Trusted AI SE

SERC / AIRC Phase I & II

Samuel Cornejo,

Alejandro Salado,

Afrooz Jalilzadeh,

Amal Yousseef,

Pratik Satam,

Zeinab Alizadeh,

Department of Systems and Industrial Engineering

THE UNIVERSITY OF ARIZONA

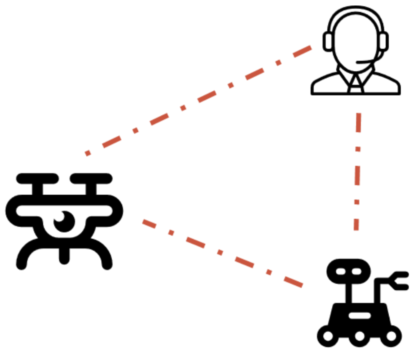
Context

- Research competition from the SERC.

How to make reliable
intelligent systems
From not-reliable
“intelligent” components?

Operational Context

- Define a system capable of clearing a mined path for a battalion to cross the mined path.
- There are 4 kind of agents involved in the system:
  - UAV: capable of surveying the area and making predictions
  - Human SME: capable of making predictions
  - UGV: rover capable of clearing the mines
  - Battalion soldiers: move towards where the operator tells them



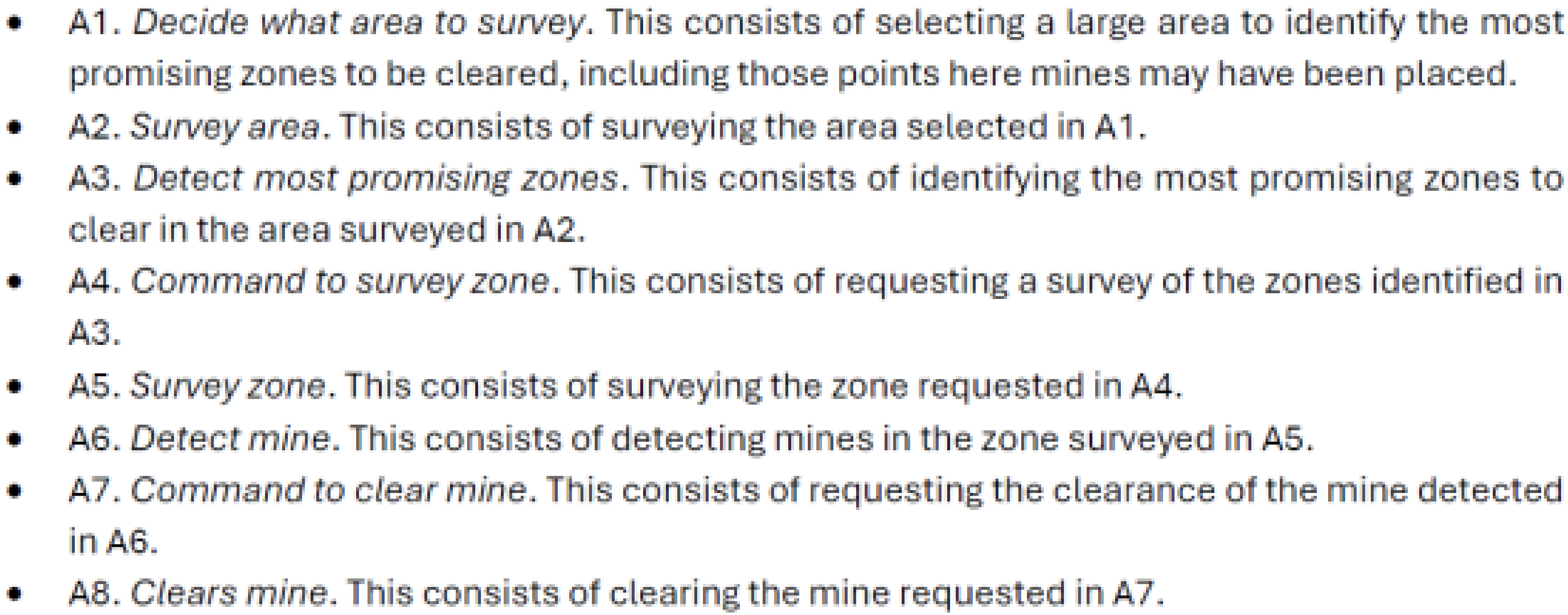
Assumptions:

- **UAV is a fast**, multi-spectral video collection system
- **UAV generates predictions** from its data.
- AI performance data corresponds to UAV
- Human SME reviews video from UAV
- **HUMAN SME generates predictions** from UAV Data.
- **Human SME gets feedback** from UGV
- There is an **ENEMY that MAY damage the system** through cyber-attacks.

Operational Solution at the Mission Level

- How do soldiers plan to traverse the cleared path?
  - **Wait** for cleared path?
  - **Walk** with the UGV?
  - **Walk X nodes** behind the UGV?

