Applying User Cognitive State to Human-Robot Interaction and Role Assignment

Jack Kolb

February 1st, 2024

Committee:

Karen Feigh (advisor)
Julie A. Adams
Sonia Chernova
Harish Ravichandar
Alan R. Wagner

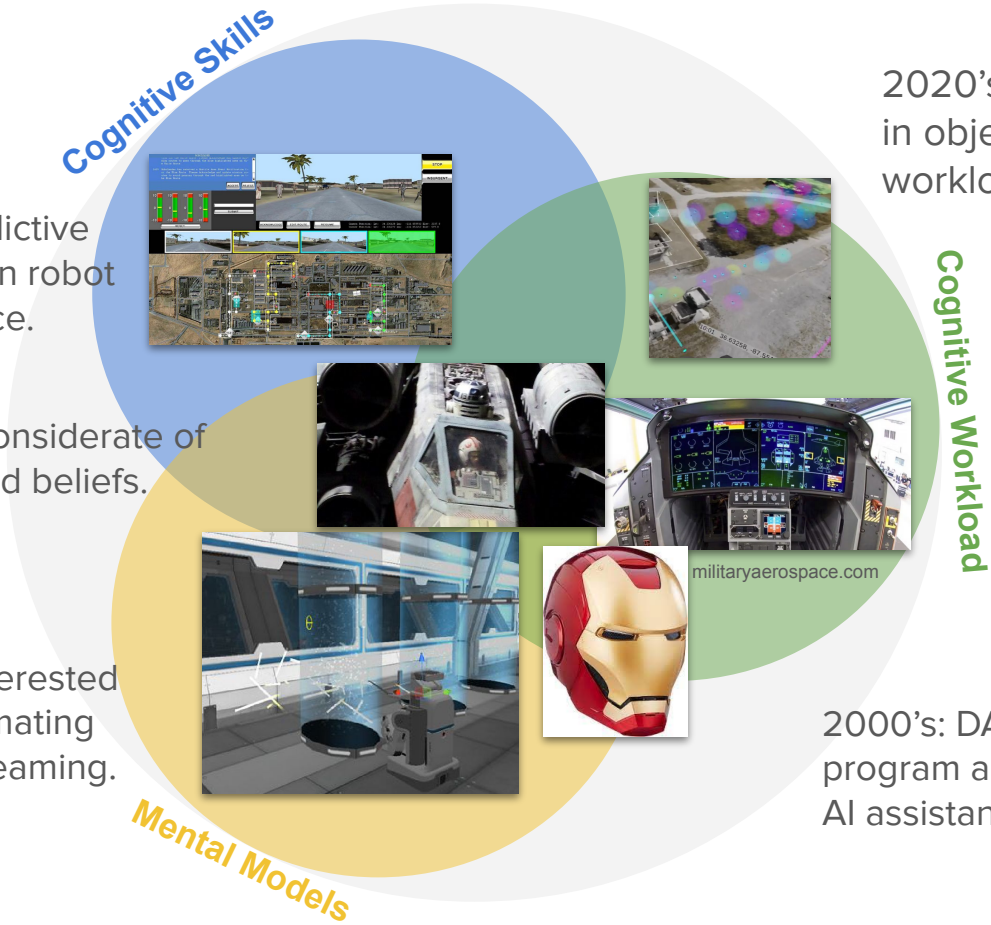
What is cognitive state...

...in the context of human-robot interaction?

2010's: ARL explored predictive power of cognitive skills on robot operation task performance.

Future: Robots considerate of user cognition and beliefs.

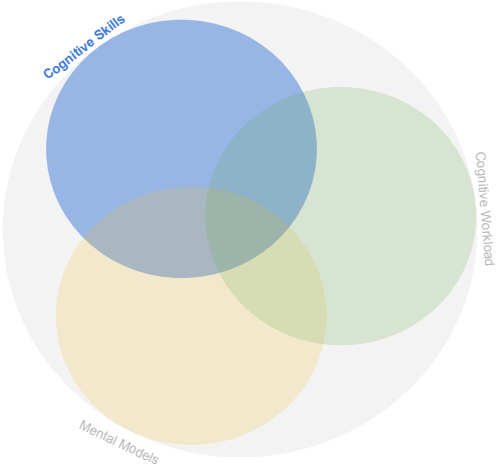
2010's, 2020's: NASA interested in robots capable of estimating user mental models for teaming.



2020's: Renewed interest in objective cognitive workload estimation.

1990's: US Air Force sought cockpits that could adapt a pilot's taskwork to moderate pilot workload. (*adaptive autonomy*)

2000's: DARPA's *Augmented Cognition* program aimed to create context-aware AI assistants for soldiers



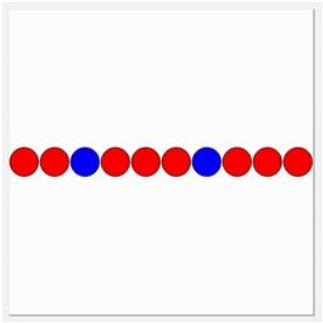
Cognitive Skills:

Researchers aim to identify distinct, testable *cognitive skills* that are used by tasks.

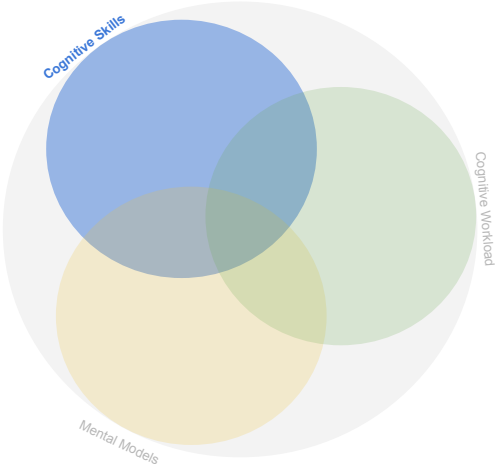
Belief that training specific cognitive skills can improve performance in related work.

Proposed relevant cognitive skills include...

- Fanout (Cummings et al. 2018)
- Visual Attention (Chen et al. 2014)
- Spatial Reasoning (Lathan & Tracey 2002)
- Network Inference (Lynn et al. 2020)
- Situation Awareness (Chen et al 2014)



Visual Attention

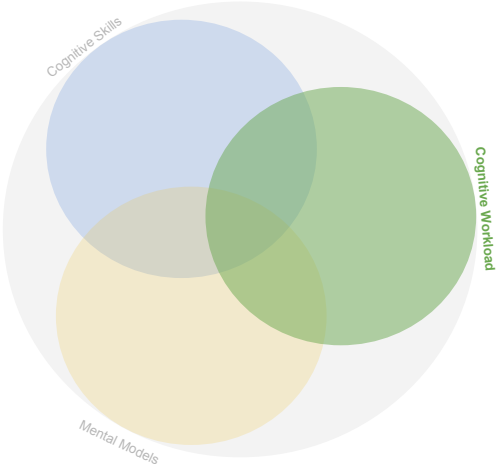


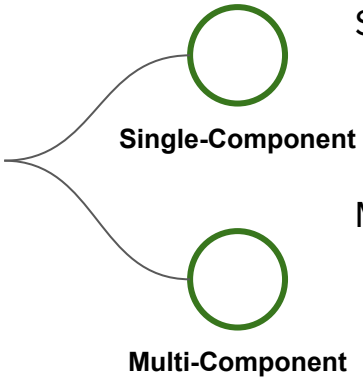
Takeaways:

- The cognitive science community has long suspected that cognitive skills transfer across similar tasks.
- Little research has operationalized cognitive skills to human-robot teaming domains.

Opportunities:

- Identify relationships between cognitive skills and future task performance.
- Apply cognitive skills as a predictive measure of performance for human-robot teaming.



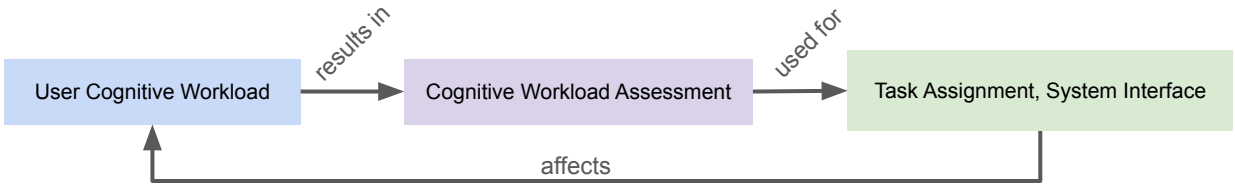


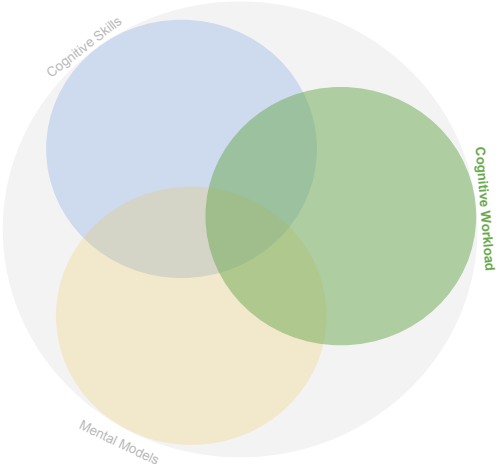
Single-Component Workload (Roscoe 1984)

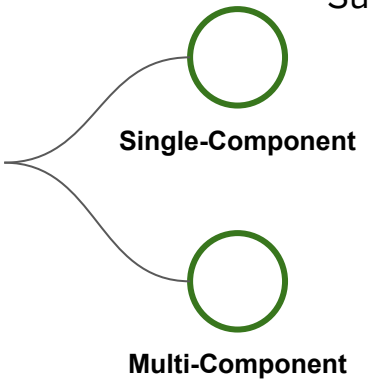
- Represent user workload as a **single** cognitive resource.
- Can moderate user workload with a task-agnostic input.

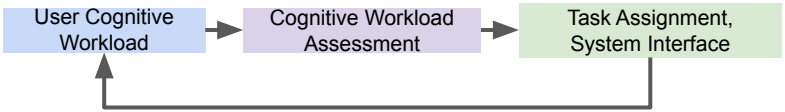
Multi-Component Workload (Wickens et al. 1983; McCraken et al. 1984)

- Workload as **multiple** subcomponents, such as:
  - *Auditory* workload
  - *Cognitive* workload
  - *Physical* Workload
  - *Speech* Workload
  - *Visual* Workload
- Can represent tasks by the relevance of subcomponents.









Subjective assessments are infeasible for adaptive autonomy, need objective metrics!

Subjective Workload Assessments:

NASA TLX (gold standard) (Hart & Staveland 1988)
- *Post-hoc* questionnaire.
- Derives workload score from six subscales.
- Single-Component.

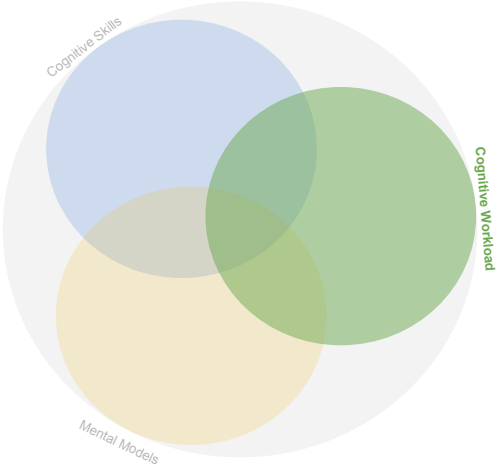
Bedford Scale (Roscoe 1984)
- *In situ* questionnaire.
- Modified from Cooper-Harper Rating Scale (for pilots).
- Decision-tree to quantify workload.
- Single-Component.

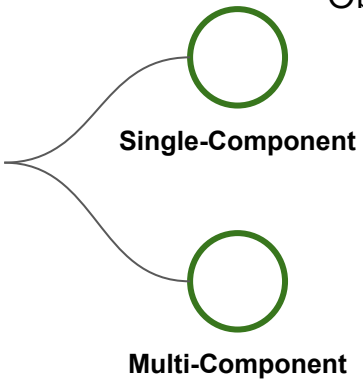
SWAT (Reid & Nygren 1988)
- *In situ* questionnaire.
- Three components.
- Not commonly used in recent work.

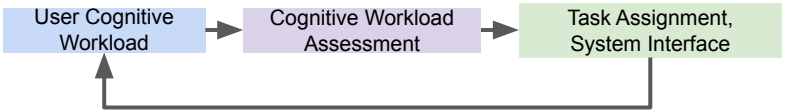
Multi-Resource Questionnaire (Boles & Aldair 2001)
- *Post-hoc* questionnaire.
- Eight components.
- Aims to follow Wickens Multi-Resource Theory

Overall Workload Scale (Hill et al. 1992)
- *Post-hoc* questionnaire.
- Single component from summing 20 scales.

Workload Profile (Tsang & Velazquez 1996)
- *Post-hoc* questionnaire.
- Single component from summing 8 scales.





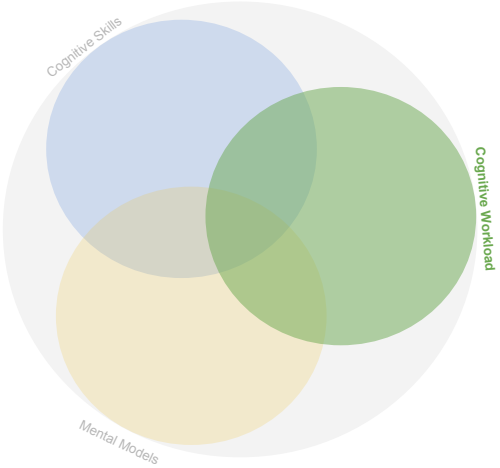


Weak correlations and conflicting results have left adaptive autonomy an open research question.

Objective Workload Assessments:

Metric	Category	Correlation	Workload Component(s)
EEG: Power Spectral Density	Response	Both	Cognitive
EEG: Event Related Potentials	Response	Increases	Cognitive
fNIRS	Response	Increases	Cognitive
Heart Rate Variability	Response	Decreases	Cognitive
Heart Rate	Response	Increases	Cognitive, Physical
Respiration Rate	Response	Decreases	Speech, Physical
Galvanic Skin Response	Response	Decreases	Cognitive, Physical
Skin Temperature	Response	Decreases	Cognitive, Physical
Blink Frequency	Response	Both	Cognitive, Visual
Pupil Dilation	Response	Increases	Cognitive
Fixation Duration	Both	Increases	Cognitive, Visual
Blink Duration	Response	Decreases	Cognitive, Visual
Blink Latency	Response	Increases	Cognitive, Visual
Speech Response Time	Response	Increases	Cognitive, Auditory, Speech
Speech Rate	Response	Increases	Cognitive, Speech
Number of Fragments	Response	Increases	Cognitive, Speech
Number of False Starts	Response	Increases	Cognitive, Speech
Number of Syntax Errors	Response	Increases	Cognitive, Speech
Filler Utterances	Response	Increases	Cognitive, Speech
Utterance Repetitions	Response	Increases	Cognitive, Speech
Utterance Length	Response	Decreases	Cognitive, Speech
Noise Level	Demand	Increases	Cognitive, Auditory
Variance in Posture	Demand	Increases	Physical
Postural Load	Demand	Increases	Physical
Vector Magnitude	Demand	Increases	Physical
Task Density	Demand	Increases	Task Dependent
Task Switches and Interruptions	Demand	Increases	Task Dependent
Secondary Task Failure Rate	Demand	Increases	Task Dependent

Heard et al. 2015

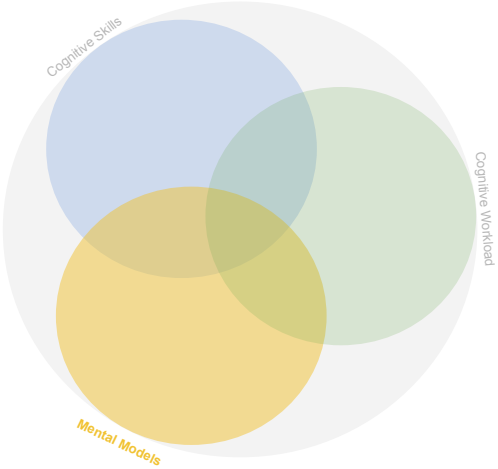


Takeaways:

- Longstanding interest in using cognitive workload as an input to human-AI systems.
- Numerous subjective metrics have been proposed and utilized.
- Very little work has successfully applied objective metrics.

Opportunities:

- Classify objective workload from physiological metrics.
- Apply objective workload to adapt system interfaces or autonomous modes.

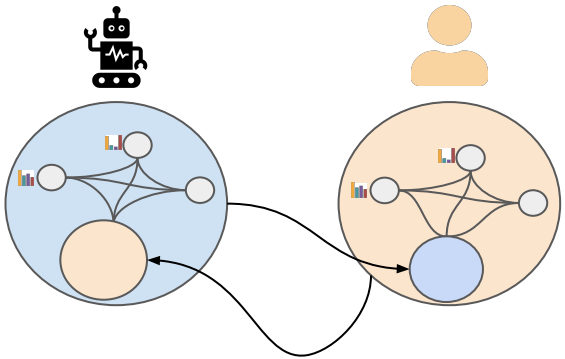


Simulation Theory

We maintain an internal simulation of the environment.

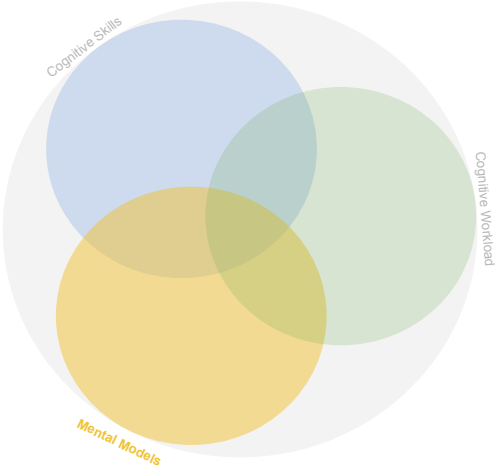
Mental Model

The data structure of the internal simulation or belief state.



Task Model		Team Model	
Equipment	**Task**	**Team Interaction**	**Team (Teammates')**
Equipment functioning	Task procedures	Roles/responsibilities	Knowledge
Operating procedures	Likely contingencies	Information sources	Skills
Likely failures	Environmental constraints	Interaction patterns	Performance history
Equipment/system limitations	Task strategiesLikely scenarios	Information flowCommunication channels	TendenciesPreferences
	Task component relationships	Role interdependencies	Attitudes

Scheutz et al. 2017

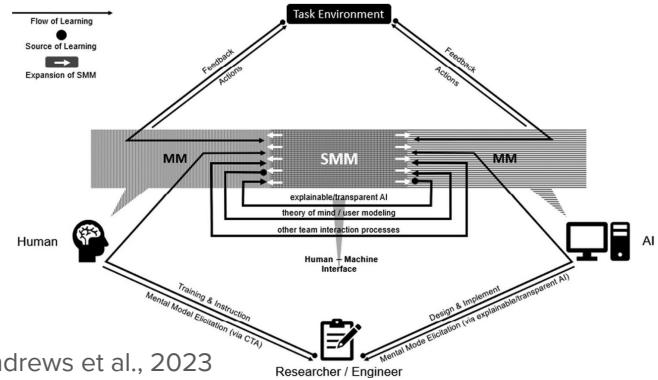


Simulation Theory

We maintain an internal simulation of the environment.

Mental Model

The data structure of the internal simulation or belief state.



Andrews et al., 2023

Team Mental Model

The collective mental model of a team.

Shared Mental Model

The overlaps/disparities of a teams' models.

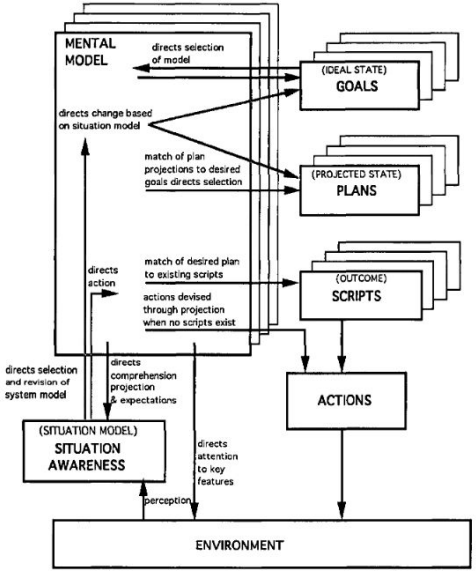


Figure 4. *Relationship of goals and mental models to situation awareness.*

Endsley 1995

Mental Model

The data structure of the internal simulation or belief state.

Situation Awareness

An task-focused representation of the mental model.

Endsley 1988

World State

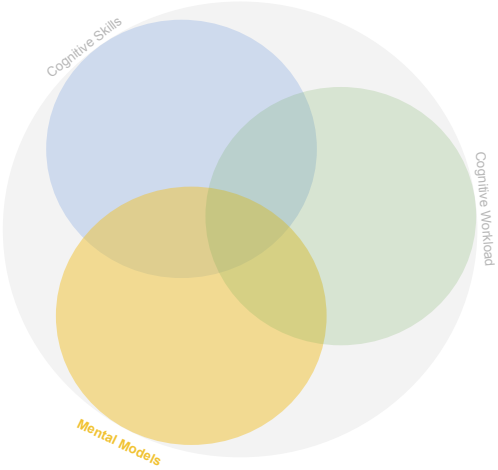
The current environment elements.

