Towards Shared Mental Models in Human-AI Teams

2023 IEEE Conference on Systems, Man, and Cybernetics

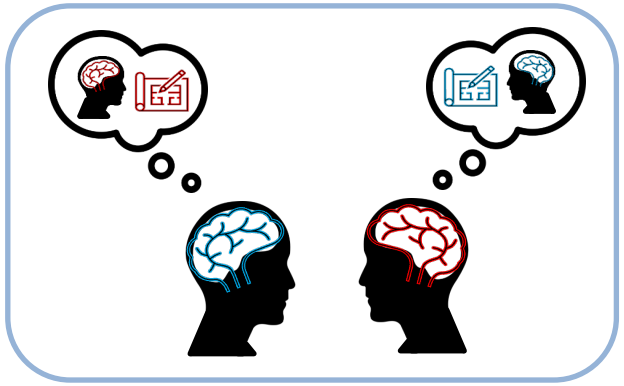
Sarah Walsh
Karen Feigh

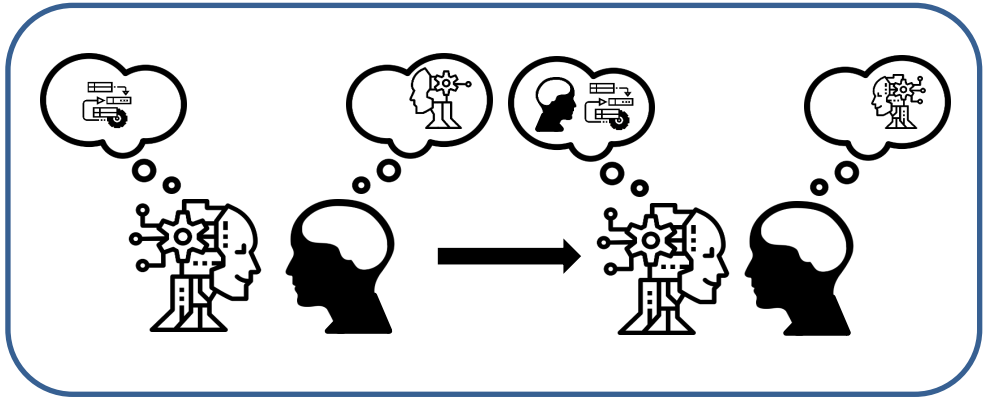
COGNITIVE ENGINEERING CENTER
GEORGIA INSTITUTE OF TECHNOLOGY

Why create shared mental models in human-AI teams?

A shared mental model is a shared perception of goals and actions through effective communication and an understanding of their fellow team members' goals and likely methods (Orasanu, 1990)

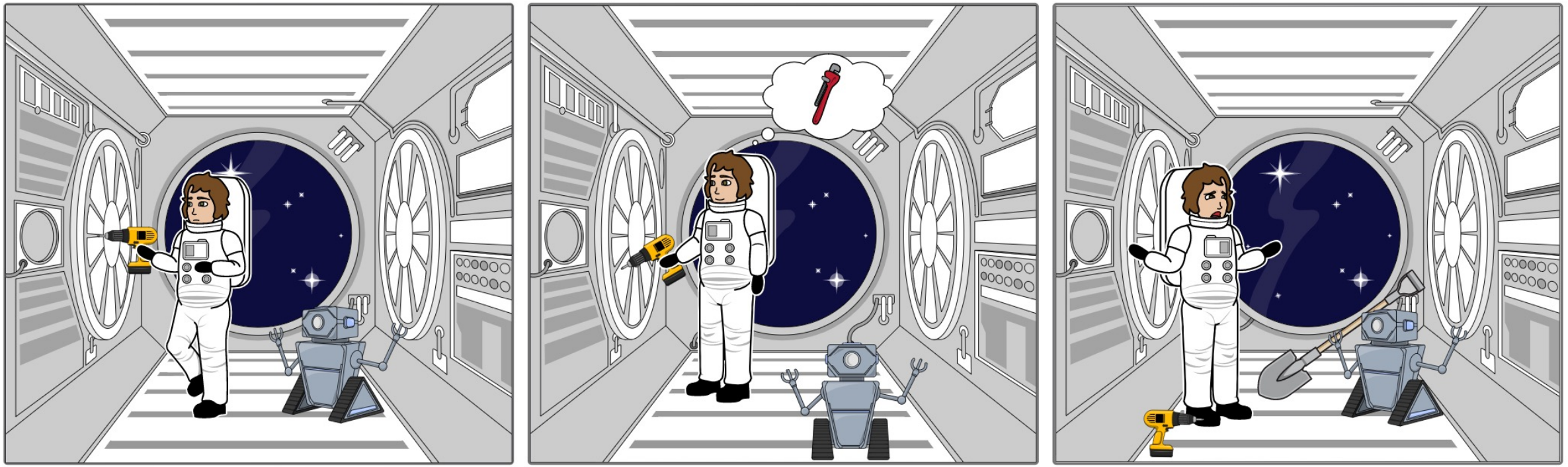
Human teams are most effective when the members of the team utilize a shared mental model (Fiore, Salas, & Cannon-Bowers, 2001)





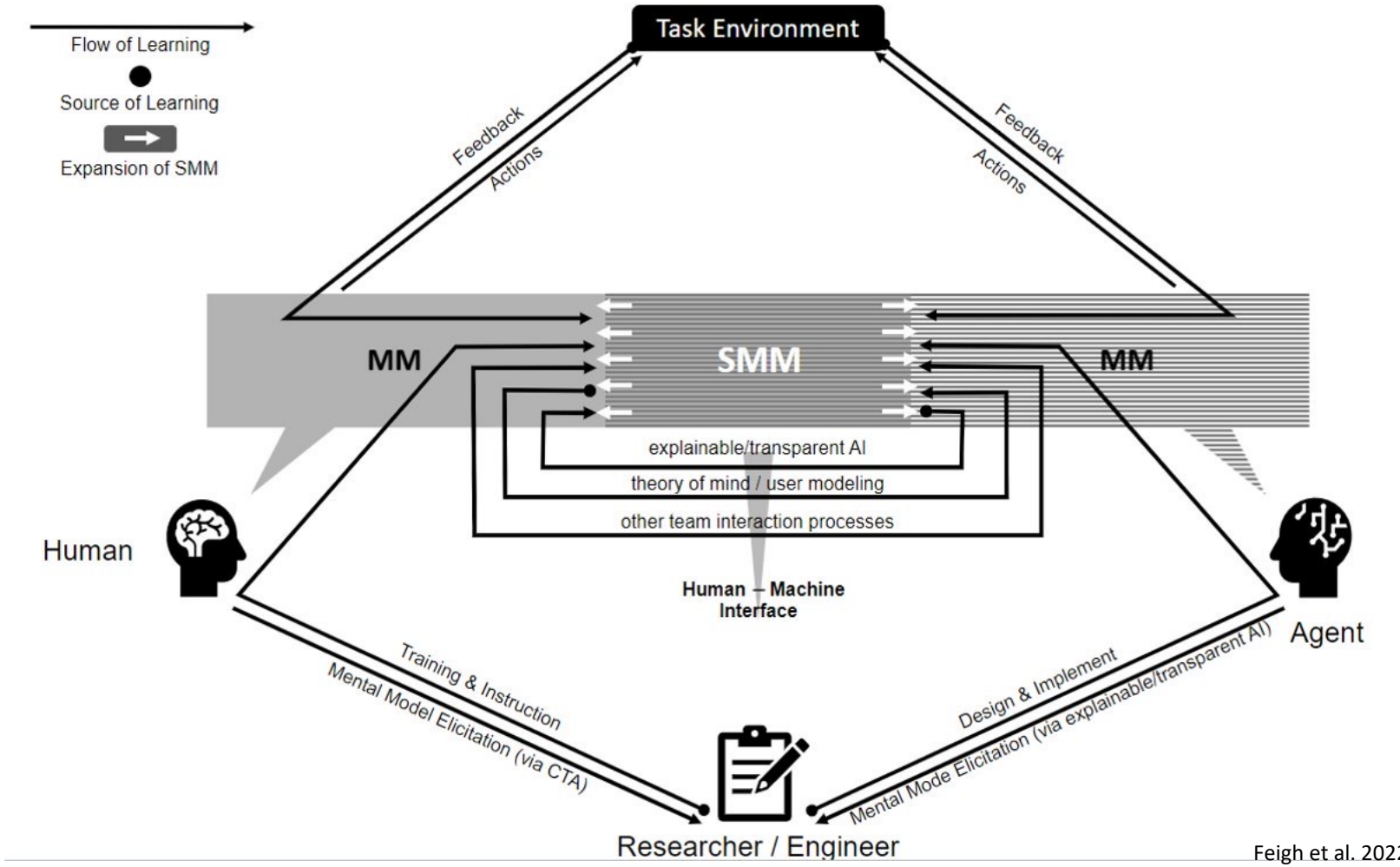
By utilizing the concept of a shared mental model, human-AI teams can become more effective, and reduce the dissonance between humans and AI systems (Human-AI SMM Hypothesis) (Scheutz, 2017)

Inconsistent Expectations in Human-Agent Teaming



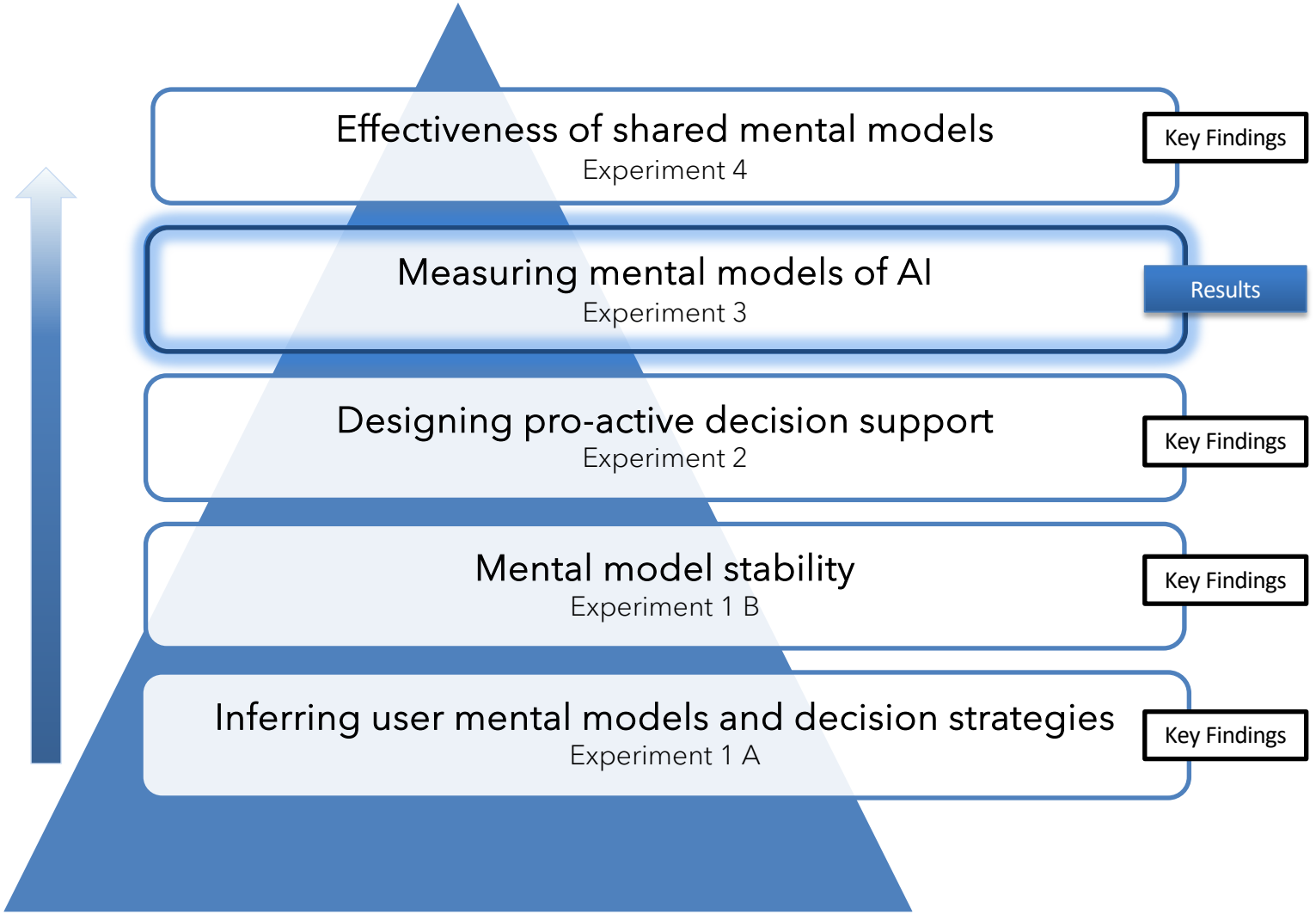
Each teammates must seek to anticipate the behavior of the other to appropriately support joint work
Expectation alignment is critical in SMMs and for team success

