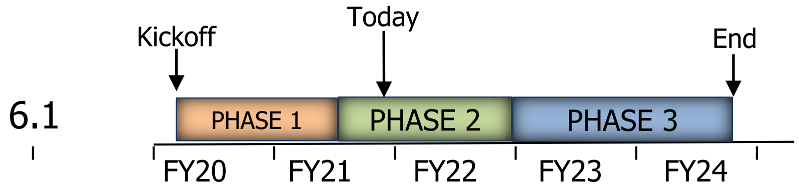
Artificial Social Intelligence for Successful Teams (ASIST)

Joshua Elliott, DARPA PM

Brian Sandberg, DARPA Technical SETA



6.1

Kickoff

Today

End

PHASE 1

PHASE 2

PHASE 3

FY20

FY21

FY22

FY23

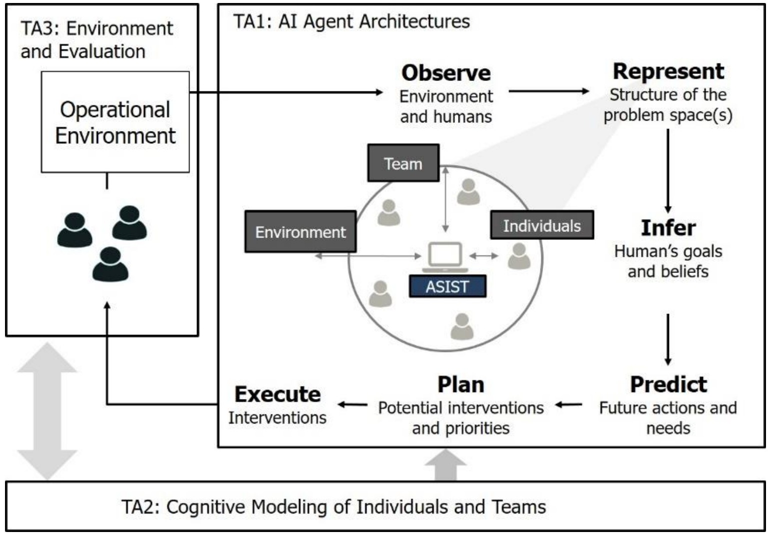
FY24

Human AI Teaming Workshop

July 29, 2021

Artificial Social Intelligence for Successful Teams (ASIST)

ASIST will develop AI theory & systems that demonstrate machine social skills needed to infer the goals and beliefs of human partners, predict what they will need, and offer context aware interventions in order to act as adaptable and resilient AI teammates


TA3: Environment and Evaluation
Operational Environment
TA1: AI Agent Architectures
Observe
Environment and humans
Represent
Structure of the problem space(s)
Team
Environment
Individuals
ASIST
Infer
Human's goals and beliefs
Predict
Future actions and needs
Plan
Potential interventions and priorities
Execute
Interventions
TA2: Cognitive Modeling of Individuals and Teams

Measures

Physiological

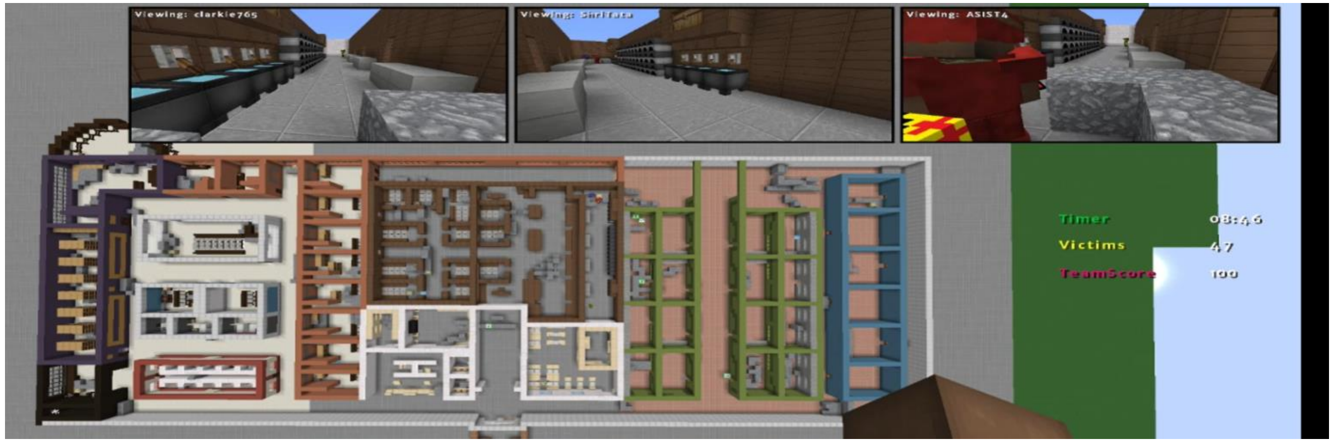
- fNIRS
- Eye tracking:
  - pupil diameter
  - gaze
- Facial expression