Context

What the elements mean for the task.

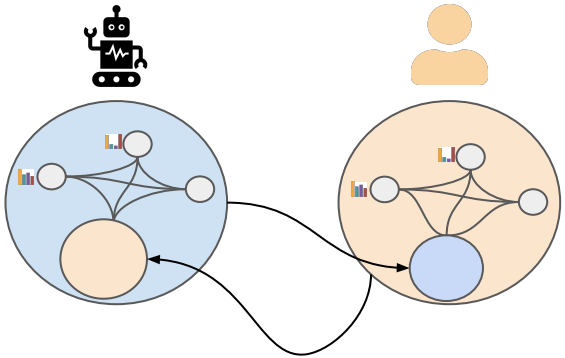
Projection

How the environment will change.



Takeaways:

- The importance of mental models is highlighted by the human factors community.
- Recent work has sought to define mental model components for human-AI teams.
- Related work has focused on passively supporting the user's mental model through system design.

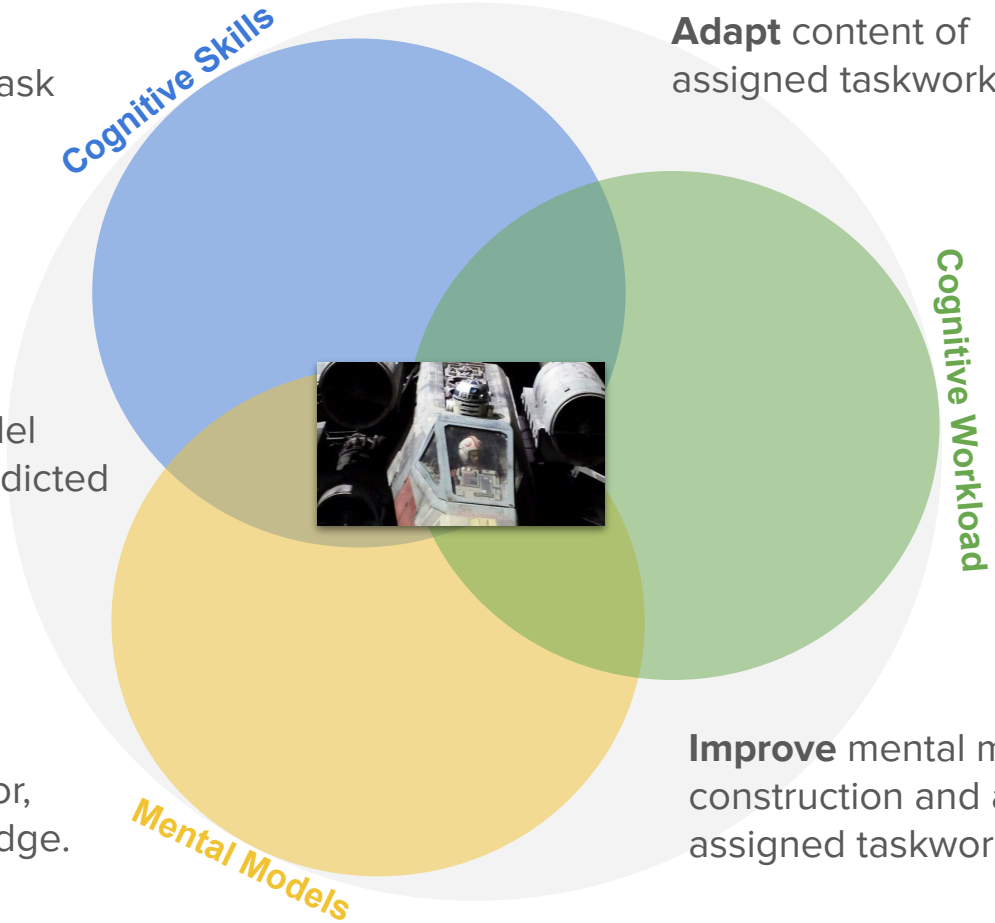


Opportunities:

- Go beyond passive support by estimating aspects of the user's mental model.
- Use the estimated mental model to actively support users in human-robot teams.

Predict user's task performance.

Adapt content of assigned taskwork.

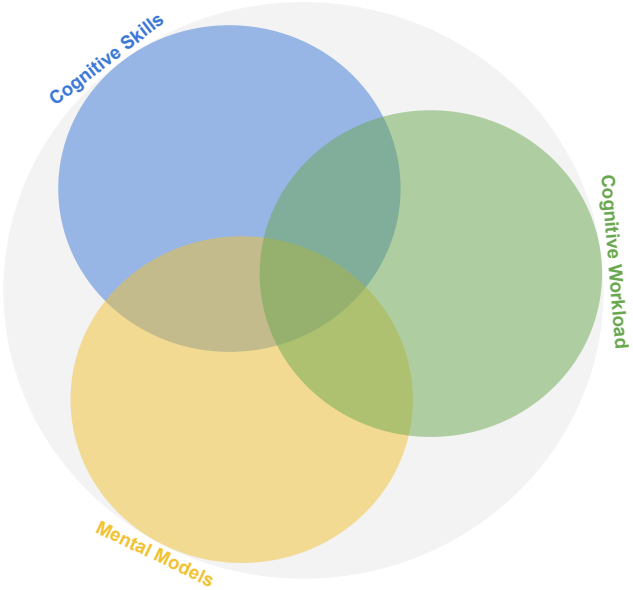


Improve mental model construction and predicted task performance.

Adapt user interactions, **adapt** volume of assigned taskwork.

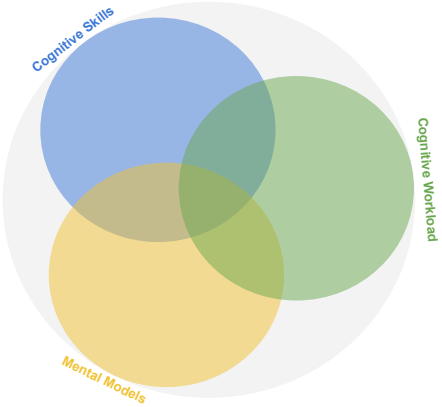
Predict user's behavior, **plan** to user's knowledge.

Improve mental model construction and adapted assigned taskwork.



Thesis Statement:

We can leverage user *cognitive skills*, *cognitive workload*, and *mental models* to enhance human-robot teams.



We can leverage user *cognitive skills*, *cognitive workload*, and *mental models* to enhance human-robot teams.

- Cognitive Skills:
  - ✓ Predict future robot operation performance using cognitive skills.
  - ✓ Demonstrate with user role assignment for human-robot teams.

  [Kolb et al. 2021, Kolb et al. 2022]

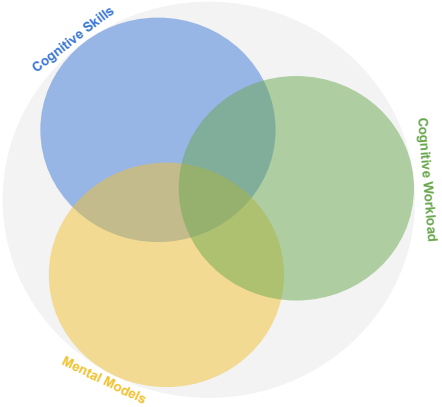
- Cognitive Workload:
  - ~ Identify off-nominal cognitive workload from physiological metrics.
  - ~ Demonstrate model transfer across human-AI team domains.

  [Agbeyibor et al. 2024]

- Mental Models:
  - ~ Estimate a human teammate's belief state from observations.
  - Personalize belief state estimation to individual users.

✓ Completed Work

~ In Progress Work



We can leverage user *cognitive skills*, *cognitive workload*, and *mental models* to enhance human-robot teams.

- **Cognitive Skills:**
  - ✓ Predict future robot operation performance using cognitive skills.
  - ✓ Demonstrate with user role assignment for human-robot teams.

  [Kolb et al. 2021, Kolb et al. 2022]

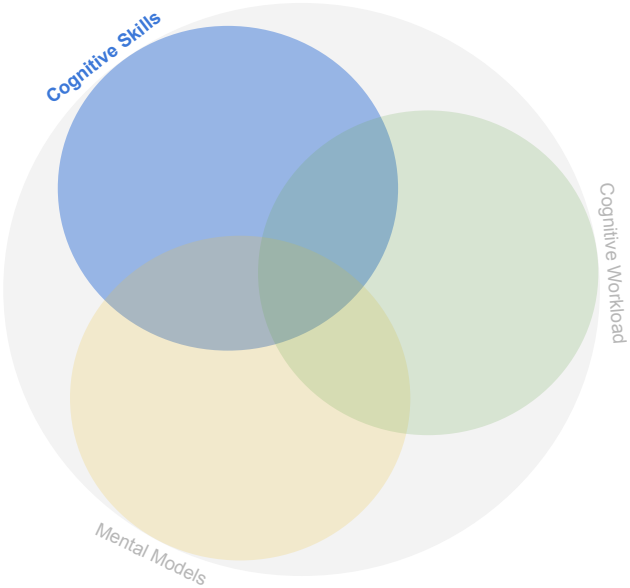
- Cognitive Workload:
  - ~ Identify off-nominal cognitive workload from physiological metrics.
  - ~ Demonstrate model transfer across human-AI team domains.

  [Agbeyibor et al. 2024]

- Mental Models:
  - ~ Estimate a human teammate's belief state from observations.
  - Personalize belief state estimation to individual users.

✓ Completed Work

~ In Progress Work

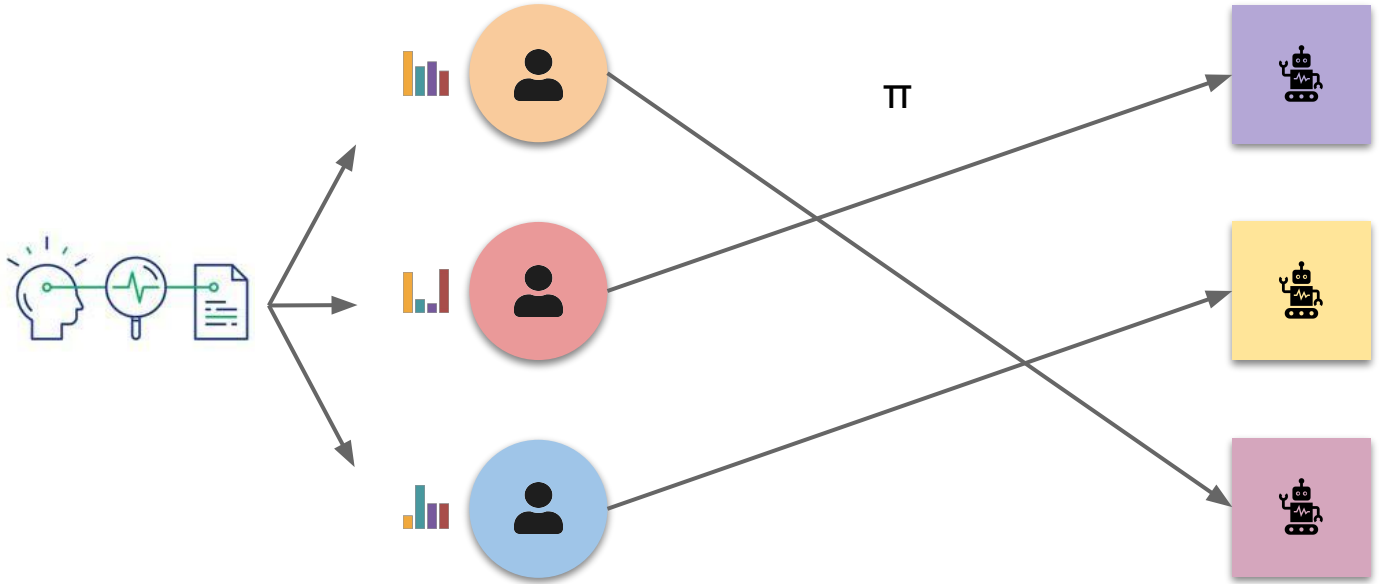


Can we ***predict user performance*** at robot teleoperation ***only with*** information about their ***cognitive skills***?

- Predict future robot operation performance using cognitive skills.
  [Kolb et al. 2021]
- Demonstrate with user role assignment for human-robot teams.
  [Kolb et al. 2022]

Novice Users

Robot Teleoperation Tasks



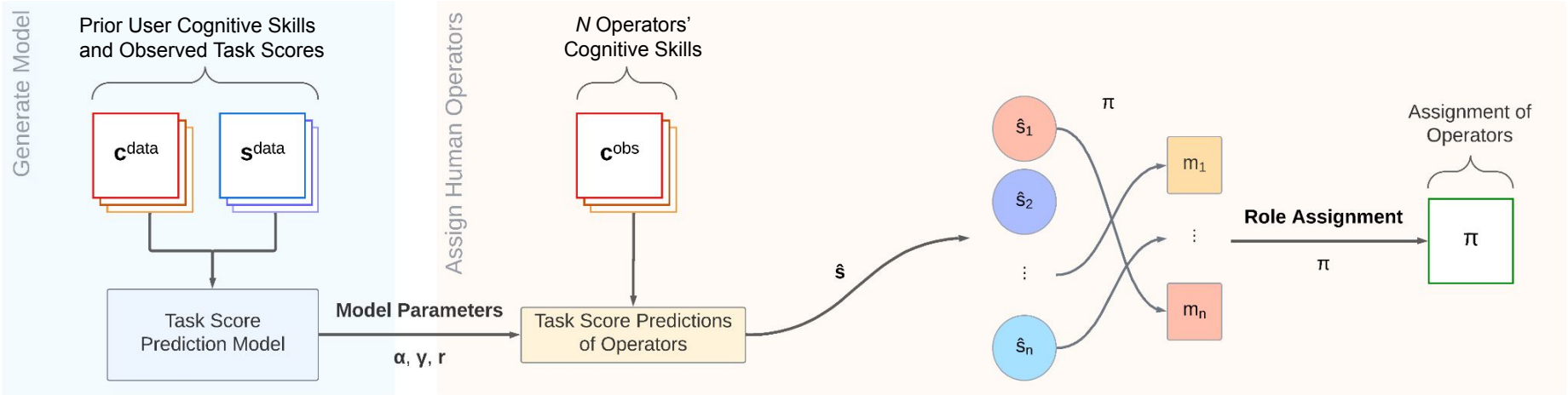
6 possible assignments...

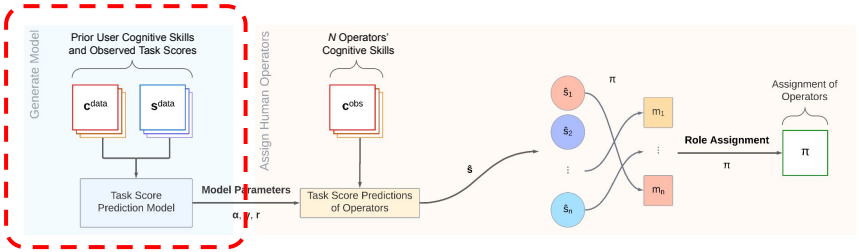
How do we choose one?

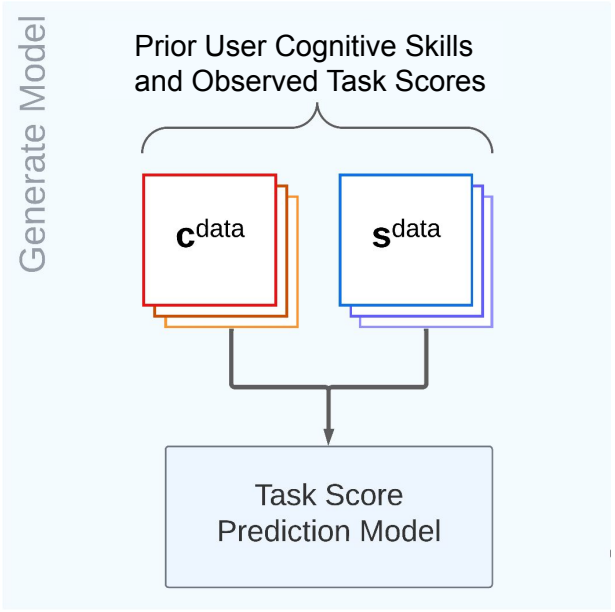
As users are novices, we only have *random assignment* to compare to.

Conducted in the past

Conducted just before role assignment

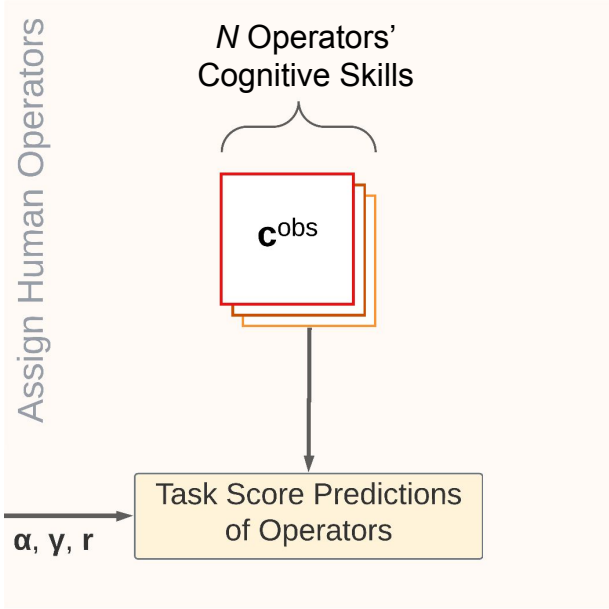


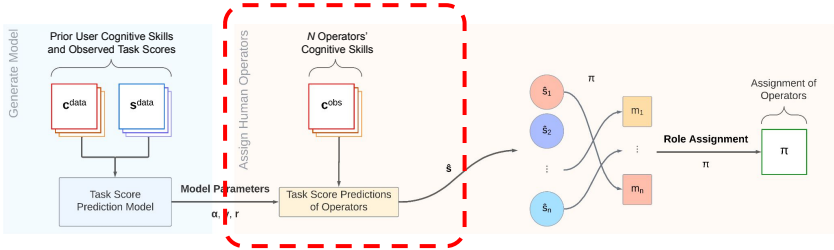




Construct a model to predict task scores:

- *In previous sessions*, collect a dataset of participants taking *U* cognitive skill tests and then *M* robot teleoperation tasks.
  - c^{data}: *U* cognitive skill tests
  - s^{data}: then *M* robot teleoperation tasks
- Fit a model (linear regression) to each teleoperation task and cognitive skill pairing (*M* x *U* regressions).
- Return the regression slopes, y-intercepts, and correlations.
  - α: regression slopes
  - γ: y-intercepts
  - r: correlations



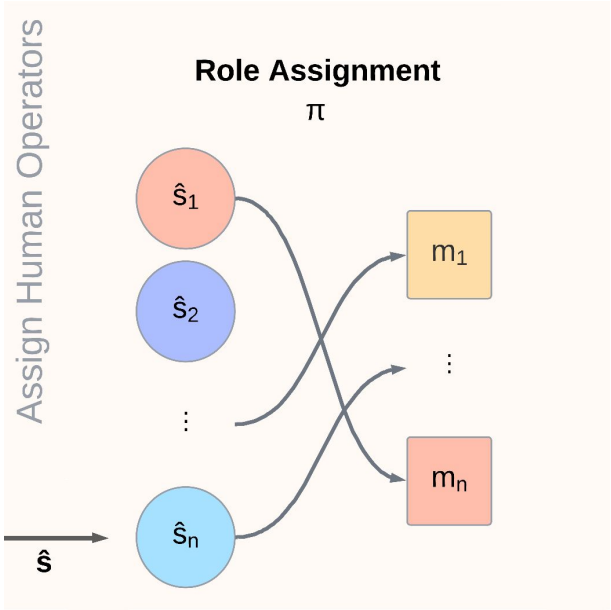


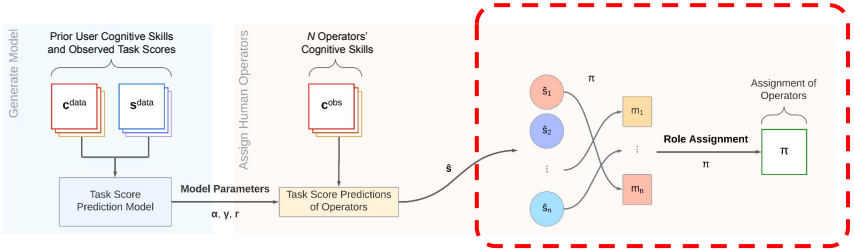
Predict the task scores of a new participant team:

- Conduct cognitive skill tests for the participants. ($\mathbf{c}^{\text{obs}}$)
- Predict the teleoperation task scores for each participant *n*.

$$\hat{\mathbf{s}}_{n,m} = \sum_{u=0}^{U} \gamma_{m,u}(\alpha_{m,u}\ \mathbf{c}_{u,n}^{\text{obs}} + \beta_{m,u}) \qquad \gamma_{m,u} = \frac{|r_{m,u}|}{\sum_{u'=0}^{U}|r_{m,u'}|}$$

- Return the predicted teleoperation task scores. ($\hat{\mathbf{s}}$)





Conduct role assignment with the predicted task scores:

- Choose the role assignment π that maximizes the team's cumulative predicted scores.

$$\mathbf{S} = \sum_{m=1}^{M} \mathbf{s}_{\pi(m),m}$$

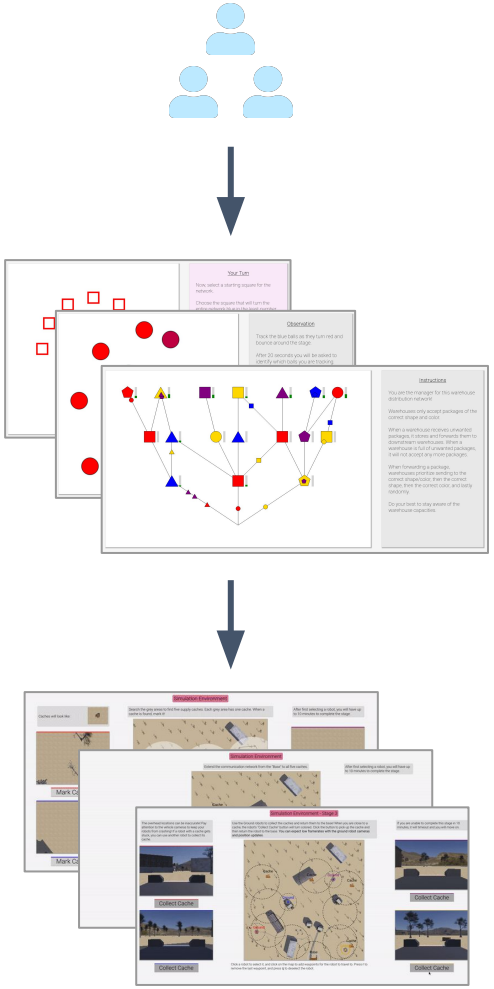
- Frame as the *Optimal Assignment Problem* also known as *"minimum cost allocation"*.