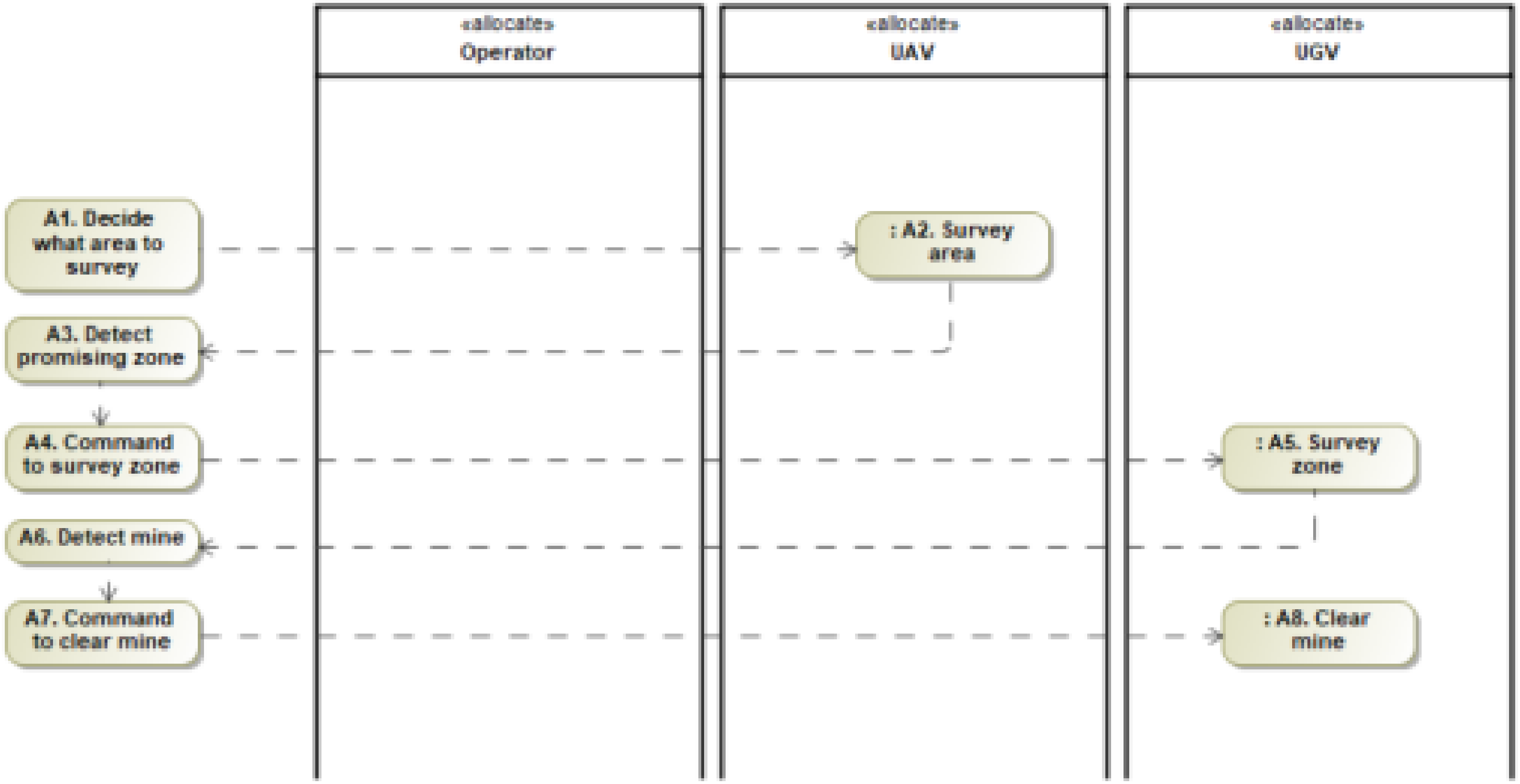
Generic Functional Flow (Not necessarily in this order)



- A1. *Decide what area to survey*. This consists of selecting a large area to identify the most promising zones to be cleared, including those points here mines may have been placed.
- A2. *Survey area*. This consists of surveying the area selected in A1.
- A3. *Detect most promising zones*. This consists of identifying the most promising zones to clear in the area surveyed in A2.
- A4. *Command to survey zone*. This consists of requesting a survey of the zones identified in A3.
- A5. *Survey zone*. This consists of surveying the zone requested in A4.
- A6. *Detect mine*. This consists of detecting mines in the zone surveyed in A5.
- A7. *Command to clear mine*. This consists of requesting the clearance of the mine detected in A6.
- A8. *Clears mine*. This consists of clearing the mine requested in A7.

