figures of merit (FoM) that assessed the proposed trajectory: velocity vs. altitude; heating rate, heat load, and acceleration vs time; latitude vs. longitude; and landing confidence. The FoM charts were shaded to indicate safe, risky, and dangerous thresholds, and the participants were instructed on how to interpret each. Using the FoMs, the agents individually evaluated whether or not to execute the landing trajectory (Final Shared Situation Awareness Check represented by yellow diamond #2). The AI Mission

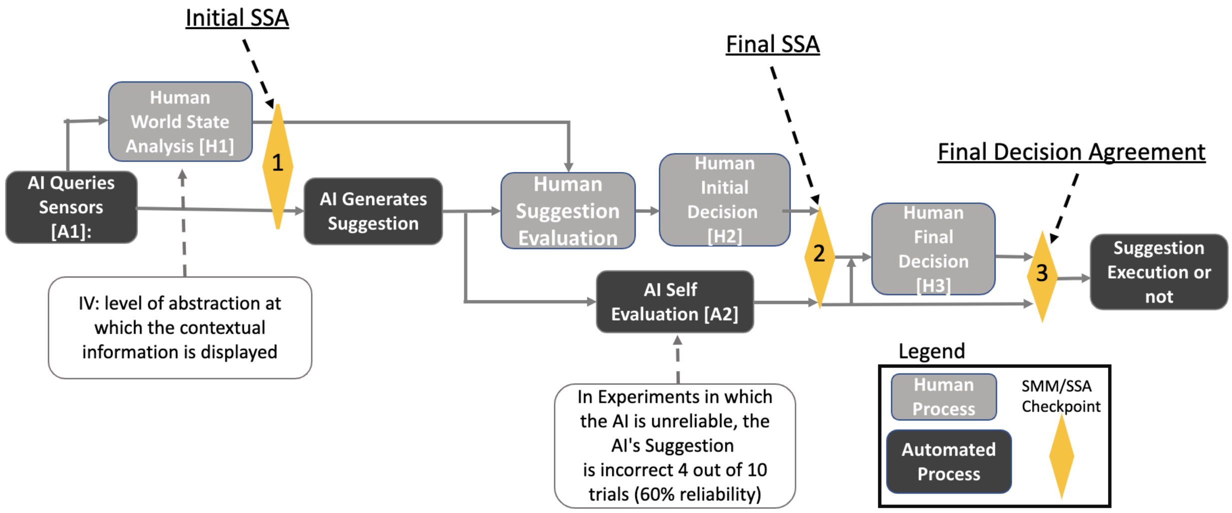

Fig. 1 Task Outline and Metrics for Entry, Descent, and Landing Trajectory Planning

Computer recommendation was made known to the participant following registration of their own. The participant made a final decision on whether to execute or abort the mission in light of the AI's recommendation (Final Decision Agreement Point represented by yellow diamond #3).

B. Experiment Design

The amount of contextual information shown to the participant is operationalized at 3 levels: Observation-Only, Interactive, or Absent. In the Interactive mode, participants view three world state information screens (Figure 2) and answer a multiple-choice question about each component. These questions are intended to increase the participant's situation awareness by forcing them to actively process each information source to minimum degree. In the Observation-Only mode, participants view the information screens, however are *not* prompted for any further engagement with the information. The Absent mode is the control group; participants do not receive any information about the world state, and start directly at the Trajectory Evaluation phase of the experiment. This lack of information mimics current black box systems.

More information about the impact of contextual information in this task domain can be found in [1].

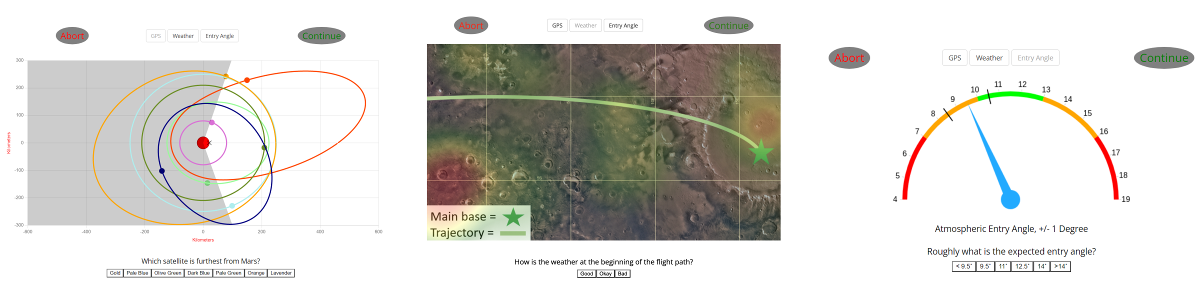

Fig. 2 World State Information from Left to Right: GPS, Atmosphere/Weather, Anticipated Entry Angle

The three levels of abstraction were:

- Low (lots of information, one level up from raw data, need-to-know information for the decision event can be extrapolated from dataset)
- Medium (decision-event relevant information is shown within ranges, need-to-know information must be extracted)
- High (need-to-know information is displayed directly and simply)

The different abstractions of each world state information can be seen in Figure 3. Participants were randomly assigned to receive all *Low*, *Medium*, or *High* levels of abstraction for the experiment.