Conceptual Model of Human-AI Shared Mental Models



Feigh et al. 2022

Overview and Technical Approach

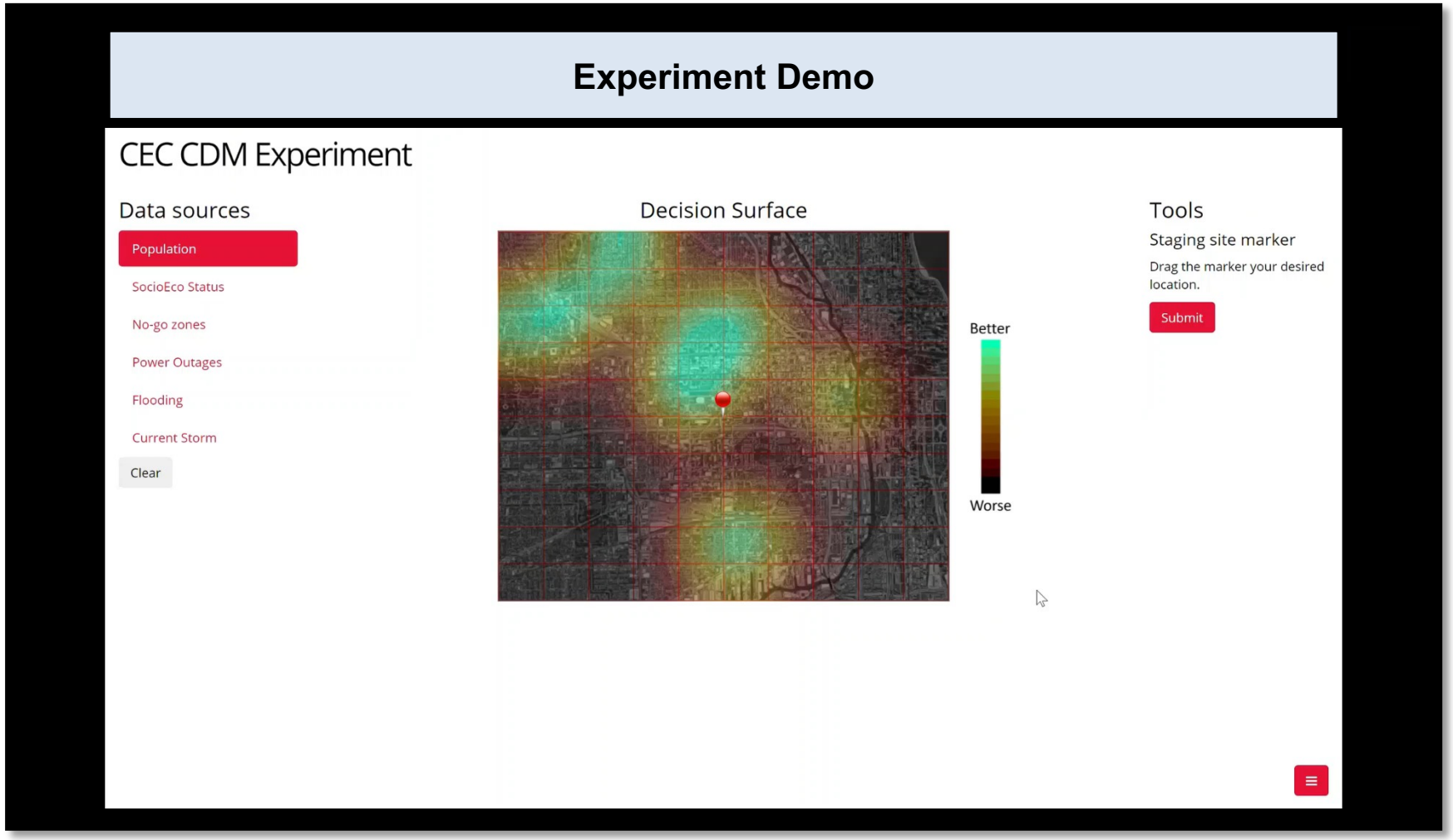


Team Model Inference: Inferring user mental models and decision strategies

1. Can we infer decision strategies from dynamic behavioral data in combination with decision accuracy?
2. How stable are people's decision strategies?
3. Can we classify these inferred decision strategies based solely on behavioral data?

COGNITIVE ENGINEERING CENTER

Experimental Interface with Geospatial, Dynamic Task



Key Findings
Experiment 1 A: Inferring Mental Models

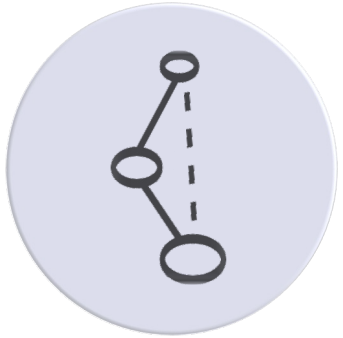
BEHAVIORAL DATA IS SUFFICIENT TO DETERMINE ARCHETYPES OF USER MENTAL MODELS THAT ARE PREDICTIVE OF PERFORMANCE

BEHAVIOR IS STABLE ENOUGH TO DETERMINE PREFERENCES AND TENDENCIES IN USER ARCHETYPES

BEHAVIOR APPEARS TO CONVERGE

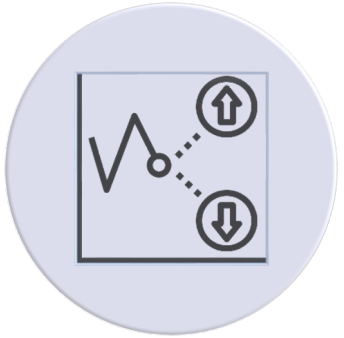
1. Walsh, S. E., & Feigh, K. M. (2022). Understanding Human Decision Processes: Inferring Decision Strategies From Behavioral Data. *Journal of Cognitive Engineering and Decision Making.*,
2. Walsh S.E. and Feigh, K.M. "Differentiating 'human in the loop 'decision process," in 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2021

Key Findings
Experiment 1 B: Stability and Predictability of Behavior



HEURISTICS AND COGNITIVE SHORTCUTS ARE USED THROUGHOUT TASKS

STABILITY (CONVERGENCE) VARIES BASED ON TASK COMPETENCY



PREDICTABILITY INCREASES WITH TASK FAMILIARITY

3. Narayanan, R., Walsh, S. & Feigh K. "Development of Mental Models in Decision-Making Tasks" *Accepted at HFES 2023*

Designing pro-active decision support

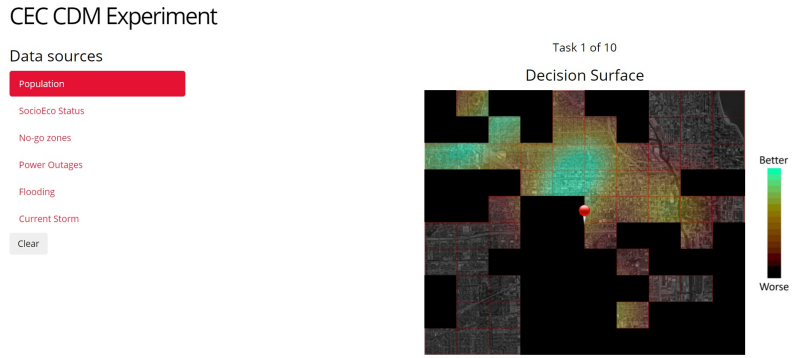
1. What form of decision support (heuristic or analytic) improves performance (accuracy, effort, time to complete)?
2. Does decision support that aligns with natural decision strategy improve performance over strategy-aid mismatch?

COGNITIVE ENGINEERING CENTER

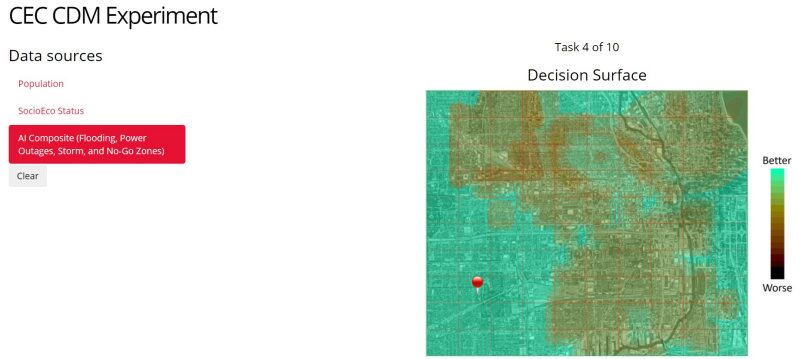
Decision Aid Breakdown

No Decision Aid-Control
- Decision Space = 600

Analytic Decision Aid
- Option space reduction from 100 to 50
- Decision Space = 300



Heuristic Decision Aid
- Attribute space reduction from 6 to 3
- Decision Space = 300



Key Findings

Experiment 2: Implementing a Decision Aid

FASTER DECISIONS

IMPROVED ACCURACY OF LOWEST PERFORMERS

LESS EFFORT- (FEWER INSTANCES OF INFORMATION ACCESS)

Heuristic decision support can lead to faster decisions with no degradation in performance

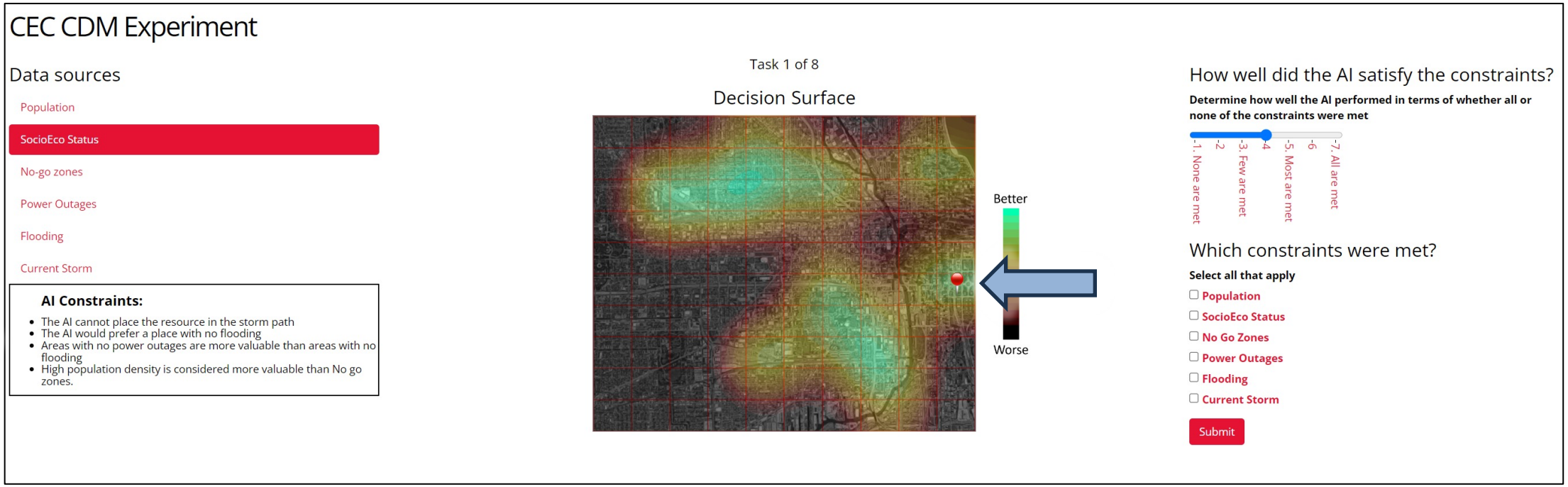
4. Walsh, Sarah E., and Karen M. Feigh. "Consideration of Strategy-Specific Adaptive Decision Support." *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*. IEEE, 2022.

Nontechnical Users Assessment of AI Performance and their Bias

1. Can an untrained user determine the accuracy of AI-decision support in a complex geospatial decision environment?
   - ✧ Can the user determine the source of the error?
2. Can an untrained user determine the accuracy of AI-decision support in a complex geospatial ***team*** decision environment?
   - ✧ In a team decision task, does the user bias towards their own goals over the team goals?