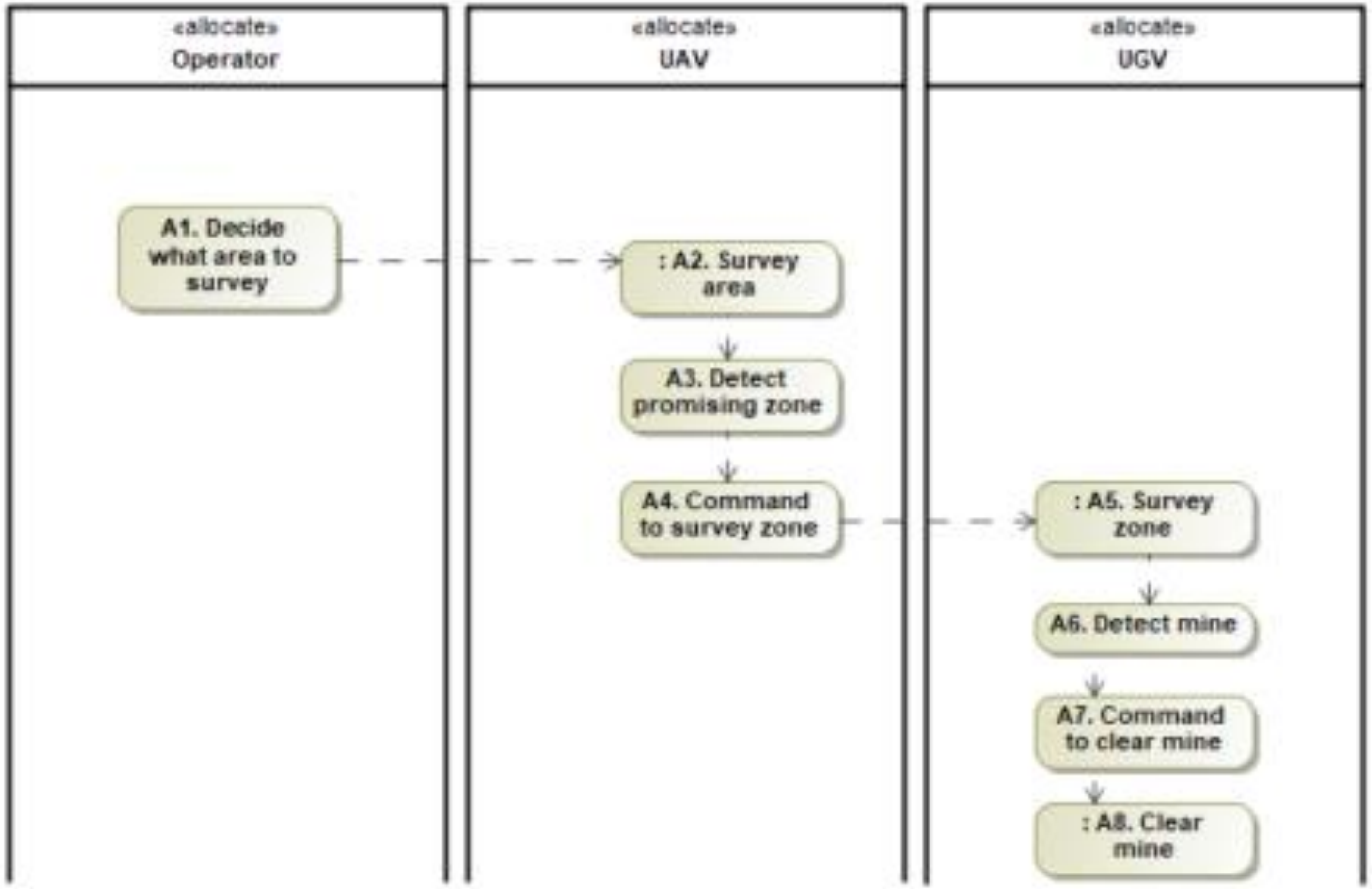
OPERATIONAL ARCHITECTURE AT THE SYSTEM LEVEL

Minimal allocation



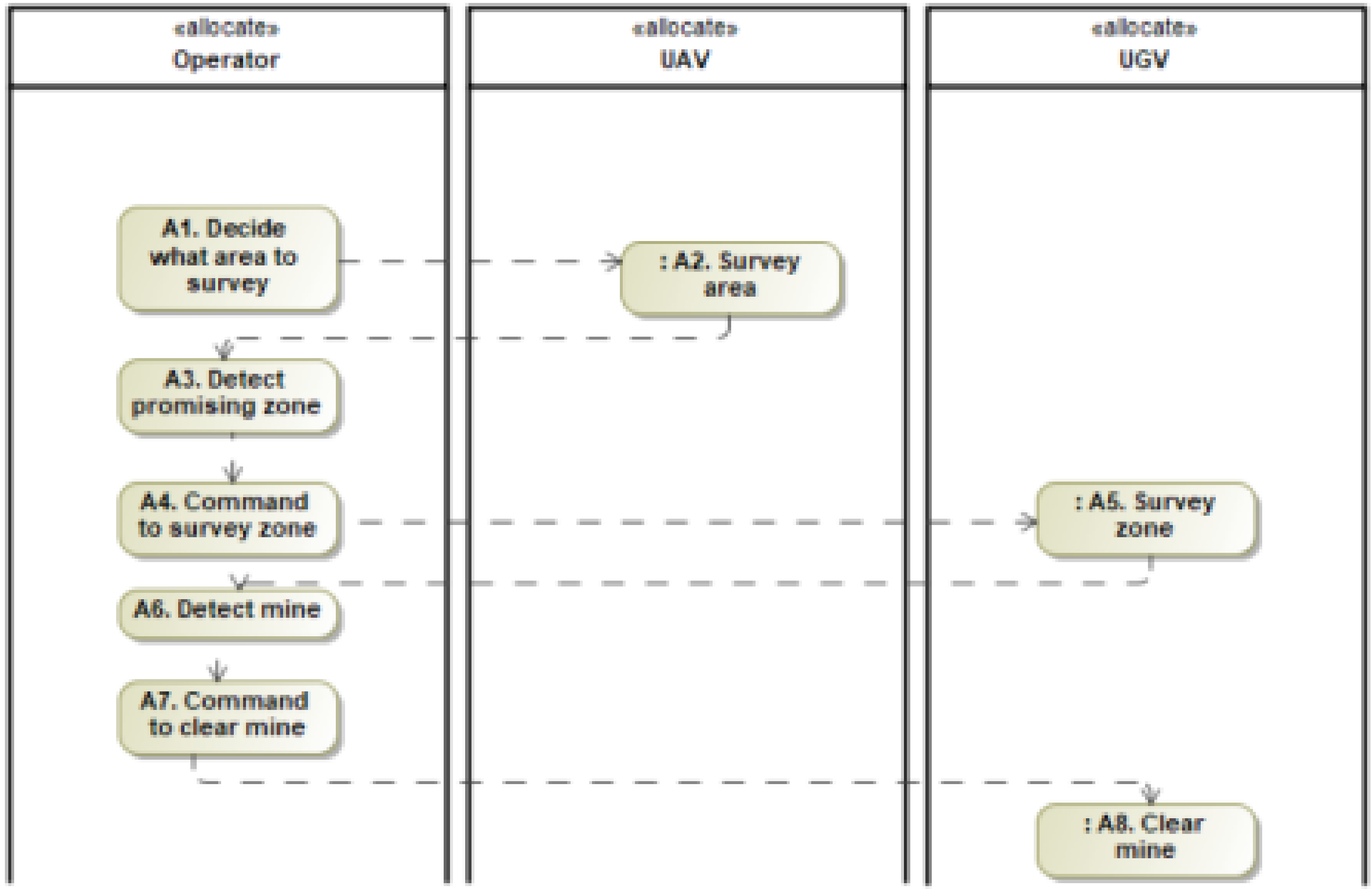
OPERATIONAL ARCHITECTURE AT THE SYSTEM LEVEL