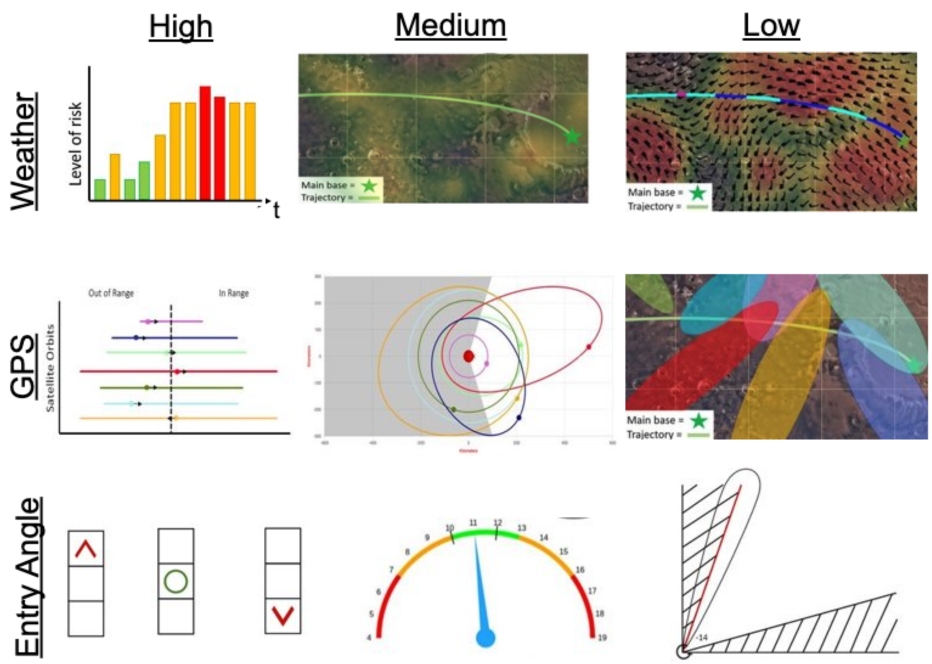

Fig. 3 World State Information (rows) displayed at different Levels of Abstraction (cols)

The AI then "generates" a trajectory based on the current world state information. The mapping of world state information to the generated trajectory was noted prior to real time data collection and implemented in a pre-hoc *Wizard of Oz* format. Figure 4 shows an example of the six figures of merit for the AI-suggested trajectory.

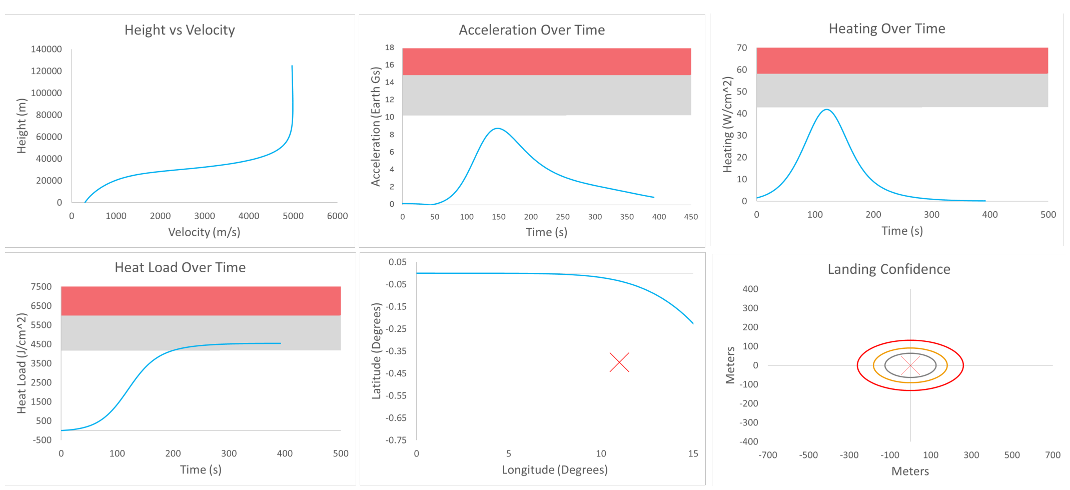

Fig. 4 Figures of Merit that characterize the AI-Generated Trajectory

The recommender system's reliability is manipulated by purposefully making the AI give an incorrect execute or abort recommendation in 4 out of 10 scenarios (60% reliability). This is done to see if the level of abstraction at which the information was displayed has an effect on the participant's ability to understand when the AI is incorrect.