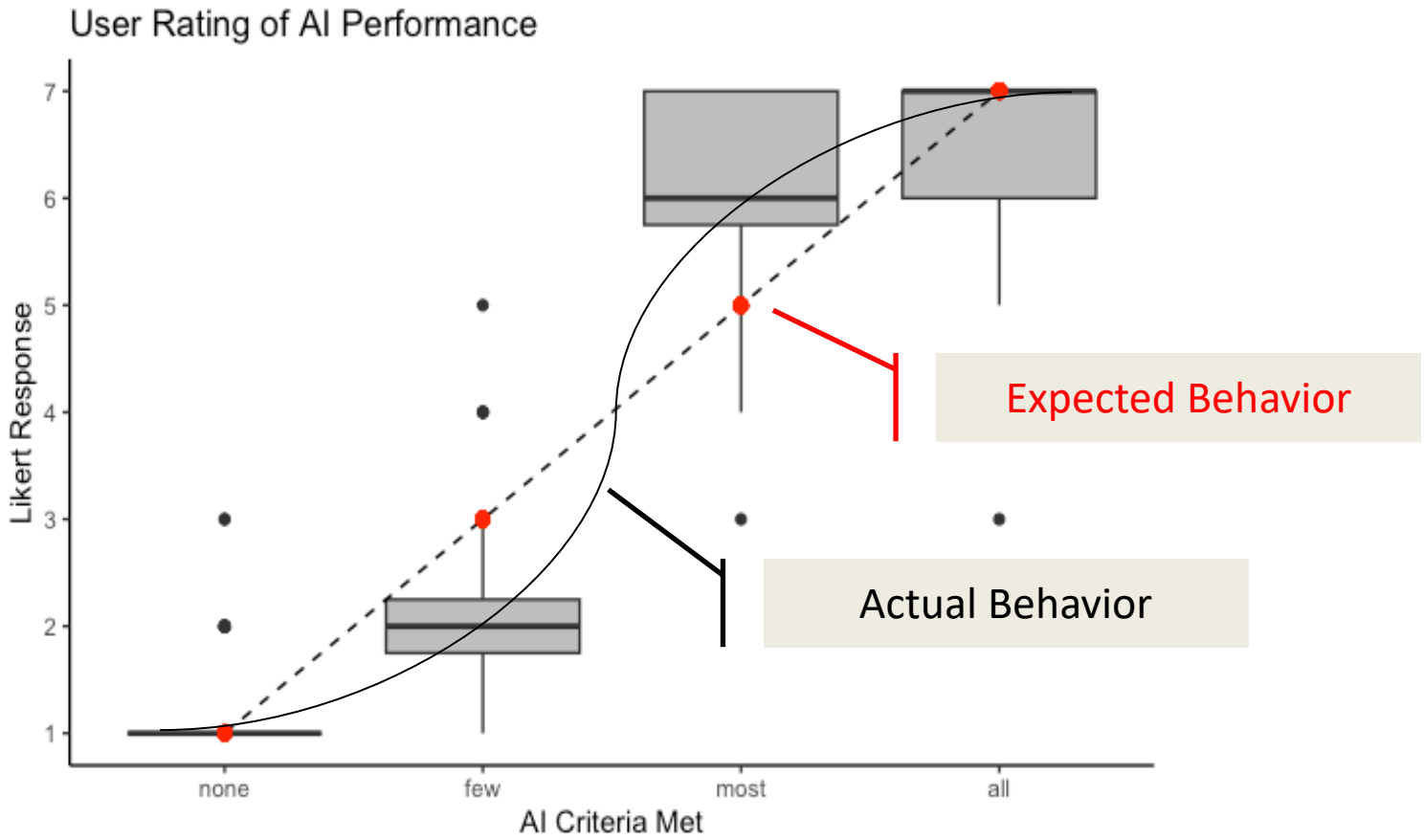
COGNITIVE ENGINEERING CENTER

Assessing the Mental Model of AI Error in Dynamic Geospatial Decision Task



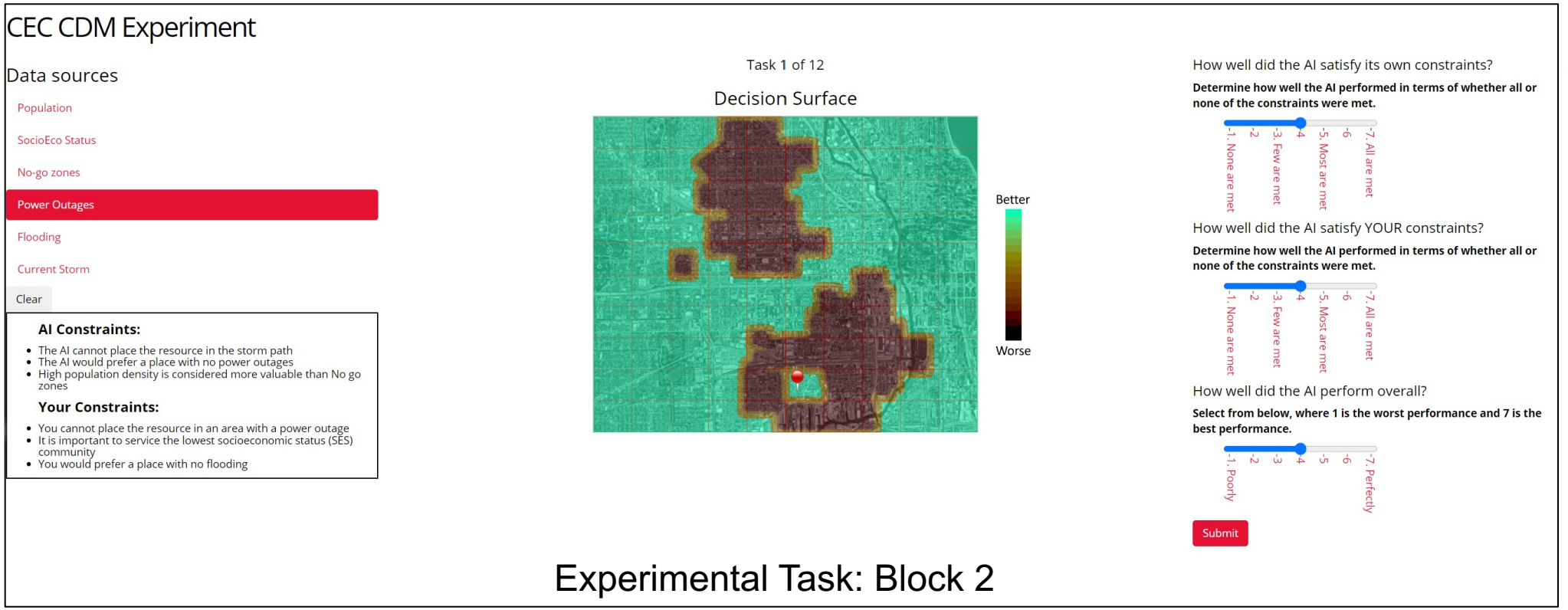
Experimental Task: Block 1

Does the accuracy of the AI impact how well users perform at assessing the AI performance?



Users tended to categorize performance into discrete bins rather than on a continuous scale (Binary Bias)

Assessing the User Bias in Dynamic Geospatial Decision Task

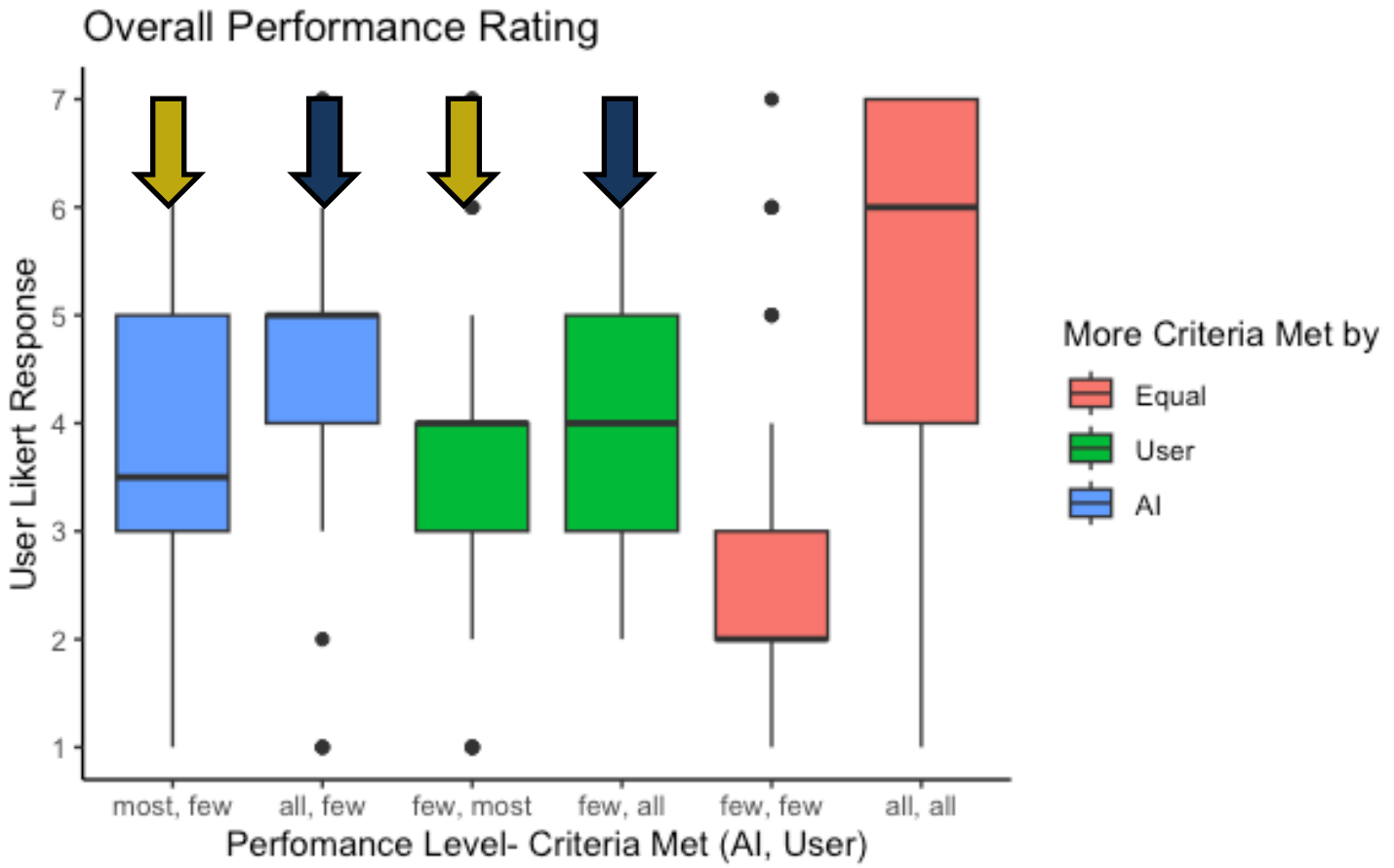


Experimental Task: Block 2

In a team decision task, does the user bias towards their own goals over the team goals?

No indication that the users bias towards one set of criteria being met over the other was found

Users are able to assess the team performance with objectivity



Key Findings

Experiment 3: Nontechnical Users Assessment of AI Performance and their Bias

USERS WERE ABLE TO ACCURATELY DETERMINE HOW WELL THE AI PERFORMED

USERS EXHIBIT A BINARY BIAS

USERS PRIORITIZE TEAM GOALS OVER THEIR OWN INDIVIDUAL GOALS

Binary biases effect leads novices and experts to create false dichotomies (Fisher, 2018). Users may tend to bin these systems as 'good' or 'bad' leading to an over reliance and trust on some systems and misuse or disuse on others.

5. Walsh S. and Feigh K. "Mental Models of AI Performance and Bias of Nontechnical Users" *Accepted at SMC 2023*

Effectiveness of a Shared Mental Model

Primary Research Question: Does a limited SMM (more accurate team model) improve the decision making metrics (performance, workload, time to complete, compliance with AI)?

1. Is there a benefit to providing a Team Model for HAT tasks?
2. Is there a benefit to having a two-way model v. a one-way?
3. Can a Team Model make up for dissonance in task understanding?

COGNITIVE ENGINEERING CENTER

Components of a Shared Mental Model

Experiment Design: HAT Effects of Shared Mental Model

		AI Model Levels	
		Version 1: Optimize Team Score	**Version 2: Optimize AI Score**
User Model Levels	**MM 1: Complete Task and Team Model**	Complete Shared Mental Model	----------
	MM 2: Complete Team Model	Bi-directional Team Model	Uni-directional Team Model (User model of AI)
	MM 3: Incomplete Task and Team Model	Uni-directional Team Model (AI model of User)	No Team Model

MM Verified with Post-instructions Quiz

Key Findings

Experiment 4: Effectiveness of a Shared Mental Model

Users score highest with a complete SMM

AI Team Model reduces time and effort

Two-way Team Model improves score compared to no Team Model

- User can compensate with additional time and effort

AI Team Model reduces the users' workload

A complete SMM decreases frustration

Towards Shared Mental Models in Human-AI Teams

2023 IEEE Conference on Systems, Man, and Cybernetics

Sarah Walsh
sewalsh@gatech.edu
Karen Feigh

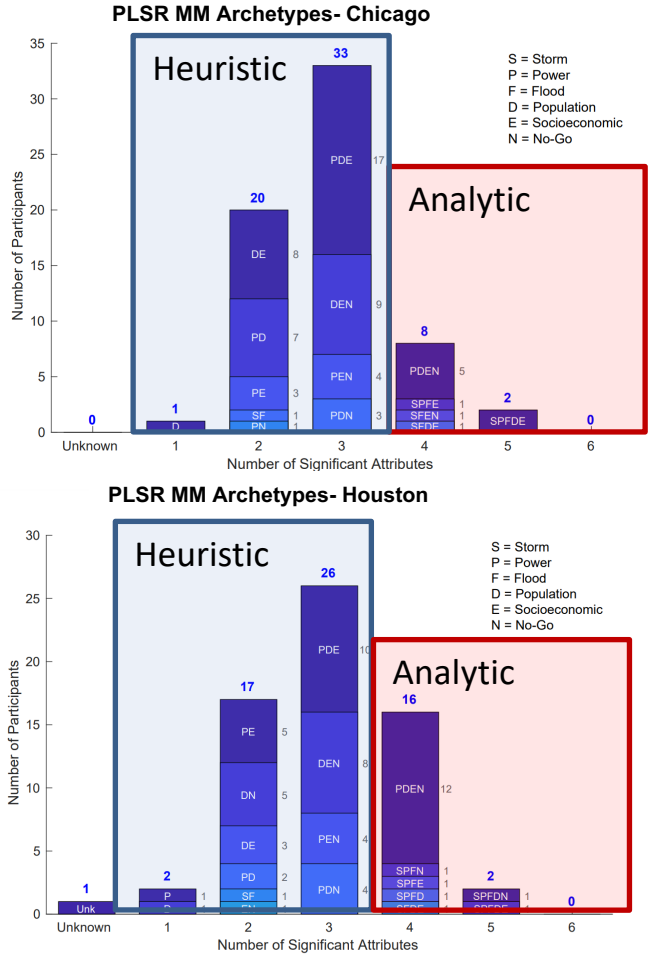
COGNITIVE ENGINEERING CENTER
GEORGIA INSTITUTE OF TECHNOLOGY

Experiment 1 Backup

COGNITIVE ENGINEERING CENTER

Results: Classification

What was the distribution of Mental Model Archetypes for each Task Domain?