AI intensive allocation



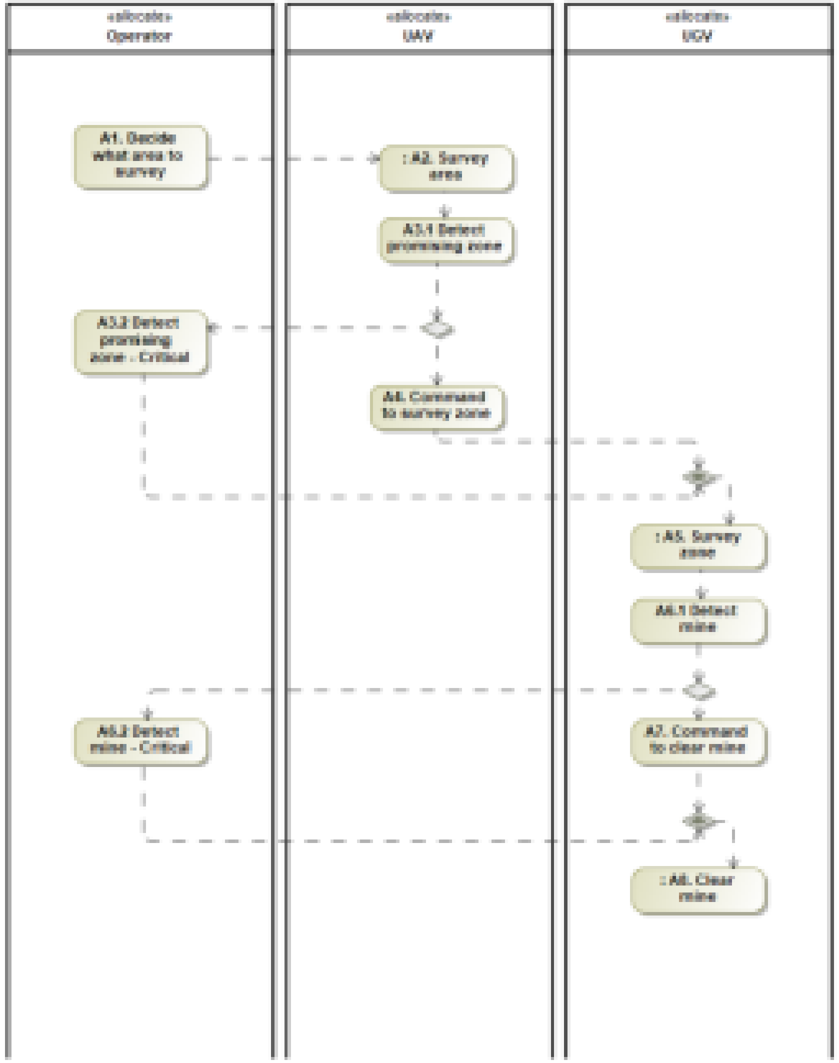
OPERATIONAL ARCHITECTURE AT THE SYSTEM LEVEL