- Return the *Individualized Role Assignment* (**IRA**).

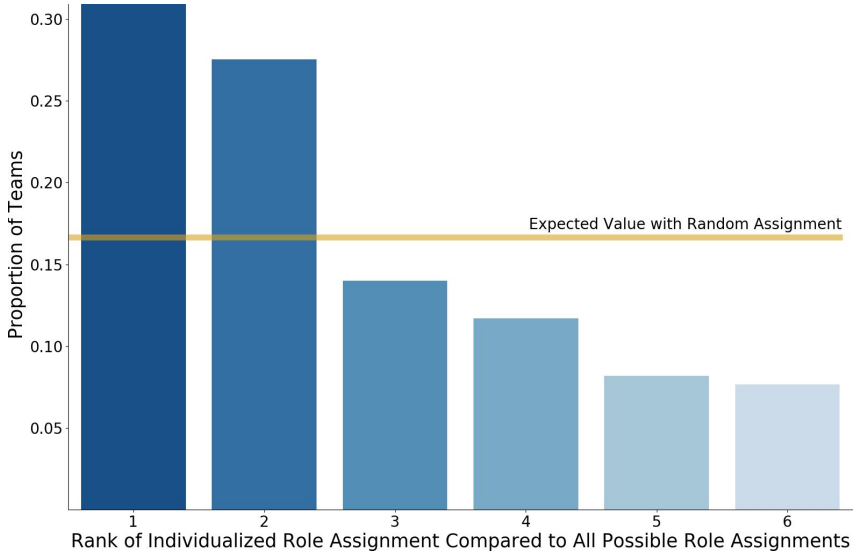
π

Conducted a user study to evaluate our architecture.

- Applied three online cognitive skill tests from the cognitive science literature.
- Developed three online robot teleoperation tasks (using 3D simulation).
- 29 participants individually completed the cognitive skill tests, and then the teleoperation tasks.
- Evaluated participant data post-hoc, using participant subsets to represent teams (29 choose 3 teams).

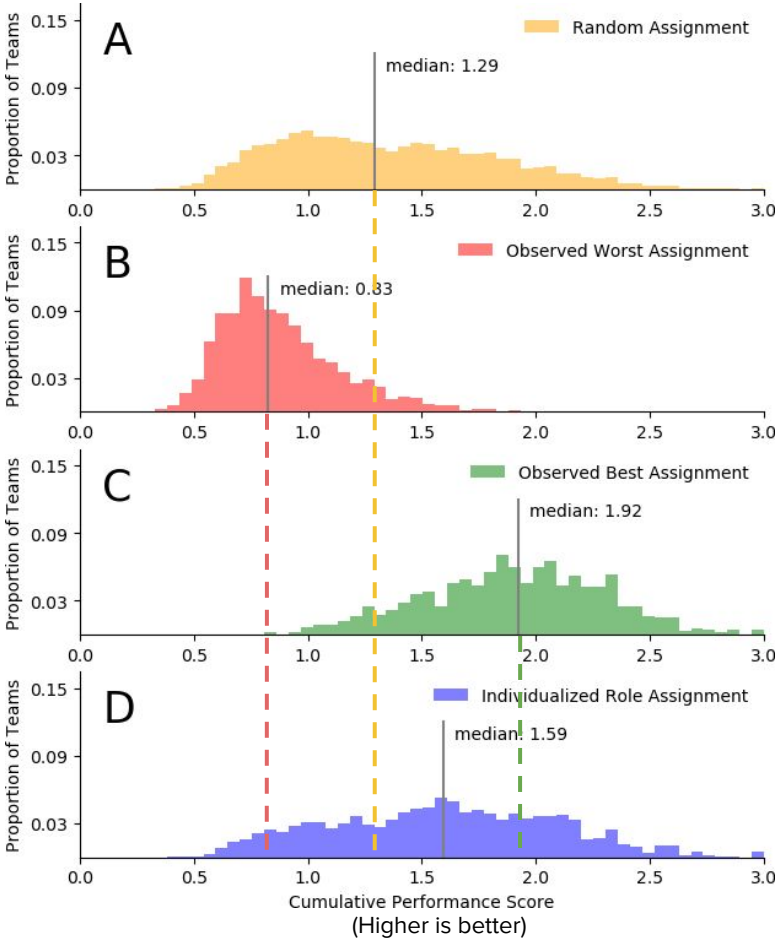


- **31%** of IRAs **matched** **observed best** assignment (*Random: 16.6%*).
- IRA **skewed toward** best possible assignment.



- **31%** of IRAs **matched observed best** assignment (*Random: 16.6%*).
- IRA **skewed toward** best possible assignment.
- IRA had a median **24%** team score **improvement** over random assignment.
- **73%** of IRAs **outperformed** random assignment. (*Random: 50%*)

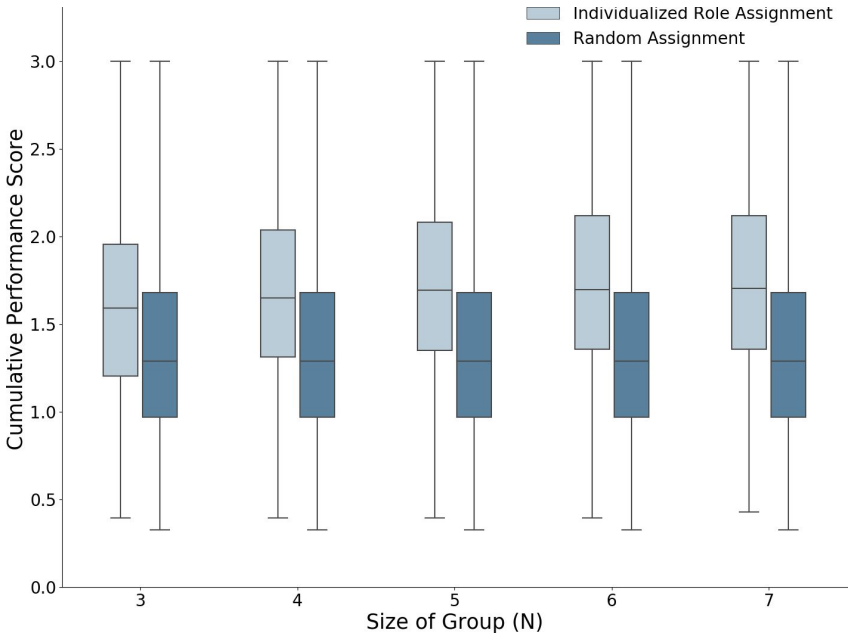
IRA



- **31%** of IRAs **matched observed best** assignment (*Random: 16.6%*).
- IRA **skewed toward** best possible assignment.

- IRA had a median **24%** team score **improvement** over random assignment.
- **73%** of IRAs **outperformed** random assignment. (*Random: 50%*)

- Results held with larger candidate pools (from 4 choose 3, from 5 choose 3, etc)

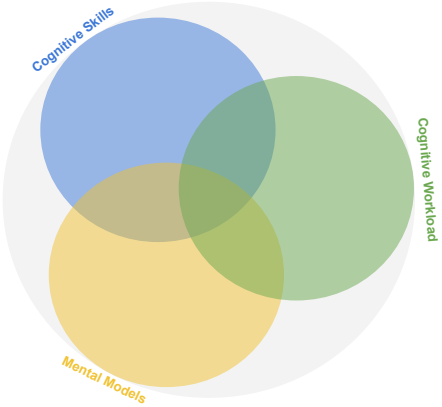


Can we ***predict user performance*** at robot teleoperation ***only with*** information about their ***cognitive state of mind***?

- Cognitive skills can predict future performance at command and control tasks.
- Applications towards role assignment and human-robot teaming are viable.
- Cognitive skill/task relationships hold between-subjects.

Kolb, Jack, et al. "Predicting Individual Human Performance in Human-Robot Teaming."
2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN). IEEE, 2021.

Kolb, Jack, et al. "Leveraging Cognitive States in Human-Robot Teaming."
2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE, 2022.



We can leverage user *cognitive skills*, *cognitive workload*, and *mental models* to enhance human-robot teams.

- **Cognitive Skills:**
  - ✓ Predict future robot operation performance using cognitive skills.
  - ✓ Demonstrate with user role assignment for human-robot teams.

  [Kolb et al. 2021, Kolb et al. 2022]

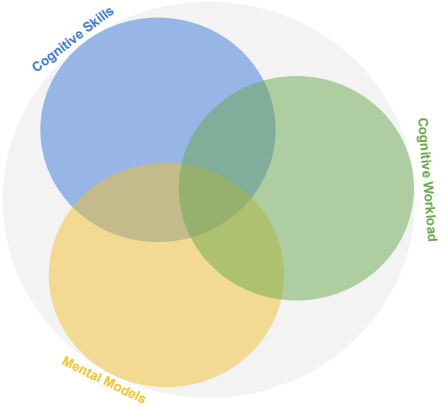
- Cognitive Workload:
  - ~ Identify off-nominal cognitive workload from physiological metrics.
  - ~ Demonstrate model transfer across human-AI team domains.

  [Agbeyibor et al. 2024]

- Mental Models:
  - ~ Estimate a human teammate's belief state from observations.
  - Personalize belief state estimation to individual users.

✓ Completed Work

~ In Progress Work



We can leverage user *cognitive skills*, *cognitive workload*, and *mental models* to enhance human-robot teams.

- Cognitive Skills:
  - ✓ Predict future robot operation performance using cognitive skills.
  - ✓ Demonstrate with user role assignment for human-robot teams.

  [Kolb et al. 2021, Kolb et al. 2022]

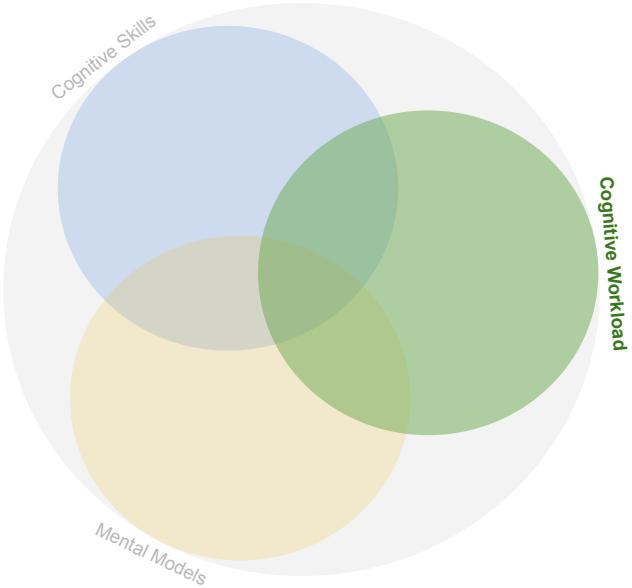
- **Cognitive Workload:**
  - ~ Identify off-nominal cognitive workload from physiological metrics.
  - ~ Demonstrate model transfer across human-AI team domains.

  [Agbeyibor et al. 2024]

- Mental Models:
  - ~ Estimate a human teammate's belief state from observations.
  - Personalize belief state estimation to individual users.

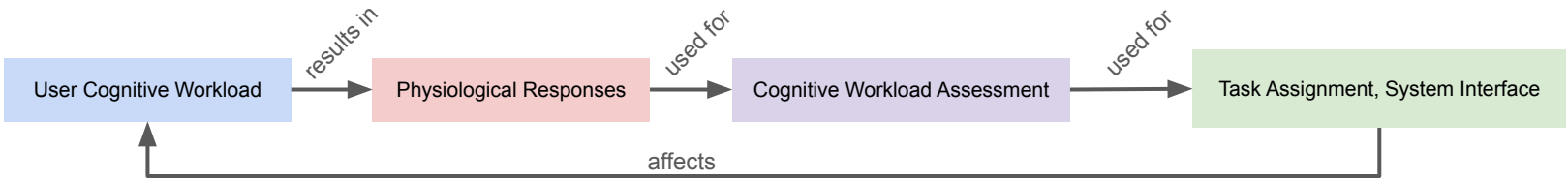
✓ Completed Work

~ In Progress Work



Can we ***monitor user cognitive workload*** in real-time for ***adaptive automation*** in coupled human-robot teams?

- Identify off-nominal cognitive workload from physiological metrics.
- Demonstrate model transfer across human-AI team domains.



Hypotheses:

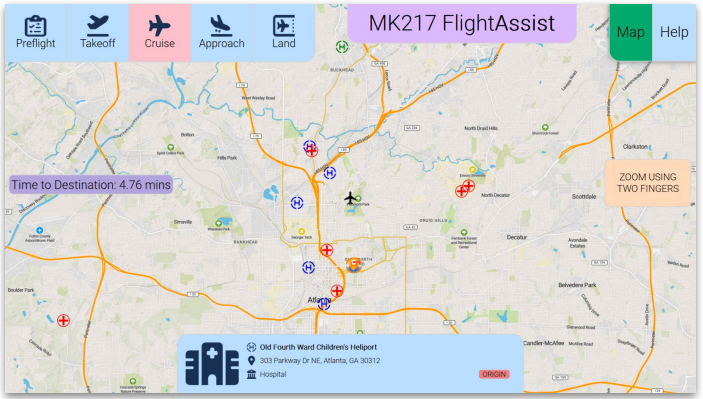
1. Off-nominal cognitive workload can be identified from **physiological responses**.
2. A human-AI system can **adapt** to the user cognitive workload to **improve** team performance.

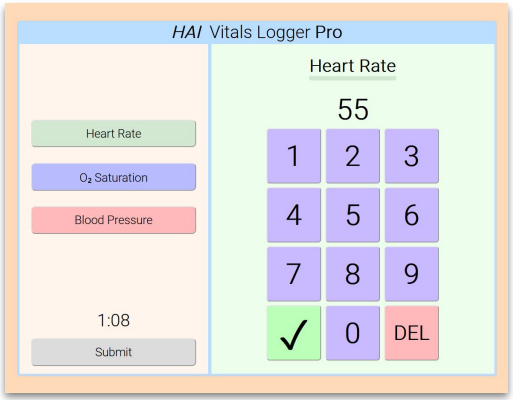
Approach:

1. Build a classifier to detect **off-nominal** workload from physiological data.
2. Adapt an AI teammate's autonomous mode to the user's workload **in a separate domain**.

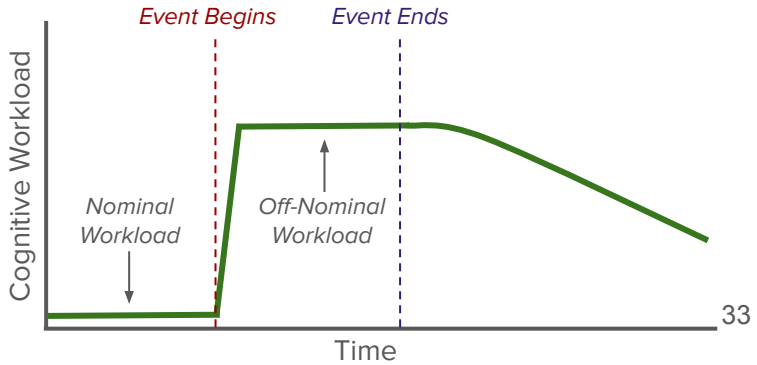
1. Build a classifier to detect **off-nominal** workload from physiological data.

MedEvac Domain





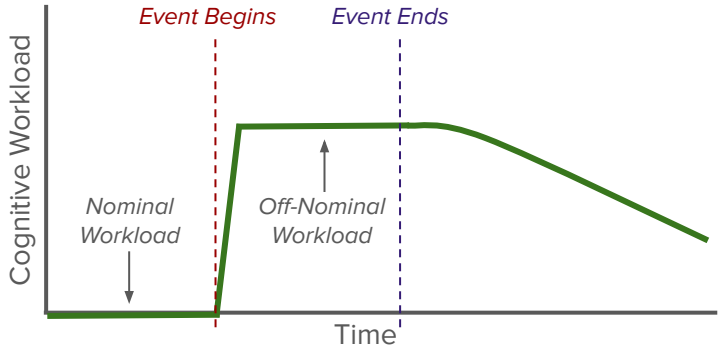
- A user is transporting an infant to a hospital in Atlanta.
- During the flight, an emergency forces a flurry of activity.
- The user interacts with the system to redirect to the nearest hospital that can support the patient.



1. Build a classifier to detect **off-nominal** workload from physiological data.

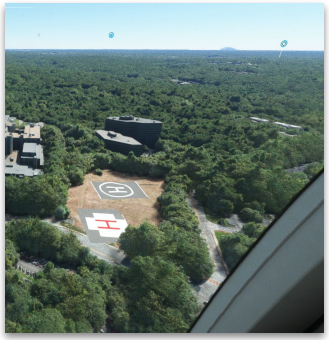
Recruited 13 participants:

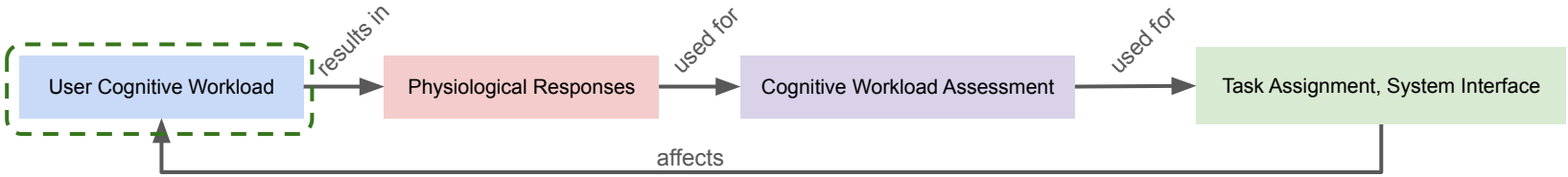
- Participants had medical experience.
- Recruited from local fire stations, hospitals, GT's EMT Club.
- Participants completed four study scenarios.

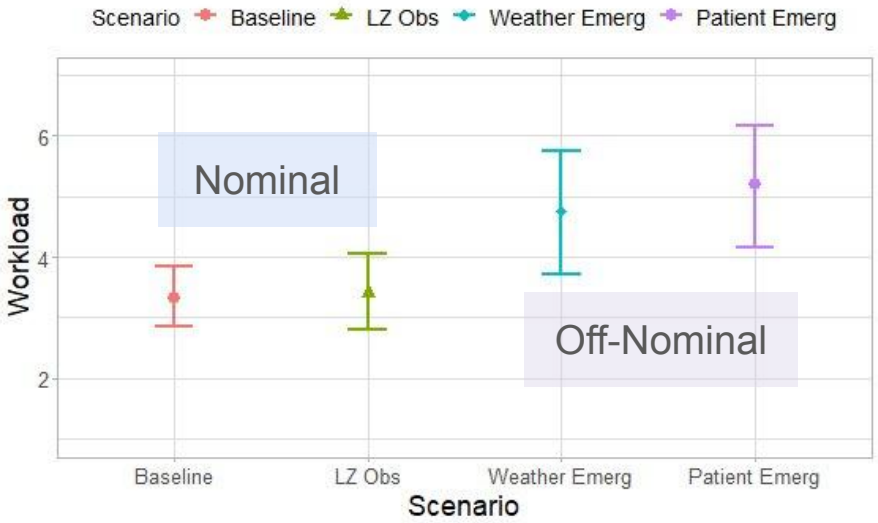


Four study scenarios:

1. ***Baseline***: no emergencies occur, flight is uneventful.
2. ***Obstructed LZ***: helipad is blocked by debris, user chooses a new destination.
3. ***Weather***: a storm forces an emergency landing, user chooses a new destination.
4. ***Medical***: patient vitals go haywire, user chooses a new destination.