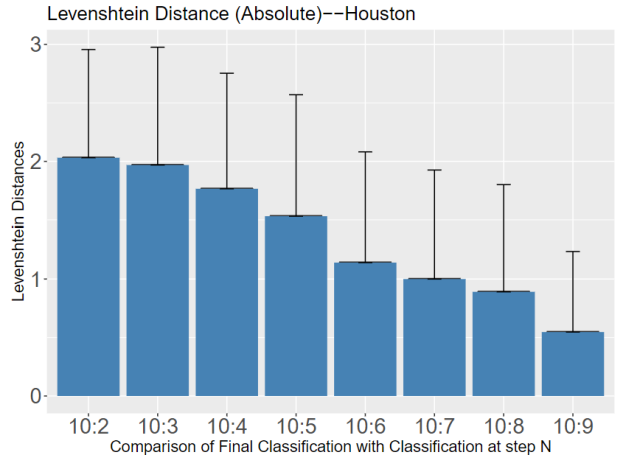
Classifying individuals into inferred archetypes based solely on observable behavioral data?

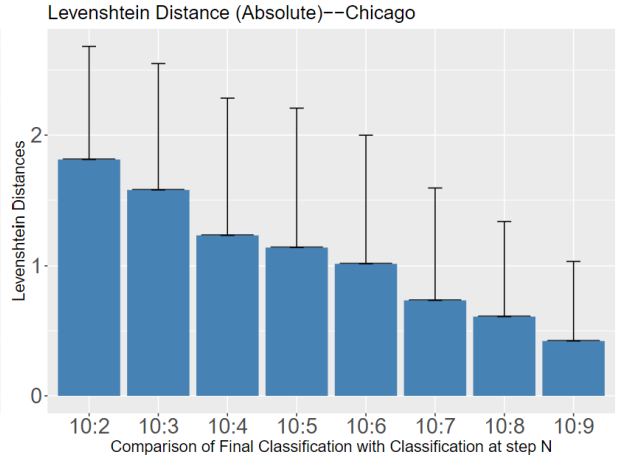
Chicago			
OOB estimate of error			7.8%
Confusion Matrix:			
	Analytic	Heuristic	Class Accuracy
Analytic	373	27	93.3%
Heuristic	38	394	91.2%

Houston			
OOB estimate of error			19.1%
Confusion Matrix:			
	Analytic	Heuristic	Class Accuracy
Analytic	319	56	85.1%
Heuristic	85	273	75.7%

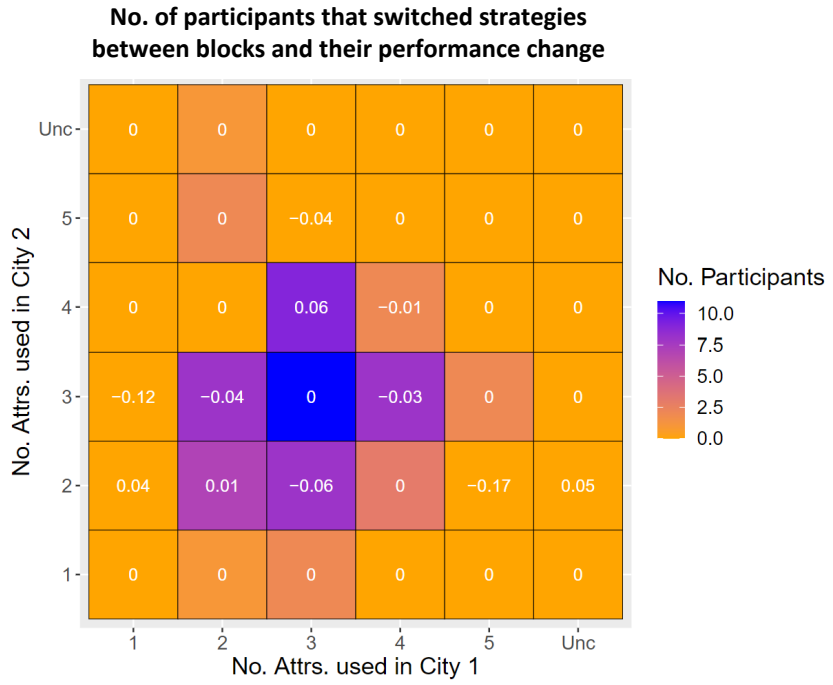
Results: How stable are the archetype groupings?

Stability within each Block (City)



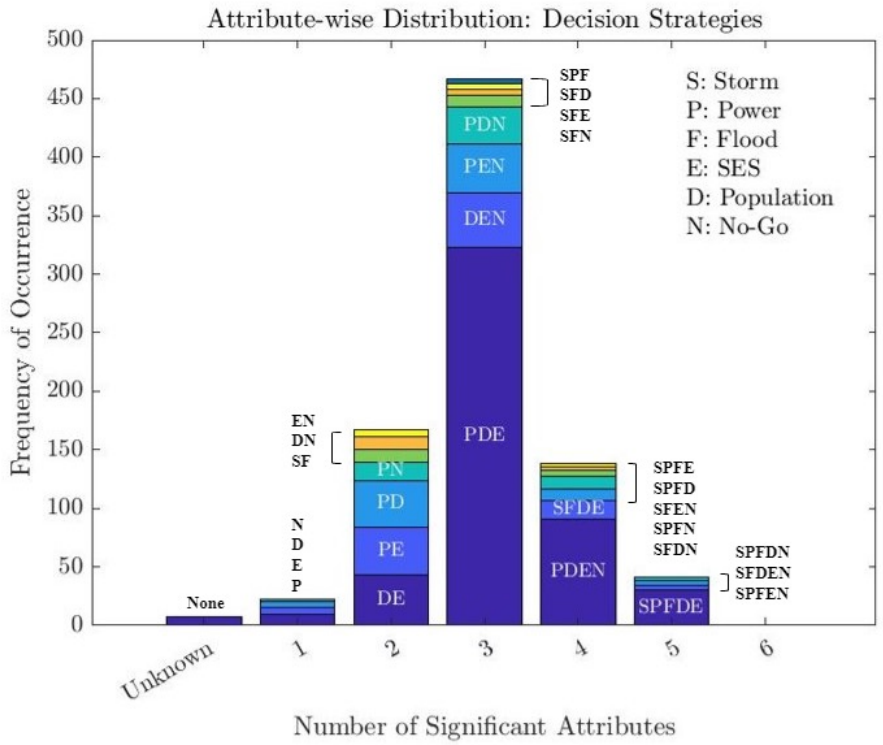


Stability between Blocks (Cities)



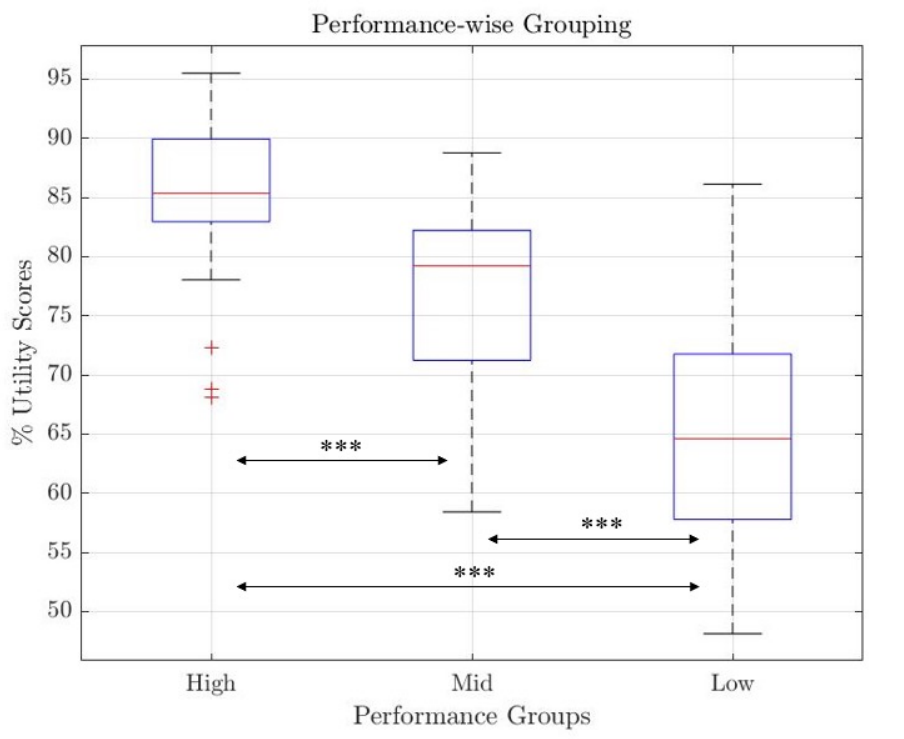
We have extended this study (Experiment 1.5) to 30 time steps to check for convergence

Inferring Decision Strategies using Behavior Data

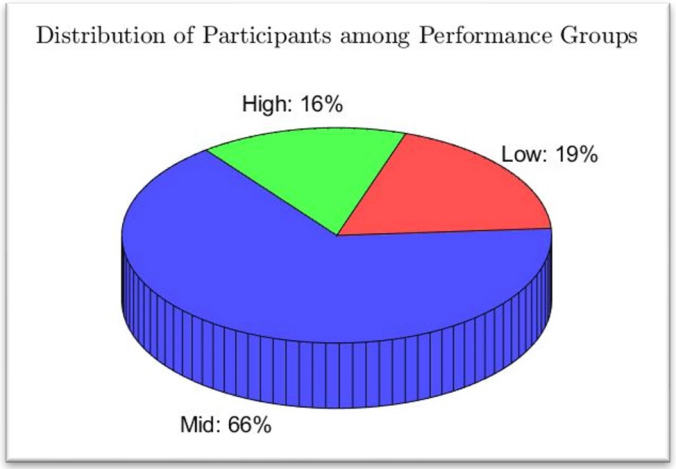


- Majority participants used 3-attributes to inform their decisions
- Followed by 2-attribute strategies
- Together, they constitute 76% of all strategies
- 3% of all strategies were ‘Take-the-Best’
- None with an equal weighting scheme
- 7 instances where participants acted arbitrarily (no strategy)
- Power (P), Population (D), and SES (E) were the most popular (visually complex)

RQ 2: Stability and Predictability Assessment



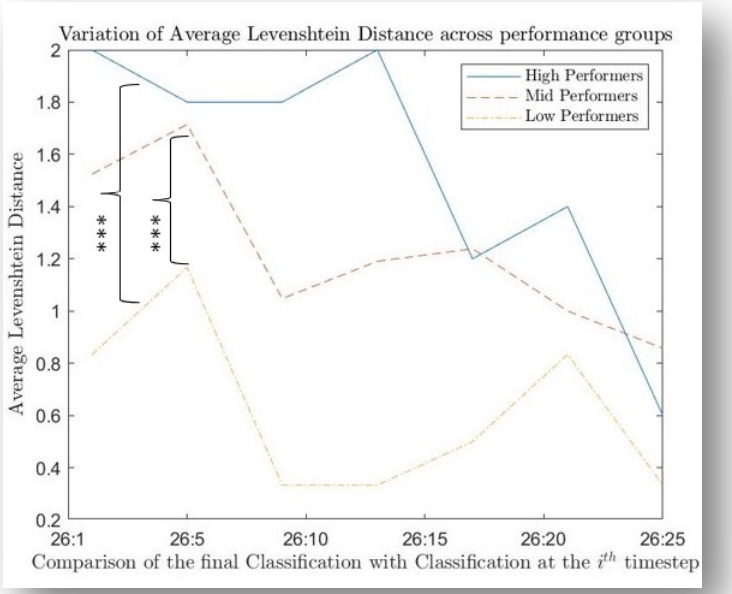
- High ≥ M + 1 SD
- M – 1 SD < Mid < M + 1 SD
- Low ≤ M – 1 SD



*** p < 0.01; ** p < 0.05; *p < 0.1

Performance groups are significantly distinct!