Human intensive allocation



OPERATIONAL ARCHITECTURE AT THE SYSTEM LEVEL

Performance-based allocation

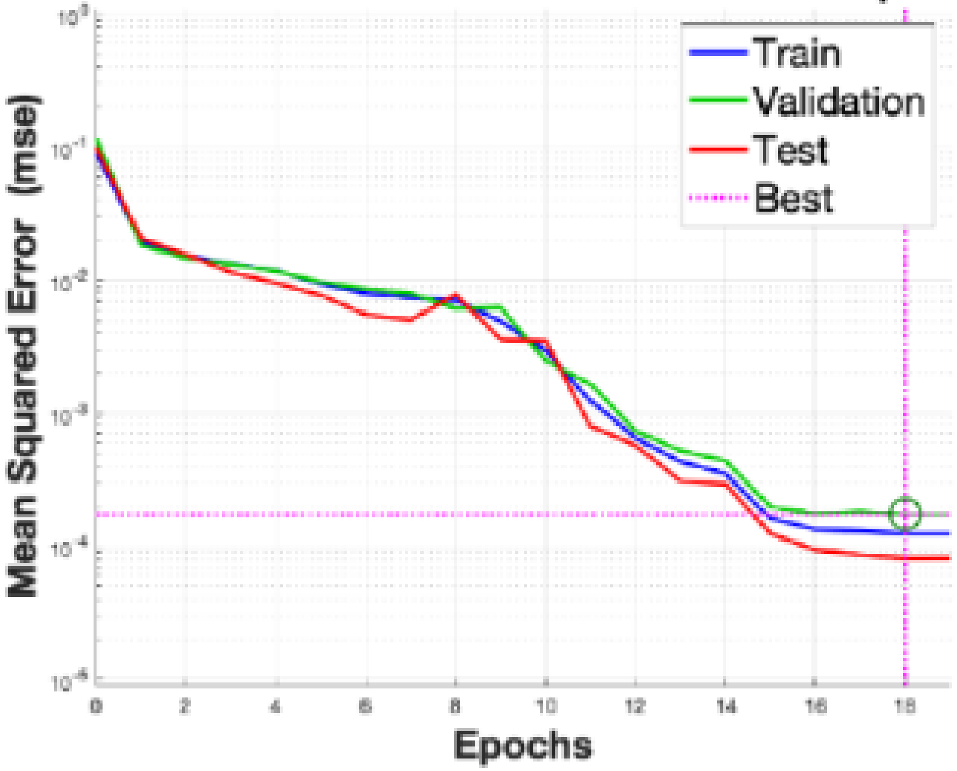

«allocate»
Operator
«allocate»
UAV
«allocate»
UGV
A1. Decide what area to survey
: A2. Survey area
A3.1 Detect promising zone
A3.2 Detect promising zone - Critical
A4. Command to survey zone
: A5. Survey zone
A6.1 Detect mine
A6.2 Detect mine - Critical
A7. Command to clear mine
: A8. Clear mine