Surveys

- Satisficing
- Workload
- Spatial ability
- Demographics
- Goals
- Task knowledge
- Strategy

Competency test score

Testbed message bus events



Minecraft USAR Testbed Environment

Challenges:

- Application of Machine Theory of Mind and Teams to identify team problems and automate coaching to improve team function and performance
- Modularization of Agent Social Intelligence (ASI) agents with asynchronous sensing and that are robust, adaptable, safe, and effective

Approach:

- Develop ASI agent as coaches to optimize team function
- Infer mental states, predict behaviors, and offer guidance that improves decisions and coordination between members
- Operate in increasingly complex and specialized environments
- Adapt to unexpected perturbations
- Contribute to a revolution in cognitive modeling with shared situational awareness for effective human-machine teaming
- Engage with the warfighter as a collaborator: *capable of understanding the attitudinal, behavioral, and cognitive components of teamwork*

Accomplishments:

- Created and prototyped experimental design and procedures
- Conducted full-scale evaluations
- Developed and released testbed; adapted for use in other DARPA programs

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Applications of Symbiotic Teams – Mission Relevance

Force Application: Design and test teams & systems in complex tactical missions

Cyber Protection Teams: network security, cyber systems engineering

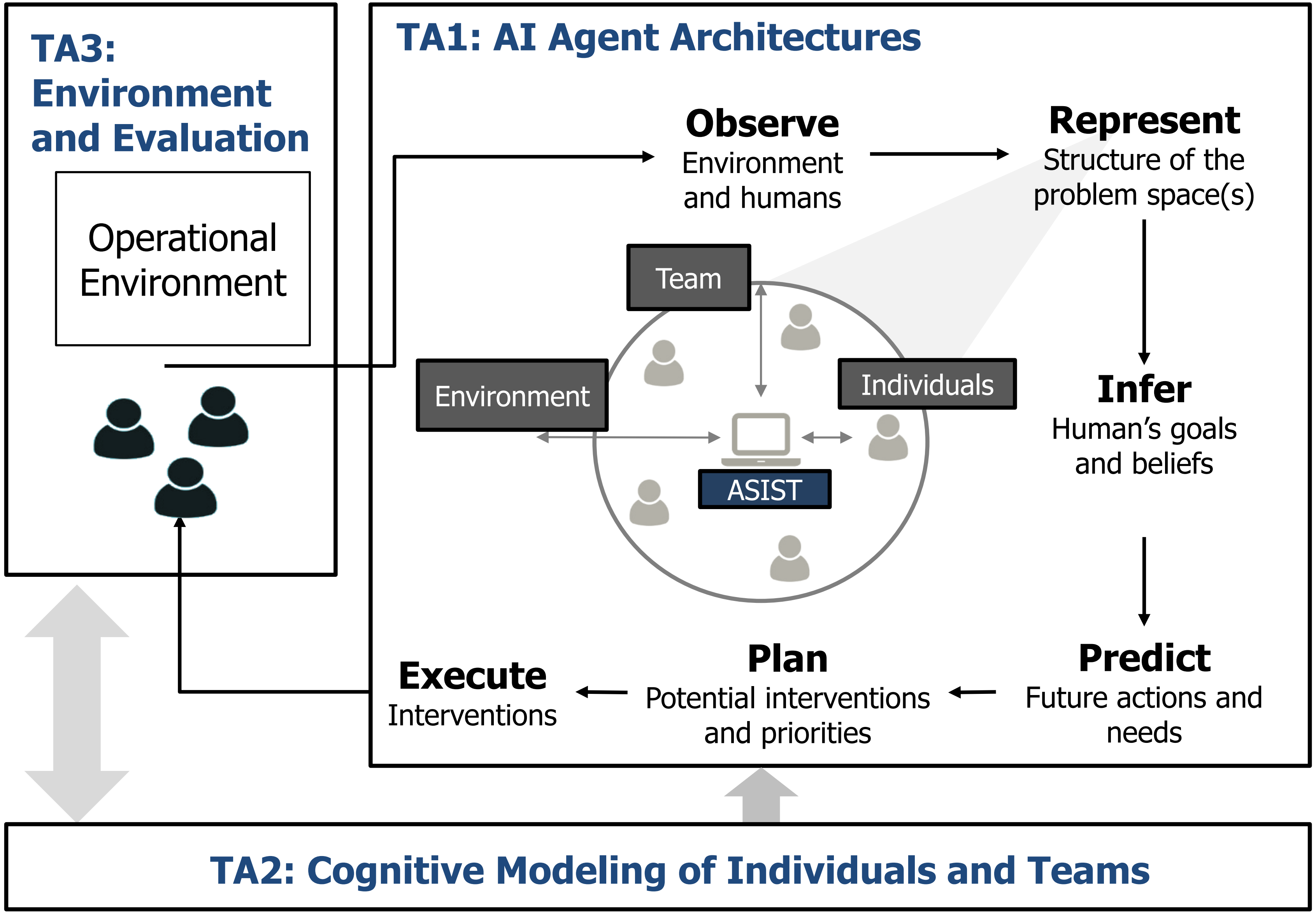
Coaching, Training and Education: Maintenance and repair, Tactical field care, Adaptive training