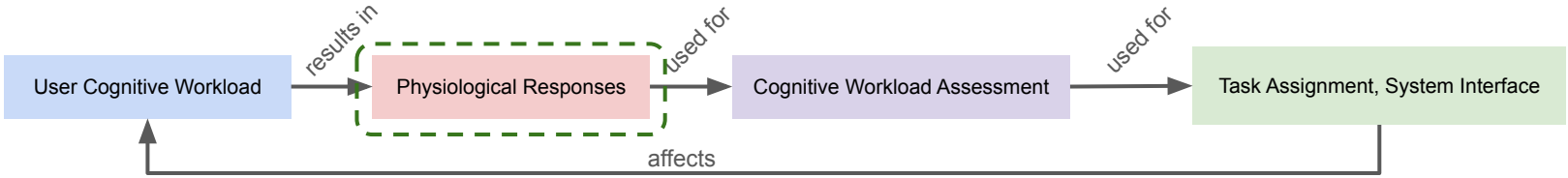




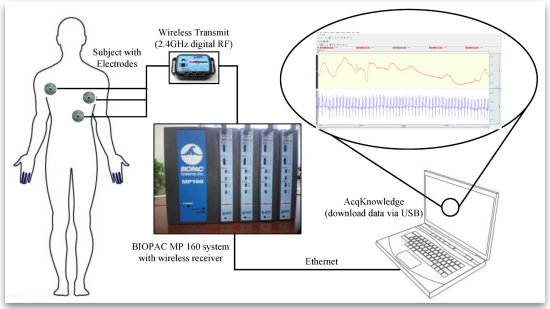


Used NASA-TLX to validate that scenarios induced **off-nominal** workload.

Can we use physiological responses to classify **nominal** vs. **off-nominal** states?



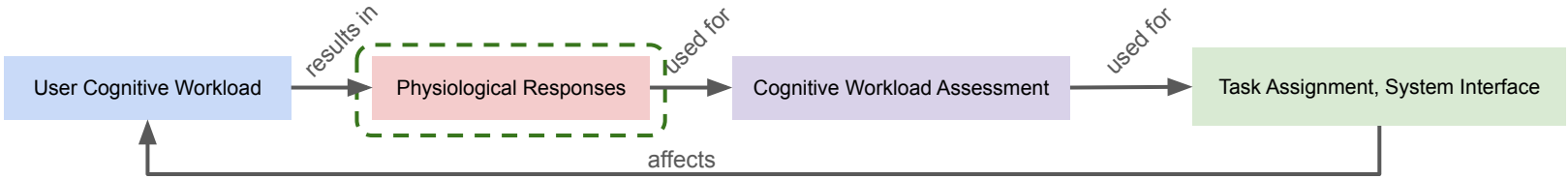
Wang et al. 2019



Physiological Sensors:

- **Heartbeat Sensor**
- **Respiration Sensor**
- **Eye Pupil Tracker**
- **Hand Joint Pose Sensor**

iMotions



From the **sensor** data...

Physiological Features:

Mean, StDev **Heart Rate**

Heart Rate Variability (SDNN: StdDev of peak-to-peak [NN] times)
Heart Rate Variability (RMSSD: mean of successive NN differences)

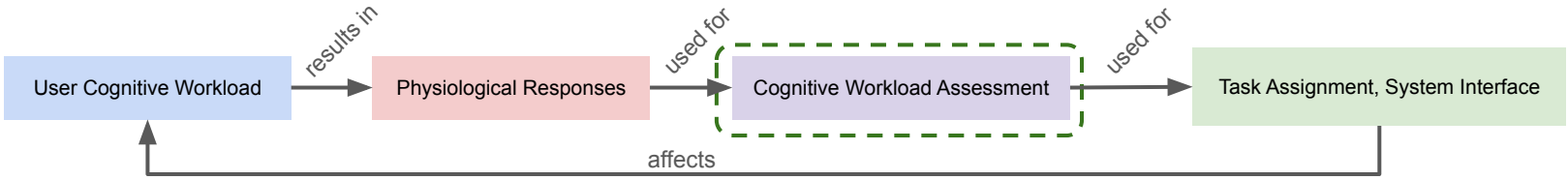
Mean, StDev **Respiration Rate**

Mean, StDev **Pupil Diameter** (left and right individually)
Mean, StDev **Saccade Length** (eye twitch)

*Mean, StDev **Hand Shakes***
***Touch Point** Accuracy*

(explored by collaborator)

Build a classifier to detect **off-nominal** workload from physiological data.

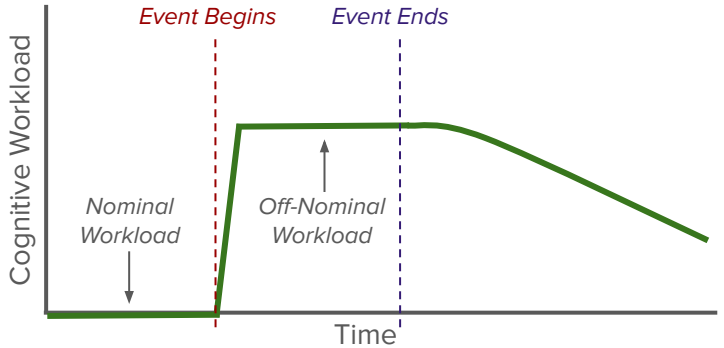


Workload Assessment:

1. *Can we classify between **nominal** and **off-nominal** scenarios?*

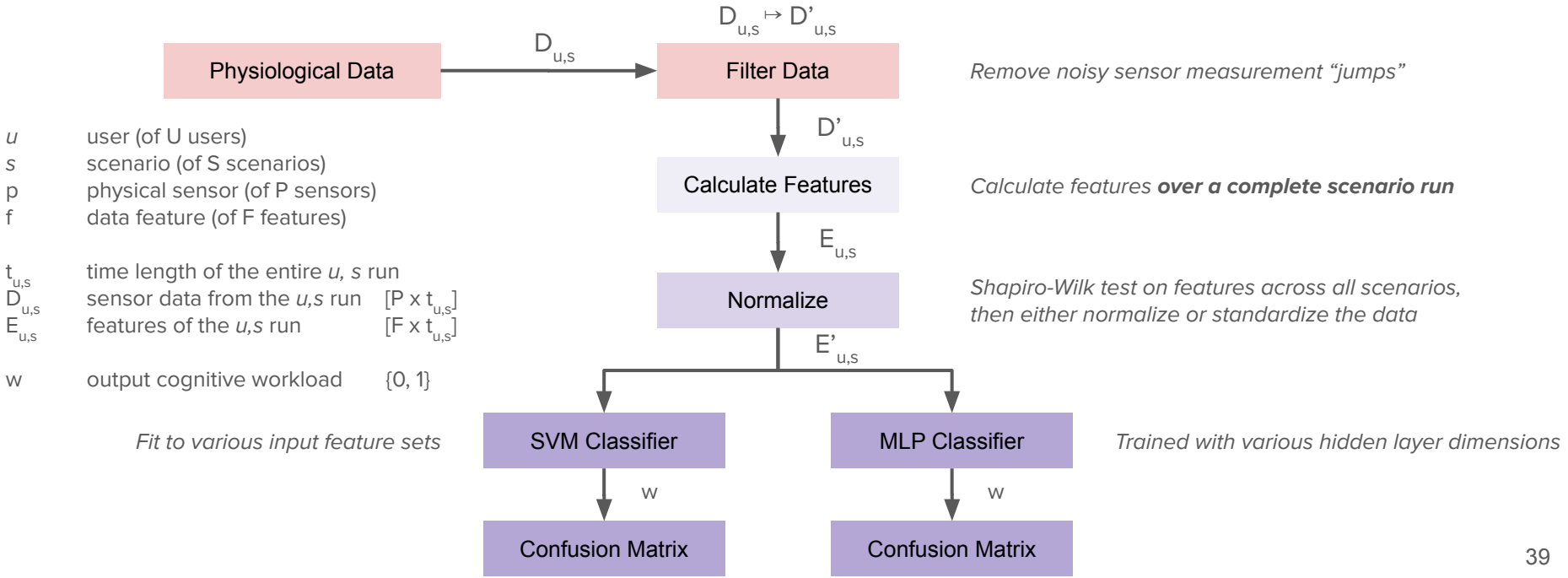
   if so...

2. *Can we identify **when** a user enters an **off-nominal** state?*



*Can we classify between **nominal** and **off-nominal** scenarios?*

Approach: motivate by classifying scenarios – ran SVM and MLP on scenario-long features.



*Can we classify between **nominal** and **off-nominal** scenarios?*

Approach: motivate by classifying scenarios – ran SVM and MLP on scenario-long features.

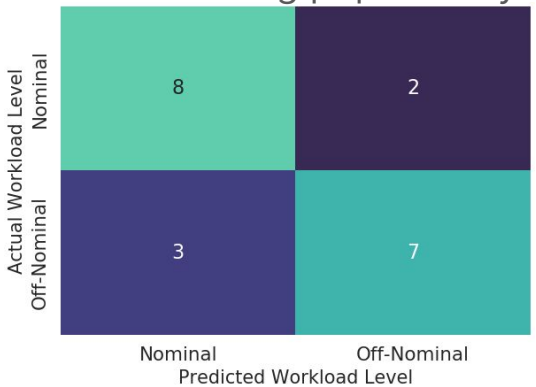
Including pupilometry

F1: .80 (16/20)

SVM, with features:

std heart rate, heart rate var RMSSD, std resp rate, std pupil diam right

Not including pupilometry



F1: .75 (15/20)

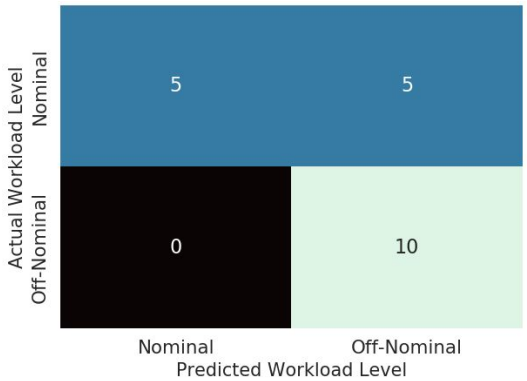
SVM, with features:

std heart rate, heart rate var RMSSD, std resp rate

*Can we classify between **nominal** and **off-nominal** scenarios?*

Approach: motivate by classifying scenarios – ran SVM and MLP on scenario-long features.

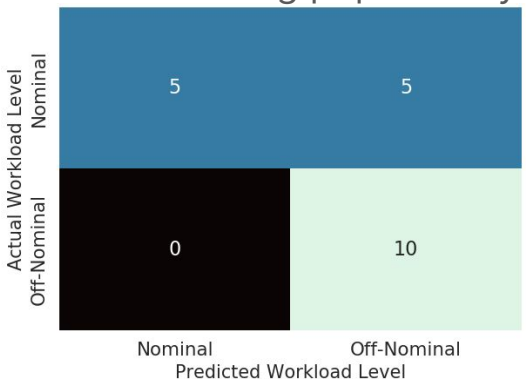
Including pupilometry



F1: .75 (15/20)

MLP, with **3** hidden layers of **8** nodes, using all features.

Not including pupilometry



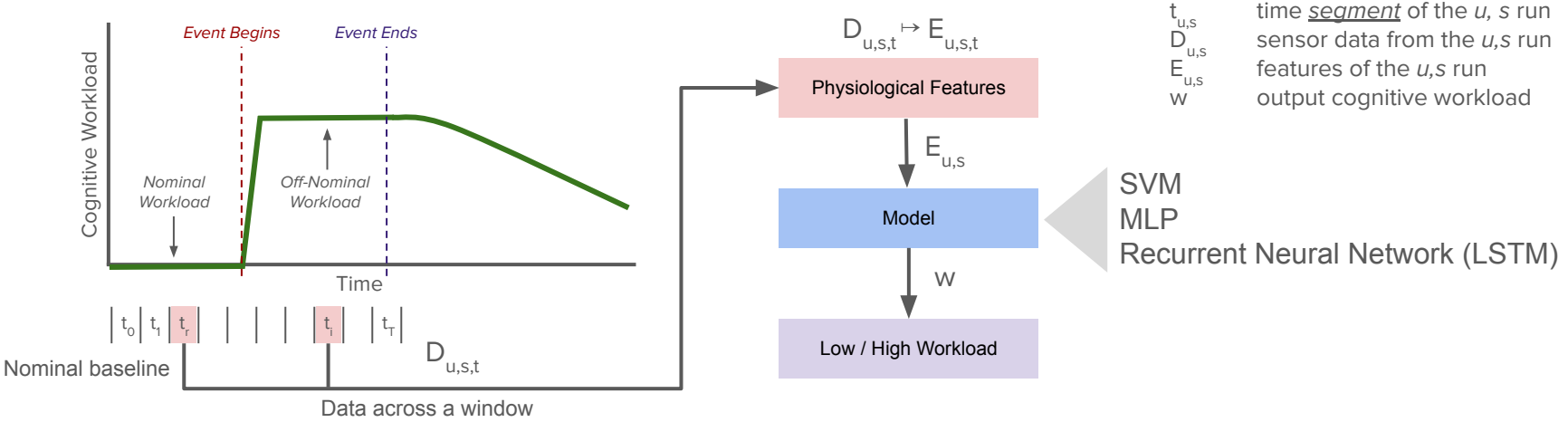
F1: .75 (15/20)

MLP, with **6** hidden layers of **256** nodes, using all features *except* pupilometry.

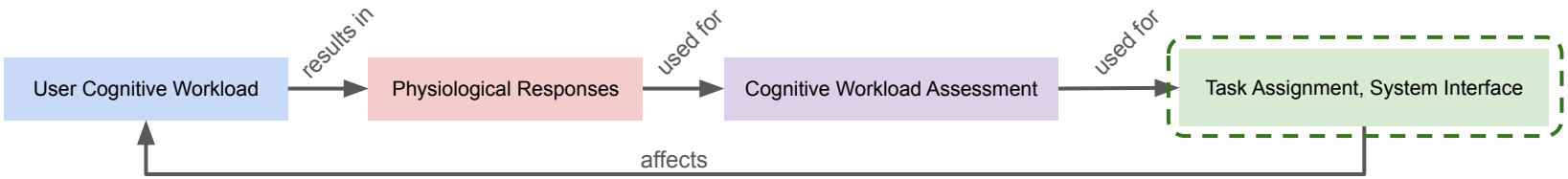
*Can we identify **when** a user enters an **off-nominal** state?*

Takeaways:

- Classification with scenario-long features can distinguish between **nominal** and **off-nominal** scenarios.
- Findings motivate classifying real-time workload state.



u	user (of U users)	
s	scenario (of S scenarios)	
p	physical sensor (of P sensors)	
f	data feature (of F features)	
$t_{u,s}$	time *segment* of the *u, s* run	
$D_{u,s}$	sensor data from the *u,s* run	$[P \times t_{u,s}]$
$E_{u,s}$	features of the *u,s* run	$[F \times t_{u,s}]$
w	output cognitive workload	{0, 1}



Hypotheses:

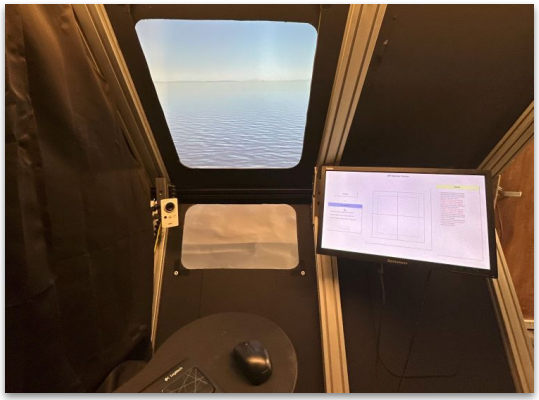
1. Off-nominal cognitive workload can be identified from **physiological responses**.
2. A human-AI system can **adapt** to the user cognitive workload to **improve** team performance.

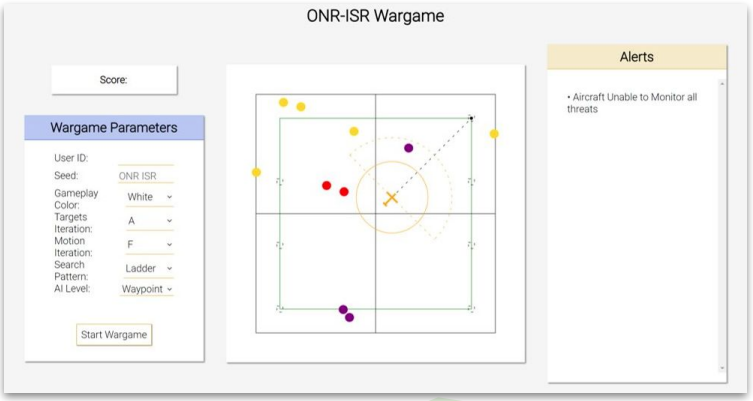
Approach:

1. Build a classifier to detect **off-nominal** workload from physiological data.
2. Adapt an AI teammate's autonomous mode to the user's workload **in a separate domain**.

Adapt an AI teammate's autonomous mode to the user's workload **in a separate domain**.

ISR Domain



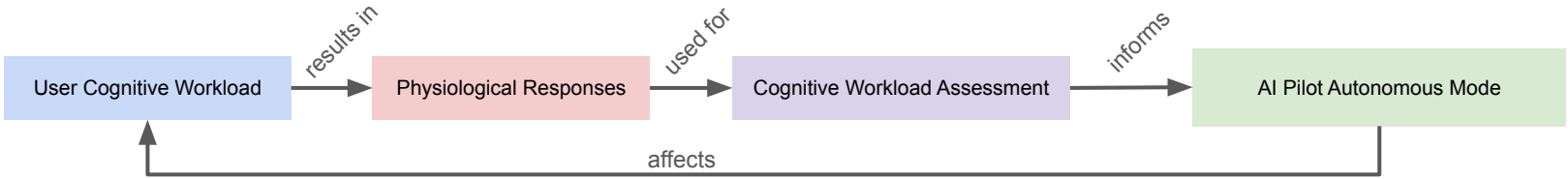


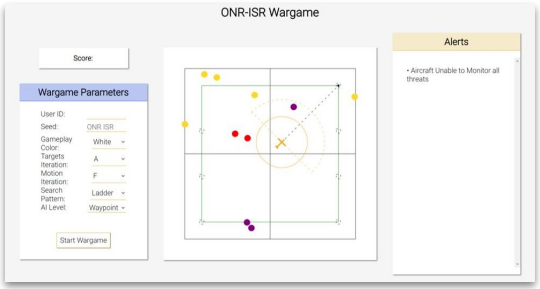
- A user scans a coastline to keep track of threat actors, while an AI pilot follows a known search pattern (e.g., *box*).
- Different threat classes require different responses, the user adds immediate waypoints to support the mission.
- The AI pilot has several autonomous modes it can select from (e.g., avoiding threat actors).

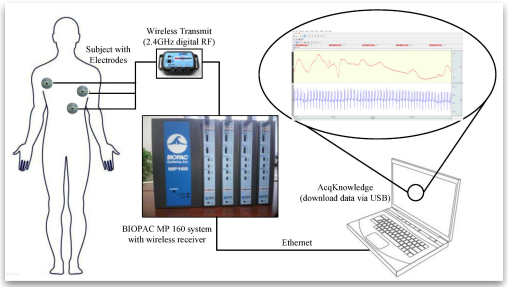
Autonomous Modes:

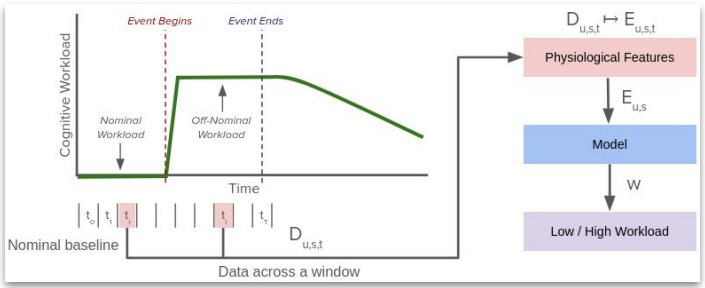
- Follow waypoints.
- Verify user waypoints.
- Suggest search pattern.
- Collision avoidance.

Adapt an AI teammate's autonomous mode to the user's workload **in a separate domain**.