© S. Cornejo, A. Salado, A. Jalilzadeh, A. Youssef, P. Satam, Z. Alizadeh

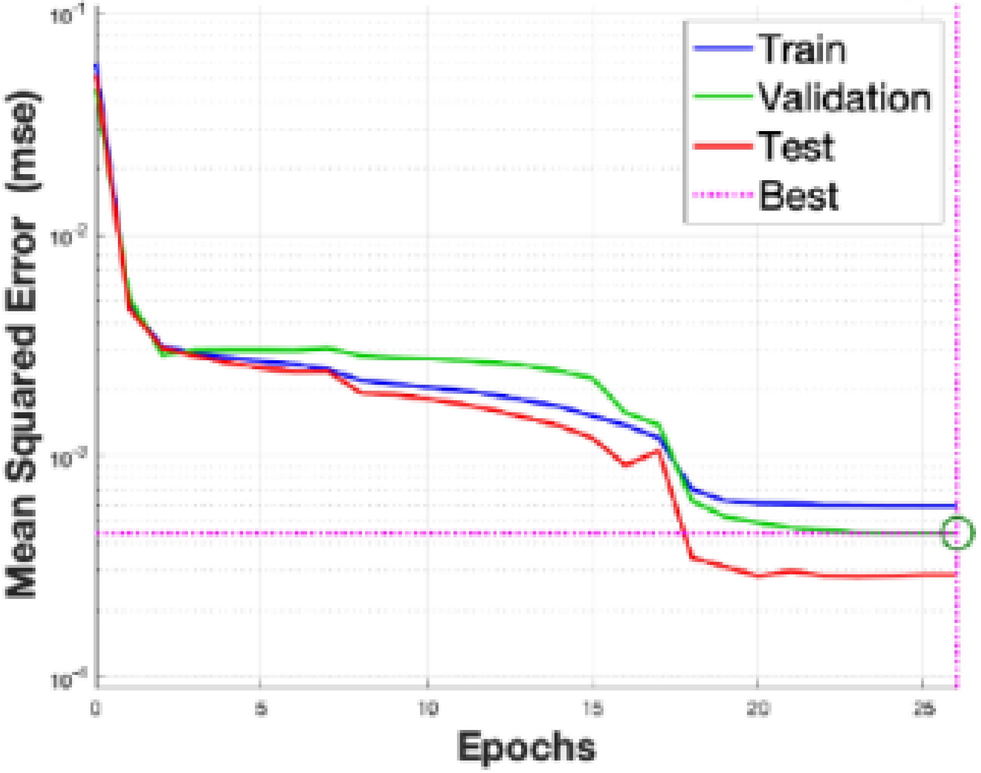
THE UNIVERSITY OF ARIZONA

CHARACTERIZING AI vs HUMAN PERFORMANCE

AI: Best Validation Performance is 0.00017419 at epoch 18



Human: Best Validation Performance is 0.00044541 at epoch 26

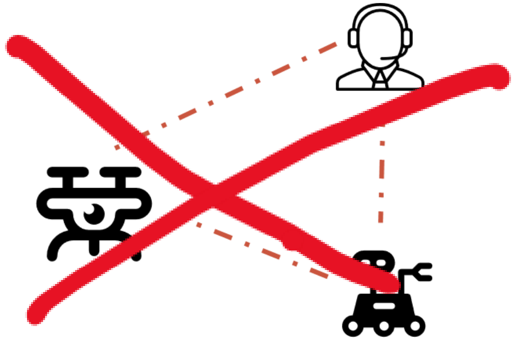


Metrics of Effectiveness Definition

- **Time** to clear Path (explicit):
  - Time needed to declare a path as clear for a battalion to move from point A to point B in less than ARG
- **Effectiveness** (not explicit):
  - The path defined as clear must have a minimum likelihood of being clear of ARG
- **Trustworthiness** (not explicit):
  - The path defined as clear must have a minimum level of trustworthiness such that the soldiers believe the path is safe. The minimum trustworthiness level is ARG.

Security Breaches

- **An enemy may distort the communication network**. Therefore, two uncertainties arise: whether **part of the message sent is distorted** (predictions, mine cleared declaring, etc) or **messages from outside the system were inserted in the system**.
- **An enemy may distort the communication service capacities**. Some of the communication links can be closed totally or intermittently.



Mission Model Formalization

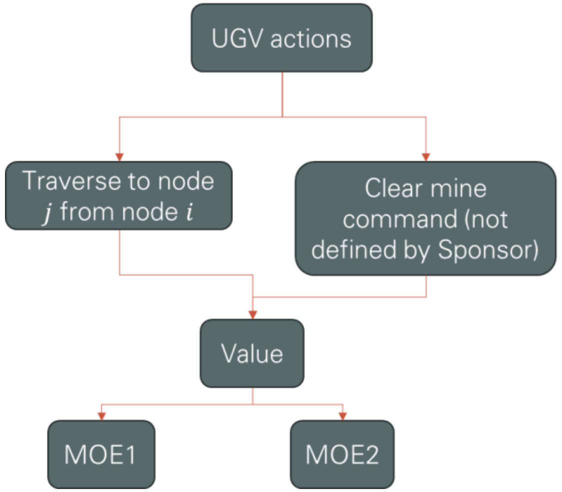
- **Formalize** how to estimate the MOEs defined above.
- Incorporate **sources of uncertainty**
- **Characterize the performance of the System** with respect to the mission at hand

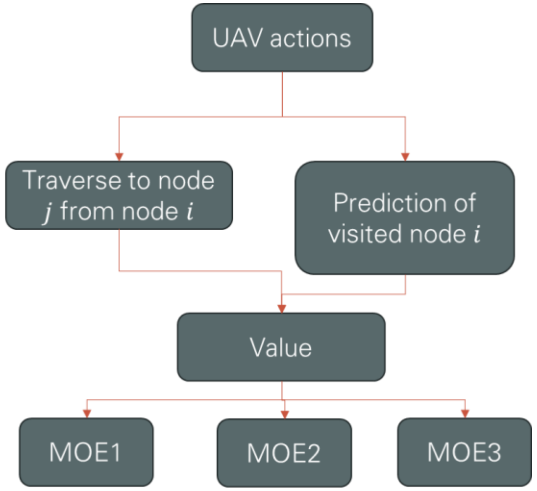
SIMULATION & ANALYSIS

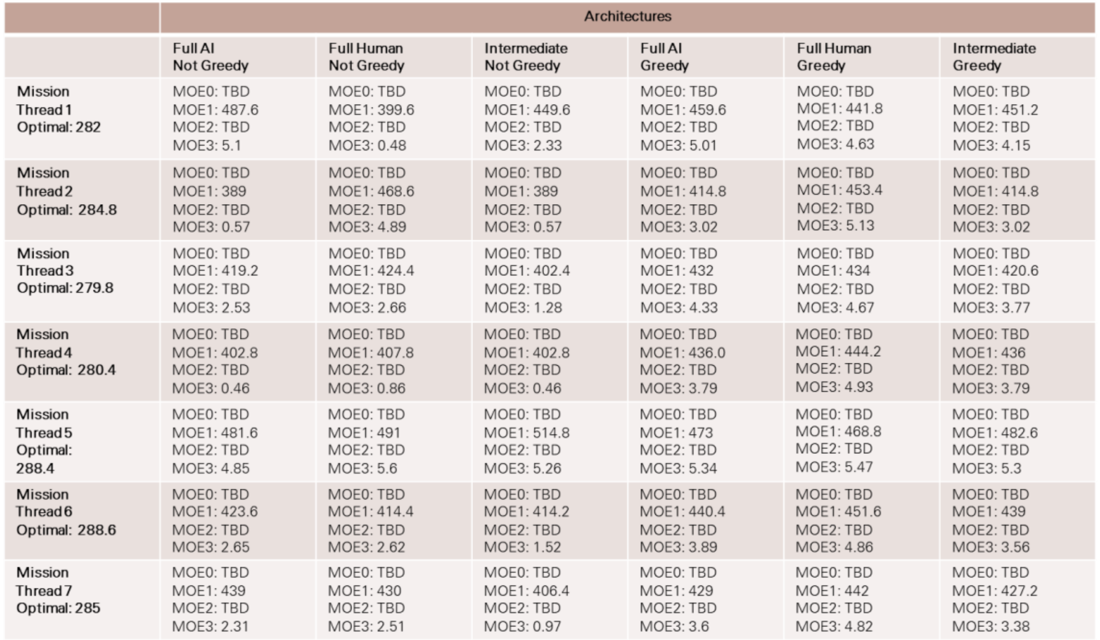
Scenarios = Mission threads X Architectures