Search And Rescue/Urban Raid: First responder, Hostage rescue

ASIST will allow AI to be a dynamic team partner (collaborative and interdependent) to understand mission objectives and intervene for effective teaming

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Program Structure



TA3: Environment and Evaluation

Operational Environment

TA1: AI Agent Architectures

Observe Environment and humans

Represent Structure of the problem space(s)

Infer Human's goals and beliefs

Predict Future actions and needs

Plan Potential interventions and priorities

Execute Interventions

TA2: Cognitive Modeling of Individuals and Teams

ASIST agents will:

- Provide artificial coaching to optimize team function
- Operate in increasingly complex and specialized environments
- Adapt to unexpected perturbations
- Contribute to a revolution in cognitive modeling for human-machine teaming

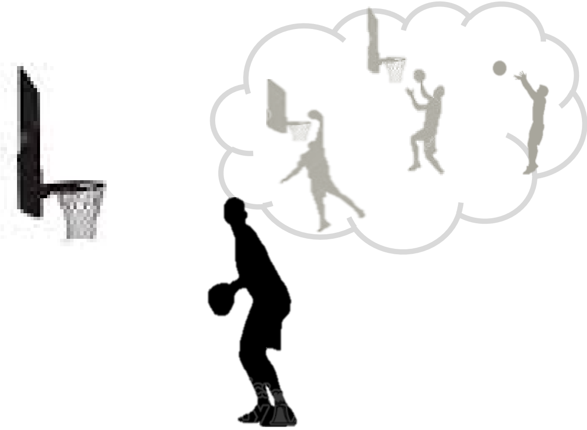
Change in Program Phases:

- First year success in programmatic and integration enabled faster pivot to more complex teaming….more profound results
- Moving from passive to active to generalization

What makes humans good teammates?

Mental Models of Environment

Humans can build robust mental models of their environment



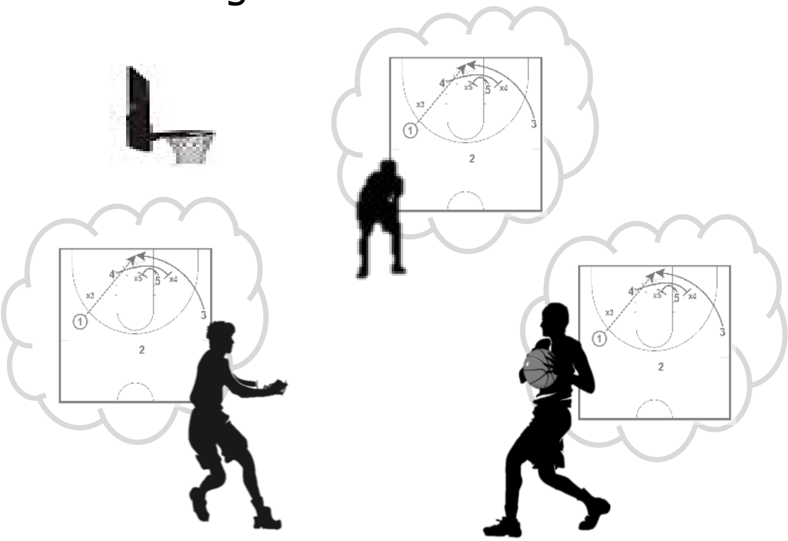
Mental Models of Others

Humans can infer, from observed actions and context, the mental states of other humans

Theory of Mind

Shared Mental Models

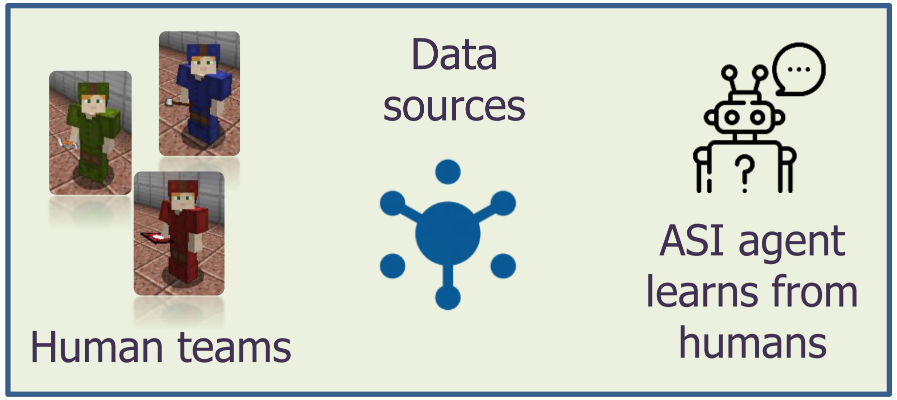
To perform in teams, humans use experience and training to align their Mental Models



Social Intelligence

Inferring Team Shared Mental Model (SMM)