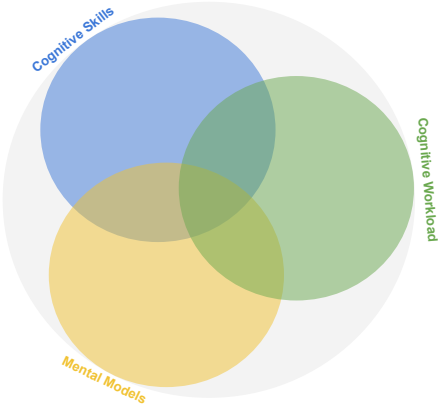
Evaluation Metrics:

Cognitive Workload: *NASA TLX post-hoc*
Performance: *Task score with adaptive AI pilot vs. non-adaptive AI pilot*

Can we ***monitor user cognitive workload*** in real-time for ***adaptive automation*** in coupled human-robot teams?

- Preliminary work suggests physiological metrics indicate cognitive workload in coupled human-robot teams.
- **Proposed work** will explore in-situ monitoring to inform adaptive automation.

Agbeyibor, Richard, et al. "Run Time Assurance and Human AI Fluency in Manned Autonomous Intelligence Surveillance and Reconnaissance." *Under review for the 2024 AIAA Aviation Forum and Exposition,*.



We can leverage user *cognitive skills*, *cognitive workload*, and *mental models* to enhance human-robot teams.

- Cognitive Skills:
  - ✓ Predict future robot operation performance using cognitive skills.
  - ✓ Demonstrate with user role assignment for human-robot teams.

  [Kolb et al. 2021, Kolb et al. 2022]

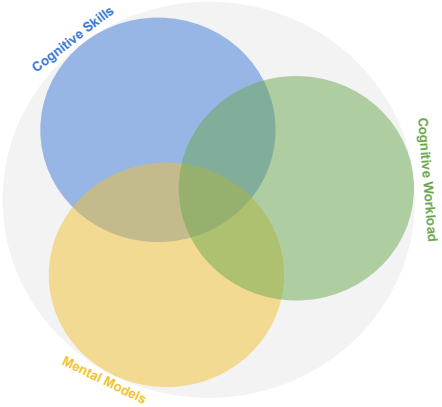
- **Cognitive Workload:**
  - ~ Identify off-nominal cognitive workload from physiological metrics.
  - ~ Demonstrate model transfer across human-AI team domains.

  [Agbeyibor et al. 2024]

- Mental Models:
  - ~ Estimate a human teammate's belief state from observations.
  - Personalize belief state estimation to individual users.

✓ Completed Work

~ In Progress Work



We can leverage user *cognitive skills*, *cognitive workload*, and *mental models* to enhance human-robot teams.

- Cognitive Skills:
  - ✓ Predict future robot operation performance using cognitive skills.
  - ✓ Demonstrate with user role assignment for human-robot teams.

  [Kolb et al. 2021, Kolb et al. 2022]

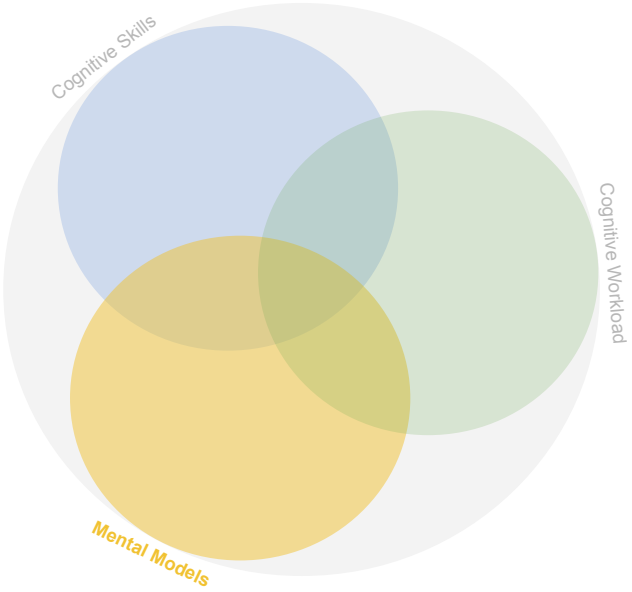
- Cognitive Workload:
  - ~ Identify off-nominal cognitive workload from physiological metrics.
  - ~ Demonstrate model transfer across human-AI team domains.

  [Agbeyibor et al. 2024]

- **Mental Models:**
  - ~ Estimate a human teammate's belief state from observations.
  - Personalize belief state estimation to individual users.

✓ Completed Work

~ In Progress Work

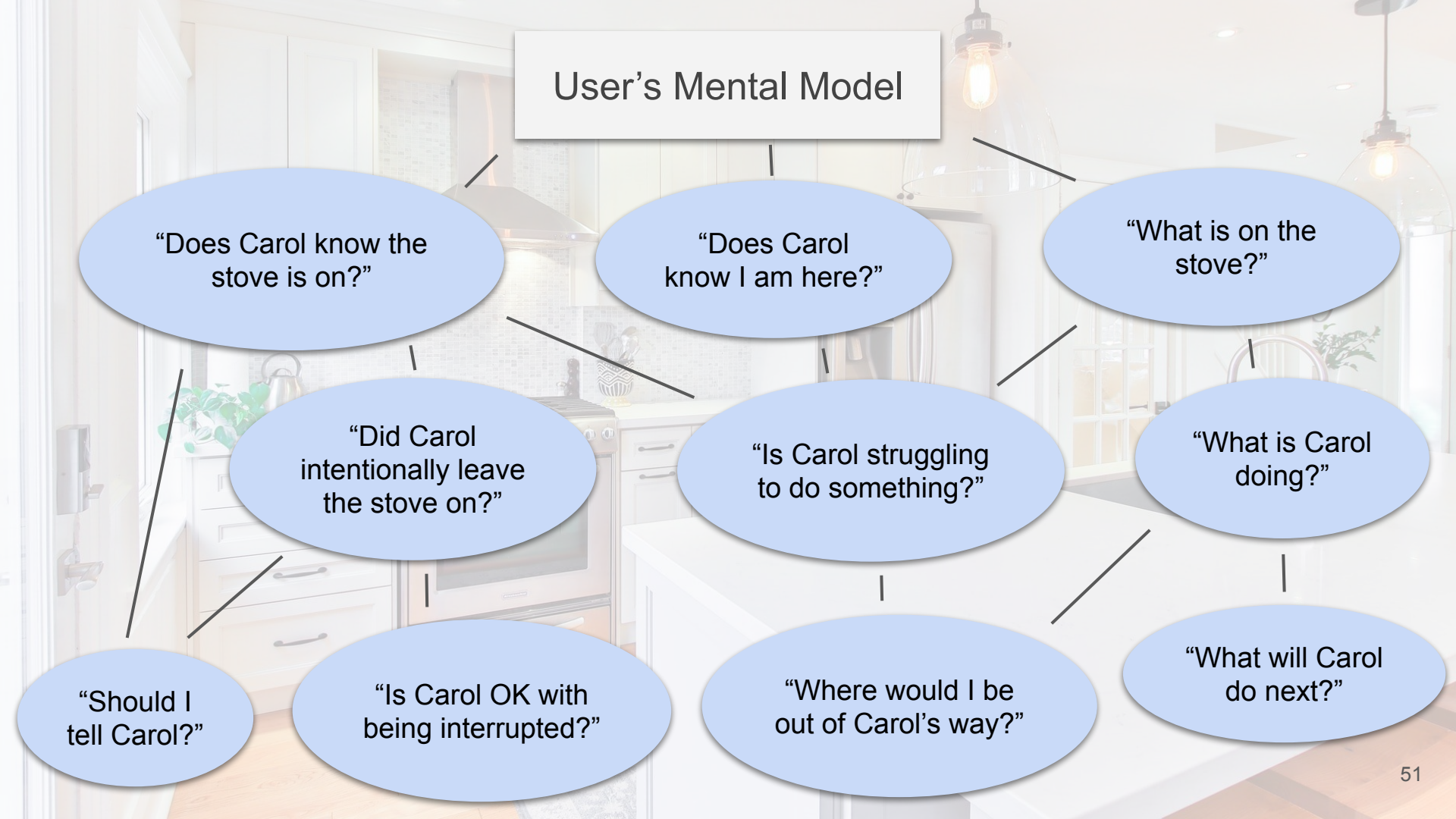


Can robots estimate their human teammate's ***mental model*** to ***reason*** about them in ***complex environments***?

- Estimate a human teammate's belief state from observations. **[Proposed work]**
- Personalize belief state estimation to individual users. **[Proposed work]**


Carol Leaves

User's Mental Model



Related work focuses on:

- End-to-end models that answer **specific** questions.
- **Logical predicate** scene graphs that are general-purpose.

*What if we **do not know** the questions we want to answer?*

*What if we want to consider **user or observational uncertainty**?*

Decisions

"Where should I hang out?"

"Is now a good time to remind Carol?"

"Should I alert the group that I am present?"

Queries

"Could you help me find my keys?"

"Does Carol know I placed milk in the fridge?"

"Could you ensure Carol takes her medicine?"

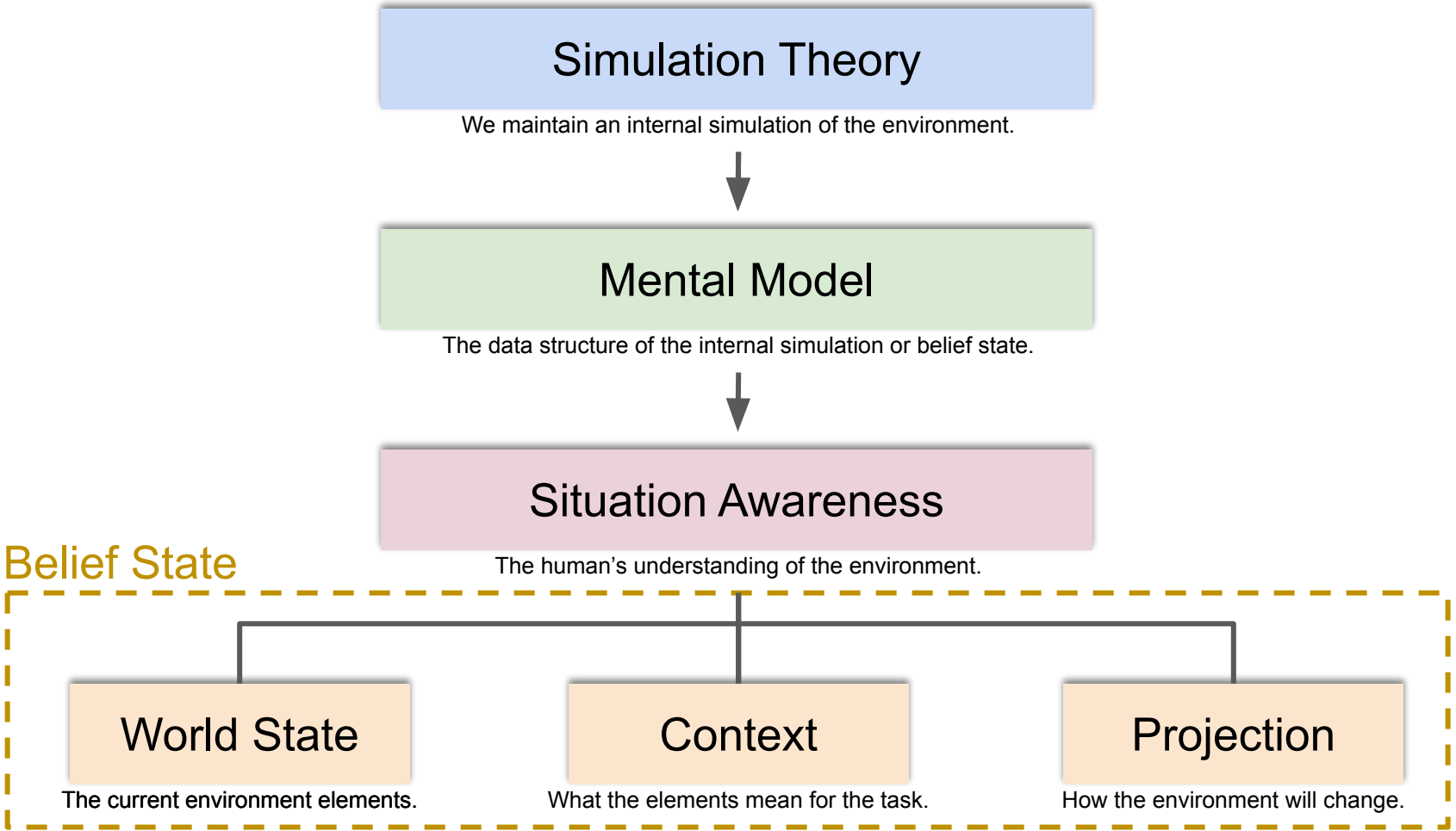
Plans

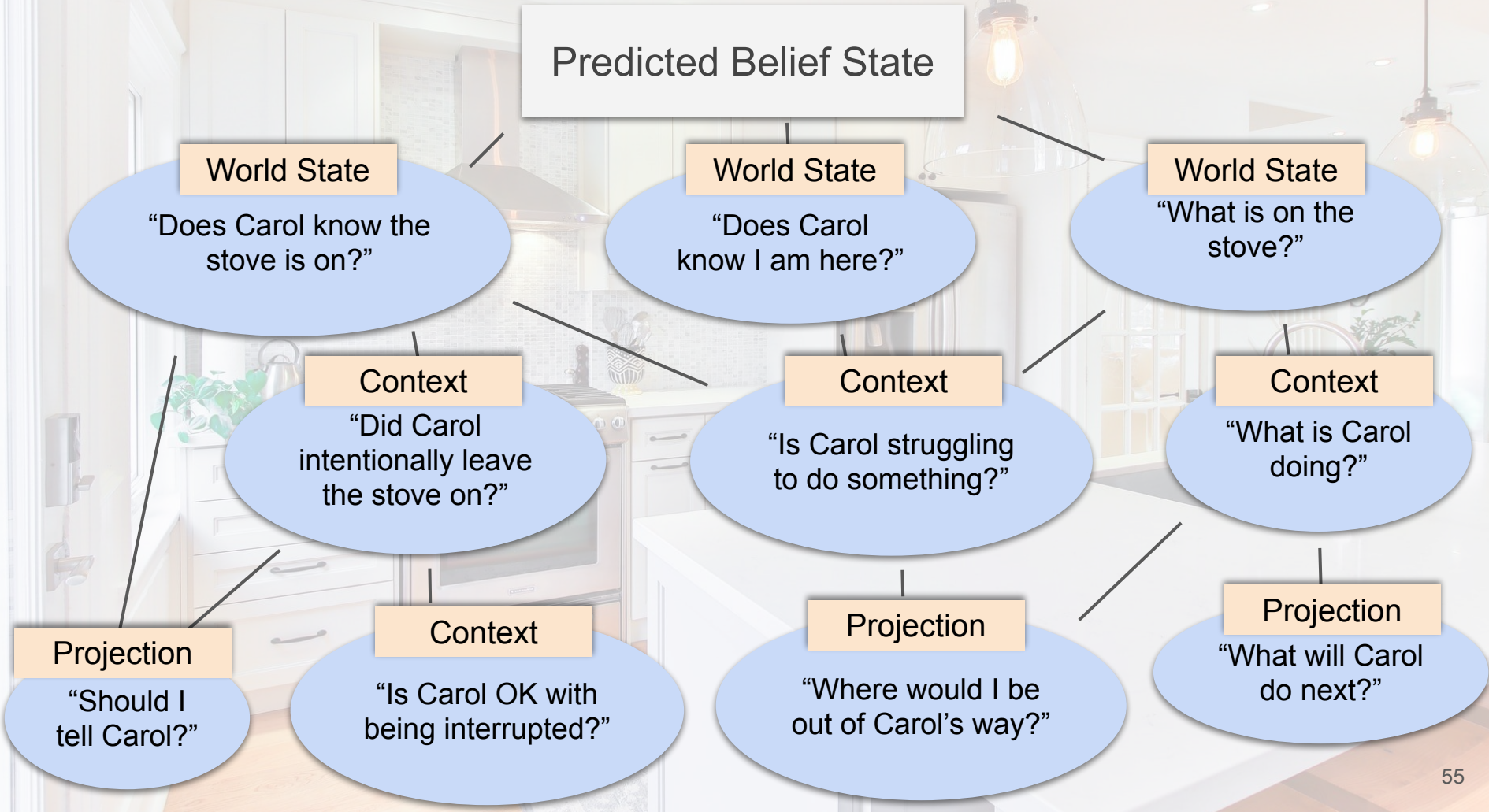
"How should I navigate around the people?"

"When would be a good time to charge?"

"Do the humans want to see me?"

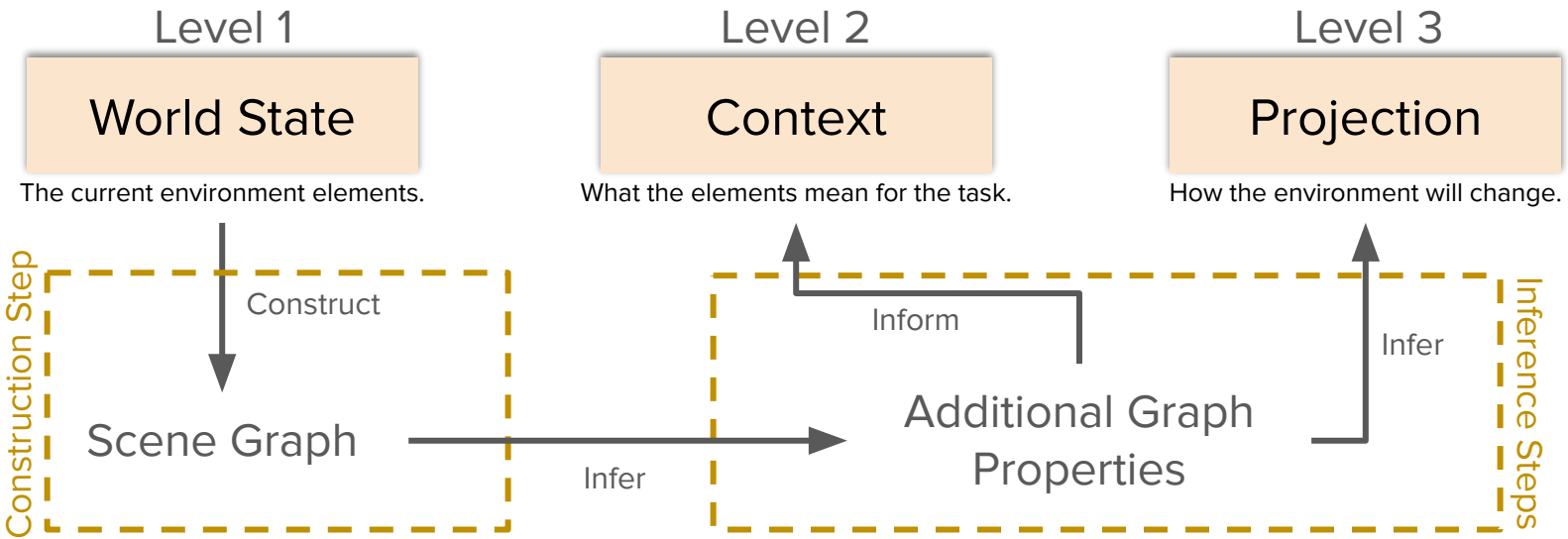
A strong **predicted mental model** of humans in the environment is useful for numerous HRI questions

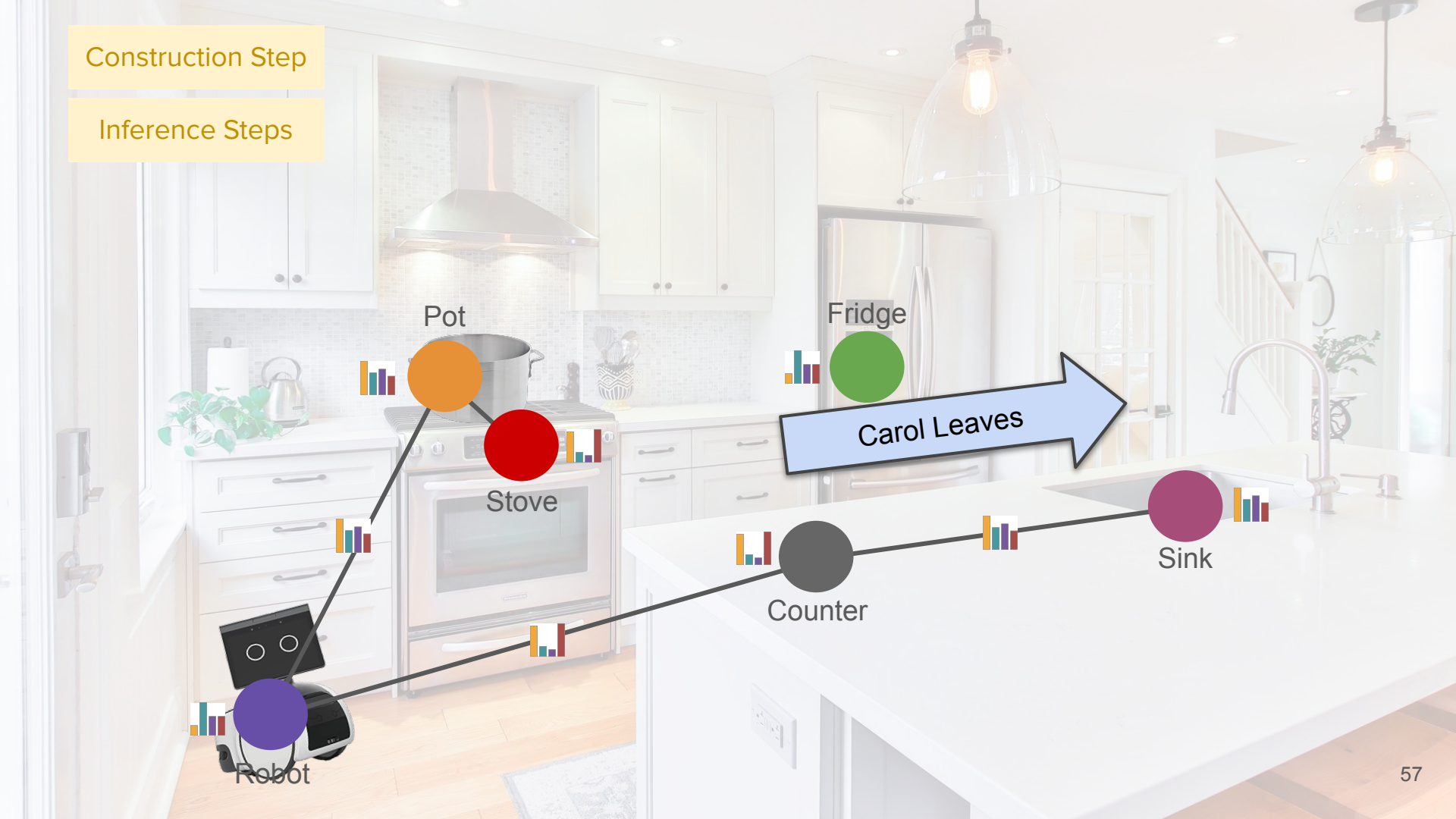

Simulation Theory
We maintain an internal simulation of the environment.
Mental Model
The data structure of the internal simulation or belief state.
Situation Awareness
The human's understanding of the environment.
Belief State
World State
The current environment elements.
Context
What the elements mean for the task.
Projection
How the environment will change.