Monte Carlo – 100 runs for each scenario

min(MOE0), min(MOE1), max(MOE2), max(MOE3)

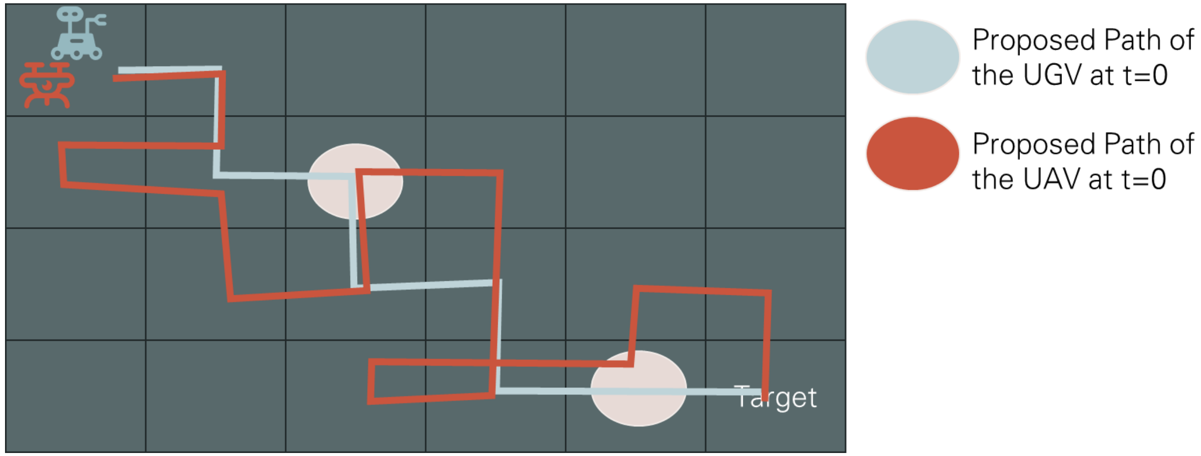

UGV actions
Traverse to node j from node i
Clear mine command (not defined by Sponsor)
Value
MOE1
MOE2


UAV actions
Traverse to node j from node i
Prediction of visited node i
Value
MOE1
MOE2
MOE3