Observable *(s)* and latent *(x)* features influence team performance



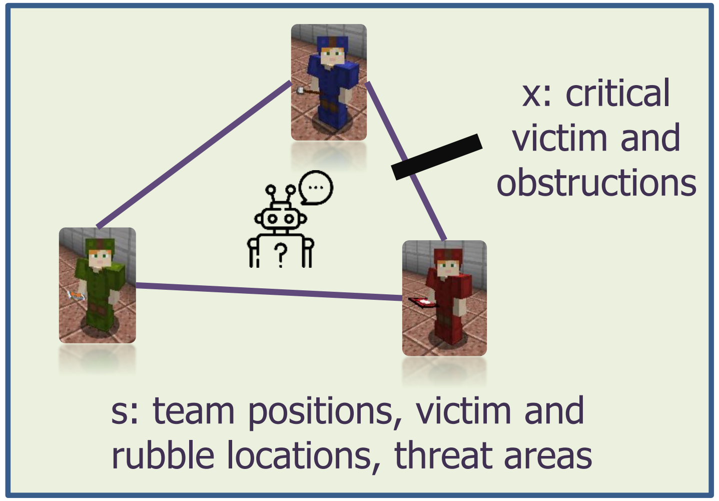
Human teams

Data sources

ASI agent learns from humans

Learn team SMM (model human teams)

Latent features can differ among team members

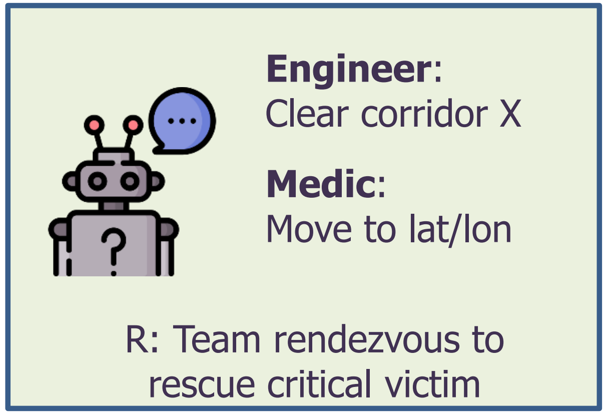


x: critical victim and obstructions

s: team positions, victim and rubble locations, threat areas

Infer lack of team SMM (model misalignment)