Stability (Convergence)



*** p < 0.01; ** p < 0.05; *p < 0.1

- Convergence towards final strategy is observed among all participant groups
- Significant correlation (p < 0.01) b/w change in strategy and performance among high performers
- Least correlation among the lowest performers
- High performers adapt then settle → reward seekers
- Low performers settle early → risk averse

Pearson product-moment

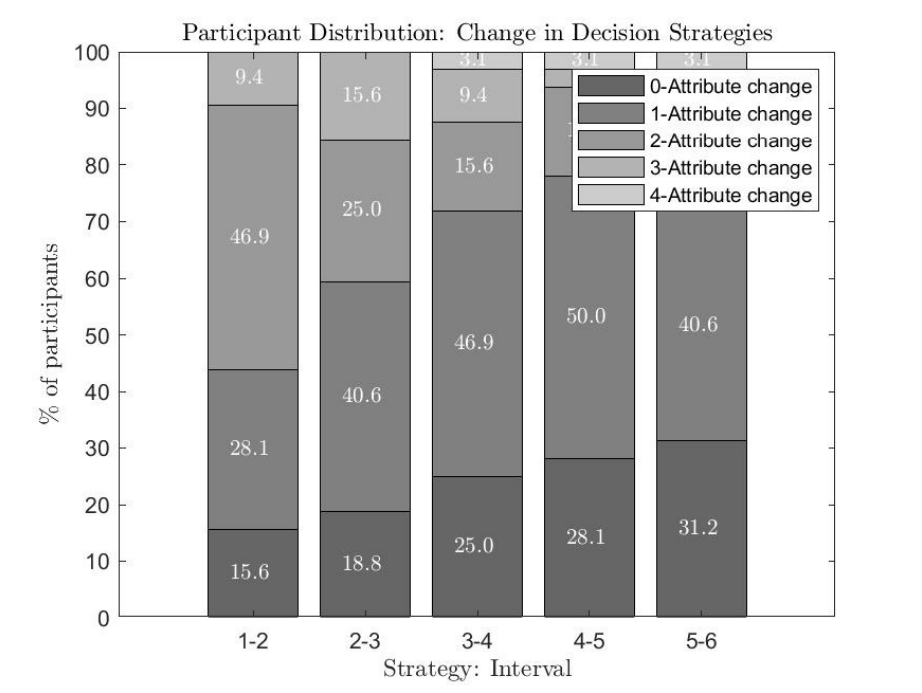
Group	R-score	P-score
Top	0.6004	0.0015***
Mid	0.5868	0.0020***
Low	0.1849	0.3761

Spearman-rank correlation

Group	R-score	P-score
Top	0.7315	3.25e-05***
Mid	0.5996	0.00153***
Low	0.2027	0.3309

Adaptability varies with competency levels

Predictability (Consistency)



- Marginal change in strategies
- Quantified by LD between consecutive strategies
- Proportion of participants with LD = 0/1 goes up monotonically over time
- Non-significant correlation with performance variation across consecutive timesteps
- Lesser variations in strategies regardless of performance improvement

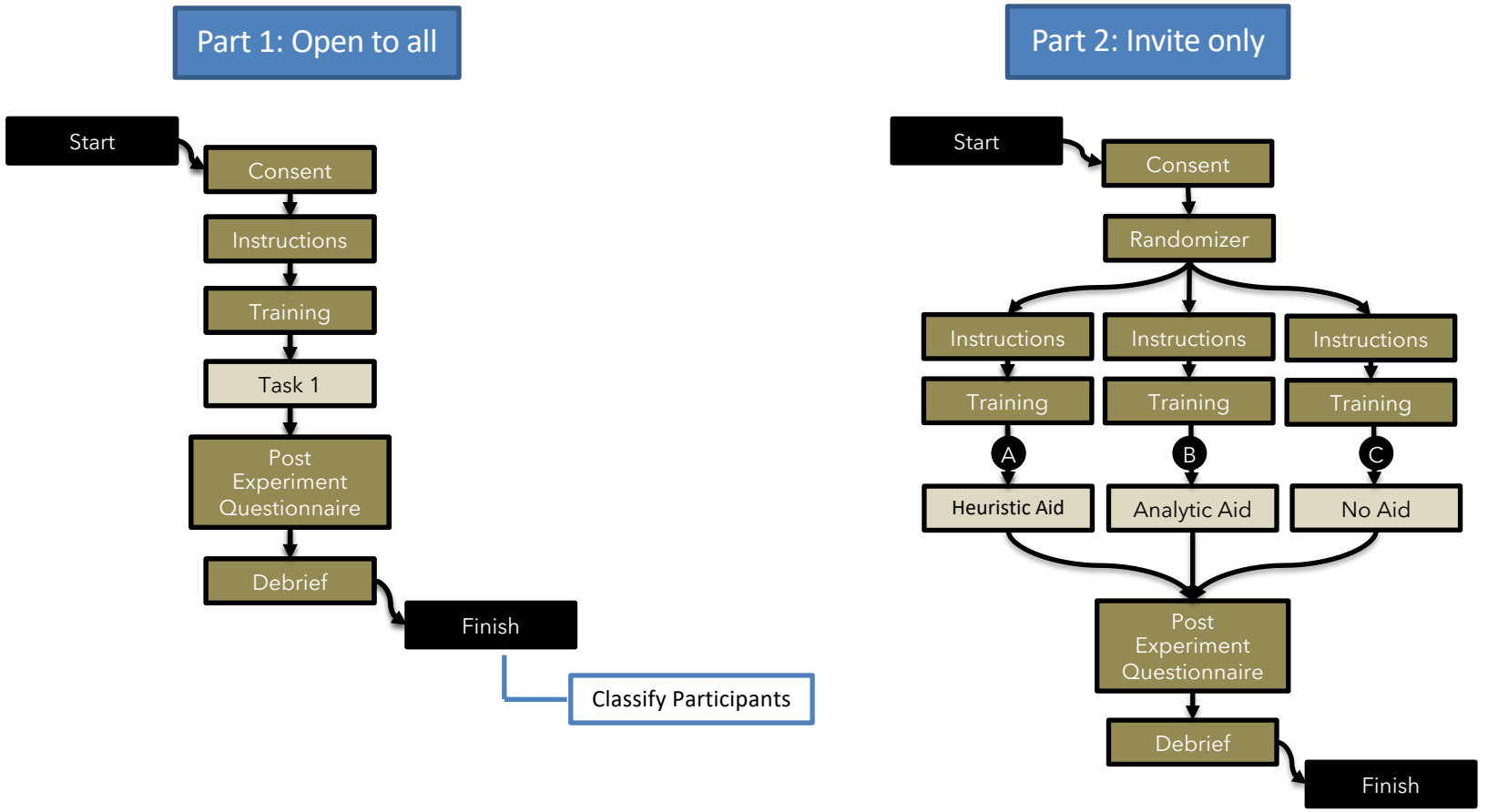
User predictability increases with task familiarity

Experiment 2 Backup

COGNITIVE ENGINEERING CENTER

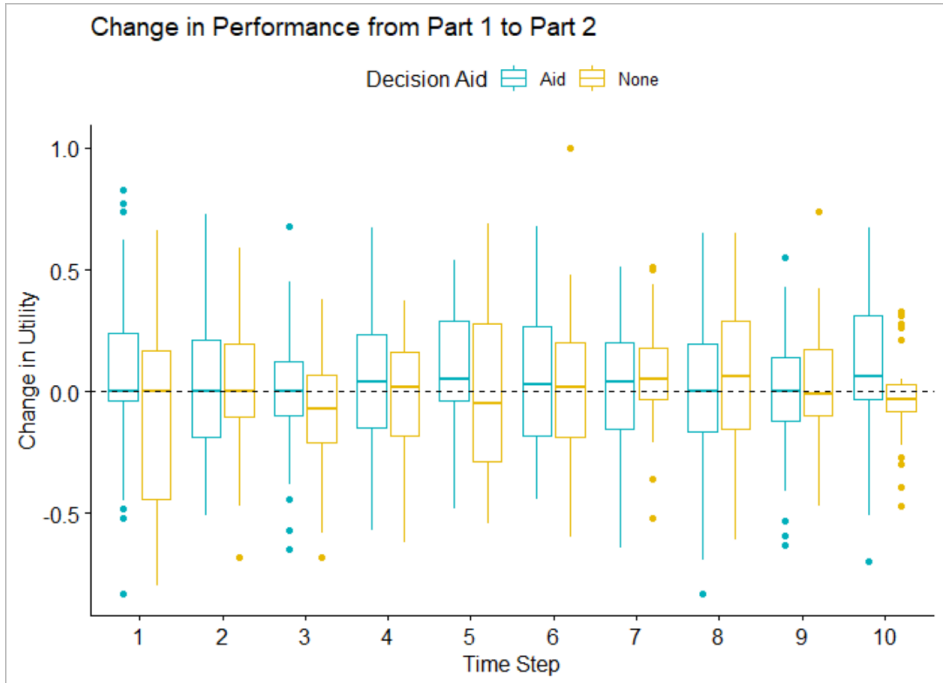
Experiment Design

Assess benefits of altering aid based through performance and workload



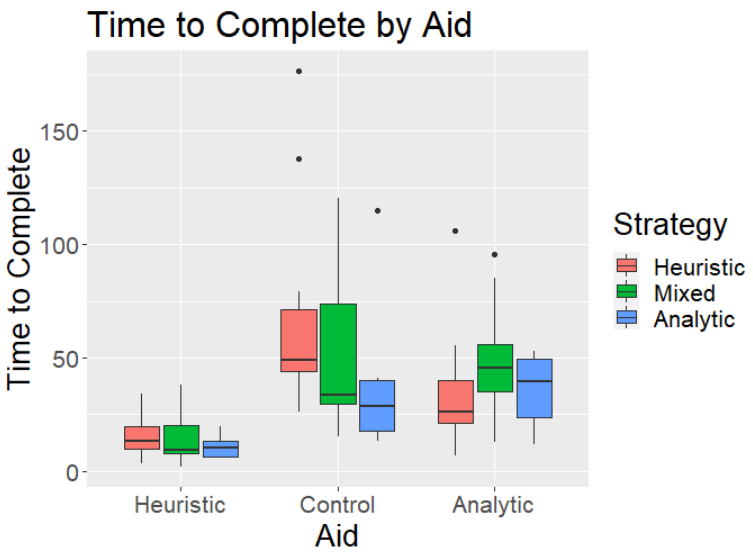
Change in Accuracy: Aid v. No Aid

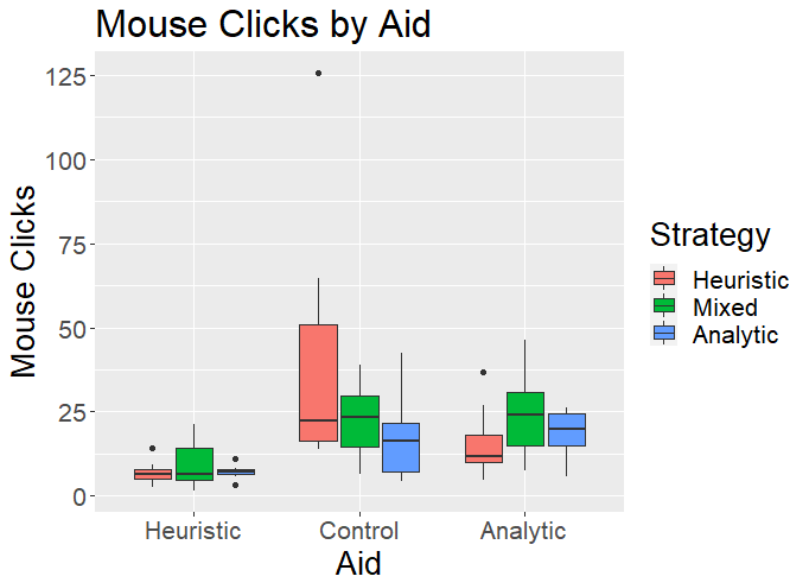
Change in Decision Making accuracy from Part 1 to Part 2



- There was no improvement (p=0.5) between Part 1 and Part 2 by participants that were not given an aid
- An ANOVA showed that there was significant improvement (p=0.0059) in decision making accuracy from those participants that were given a decision aid in Part 2

Effect on Effort (Time, Mouse Clicks)

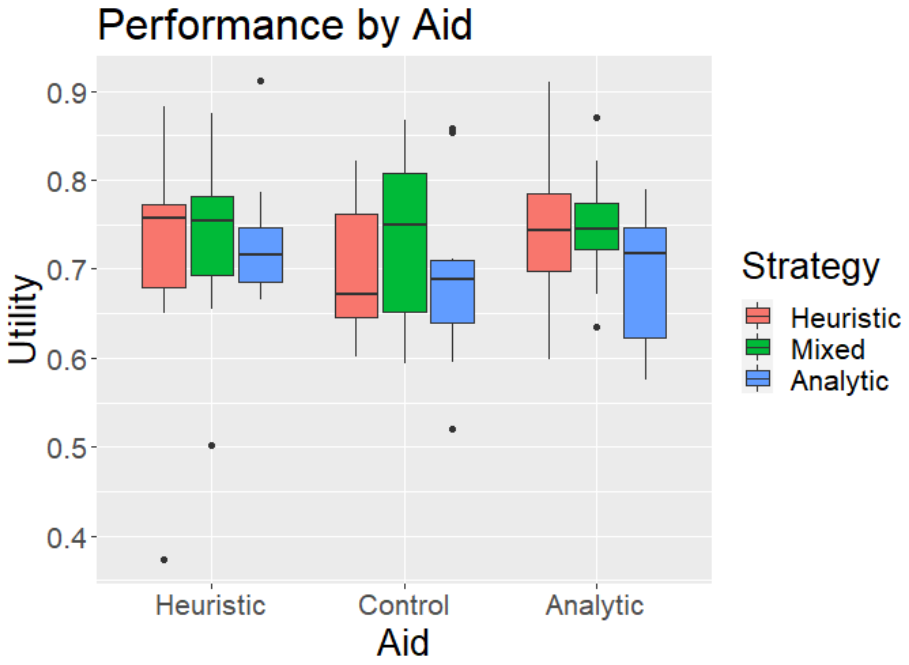




- **Time to Complete:** An ANOVA showed decision aid does impact (p=1.7e-6) time to complete
- **Mouse Clicks:** An ANOVA showed decision aid does impact (p=3.99e-5) number of mouse clicks