C. General Procedure

We conducted a between-subjects study in which participants and the recommender system jointly make a high-stakes decision. We manipulated the amount of contextual information the participant had, the level of abstraction at which that contextual information was displayed, and the reliability of the recommender system. We measured the effects on final task performance, trust in the recommender system, user workload, and the shared situation awareness between participants and the recommender system.

Participants were recruited through an online recruitment platform (Prolific, www.prolific.co). Individuals who were under 18, located outside of the USA, not proficient in English, and/or did not have normal or corrected-to-normal vision were excluded from the studies. The experiments were conducted online and collected data from 90 participants in each study. Participants were assigned randomly to a treatment group, and the order of scenarios shown to participants was balanced.

Participants were given briefing documents and asked to complete a consent form upon starting the study. Each study had four main components: pre-experiment questionnaires, training session, data collection, and post-experiment questionnaires. The pre-experiment questionnaires measured dispositional trust of the participant in two areas: Faith in Technology and Faith in Persons. These two factors were measured to see if they could predict the accuracy of the human's mental model and overall task performance. Participants were assigned randomly to a treatment group, and the order of scenarios shown to participants was balanced. Instructional videos, tailored to the specific treatment, were used to train participants on the task and the AI Mission Computer. No task-relevant experience was assumed. Participants then completed 6 practice rounds of the mission planning task. Participants who had correct task performance in 5 out of 6 practice trials (and for those in the interactive world state condition, 5 out of 6 correct judgements of the world state) were allowed to proceed past the practice rounds. Passing participants then proceeded to complete 10 trials of the mission planning task. Participants were given feedback on whether their decision was correct or not during the practice rounds, but they were not given any feedback on their performance during the actual data collection rounds. At the end of the experiment, participants completed two final questionnaires: 1) NASA TLX [6] and 2) i-THAU trust assessment [7]. The post-experiment questionnaires measured cognitive workload, participant frustration, perceived performance, and situational and learned trust. Cognitive workload frustration, and perceived performance spoke to the human's experience in completing the AI-advised decision-making task. Situational and learned trust spoke to the human's impression and mental model of the AI.

D. Measures

We first recorded the agreement between the participant's and the AI's initial judgment of the world state conditions (yellow diamond #1 in Fig. 1) and between their initial decisions of whether or not to execute the proposed trajectory (yellow diamond #2 in Fig. 1). This served as an approximation of the final shared situation awareness (*Final SSA*) between the participant and the AI. We computed this on a per participant basis across all 10 scenarios they experienced.

Second, we recorded the *final decision agreement* between the human and the AI after the AI's recommendation to execute/abort was revealed. (yellow diamond #3 in Fig. 1)

Third, we recorded the number of decisions per participant that changed from initial to final decisions and from there computed the percentage of initial disagreements that were resolved. We call this metric *Sway*. High *sway* can indicate overtrust in the AI, while low *sway* can indicate undertrust.

Additionally, since the AI was not 100% reliable, we measured the participant's judgements and decisions against a *"ground truth"*. The Ground Truth was determined using a set of heuristics developed by the researchers to evaluate the figures of merit characterizing the proposed trajectory. We verified these Ground Truth assignments through several pilot tests and found that the general population agreed with our assessments. Thus using the Ground Truth, we recorded the *human's initial judgement* on whether they think the proposed trajectory should be executed or aborted, and their final *task performance*.

Additionally, responses were recorded from the three subjective questionnaires given to participants. The pre-experiment questionnaire was the Pre-Trial section of the Interdependent Trust For Humans and Automation Survey (i-THAu) [7] in which participants were asked to answer a series of statements about their Faith in Persons and Faith in Technology on a seven-point Likert scale from [-3:3]. A composite average of their answers informs their overall *dispositional trust* in these two categories. A rating of -3 indicated a lack of faith - the participant did not tend to trust people or technology. Conversely, a rating of 3 indicated a high level of faith in people or technology.

One of the post-experiment questionnaires was the Post-Trial section of i-THAu trust assessment. Participants responded to a series of statements about their experience of working with the automated system on a seven-point Likert scale from [-3:3]. A rating of -3 indicated a lack of trust – the subject didn't depend on the AI to help them with the task, or they didn't understand the role of the AI. Conversely, a rating of 3 indicated a high trust in and understanding of the AI.