Decide when and how to intervene



Engineer: Clear corridor X

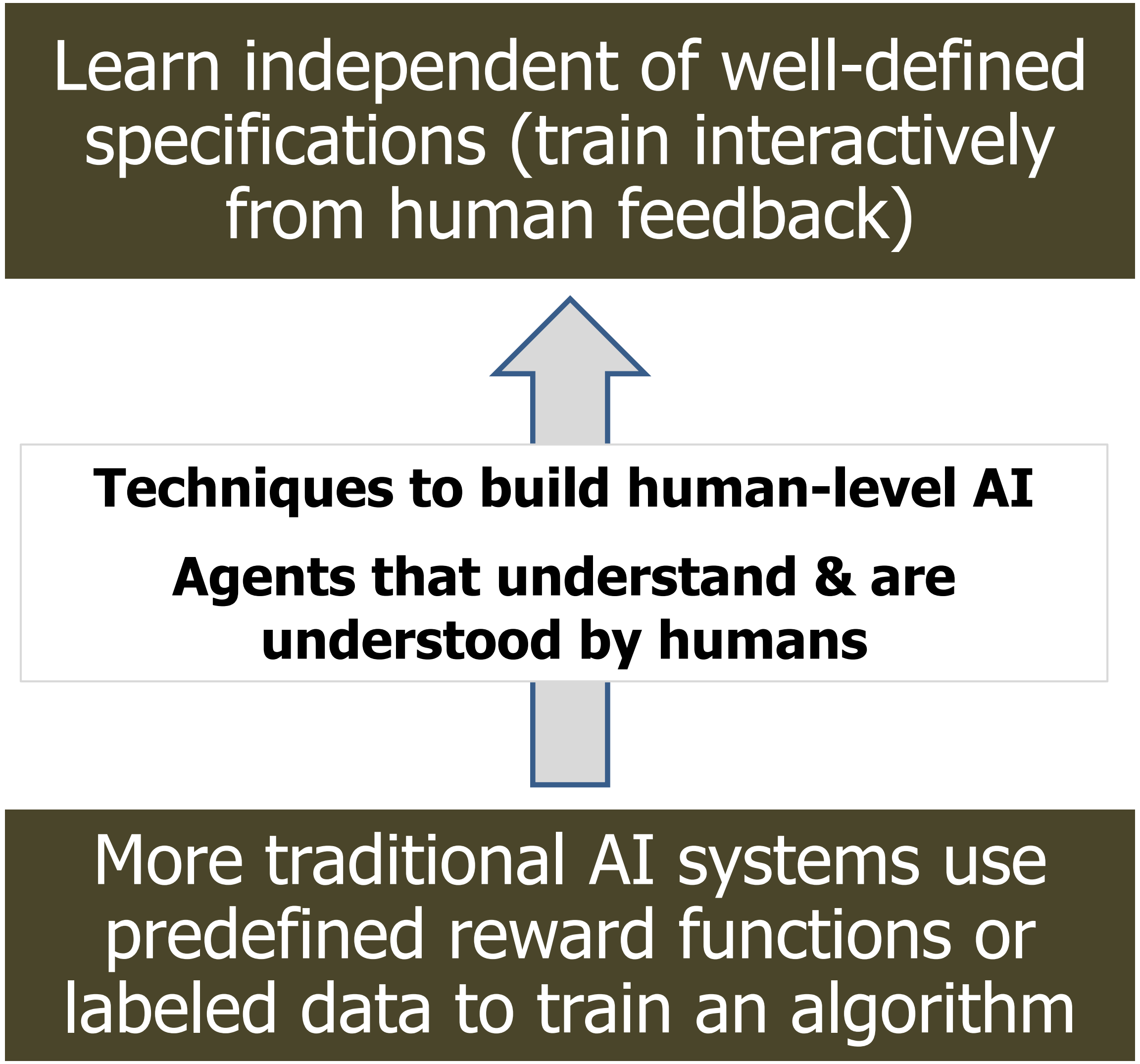
Medic: Move to lat/lon

R: Team rendezvous to rescue critical victim

Intervene (correct errors or improve team performance)

Hard problem to sense all the observable and latent features necessary to create Team SMM

Challenges for agents that collaborate with humans for effective HMT



Learn independent of well-defined specifications (train interactively from human feedback)

Techniques to build human-level AI

Agents that understand & are understood by humans

More traditional AI systems use predefined reward functions or labeled data to train an algorithm

- Learn human mental models to better coordinate with them
- Learn true signal derived from human behaviors or explicit feedback (NL, demonstrations, preferences, etc.)
- Human feedback as policy or reward shaping

Satisfying literal specification, but not achieving intended outcome … just reward is often not great for evaluation

Source: (Amodei & Clark, 2016)

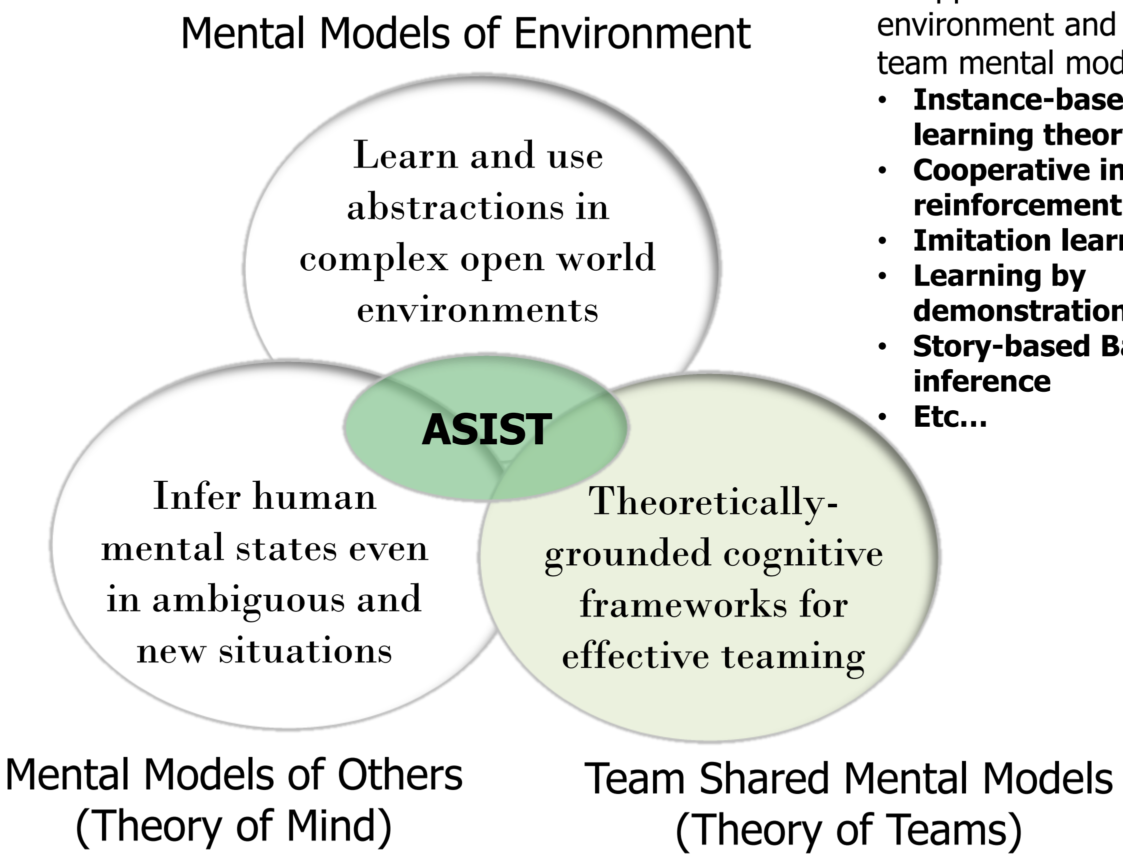
How it is done today?
Model Theory of Mind (ToM) of human participants and team