Cognitive Engineering Center
Georgia Institute of Technology

Effect on Performance

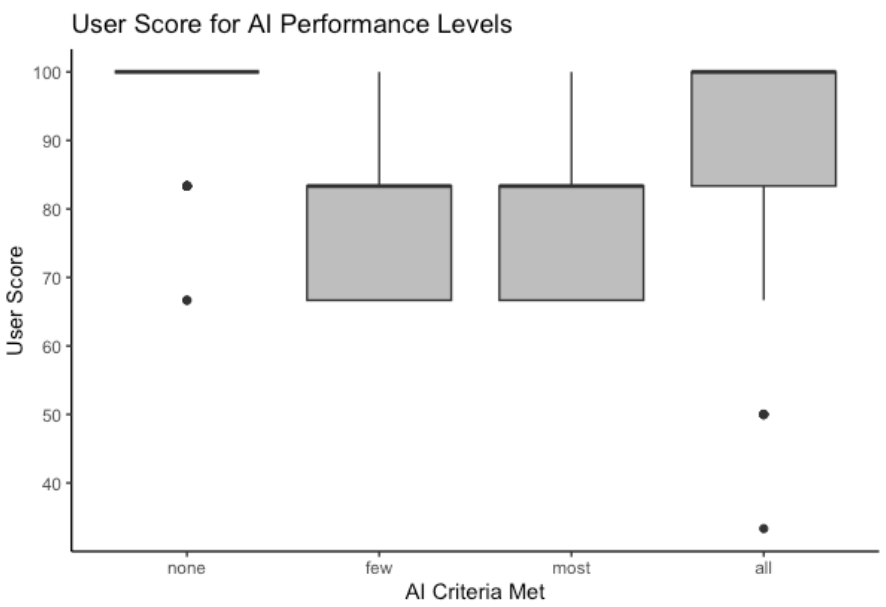


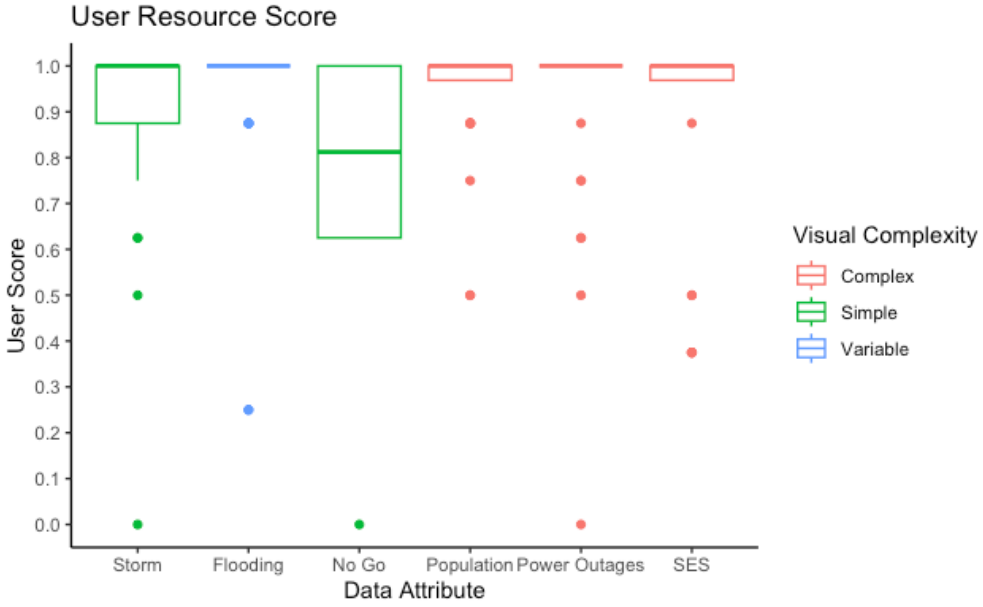
- **Performance:** 'mixed' strategy participants performed significantly better by over 8% (p=0.0485) between trials compared to the 'analytic' strategy when no aid was given
- This indicates that the decision aid can boost performance of the lowest performers to bring them up to the performance standard of the other strategy groups

Experiment 3 Backup

COGNITIVE ENGINEERING CENTER

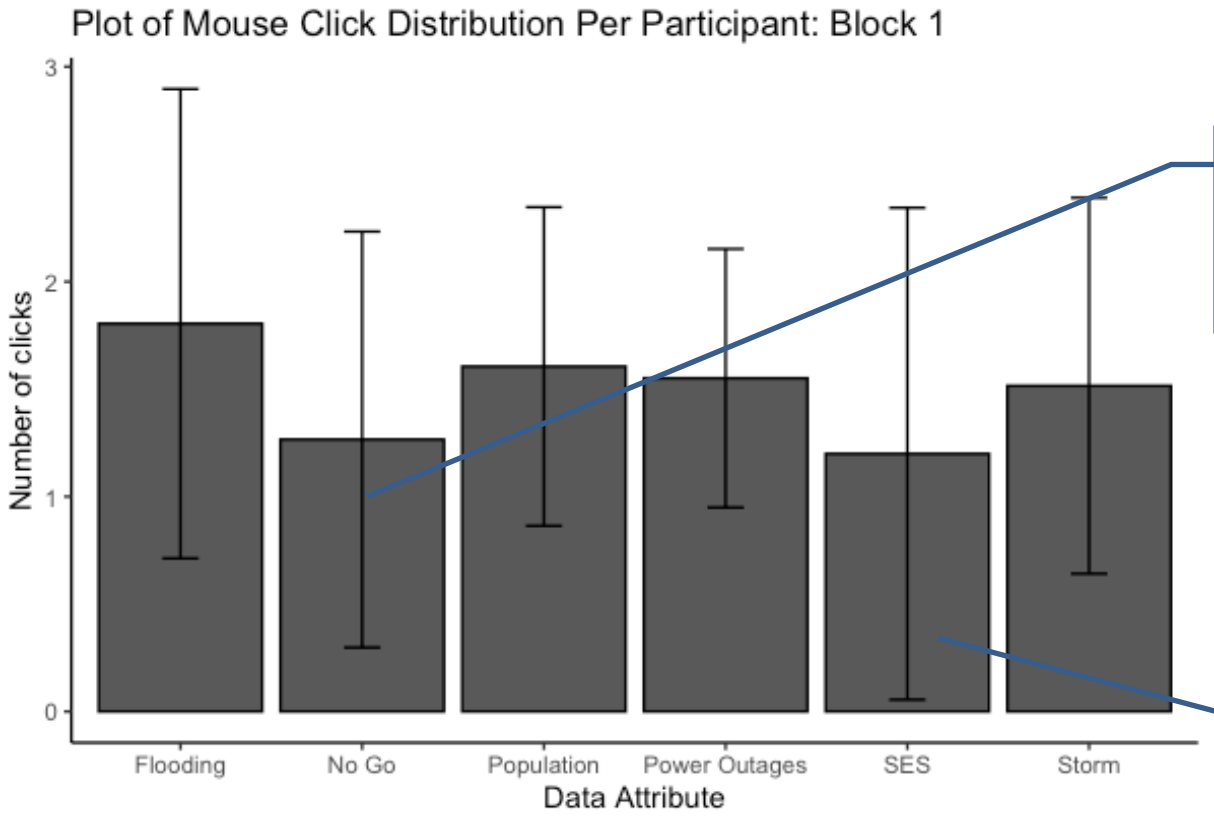
How accurate is the User's Mental Model?





- Finding 1: Users are most accurate in the all or nothing cases
- Finding 2: Lowest performance on No Go Zones attribute
- Users may be adopting a heuristic in their mental model that the lowest weighted constraint can be ignored

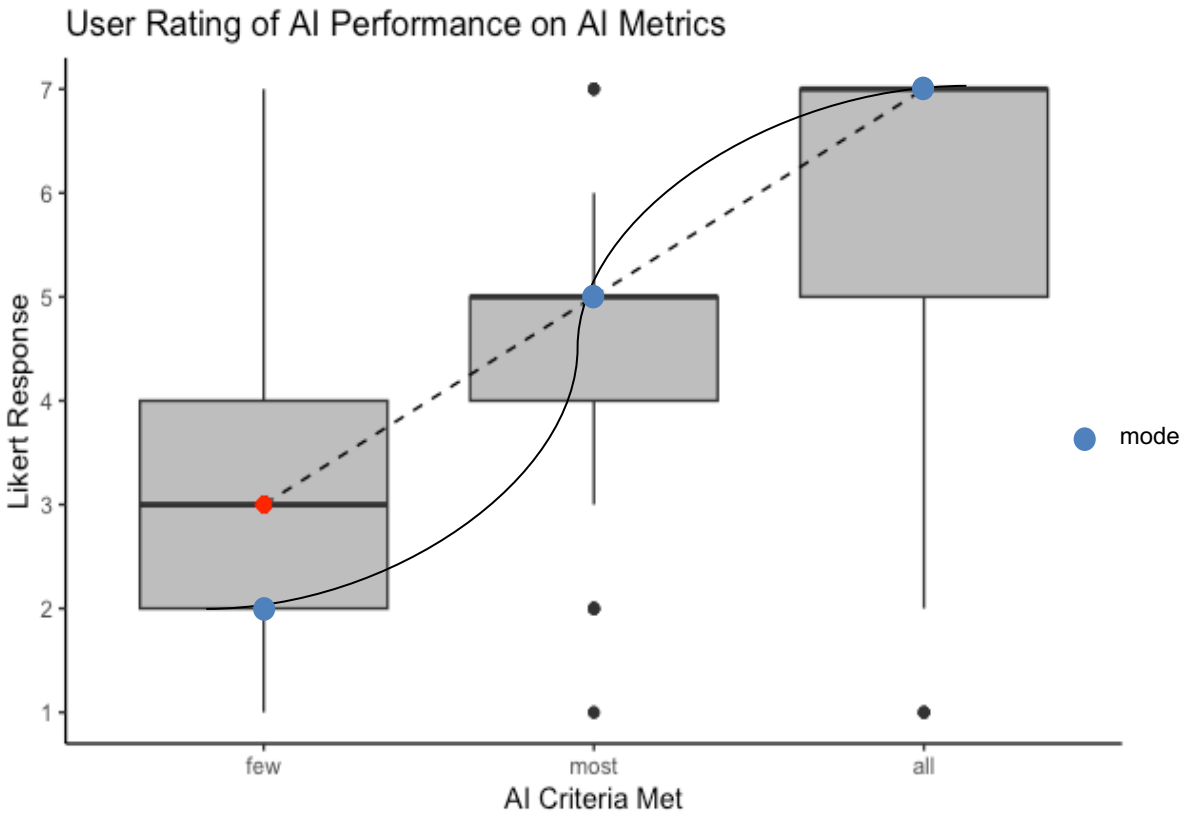
How accurate is the User's Mental Model?



Low Mouse Clicks on No Go Zones may indicate that Users were ignoring this low weighted resource

Some recognized that SES did not matter, some did not

Does the accuracy of the AI impact how well humans perform at assessing the AI?



- Yes, user can accurately assess AI performance

AI Constraints:

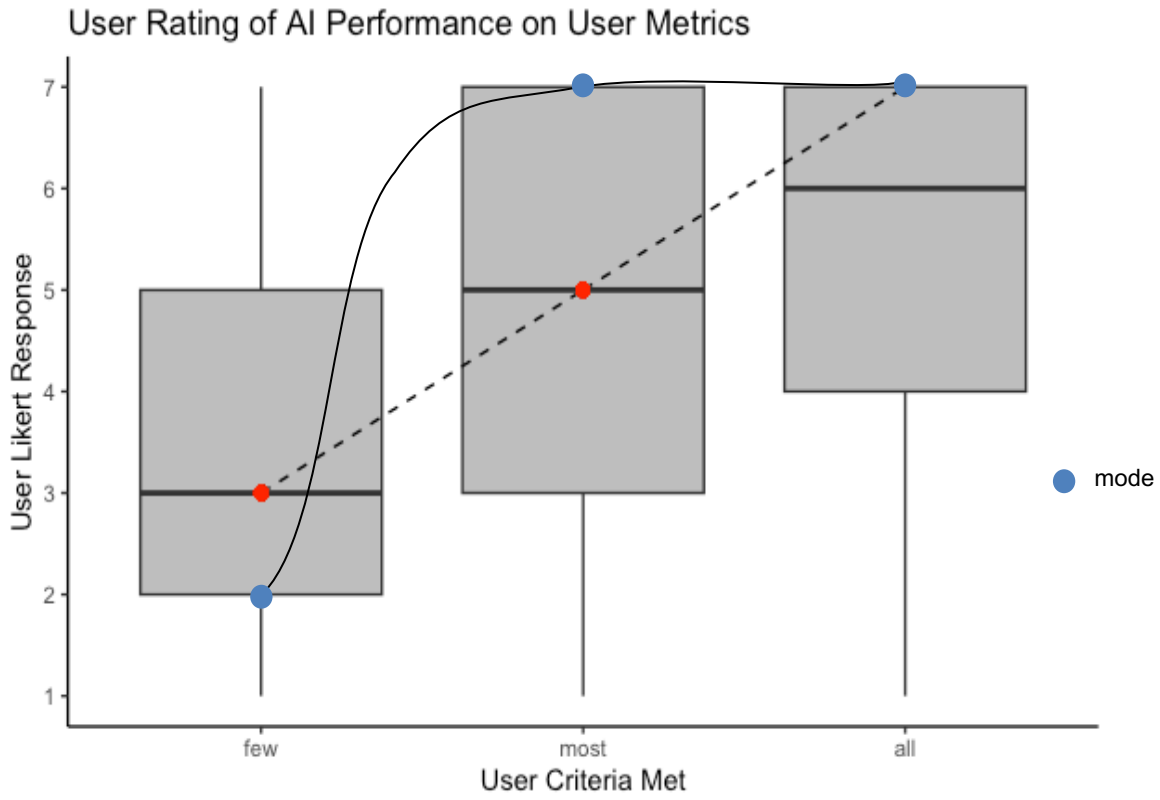
- The AI cannot place the resource in the storm path
- The AI would prefer a place with no power outages
- High population density is considered more valuable than No go zones

Your Constraints:

- You cannot place the resource in an area with a power outage
- It is important to service the lowest socioeconomic status (SES) community
- You would prefer a place with no flooding

```
data:  df$Likert and df$treatmentAI

     few     most
most 1.6e-10 -
all  < 2e-16 5.7e-08

P value adjustment method: bonferroni
```

Does the accuracy of the AI on the *user metrics* impact how well humans perform at assessing the AI?