The other post-experiment questionnaire was the NASA Task Load Index (TLX) [6] which measured overall cognitive workload. Specifically, we recorded the participants' Mental Demand, Effort, Frustration, and Perceived Performance. For all metrics except Perceived Performance, higher values indicate a higher workload on the participant, which indicates a poorer experience. For Perceived Performance, higher values indicate a better perception of their own performance, which indicates a better experience. Subscales regarding physical demand and temporal demand were excluded as the experiment was conducted online with no time constraints, and as such, these metrics do not speak to anything regarding the participant's mental model or experience as it relates to the task.

IV. Results

This section details the results of objective and subjective team decision-making metrics.

A. Final Shared Situation Awareness

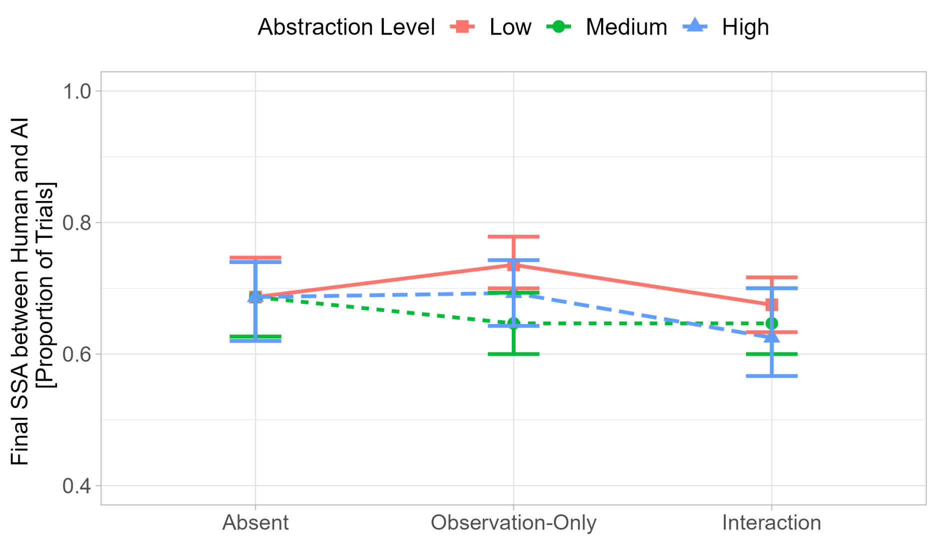

Fig. 5 Final Shared Situation Awareness [%] Between Human and 60% Accurate AI

We first assessed the final shared situation awareness (Final SSA) of the two teammates (yellow diamond #2 in Figure 1). Ideally the human would only agree with their AI partner 60% of the time (when the AI was correct). This would indicate that the human can recognize a limitation in their AI partner and adjust their mental model and actions accordingly to ensure successful team performance. Figure 5 shows the percentage of the Final SSA between the participant and the AI during the World State Assessment and Trajectory Evaluation phases. We see that having contextual information brings the human and AI closer to that 60% agreement at the medium level of abstraction. But when only observing the world state information, providing too much or not enough information can weaken shared understanding of the world state. Providing interaction with the world state information brought the Final SSA closer to the ideal 60% agreement. This indicates that, if added interaction to information is not possible, then care should be given to provide the most effective abstraction of information, and vis versa.

Table 1 presents the ANOVA for the fitted linear mixed-effects model of the Final Shared Situation Awareness between the participant and the AI Mission computer before the AI's decision was revealed (yellow diamond #2 in Figure 1). The results indicate that the World State Awareness significantly impacts the Final SSA between the participant and AI. The level of abstraction at which the world state information is displayed in not statistically significant in influencing Final SSA. Neither of the covariates (*Faith in People* or *Faith in Technology*) were statistically significant.