Current methods of ToM and limitations

Challenge	Approach	Limitations and difficulties modeling human psychology
Mental Models of Environment	Decision theory	Rely on assumptions of rationality that people constantly violate
	Cognitive architectures	Agents not equipped with a general cross-domain knowledge (only used in narrow tasks); does not support multiple conflicting goals
	Formal systems	Not able to deal with complex knowledge structures of humans; representations insensitive to distinctions among conflicting goals
Mental Models of Others	Game theory	Rely on concepts of equilibria that people rarely achieve in an unstructured social setting
	Bayesian	Performs Bayesian inference over beliefs and desires simultaneously; does not support multiple conflicting goals
	Neural Network	Model single agent (not human) in simple environment; high training cost; not scalable to complex domains
Team Shared Mental Model	New research in real world human-AI teaming (e.g. learning from human teams, modeling teams with no explicit reward specification)	

- Baker, Chris, Jara-Ettinger, Julian, Saxe, Rebecca, and Tenenbaum, Joshua. **Rational quantitative attribution of beliefs, desires and percepts in human mentalizing**. Nature Human Behaviour, 1(4):0064, 2017
- Rabinowitz, N. C., Perbet, F., Song, H. F., Zhang, C., Eslami, S. M. A., & Botvinick, M. (2018). **Machine Theory of Mind**. ArXiv:1802.07740

ASIST models of ToM and Theory of Teams


Mental Models of Environment
Learn and use abstractions in complex open world environments
ASIST
Infer human mental states even in ambiguous and new situations
Theoretically-grounded cognitive frameworks for effective teaming
Mental Models of Others (Theory of Mind)
Team Shared Mental Models (Theory of Teams)

AI approaches to learning environment and human team mental models:

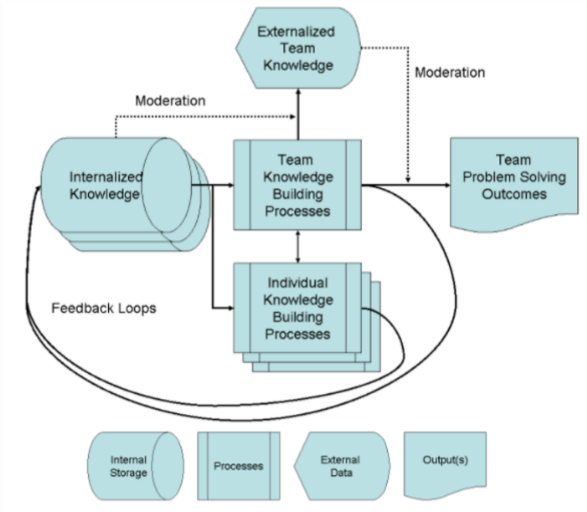
- **Instance-based learning theory**
- **Cooperative inverse reinforcement learning**
- **Imitation learning**
- **Learning by demonstration**
- **Story-based Bayesian inference**
- **Etc…**

- Nguyen, T., Gonzalez, C., **Cognitive Machine Theory of Mind**, 42nd Annual Meeting of the Cognitive Science Society (CogSci 2020), pp. 2560-2566
- Jain, V., Jena, R., Li, H., Gupta, T., Hughes, D., Lewis, M., & Sycara, K., "**Predicting Human Strategies in Simulated Search and Rescue Task**," NeuRIPS, AI+HADR Workshop, 2020.

Agent Social Intelligence (ASI) enables Superhuman Teams

Apply variety of research dimensions to human-AI symbiosis

Integrating Theories of Teams for ASI


Externalized Team Knowledge
Moderation
Moderation
Internalized Knowledge
Team Knowledge Building Processes
Team Problem Solving Outcomes
Individual Knowledge Building Processes
Feedback Loops
Internal Storage
Processes
External Data
Output(s)

Macro-Cognition in Teams

- Shared Mental Models
- Macro-Cognition in Teams
- Charnov’s Marginal Value Theorem (MVT) and Optimal Foraging Theory
- Trust to Strategy Alignment
- Collective Intelligence
- Etc…

Test hypotheses derived from grounded social science theory (pre-registration for rigorous and reproducible research), produce useful modular **analytics** to improve ASI