- Yes, users can on average assess AI performance, but with more variation than we saw in AI criteria

AI Constraints:

- The AI cannot place the resource in the storm path
- The AI would prefer a place with no power outages
- High population density is considered more valuable than No go zones

Your Constraints:

- You cannot place the resource in an area with a power outage
- It is important to service the lowest socioeconomic status (SES) community
- You would prefer a place with no flooding

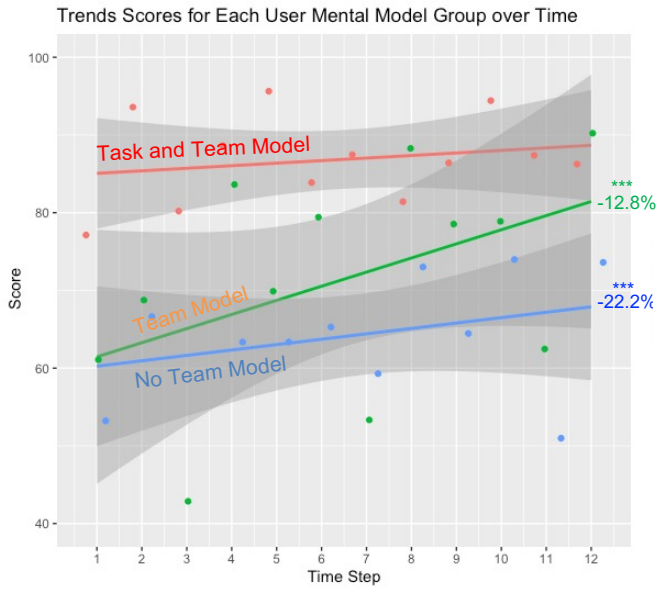
```
data:  df$Likert2 and df$treatmentUser

     few     most
most 1.6e-05 -
all  < 2e-16 0.015

P value adjustment method: bonferroni
```

Experiment 4 Backup

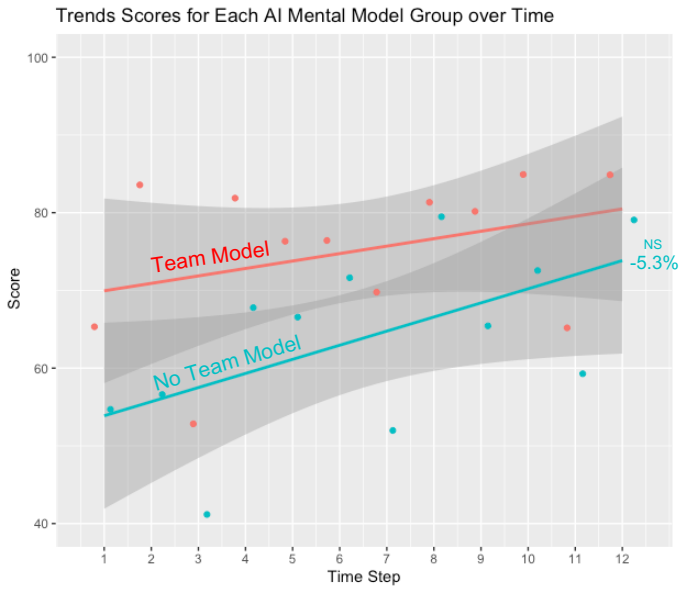
COGNITIVE ENGINEERING CENTER

Group	AI has accurate Team Model?	User has accurate Team Model?	AI and User have accurate Task Model?
1 “baseline” **X**	✓	✓	✓
2 ”both team models” **S**	✓	✓	✗
3 ”no MM of the user” **UoA**	✗	✓	✗
4 “no MM of the AI” **AoU**	✓	✗	✗
5 “no team models” **N**	✗	✗	✗

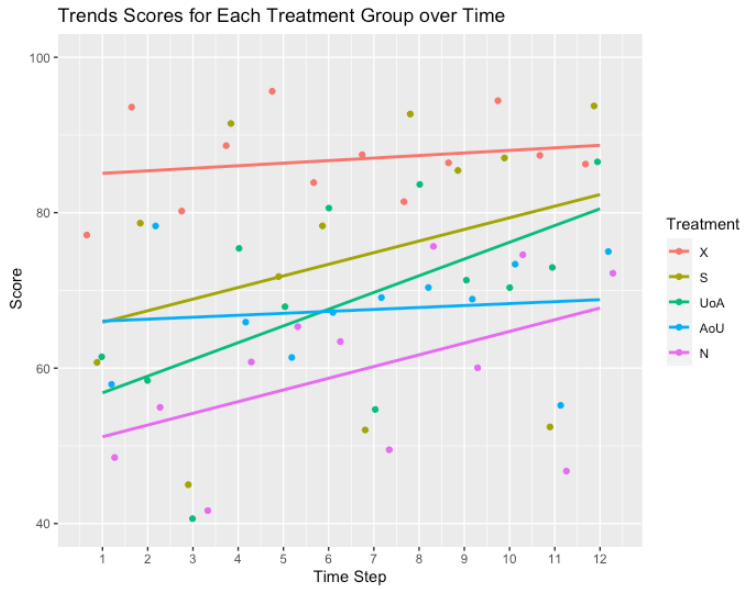
User Mental Model Groups



AI Mental Model Groups



Treatments Groups

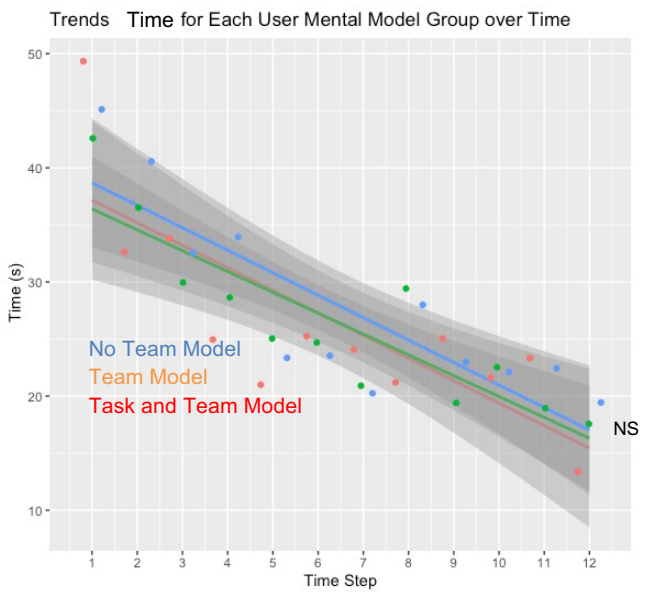


How Mental Models affect Performance

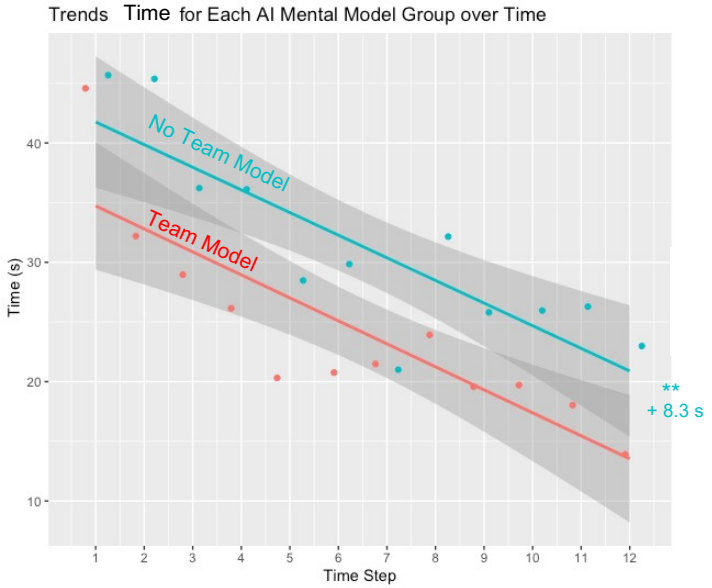
LMER Results

- The user's mental model affected overall performance ($X^2(2) = 19.076$, p = 7.207e-05), by lowering it 12.82% ± 4.986 (se) when the task model is incomplete and 22.17% ± 4.955 when the task and team models are incomplete
- An AI with a more accurate team model increased performance by 5.31% ± 3.797, however these results were not statistically significant

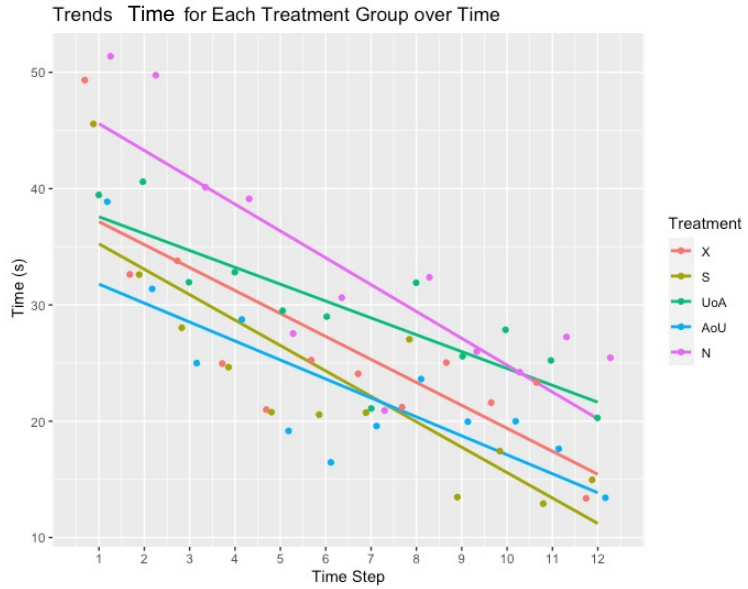
User Mental Model Groups



AI Mental Model Groups



Treatments Groups

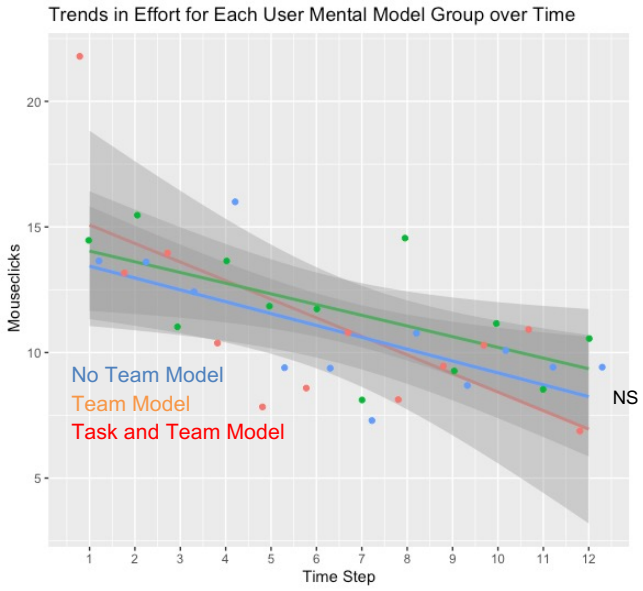


How Mental Models affect Task Speed

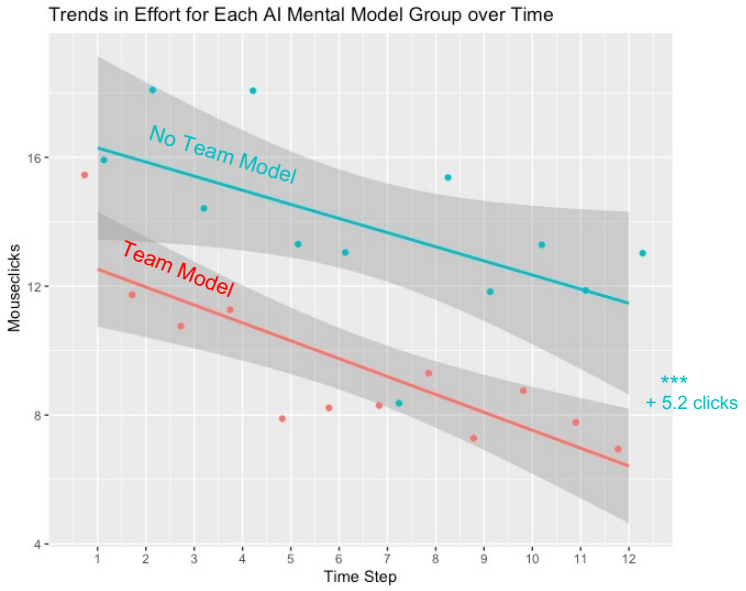
LMER Results

- The AI's model of the user affected time to complete ($X^2(1) = 7.7763$, $p = 0.005294$), increasing it by ~8.286 s when the team model is incomplete
- The user's mental model affected task speed by lowering it 4.002 when the task model is incomplete and and additional 2.585 when the task and team models are incomplete, however these results were not statistically significant