Table 1 ANOVA for Final Shared Situation Awareness

	df, error	F	P
WSAwareness	2, 120	3.485	0.0338
Abstraction	2, 120	1.134	0.3252
Faith in Tech	1, 120	0.529	0.4682
Faith in Persons	1, 120	1.114	0.2933
WS-A:Abs	4, 120	1.332	0.2621

B. Final Decision Agreement

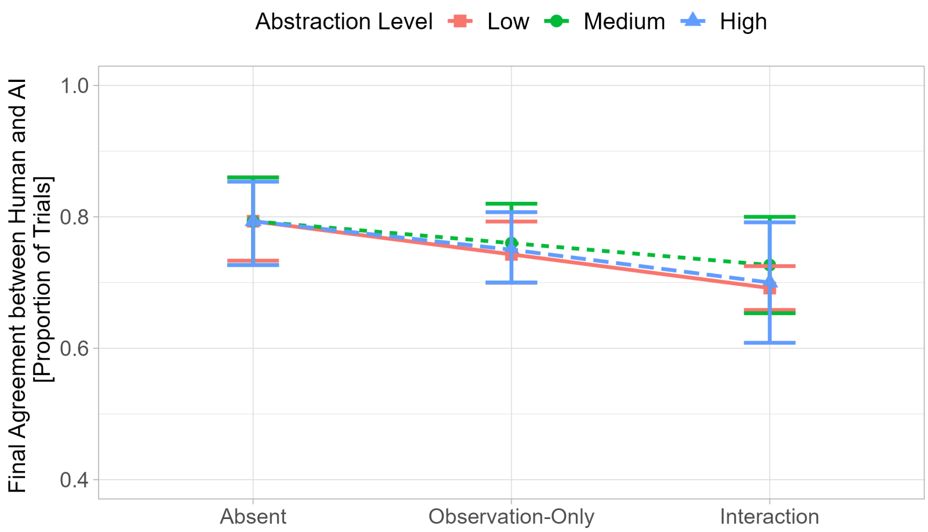

Fig. 6 Final Agreement [%] Between Human and 60% Accurate AI

Figure 6 shows the team's Final Decision Agreement (yellow diamond #3 in Figure 1). Because the AI in this experiment has a reliability of 60%, we would expect final team agreement to be 60% if the human was 100% correct in their judgement. From Figure 6, we see that having contextual information brings the human and AI closer to that 60% agreement at every abstraction level. A linear mixed effects analysis for this metric revealed that World State Awareness significantly impacts the Final Decision Agreement between the participant and AI. The level of abstraction at which the world state information is displayed in not statistically significant in influencing Final Decision Agreement (Table 2). Neither of the covariates (*Faith in People* or *Faith in Technology*) were statistically significant.

Table 2 ANOVA for Final Agreement

	df, error	F	P
WSAwareness	2, 120	5.439	0.0055
Abstraction	2, 120	0.435	0.6483
Faith in Tech	1, 120	1.598	0.2087
Faith in Persons	1, 120	0.031	0.8605
WS-A:Abs	4, 120	0.339	0.8511

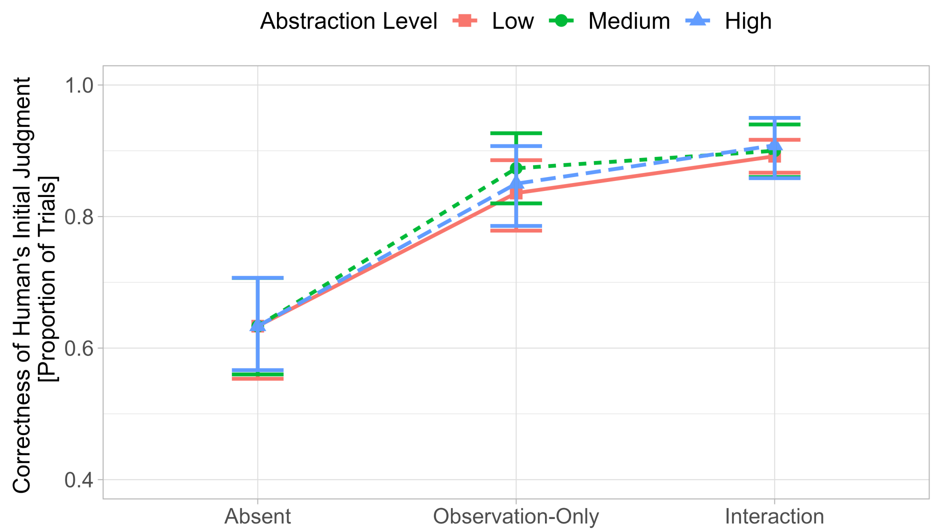

Fig. 7 Human's Initial Judgement [%] on Decision Before Seeing AI's Suggestion

C. Human's Initial Judgement

Figure 7 shows the Human's Initial Judgement of what decision to take before seeing the AI's suggestion. Participants who had access to world state information had much higher correct initial judgement (around 85-90% of trials) than those who had no contextual information (~ 65% of trials), regardless of the abstraction level at which the world state information was displayed.