Operational pipelines

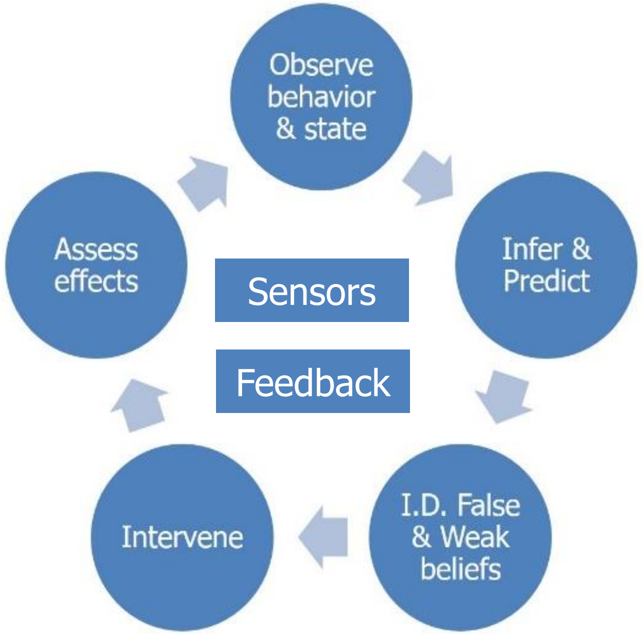
ASI agent capabilities to observe human & team behaviors, infer their mental states, predict future behaviors, and intervene

ASI Architecture

- **Open**
- **Generalizable**
- **Modular**
- **Scalable**

ASI Agents

- **Robust**
- **Adaptable**
- **Safe**
- **Effective**



Result: Machine ToM, along with operationalized theories from team science enable automated identification of risks to team and mission and coaching to improve team function and performance

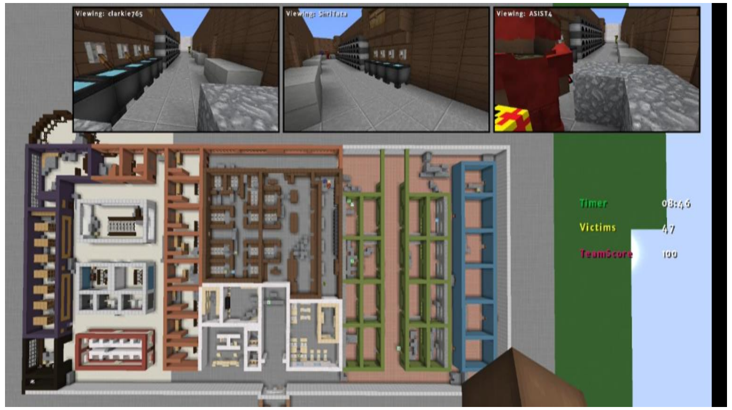
Testbed with operational complexity in a controlled environment

- DoD-Owned Testbed & Experimentation
  - Automated, replicable HSR, flexible, containerized, cloud-based
  - Supports real-time AI agent interaction with human
  - Distributed team environment with robust message scheme
  - Evaluations of increasing complexity and scale, common baseline measures, objective and subjective data
- Studies inform focus of future studies
  - Study 1: Individual, passive ASI
    - Focused on broad territory of AI-human interaction; learning and baselining
      - Infer the training condition (state) of the player
  - Study 2: Team, passive ASI
    - Focused on measures relevant to Theory of Mind; 68 human teams (204 participants)
      - Infer participant mental model
      - Predict participant performance and action
  - Study 3: Team, active ASI
    - Focus on effective interventions
- ASI Agents
  - Understand human social intelligence in a team context
  - Predict what is needed and intervene as an effective partner
  - **Handle known and unknown perturbations** in task, team, mission, or environment for fast adaptation and team resilience

> "No battle plan survives first contact with the enemy"
>
> -- Helmuth von Moltke

Agent orchestration, execution, and analytic results (complexity, accuracy, speed of inference)

Minecraft
USAR
Testbed
Environment



Measure	Study 1	Study 2
Complexity of mission/testbed	1 player, 3 training conditions	team of 3; 2 planning conditions; variation in knowledge (maps, blocks)
Volume of messages/data	200GB/2000 files	>500GB/3000 files
Common metrics	2: team peformance, mental model	4: team performance, mental models (2), ToM
ASI capabilities pre-registered (TA1)	24	30 (>50)
Hypotheses pre-registered (TA2)	47	58 (>100)
Number of analytic agents	2	5 (>20)

Study 3-6 will increase team and task complexity and produce agents that are more general, modular, scalable, useful, and trustworthy

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Modeling a USAR team collaboration task and experiment

Application: Computer simulations of USAR collaborative task inspired by real world scenarios

Team: 3-member team with role selection and skill

Search Specialist (Searcher)

Stretcher

Medical Specialist (Medic)

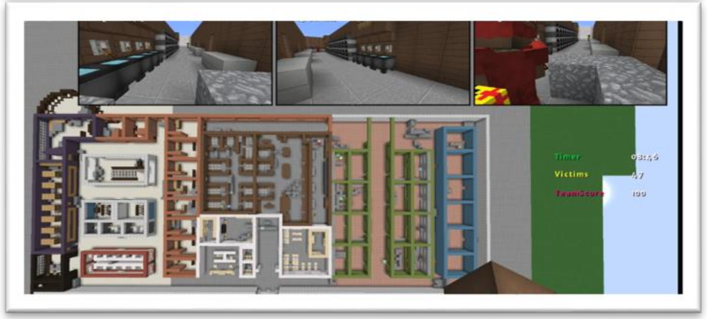
First-Aid Kit

Heavy Equipment Specialist (Engineer)