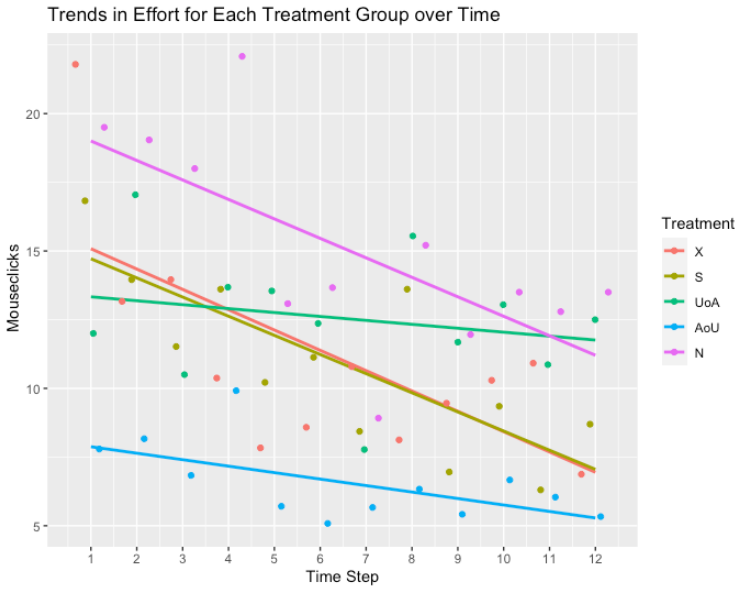
User Mental Model Groups



AI Mental Model Groups



Treatments Groups



How Mental Models affect Effort (active information access)

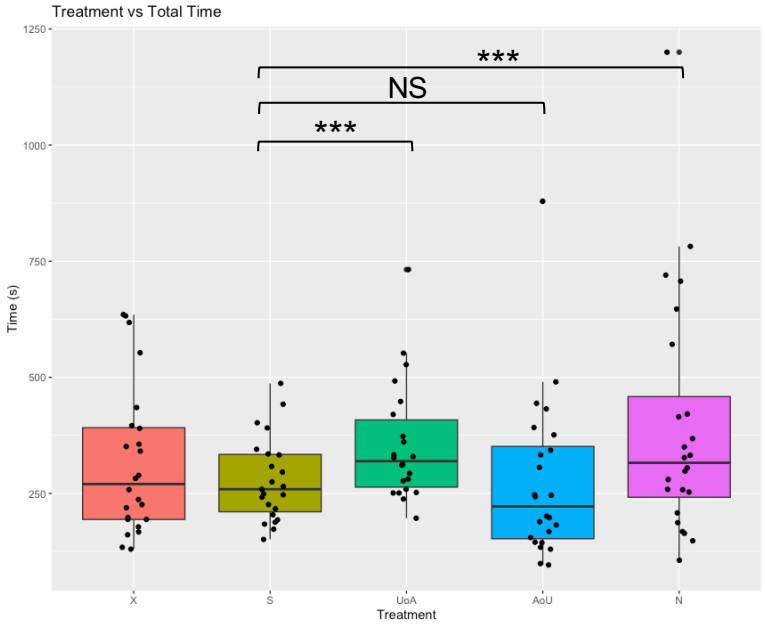
LMER Results

- The AI's model of the user affected information access ($X^2(1) = 12.604$, $p = 0.000385$), increasing it by ~5.204 clicks when the team model is incomplete
- The user's mental model affects on information access were small and not statistically significant

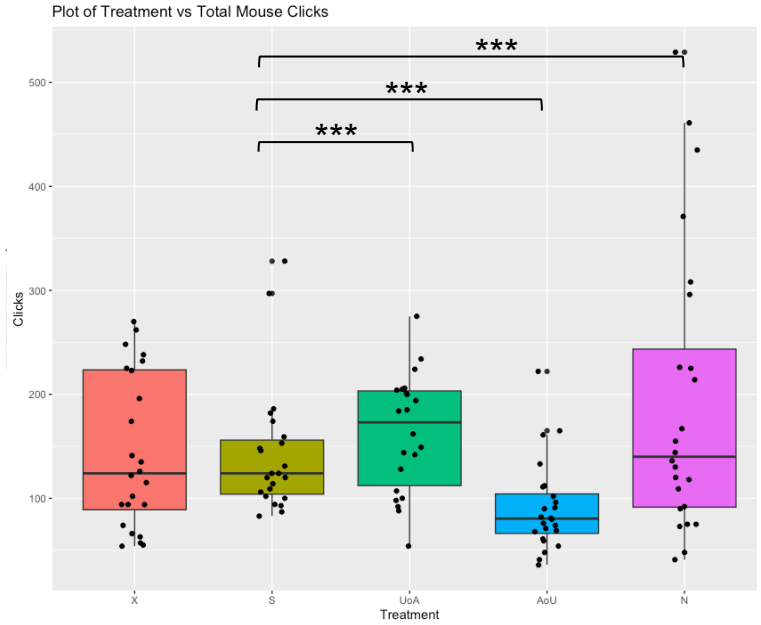
How important is a two-way model?

- The two-way model significantly improved decision-making accuracy while decreasing time and effort on tasks compared to the those who had an inaccurate Team Model
- The users with a one-way model (MM of AI) were able to compensate for an AI without a user model but spent significantly longer on tasks with much higher effort
- Users that did not have a model of the AI, but were teamed with an AI that had a model of the user, performed significantly worse than those with a two-way model. They had significant low effort than the other groups and had a much higher SD in score.

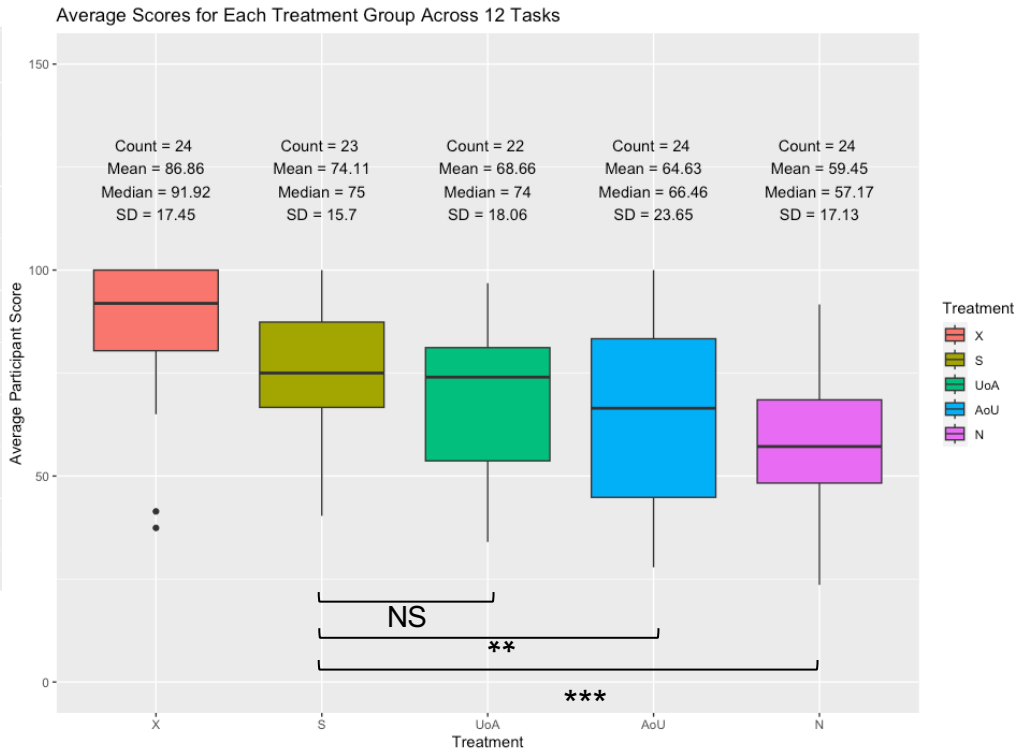
Time



Effort

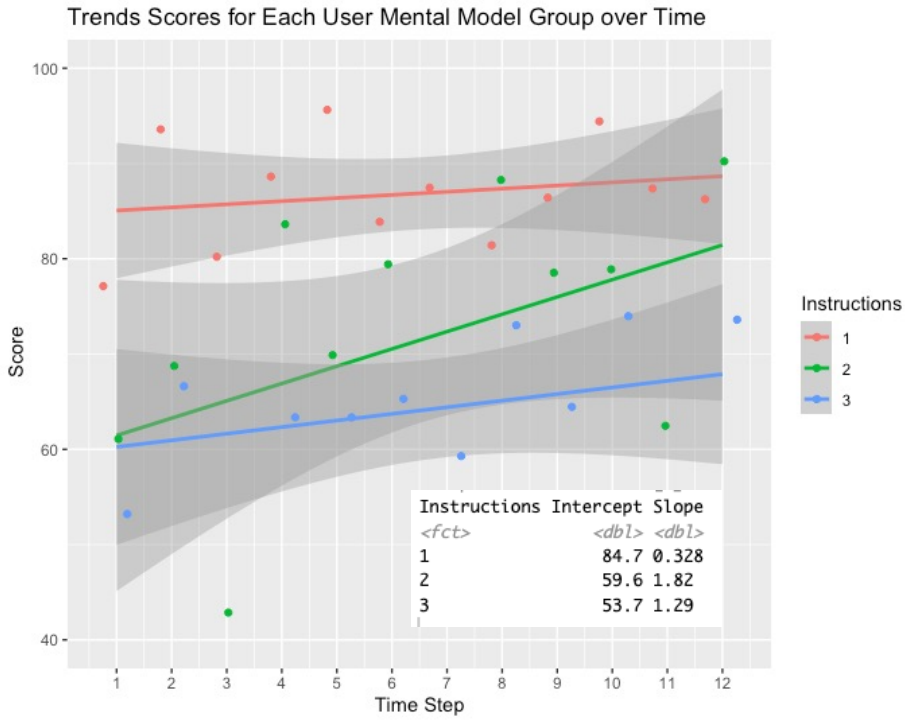


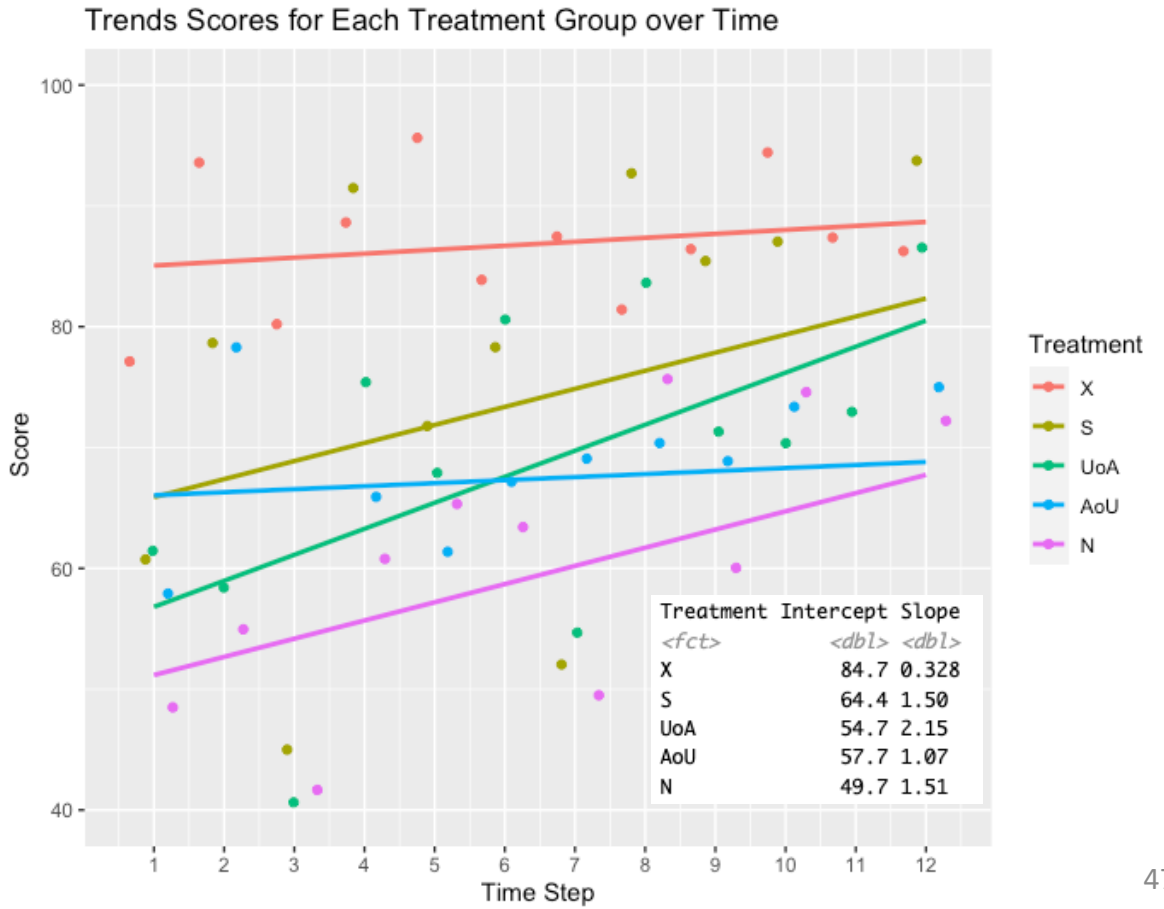
Performance



Learning: How do mental models affect performance over time?

- As expected, users with complete Task and Team Models do not have significant improvements over time because they go into the task with all the information to do well.
- However, the groups in which the users had a Team Model had much stronger improvements over time compared to those without an inaccurate Team Model
- **We find that over time users with a were largely able to compensate for inaccuracies in the Task Model**





Does a SMM reduce the User's Workload?

NASA TLX

- Users that worked with an AI with a Team Model experienced less effort, lower mental demand, and higher perceived performance
- Users that had a complete task model experience less frustration than those with an incomplete task and team model, but no difference in frustration from those with a complete Team Model.
- Users with a complete Task Model perceived higher performance than those with an incomplete Task Model– this perception agrees with objective measures of performance