Table 3 ANOVA for Human's Initial Judgement

	df, error	F	P
WSAwareness	2, 120	68.300	<0.0001
Abstraction	2, 120	0.204	0.8159
Faith in Tech	1, 120	10.584	0.0015
Faith in Persons	1, 120	3.399	0.0677
WS-A:Abs	4, 120	0.254	0.9065

The results from the ANOVA for this metric (Table 3) indicate that the World State Awareness is statistically significant in predicting how accurate the human's initial judgement is before seeing the AI's suggestion. *Faith in Technology* was also weakly significant in influencing the Human's Initial Judgement. No other effects were significant.

D. Task Performance

Figure 8 shows the final *Task Performance*. Participants who had access to world state information made the correct final decision a higher proportion of the time than those who had no additional information. This indicates that providing contextual information to the human improves their decision making and the team's overall performance, in spite of an unreliable AI teammate, and regardless of the abstraction level at which the information is displayed.

The results from the ANOVA for this metric (Table 4) indicate that the World State Awareness is statistically significant in influencing how correct the participant's final decision is. Our variable of interest (abstraction level) was not statistically significant, nor were the covariates (*Faith in People* or *Faith in Technology*).

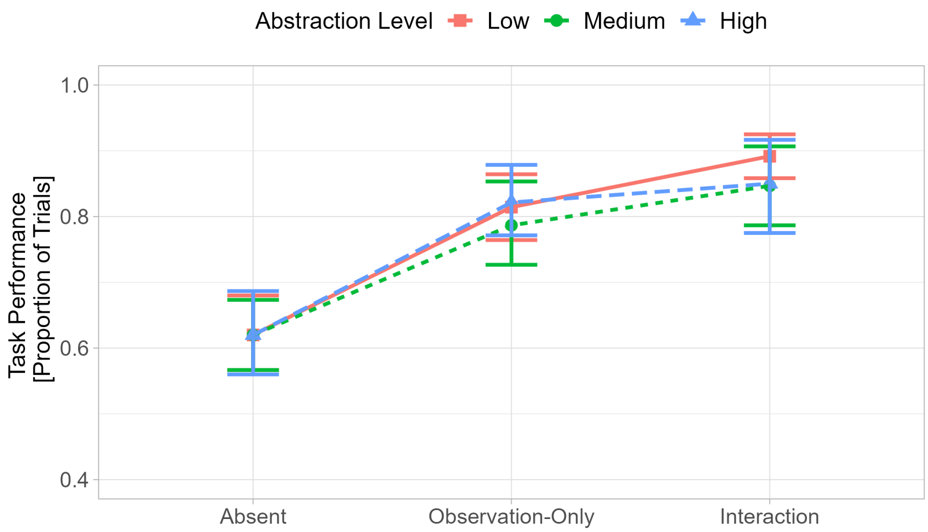

Fig. 8 Task Performance [%]

Table 4 ANOVA for Task Performance

	df, error	F	P
WSAwareness	2, 120	44.778	<0.0001
Abstraction	2, 120	0.193	0.8244
Faith in Tech	1, 120	0.008	0.9289
Faith in Persons	1, 120	0.038	0.8467
WS-A:Abs	4, 120	0.144	0.9654

E. Sway

Figure 9 depicts the percentage of scenarios in which the participant reconsidered their initial decision to align with the AI after seeing what the AI suggested during the evaluation phases. We see that the low abstraction level is most effective in yielding less team disagreements, probably because participants have all of the information they could possibly need to make an informed decision and therefore have sound judgement in the first place. The results from the ANOVA for this metric (Table 5) indicate that no main effects were statistically significant in predicting when the participant will change their decision to align with the AI's suggestion. However, *Faith in Technology* was weakly statistically significant in influencing how likely the participant was to be swayed by the AI's suggestion.

Table 5 ANOVA for Resolved Team Disagreements

	df, error	F	P
WSAwareness	2, 120	2.11626	0.1250
Abstraction	2, 120	1.77348	0.1742
Faith in Tech	1, 120	4.65578	0.0329
Faith in Persons	1, 120	0.65507	0.4199
WS-A:Abs	4, 120	0.144	0.9654