Sledgehammer

Environment: Minecraft representation of the USAR open world environment

Regular and Critical Victims



Task: Leverage team roles to find and save regular and critical victims; recover from perturbations (e.g. threats) in the environment

Goal: Save as many victims as possible in 15 min

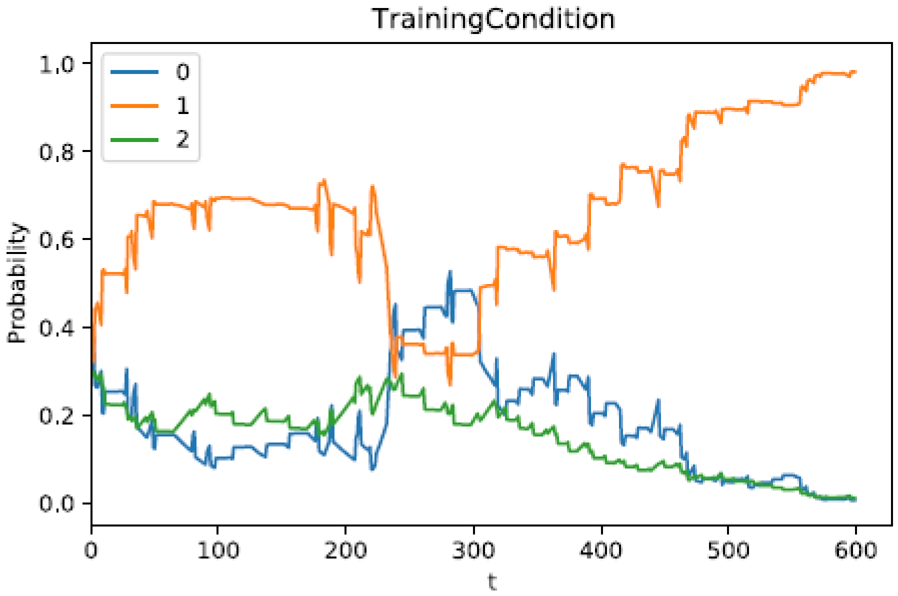
Sub-goal: Support team member under threat

Distribution Statement “A” (Approved for Public Release, Distribution Unlimited)

Agent accurately infers mental state of human

- Training induces a complex cognitive state (goals, strategy, beliefs, etc.) that drives a player's behavior (choices, reactions, etc.), making this as close as possible to measuring a true "Theory of Mind"
- Human players were given 1 of 3 training conditions at start of play, including information on tools and strategies
- Agents continuously update inference of training conditions over the course of a mission
- Agents inferred training conditions from data (80% accuracy)

Agent **infers mental states** based only on observable behaviors



Predict Actions and Performance (e.g. triage of a victim, navigation):
Accuracy of action predictions ranged from 74.7-99.2%

Infer Mental Models (e.g. knowledge condition):
69-80% accuracy

Program Metrics

Measure	Phase 1 Passive ASI		Phase 2 Active ASI		Phase 3 Generalizations	
Level of complexity	Single Human		Multiple Humans, new mission		Multiple Humans, Non-Human Teammate, New environments/missions	
Levels of team complexity	Multiple Humans					
Levels of mission complexity	Simple	Medium	Medium	Complex	Complex	Complex
Direct: Accuracy of state/action prediction Time to generate initial inference Predict team performance ToM: Infer training/information	Comparable to human (Coach)		Comparable to human (Coach)		Comparable to human (Coach)	
Survey measure: Usefulness, Trust	--		>75% of users: •Use Agent •Report trusting agent		>95% of users: •Use Agent •Report trusting agent	
Adaptation time Resilience time Coordination after perturbation (friction)	–		30% reduction compared to human team		60% reduction compared to human team	
# integrated	>20		>40		>60	
# used by TA1 agents	--		75%		90%	
# hypotheses tested/published	>50		>100		>200	

Mission Complexity:

- Victim types and location by region and room
- Number and location of blockages
- Threat rooms
- Team planning conditions

Common Metrics:

- Predict final score/trial (team performance)
- Infer map type (mental model)
- Infer block meaning (mental model)
- Predict action given false belief (ToM)

With Successful ASI, machines can be effective partners with humans

ASIST Summary

- ASIST will deliver advanced AI agents that can infer individual and team mental states, identify risks to team cohesion, and effectively intervene to optimize team performance

- Exciting area of fundamental AI theory and application…
  - to autonomously learn rich models of humans and teams
  - enable highly effective teams in complex multi-agent tasks
  - with potential for wide range of DoD applications (search and rescue, information operations, cyber teams, planning, training, force application)
  - ASI is a critical building block to achieve AGI

Creating ASI agents for highly effective human machine teaming

DARPA

www.darpa.mil