Predicted Belief State
World State
“Does Carol know the stove is on?”
World State
“Does Carol know I am here?”
World State
“What is on the stove?”
Context
“Did Carol intentionally leave the stove on?”
Context
“Is Carol struggling to do something?”
Context
“What is Carol doing?”
Projection
“Should I tell Carol?”
Context
“Is Carol OK with being interrupted?”
Projection
“Where would I be out of Carol’s way?”
Projection
“What will Carol do next?”

Objective: Estimate a human teammate's ***belief state*** as aligned to situation awareness.

Approach: Leverage scene graphs and foundation models to enable open-ended inference.



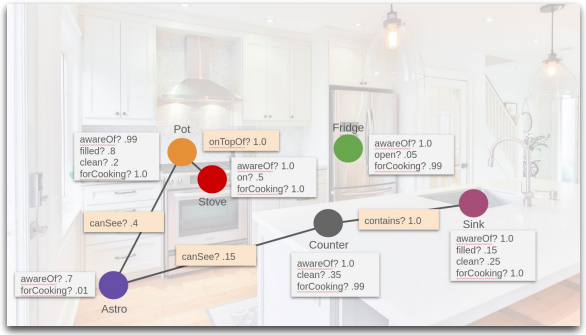

Construction Step
Inference Steps
Pot
Fridge
Carol Leaves
Stove
Sink
Counter
Robot

Latent Representation ⟷ Logical Representation

Fuzzy Logic Model

Logical Predicates Model

Indirect Observation → Update Predicted Scene Graph {0, 1}



Construction Step

Replacing binary beliefs with probabilistic beliefs

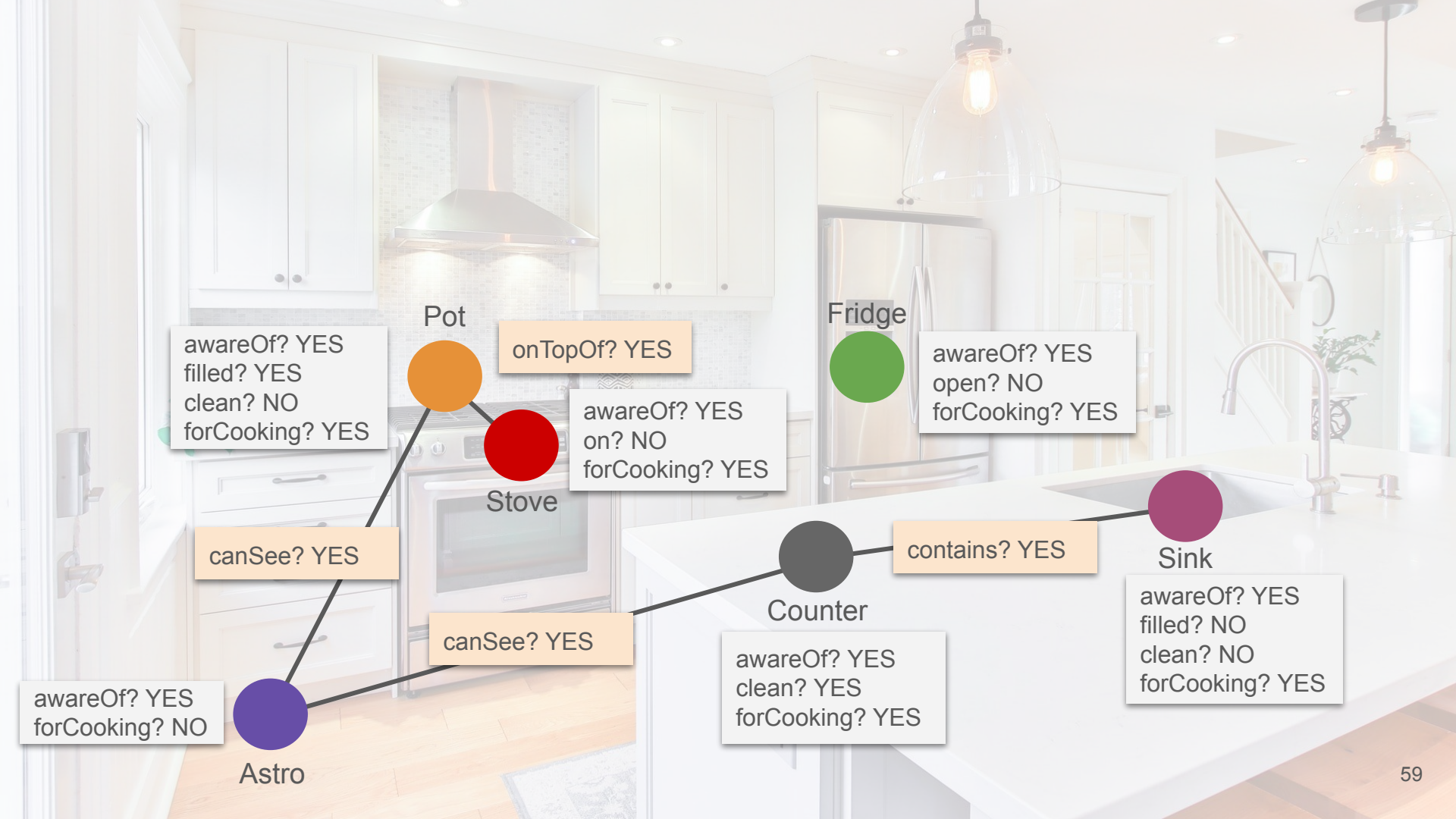
Fuzzy Logic Model

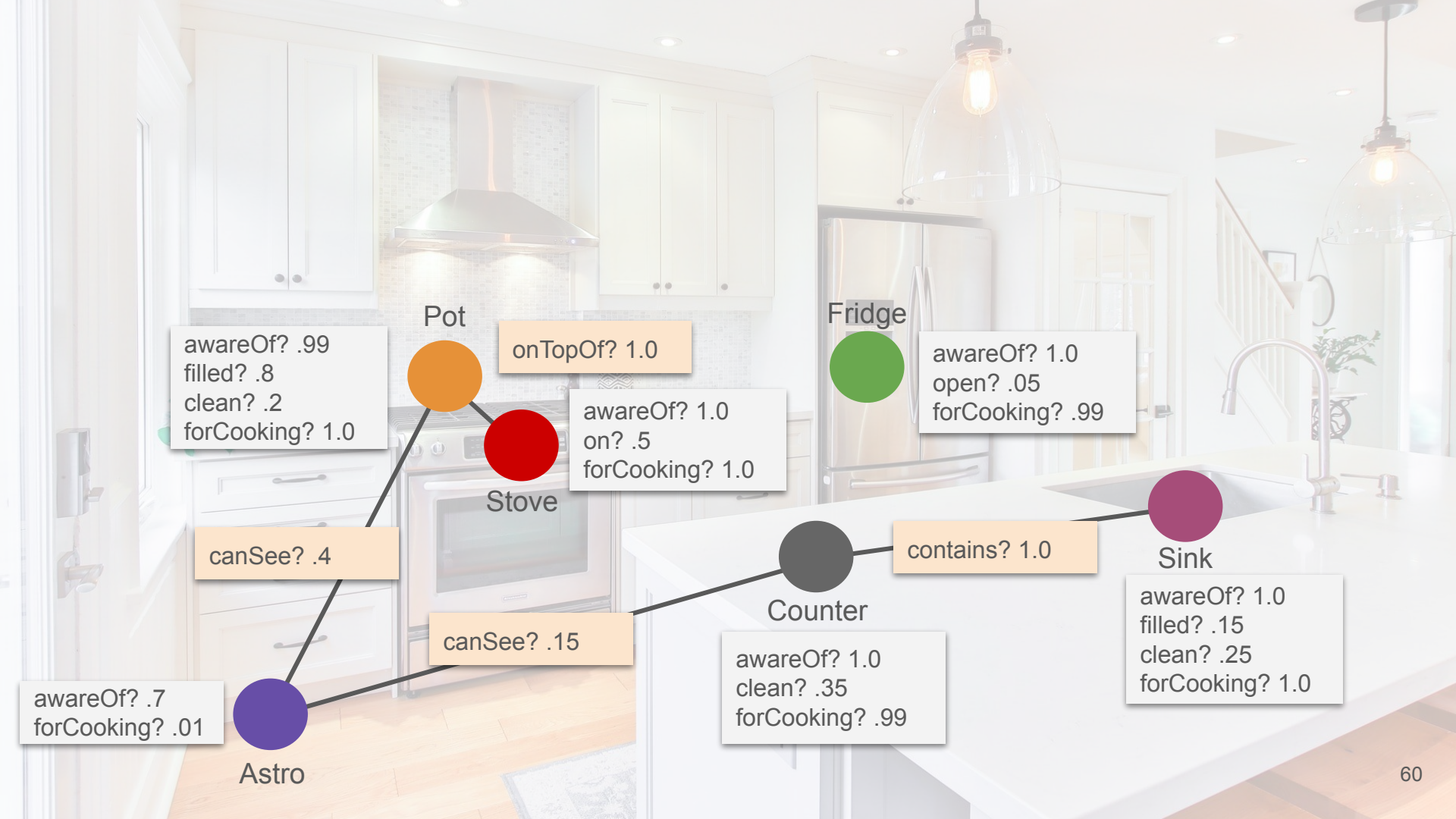
Indirect Observation → User Perception → User Comprehension → Update Predicted Scene Graph ℝ

What could the user have perceived?

How could the user have interpreted the event?

Bolton et al. 2023


Pot
awareOf? YES
filled? YES
clean? NO
forCooking? YES
onTopOf? YES
Stove
awareOf? YES
on? NO
forCooking? YES
Fridge
awareOf? YES
open? NO
forCooking? YES
canSee? YES
contains? YES
Sink
awareOf? YES
filled? NO
clean? NO
forCooking? YES
canSee? YES
Counter
awareOf? YES
clean? YES
forCooking? YES
awareOf? YES
forCooking? NO
Astro


Pot
awareOf? .99
filled? .8
clean? .2
forCooking? 1.0
onTopOf? 1.0
Stove
awareOf? 1.0
on? .5
forCooking? 1.0
Fridge
awareOf? 1.0
open? .05
forCooking? .99
canSee? .4
contains? 1.0
Sink
awareOf? 1.0
filled? .15
clean? .25
forCooking? 1.0
Counter
awareOf? 1.0
clean? .35
forCooking? .99
canSee? .15
awareOf? .7
forCooking? .01
Astro

Indirect Observation → User Perception → User Comprehension → Update Predicted Scene Graph

What could the user have perceived?

How could the user have interpreted the event?

α

The likelihood that the user perceived an event.

$$\Sigma \rightarrow I$$

- Σ finite set of input events
- I likelihood that each input event was perceived

Bao, J.; Jia, Y.; Cheng, Y.; Xi, N. Saliency-Guided Detection of Unknown Objects in RGB-D Indoor Scenes. *Sensors* 2015.

Indirect Observation → User Perception → User Comprehension → Update Predicted Scene Graph

What could the user have perceived?

α

How could the user have interpreted the event?

ϕ

Shutterstock

"Which burner was ignited?"

What the user perceived the event as.

$$\Sigma, Q \rightarrow M$$

- Σ finite set of input events
- Q possible state changes
- M belief state of the user

Indirect Observation → User Perception → User Comprehension → Update Predicted Scene Graph

What could the user have perceived?

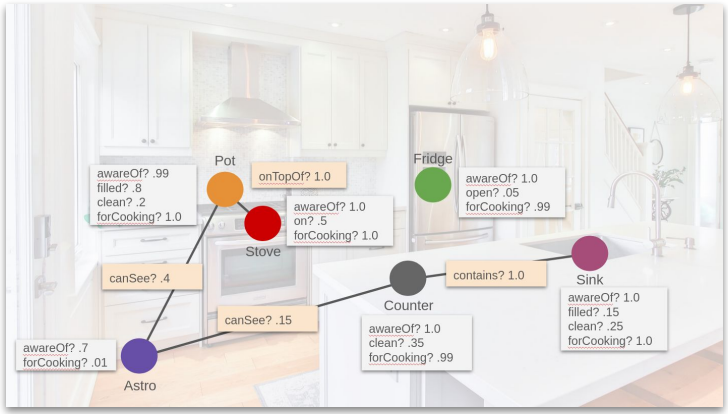
α

How could the user have interpreted the event?

ϕ

How does that update the user's belief state?

δ

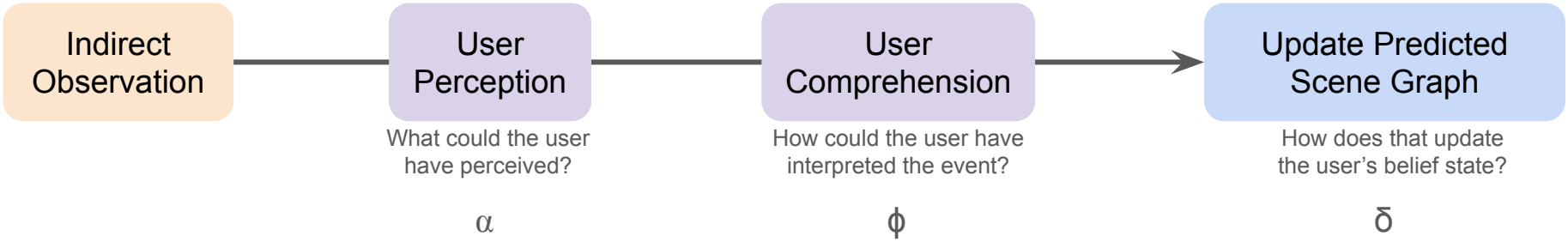


What the user's new belief state is.

$M, \alpha, \phi \rightarrow M'$

M belief state of the user

Updates the robot's prediction of the user's belief state.



Can **personalize** α, ϕ, and δ to the **human** and the **task** domain.

Can explore **different models** for α, ϕ, and δ (classical, deep learning).

Encapsulates logical predicate and probabilistic scene-graph models.

Experiment 1

"How do we compare to the logical predicates model?"

In a **simulated environment**, create a dataset on user **situation awareness** as they complete a co-located task.

Study Hypotheses:

1. In the construction step, the fuzzy logic models will outperform the logical predicates model across all observability types.
2. In the inference step, tradeoffs exist between each model's performance and compute.

Construction Step

- Logical Predicates Model (baseline)
- Fuzzy Logic (novel)

Inference Steps

- Logical Predicates Model (baseline)
- Large Language Model (novel)
- Graph Neural Network (novel)

Observability

- Full-Observability
- Frontal-Observability (all visible space)
- Proximal-Observability (immediate visible)

Experiment 1

"How do we compare to the logical predicates model?"

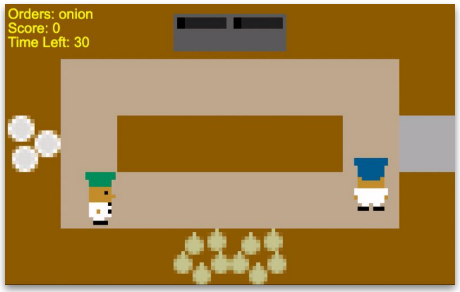
In a **simulated environment**, create a dataset on user **situation awareness** as they complete a co-located task.

Study Hypotheses:

1. In the construction step, the fuzzy logic models will outperform the logical predicates model across all observability types.
2. In the inference step, tradeoffs exist between each model's performance and compute.

Environments

- Overcooked-AI: proof of concept



- VirtualHome: closer to reality

Transfer methods to

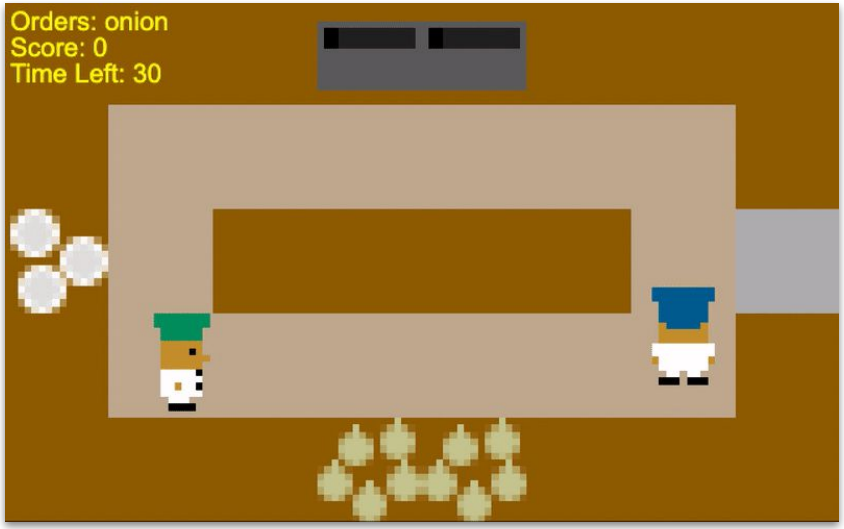
Experiment 1

"How do we compare to the logical predicates model?"

In a **simulated environment**, create a dataset on user **situation awareness** as they complete a co-located task.

Started with CHAI's *Overcooked-AI* environment, which features:

- **AI teammate** has an active role in the team.
- **Human teammate** controls their character
- Task has clear objectives and performance criteria.
- Game rounds each take a few minutes.



Recipe:

1. Place three ingredients in a pot to cook soup.
2. Place the cooked soup onto a dish.
3. Place the plated soup onto a serving station.
4. Repeat until all ingredients are cooked, or time expires.

Experiment 1

"How do we compare to the logical predicates model?"

In a **simulated environment**, create a dataset on user **situation awareness** as they complete a co-located task.

Started with CHAI's *Overcooked-AI* environment, which features:

- **AI teammate** has an active role in the team.
- **Human teammate** controls their character
- Task has clear objectives and performance criteria.
- Game rounds each take a few minutes.

Environment Modifications:

- **Partial observability.**
- SAGAT-style interruptions.
- Piped world state to a belief state module.

Experiment 1

"How do we compare to the logical predicates model?"

In a **simulated environment**, create a dataset on user **situation awareness** as they complete a co-located task.

Key modifications:

- Partial observability.
- **SAGAT-style interruptions.**
- Piped world state to a belief state module.

Where is the nearest dish?

Is the top-left pot done cooking?

What is your teammate doing?

What do you plan to do next?