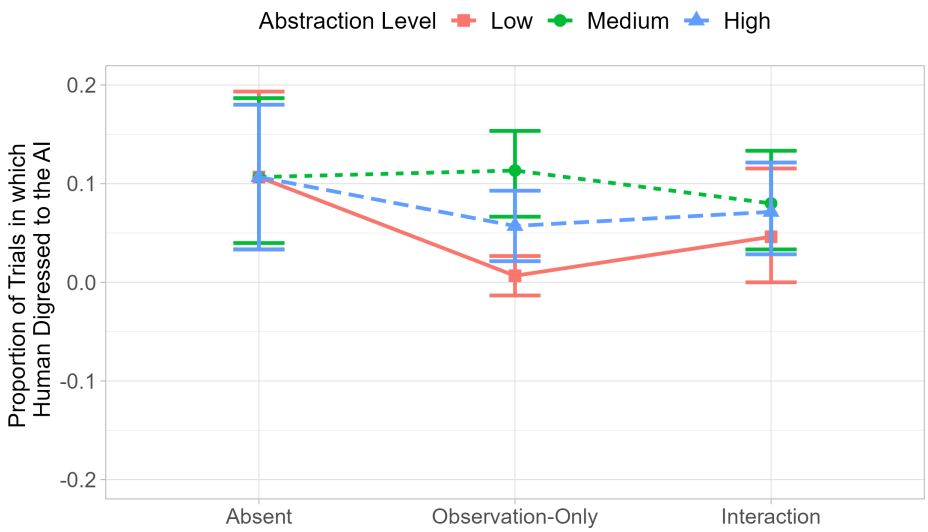

Fig. 9 Resolved Team Disagreements [%] Post-AI Suggestion

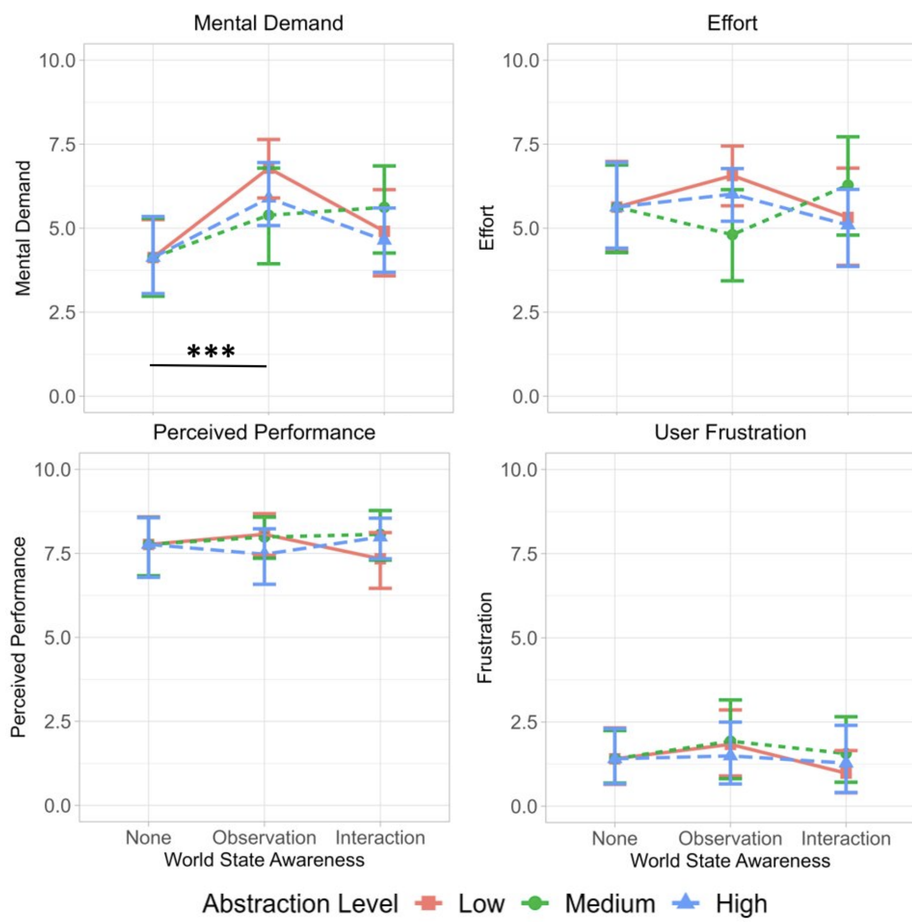

Fig. 10 Composite Results of NASA TLX Questionnaire

F. NASA TLX

We assessed the subjective mental workload of each participant using the NASA TLX assessment at the end of the experiment (Figure 10). While there is some variation in effort and mental demand between the levels of abstraction, both frustration and perceived performance showed very similar trends between abstraction levels. There was a statistically significant difference in the reported mental demand of participants who had only observed the world state information (F value = 7.3187) as opposed to having no world state information.