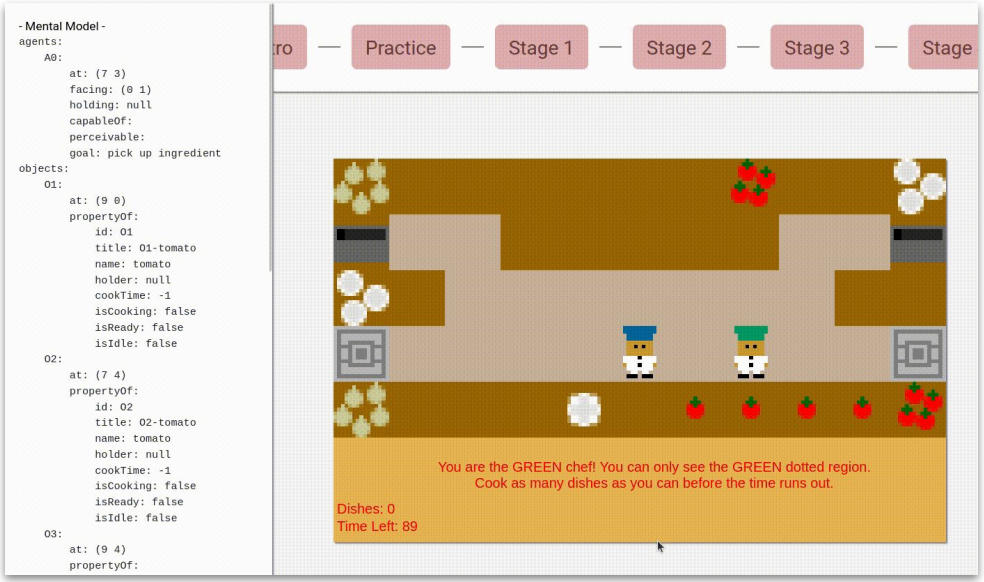
Experiment 1

"How do we compare to the logical predicates model?"

In a **simulated environment**, create a dataset on user **situation awareness** as they complete a co-located task.

Key modifications:

- Partial observability.
- SAGAT-style interruptions.
- **Piped world state to a belief state module.**



Left: The logical predicates baseline updating in real-time.

We set the AI's observability shape and range to explore how belief state models perform at different observability.

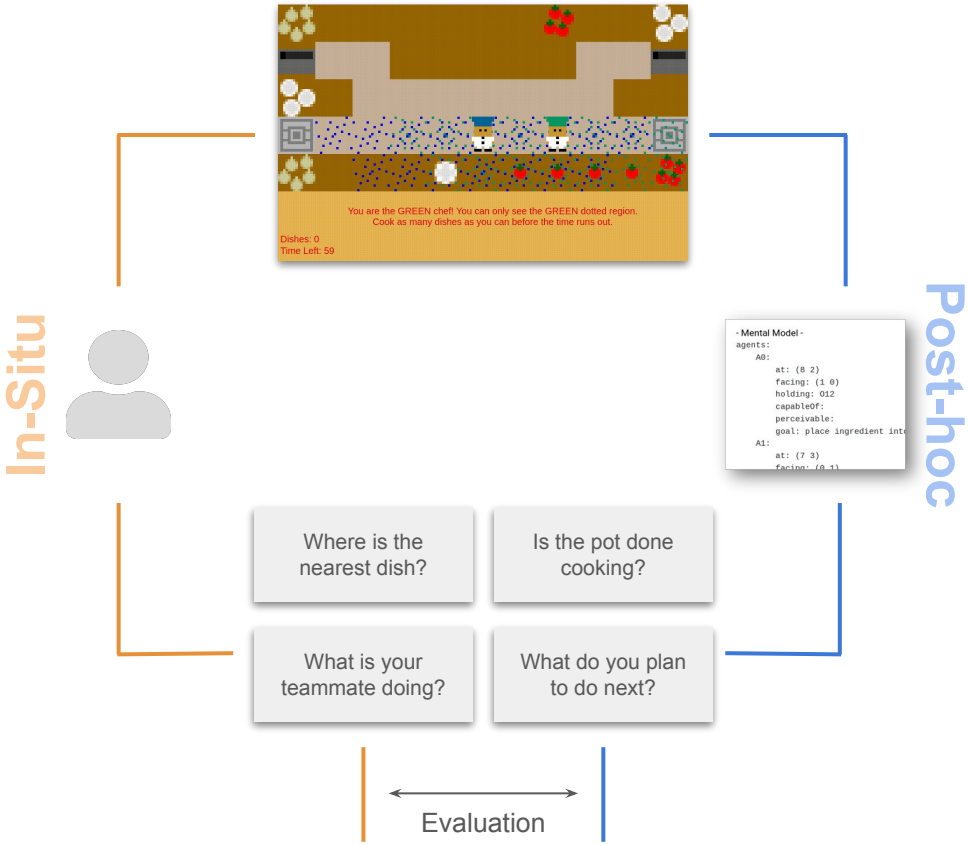
Experiment 1

"How do we compare to the logical predicates model?"

In a **simulated environment**, create a dataset on user **situation awareness** as they complete a co-located task.

Study Hypotheses:

1. In the construction step, the fuzzy logic models will outperform the logical predicates model across all observability types.
2. In the inference step, tradeoffs exist between each model's performance and compute.



Inference Steps

Where is the nearest dish?

Is the pot done cooking?

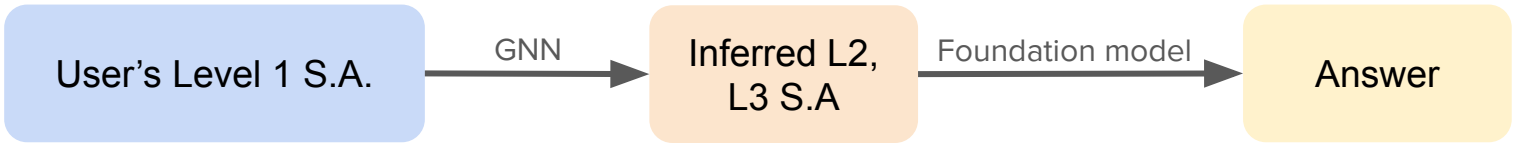
What is your teammate doing?

What do you plan to do next?

- *Logical Predicates Model* (baseline)

- *Large Language Model* (novel)

- *Large Language Model + Graph Neural Network (novel)*



Experiment 1

"How do we compare to the logical predicates model?"

In a **simulated environment**, create a dataset on user **situation awareness** as they complete a co-located task.

Study Hypotheses:

1. In the construction step, the fuzzy logic models will outperform the logical predicates model across all observability types.
2. In the inference step, tradeoffs exist between each model's performance and compute.

Evaluate accuracy (**F1**) across S.A. questions:

- **State**, locations of objects and agents (Level 1)
- **Context**, meaning of objects and agents (Level 2)
- **Future** plans for user and teammate (Level 3)

Experiment 2

"How can we personalize this?"

Improve the performance of the *Fuzzy Logic* and *Graph Neural Network* combination in *VirtualHome*, without requiring additional user data.

Approach:

Use generalized data to pre-train the model, simulation to fully train, and the individual's data to tune.

Study Hypothesis:

The personalized model will outperform the generalized models for individual users.

Construction Step – Fuzzy Logic Model

1. Use the base fuzzy logic model for ϕ and δ.
2. Simulate a user agent with a known height.
3. Construct model from the simulated agent's camera for α (perception).

Inference Steps – Graph Neural Network

1. Use the base GNN model.
2. Simulate cooking taskwork with the user agent.
3. Train the GNN with the simulated data, using the logical predicates model as a ground truth.
4. Tune the GNN from the user's responses.

Can robots estimate their human teammate's ***mental model*** to ***reason*** about them in ***complex environments***?

Experiment 1

Assess the viability of several methods for estimating user belief state in human-robot teaming.

Status: Developed *Overcooked* environment, Logical Predicates model.

Experiment 2

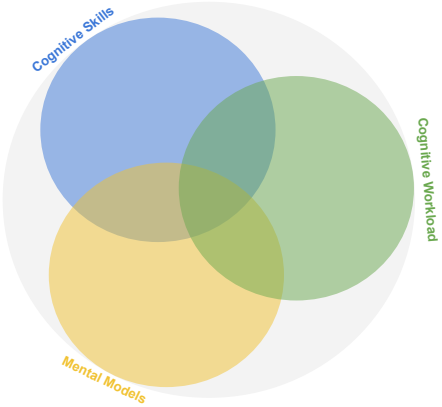
Explore how data-driven methods can be personalized to individuals.

Status: Not yet started.

Limitations

This work only considers the forward-process through the construction step and the inference step.

People also use the reverse process to predict how hidden objects will change.



We can leverage user *cognitive skills*, *cognitive workload*, and *mental models* to enhance human-robot teams.

- Cognitive Skills:
  - ✓ Predict future robot operation performance using cognitive skills.
  - ✓ Demonstrate with user role assignment for human-robot teams.

  [Kolb et al. 2021, Kolb et al. 2022]

- Cognitive Workload:
  - ~ Identify off-nominal cognitive workload from physiological metrics.
  - ~ Demonstrate model transfer across human-AI team domains.

  [Agbeyibor et al. 2024]

- Mental Models:
  - ~ Estimate a human teammate's belief state from observations.
  - Personalize belief state estimation to individual users.

✓ Completed Work

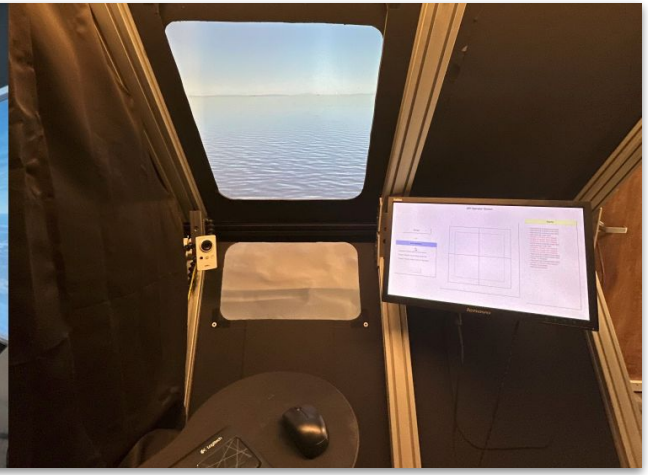
~ In Progress Work

Backup Slides

Domain 1: **Medical Evacuation**

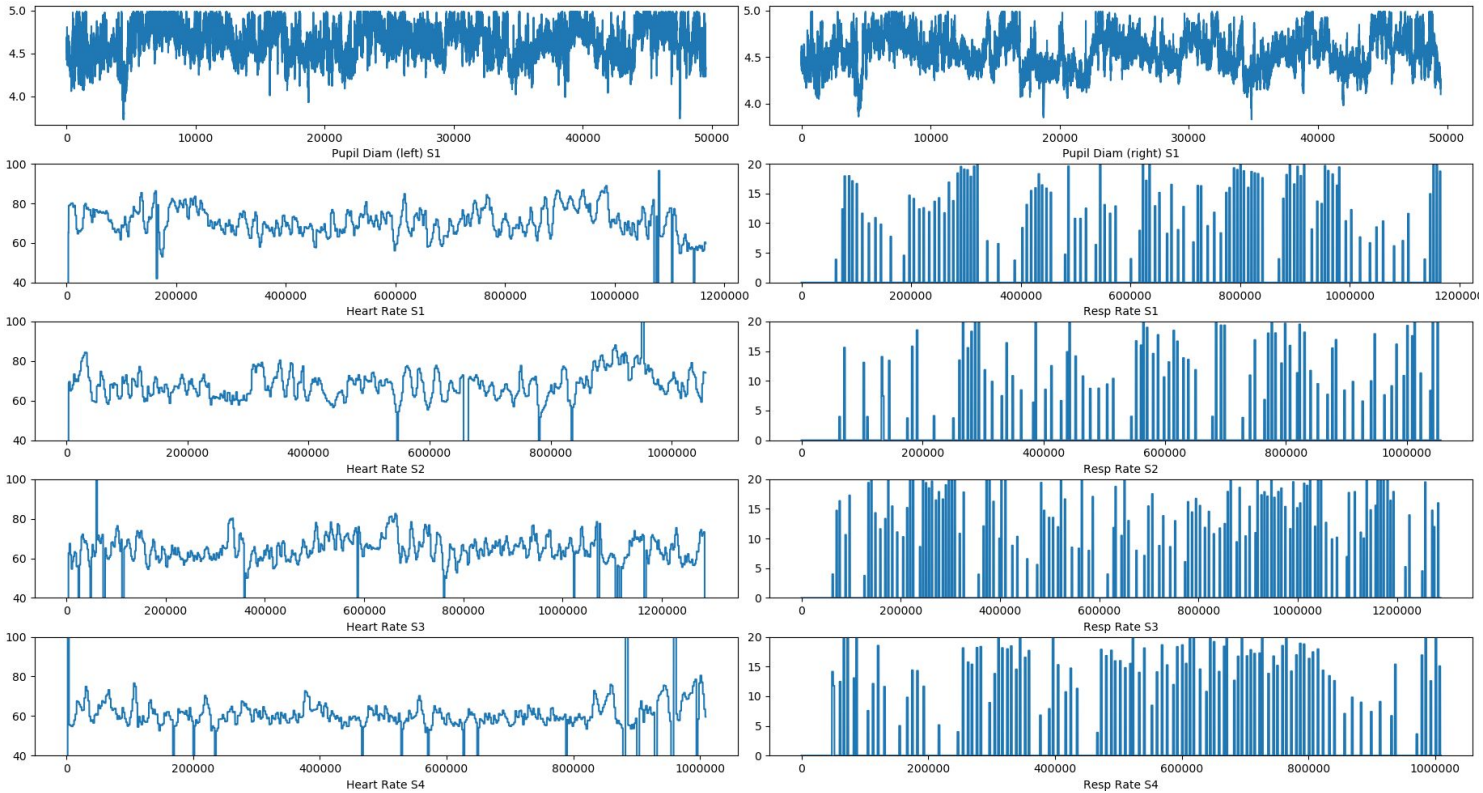
- A medical person is transporting an infant to a hospital in Atlanta.
- During the flight, an emergency forces a flurry of activity.
- The medical person must interact with the AI pilot to redirect to the nearest facility that can support the patient

Domain 2: **ISR**

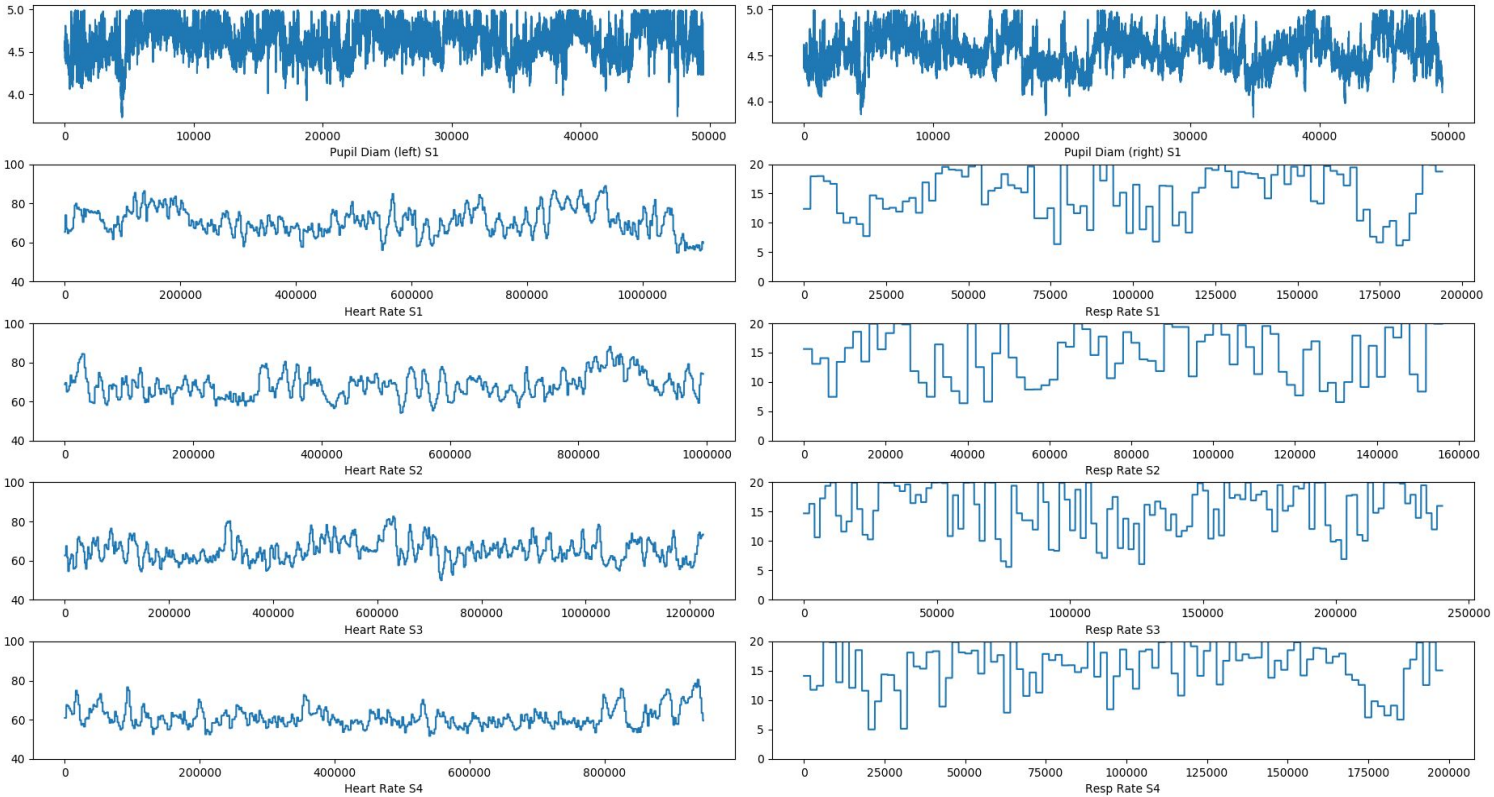


- An operator is scanning target ships in the ocean for threats.
- Different threat classes are handled uniquely (avoidance radius, observation priority)
- The AI pilot follows preset scanning paths, which the user overrides as needed.

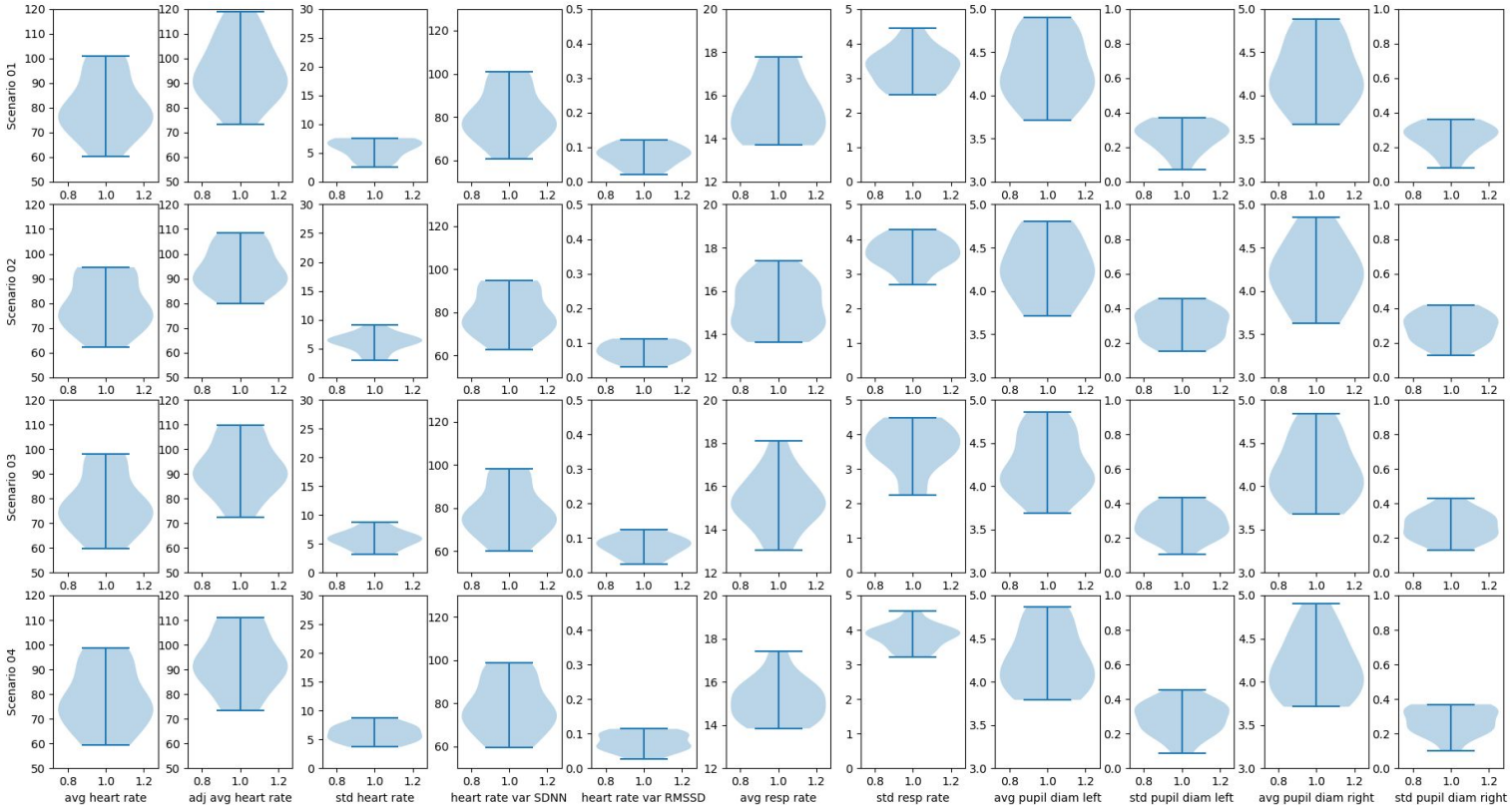
*Physiological data, **before** filtering:*



*Physiological data, **after** filtering:*

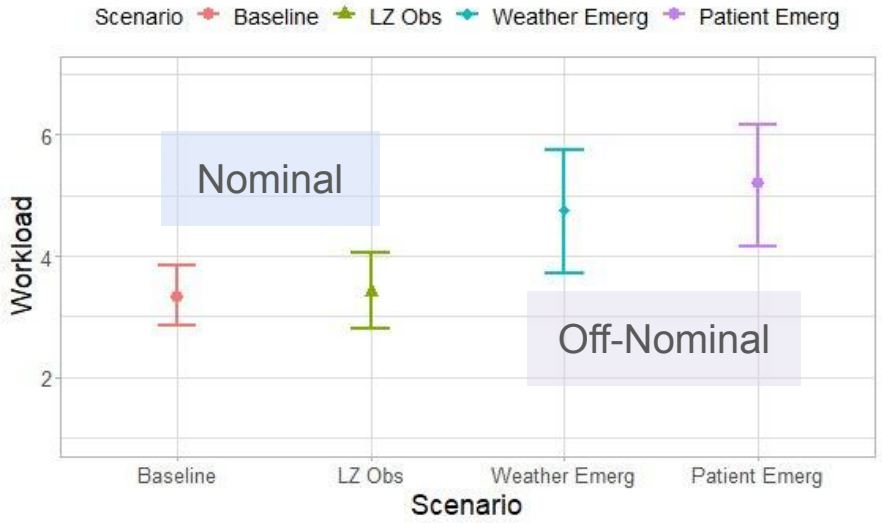


Physiological features:

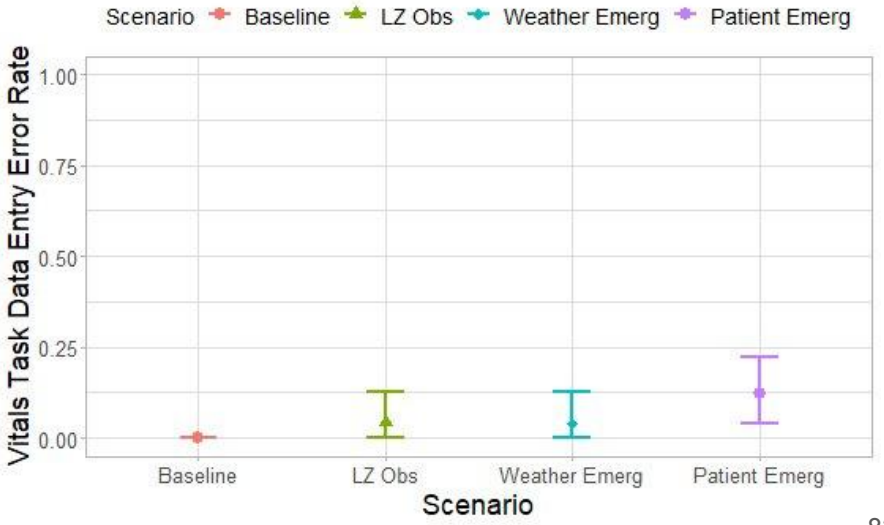


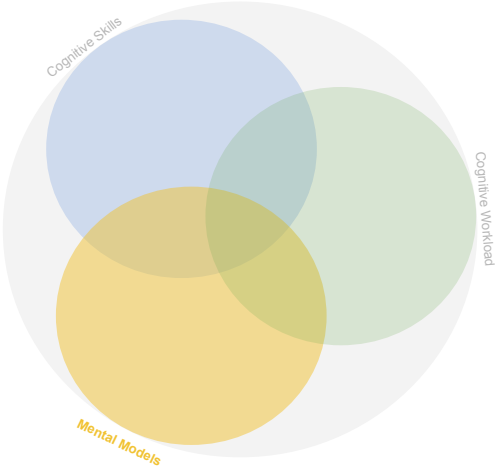
Secondary Task Performance

Used NASA-TLX to validate that scenarios induced **off-nominal** workload.



Used the secondary task performance to show that users **underperformed** in the emergency scenarios.





Simulation Theory

We maintain an internal simulation of the environment.

Mental Model

The data structure of the internal simulation or belief state.

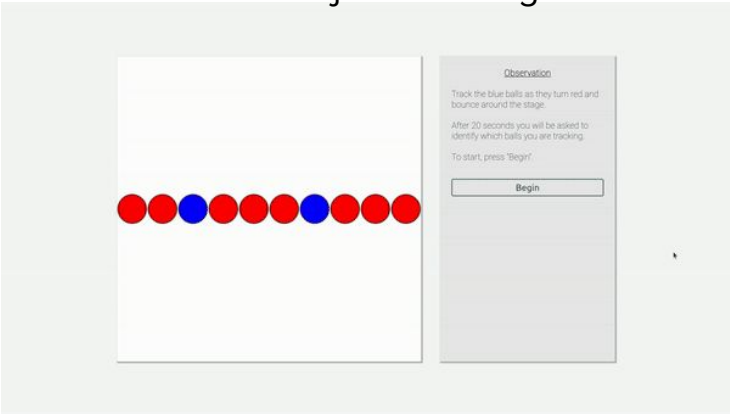
Related work focuses on ***supporting*** the user's mental model through:

- Interface/interaction design
- Explainable AI
- Decision support systems

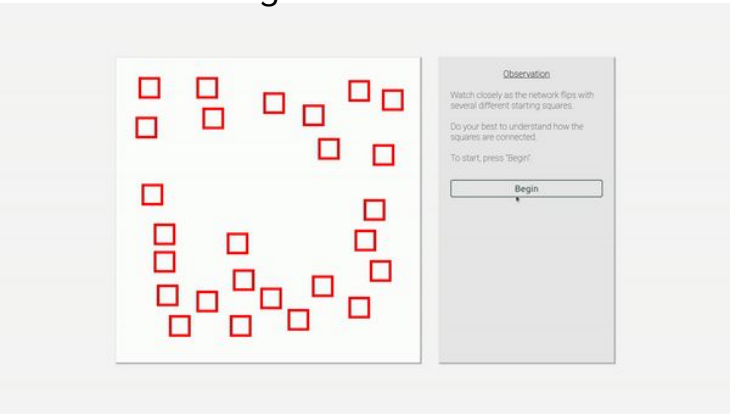
Very few works have sought to ***estimate*** the user's mental model.

- Based on the cognitive science literature.
- Each test lasts 5-10 minutes.
- Skill tests <u>do not</u> mimic teleoperation tasks.

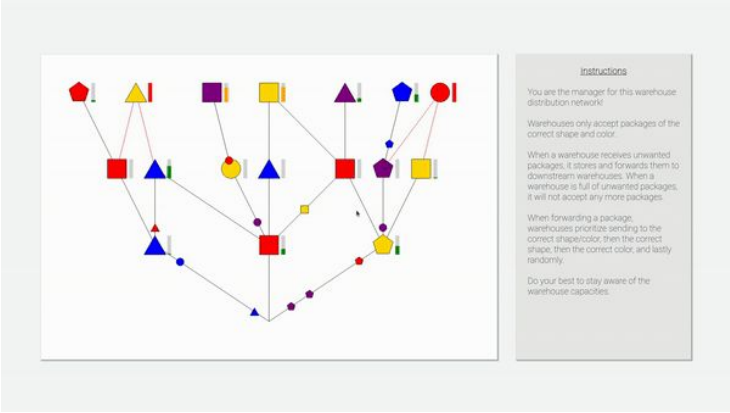
Multi-Object Tracking



Inferring Network Structure

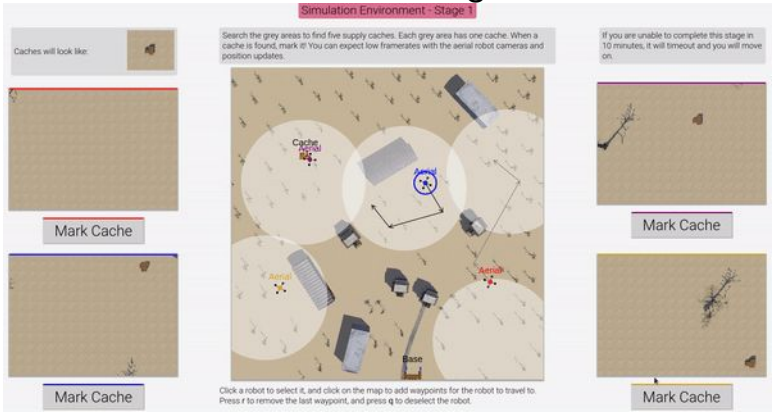


Level 1 & 2 Situational Awareness

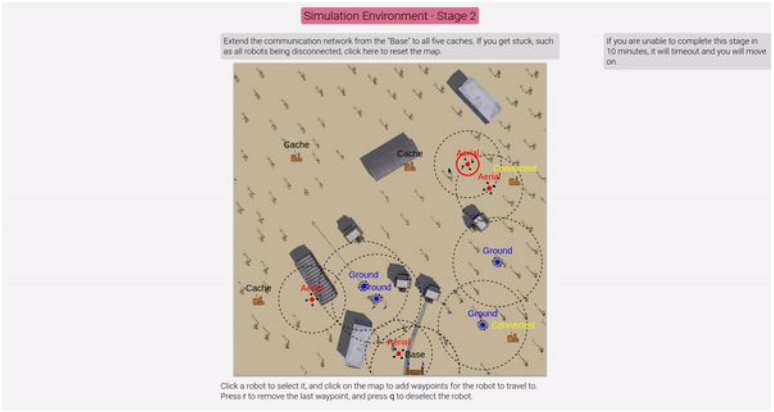


- *Tasks are independent.*
- Tasks represent a variety of real-world teleoperation tasks.
- Each task lasts 10 minutes, users are scored by their progress and time.

Search for Targets



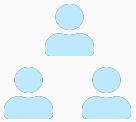
Ad-Hoc Network Construction



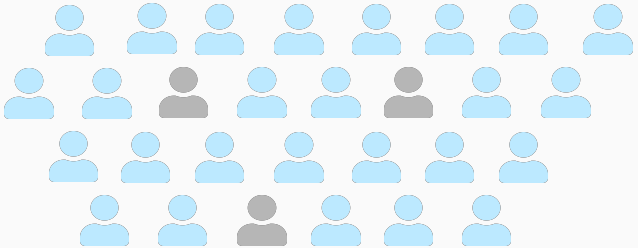
Navigate and Retrieve

- From 29 users we evaluated $\binom{29}{3} = 3654$ teams post-hoc.
- For each team, data from the remaining 26 users was used to fit the regression models.
- This lets us to compare our *individualized role assignment* (**IRA**) against what other assignments could have scored.

3 users selected for the team
(*leave-three-out sampling)*

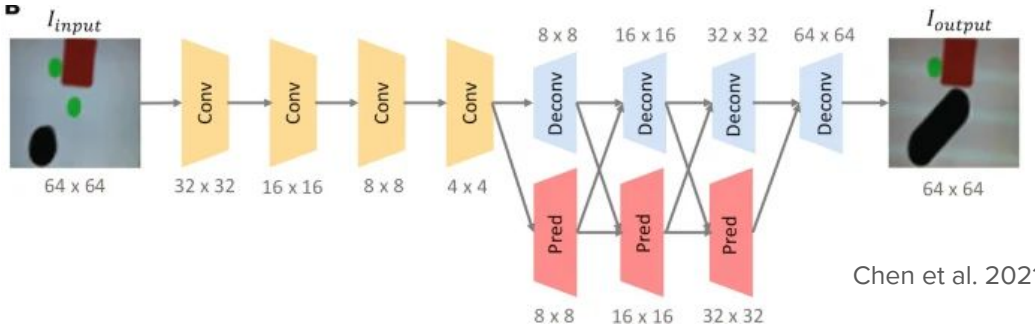


26 remaining users fit the models



Latent Representation ⟷ **Explicit Representation**

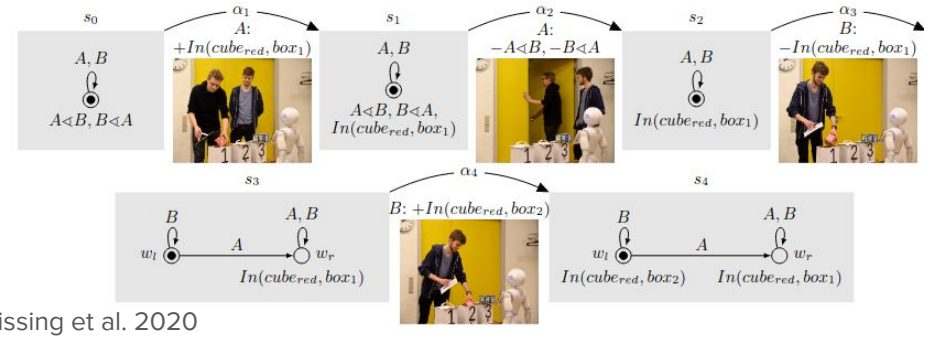
- *End-to-end models* (Chen et al. 2021, Ramachandruni et al. 2021, Szot et al. 2023)
  - Deep-learning architectures, typically assume perfect observability.
  - Some works pre-process vision into a semantic scene graph.
  - Aim to answer specific task questions (e.g., *which item will the user reach for?*).
  - Challenging to extract the latent belief state representation for other questions.



Chen et al. 2021

Latent Representation ⟷ **Explicit Representation**

- *Logical truth scene graphs* (Dissing et al. 2020, Scheutz et al. 2017) (logical predicates model)
  - Uses an **object-based graph structure** of properties (Level 1).
  - Can be used to broadly inform Level 2, Level 3 SA questions.
  - Integrates well with planning systems (e.g., PDDL)
  - Does not represent user uncertainty – *opinionated*.



Dissing et al. 2020

$\text{HASROLE}(x,searcher) :\leftrightarrow$
$\text{GOAL}(x,searchArea_7) \wedge$
$\text{GOAL}(x,reportAllResults) \wedge \ldots$ [add all goals]
$\text{REQUIRES}(\text{x, map}) \wedge$ [additional requirements]
$\exists y\, \text{CAPABLE}(x,y) \wedge \text{ACTIONTYPE}(x,door-opening) \wedge \ldots$ [add. actions]
$\forall z\, \text{SUPERIOR}(z,x) \rightarrow \text{OBLIGATED}(x,informed(z)) \wedge \ldots$ [add. obligations]
$\text{PERMISSIBLE}(a,\phi_1) \wedge \ldots$ [additional permissible states and actions]
$\neg\, \text{PERMISSIBLE}(a,\psi_1) \wedge \ldots$ [additional impermissible states and actions].

Scheutz et al. 2017

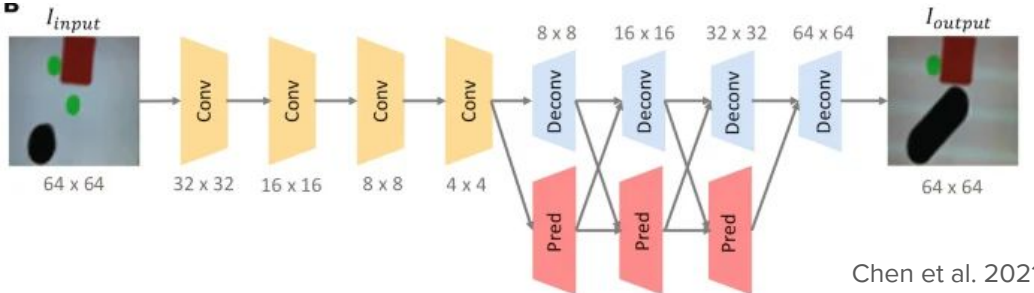
Latent Representation ⟷ **Explicit Representation**

(Chen et al. 2021)
(Ramachandruni et al. 2021)

(Carroll et al. 2019)
(Bolton et al. 2022)
(Szot et al. 2023)
(Nikolaidis et al. 2023)

(Dissing et al. 2020)
(Scheutz et al. 2017)
(Gervits et al. 2020)

- *End-to-end models*
  - Aim to answer a specific task questions (e.g., *which item will the user reach for?*).
  - Challenging to apply for other questions.



Chen et al. 2021

Latent Representation ⟷ **Explicit Representation**

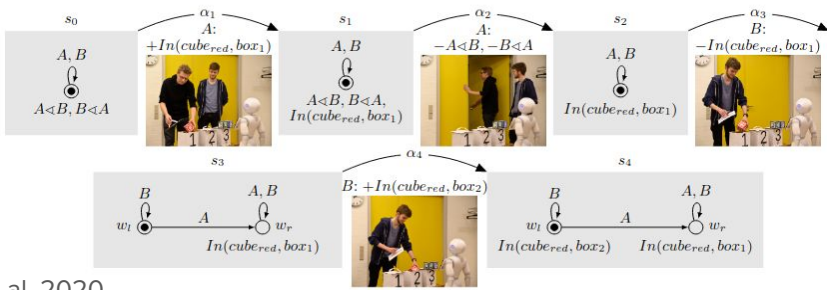
(Chen et al. 2021)
(Ramachandruni et al. 2021)

(Carroll et al. 2019)
(Bolton et al. 2022)
(Szot et al. 2023)
(Nikolaidis et al. 2023)

(Dissing et al. 2020)
(Scheutz et al. 2017)
(Gervits et al. 2020)

- *End-to-end models*
  - Aim to answer a specific task questions (e.g., *which item will the user reach for?*).
  - Challenging to apply for other questions.

- *Logical predicate models*
  - Integrates well with planning systems (e.g., PDDL)
  - Does not represent user uncertainty – *opinionated*.



Dissing et al. 2020

Latent Representation ⟷ **Explicit Representation**

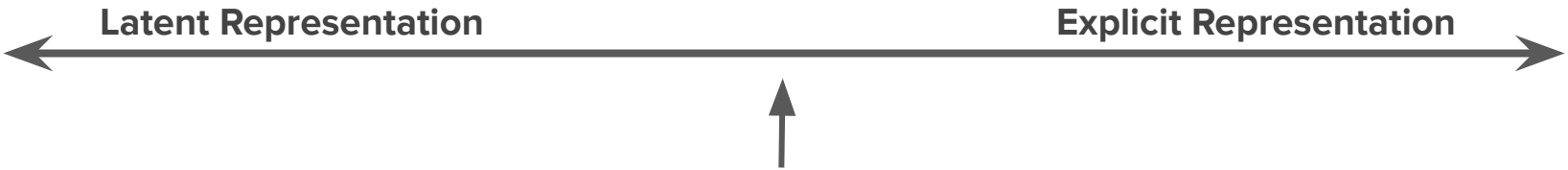
(Chen et al. 2021)
(Ramachandruni et al. 2021)

(Carroll et al. 2019)
(Bolton et al. 2022)
(Szot et al. 2023)
(Nikolaidis et al. 2023)

(Dissing et al. 2020)
(Scheutz et al. 2017)
(Gervits et al. 2020)

- *End-to-end models*
  - Aim to answer a specific task questions (e.g., *which item will the user reach for?*).
  - Challenging to apply for other questions.
- *Logical predicate models*
  - Integrates well with planning systems (e.g., PDDL)
  - Does not represent user uncertainty – *opinionated*.

- *Fuzzy Logic* (Bolton et al. 2023)
  - Uses fuzzy membership functions to inform user belief states from observations.
  - Challenged by difficulty of hand-crafting membership functions.



- *Markov Decision Processes, Partially-Observable MDPs* (Carroll et al. 2019, Nikolaidis et al. 2012)
  - Uses a state-based graph instead of an object-based graph – Level 1 situation awareness.
  - Models belief state through state-action transition probabilities and rewards.
  - POMDPs also include a probabilistic belief distribution across all states.
  - Useful for action-oriented planning – *"What will the user do next?"*
  - Challenged by intractability in complex state spaces.

- *Fuzzy Logic* (Bolton et al. 2023)
  - Uses fuzzy membership functions to inform user belief states from observations.
  - Able to consider user uncertainty, and use an **object-based graph structure**.
  - Challenged by difficulty of hand-crafting membership functions.

Inference Steps

Where is the nearest dish?

Is the pot done cooking?

What is your teammate doing?

What do you plan to do next?

- *Logical Predicates Model* (baseline)
  - Handwritten rules.
  - Does not guess how users will predict changes in objects while unseen.
  - Does not consider prior history.

- *Large Language Model* (novel)
  - Prompt engineering to GPT4, Claude.
  - System generates a prompt using the game description, scene graph, and SA question.
  - *"You are playing a game… The environment is this grid… Does player A think the pot has completed cooking?"*
  - Does not consider prior history.

- *Graph Neural Network* (novel)
  - Object-Object pairs in the scene graph are passed into a multi-layer perceptron (MLP) to estimate context (Level 2) properties.
    MLP Inputs:
    - Object Classes (one-hot)
    - Level 1 SA properties
  - Resulting properties are added to the graph.
  - A second MLP estimates projection (Level 3).
    MLP Inputs:
    - Object Classes
    - Level 1 SA properties
    - Level 2 SA properties
  - The Level 2 & Level 3 SA properties target the questions asked to users.
  - The MLPs are fit to user SA responses.