G. i-THAu

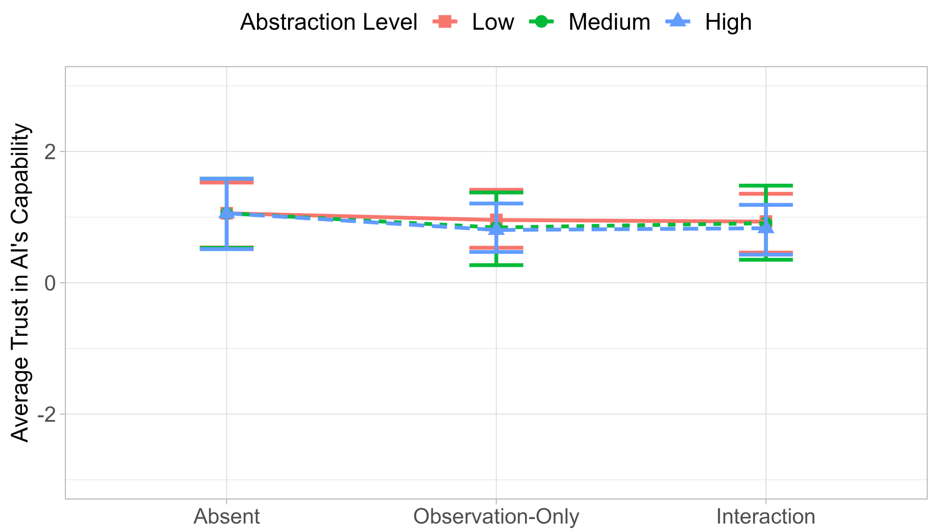

Fig. 11 Post-Experiment i-THAu Metric of Human's Trust in AI's Capabilities

Figure 11 shows the human's trust in the AI's capabilities to complete the task measured in the post-experiment *i-THAu* assessment. There is virtually no variation in trust in the AI's capability between the abstraction levels.

V. Concluding Discussion

The purpose of this experiment was to evaluate whether observed performance benefits [1] of structuring a human-AI shared decision-making process hold when the information is displayed to the participant at various levels of abstraction.

The same trends appear for the team's final decision agreement, the human's initial judgement, and their final task performance (Figures 6 - 8) when shown (or additionally having to interact with) contextual information despite the level of abstraction at which information is displayed. Additionally, the level of abstraction was not statistically significant in predicting any objective metrics.

Based on the team's final decision agreement and their resolved disagreements (Figures 6 and 9), we can see that those who did not have access to the World State conditions had a tendency to be swayed by their AI partner to make the correct decision, while those who did observe or interact with the world state conditions had higher initial agreement with the AI. While there seems to be some variation between abstraction levels amongst those who only observed the world state information, the overall proportion of trials in which participants were swayed to (correctly or incorrectly) agree with the AI is still under 10%.

Holistically, the results indicate that crucial metrics in team decision-making are robust to the level of abstraction at which information is displayed, and that previous findings from Srivastava et al. [1] hold. This indicates that the technique of providing contextual information upfront to the human decision maker can be useful in increasing their situation awareness and judgment, and in calibrating trust in their AI partner irrespective of the level of information. This can prove useful in time-sensitive situations in which there is insufficient time to parse through lots of information to figure out what is salient to users (low level of abstraction), and in well-characterized situations in which the human needs only a limited amount of knowledge to boost the results of their decision-making (high level of abstraction).

Acknowledgments

This work was funded by Sandia National Laboratory (SNL) with Dr. Paul Schutte serving as Program Manager. This work is solely that of the authors and does not represent an official SNL position.

References

[1] Srivastava, D. K., Lilly, J. M., and Feigh, K. M., "Improving Human Situation Awareness in AI-Advised Decision Making," *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, 2022, pp. 1–6. https://doi.org/10.1109/ICHMS56717.2022.9980783.

[2] Hendrick, D., "Complexity Theory and Conflict Transformation: An Exploration of Potential and Implications," 2009.

[3] Wing, J., "Computational thinking and thinking about computing," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, Vol. 366, 2008, pp. 3717–25. https://doi.org/10.1098/rsta.2008.0118.

[4] Shih, W. C., "Increasing the Level of Abstraction as a Strategy for Accelerating the Adoption of Complex Technologies," *Strategy Science*, Vol. 6, No. 1, 2021, pp. 54–61. https://doi.org/10.1287/stsc.2020.0116, URL https://ideas.repec.org/a/inm/orstsc/v6y2021i1p54-61.html.

[5] Kolb, J., Srivastava, D., and Feigh, K. M., "The Effects of Inaccurate Decision-Support Systems on Structured Shared Decision-Making for Human-Robot Teams," *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, IEEE, 2023.

[6] Hart, S. G., *NASA-task load index (NASA-TLX); 20 years later*, 2006.

[7] Razin, Y., "Interdependent Trust for Humans and Automation Survey," ???? URL https://sites.gatech.edu/feigh-lab/publications/.