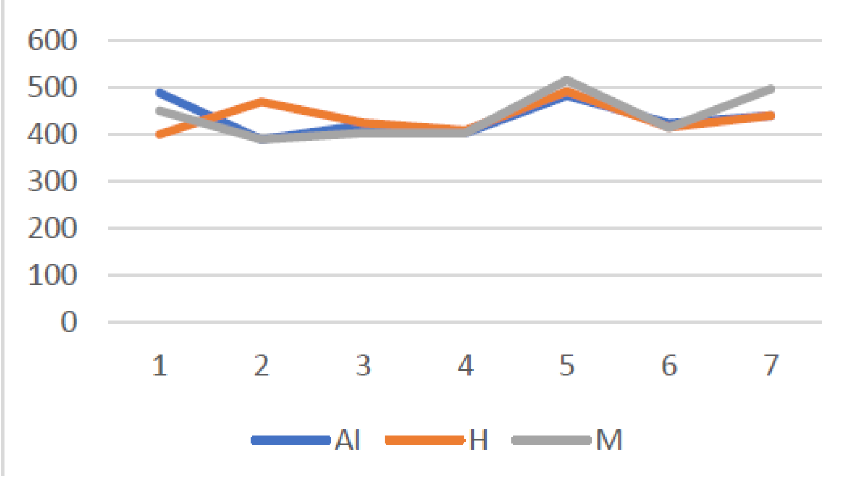


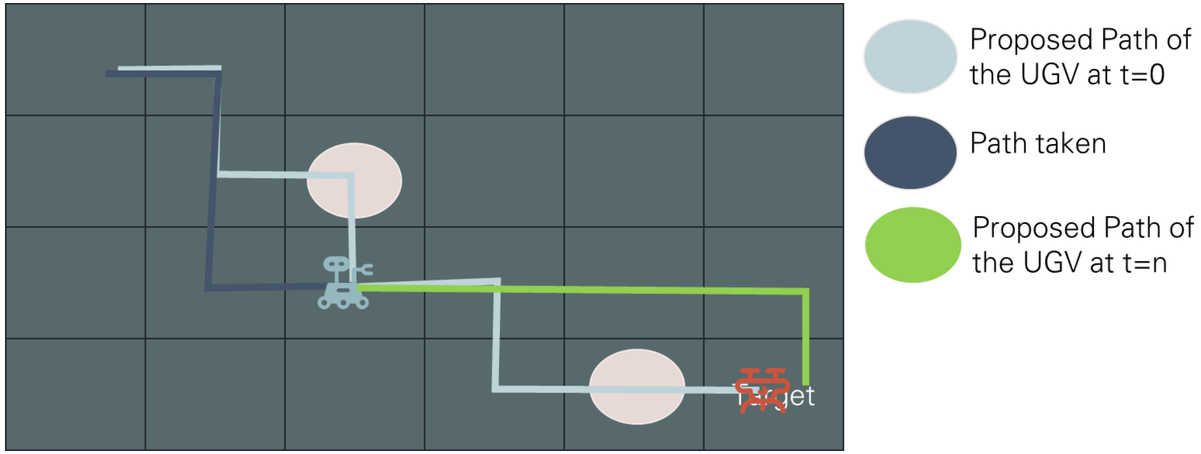
	Architectures					
	Full AI Not Greedy	Full Human Not Greedy	Intermediate Not Greedy	Full AI Greedy	Full Human Greedy	Intermediate Greedy
Mission Thread 1 Optimal: 282	MOE0: TBD MOE1: 487.6 MOE2: TBD MOE3: 5.1	MOE0: TBD MOE1: 399.6 MOE2: TBD MOE3: 0.48	MOE0: TBD MOE1: 449.6 MOE2: TBD MOE3: 2.33	MOE0: TBD MOE1: 459.6 MOE2: TBD MOE3: 5.01	MOE0: TBD MOE1: 441.8 MOE2: TBD MOE3: 4.63	MOE0: TBD MOE1: 451.2 MOE2: TBD MOE3: 4.15
Mission Thread 2 Optimal: 284.8	MOE0: TBD MOE1: 389 MOE2: TBD MOE3: 0.57	MOE0: TBD MOE1: 468.6 MOE2: TBD MOE3: 4.89	MOE0: TBD MOE1: 389 MOE2: TBD MOE3: 0.57	MOE0: TBD MOE1: 414.8 MOE2: TBD MOE3: 3.02	MOE0: TBD MOE1: 453.4 MOE2: TBD MOE3: 5.13	MOE0: TBD MOE1: 414.8 MOE2: TBD MOE3: 3.02
Mission Thread 3 Optimal: 279.8	MOE0: TBD MOE1: 419.2 MOE2: TBD MOE3: 2.53	MOE0: TBD MOE1: 424.4 MOE2: TBD MOE3: 2.66	MOE0: TBD MOE1: 402.4 MOE2: TBD MOE3: 1.28	MOE0: TBD MOE1: 432 MOE2: TBD MOE3: 4.33	MOE0: TBD MOE1: 434 MOE2: TBD MOE3: 4.67	MOE0: TBD MOE1: 420.6 MOE2: TBD MOE3: 3.77
Mission Thread 4 Optimal: 280.4	MOE0: TBD MOE1: 402.8 MOE2: TBD MOE3: 0.46	MOE0: TBD MOE1: 407.8 MOE2: TBD MOE3: 0.86	MOE0: TBD MOE1: 402.8 MOE2: TBD MOE3: 0.46	MOE0: TBD MOE1: 436.0 MOE2: TBD MOE3: 3.79	MOE0: TBD MOE1: 444.2 MOE2: TBD MOE3: 4.93	MOE0: TBD MOE1: 436 MOE2: TBD MOE3: 3.79
Mission Thread 5 Optimal: 288.4	MOE0: TBD MOE1: 481.6 MOE2: TBD MOE3: 4.85	MOE0: TBD MOE1: 491 MOE2: TBD MOE3: 5.6	MOE0: TBD MOE1: 514.8 MOE2: TBD MOE3: 5.26	MOE0: TBD MOE1: 473 MOE2: TBD MOE3: 5.34	MOE0: TBD MOE1: 468.8 MOE2: TBD MOE3: 5.47	MOE0: TBD MOE1: 482.6 MOE2: TBD MOE3: 5.3
Mission Thread 6 Optimal: 288.6	MOE0: TBD MOE1: 423.6 MOE2: TBD MOE3: 2.65	MOE0: TBD MOE1: 414.4 MOE2: TBD MOE3: 2.62	MOE0: TBD MOE1: 414.2 MOE2: TBD MOE3: 1.52	MOE0: TBD MOE1: 440.4 MOE2: TBD MOE3: 3.89	MOE0: TBD MOE1: 451.6 MOE2: TBD MOE3: 4.86	MOE0: TBD MOE1: 439 MOE2: TBD MOE3: 3.56
Mission Thread 7 Optimal: 285	MOE0: TBD MOE1: 439 MOE2: TBD MOE3: 2.31	MOE0: TBD MOE1: 430 MOE2: TBD MOE3: 2.51	MOE0: TBD MOE1: 406.4 MOE2: TBD MOE3: 0.97	MOE0: TBD MOE1: 429 MOE2: TBD MOE3: 3.6	MOE0: TBD MOE1: 442 MOE2: TBD MOE3: 4.82	MOE0: TBD MOE1: 427.2 MOE2: TBD MOE3: 3.38

EXAMPLE OF SOLUTION STRATEGIES & ASSESSMENT

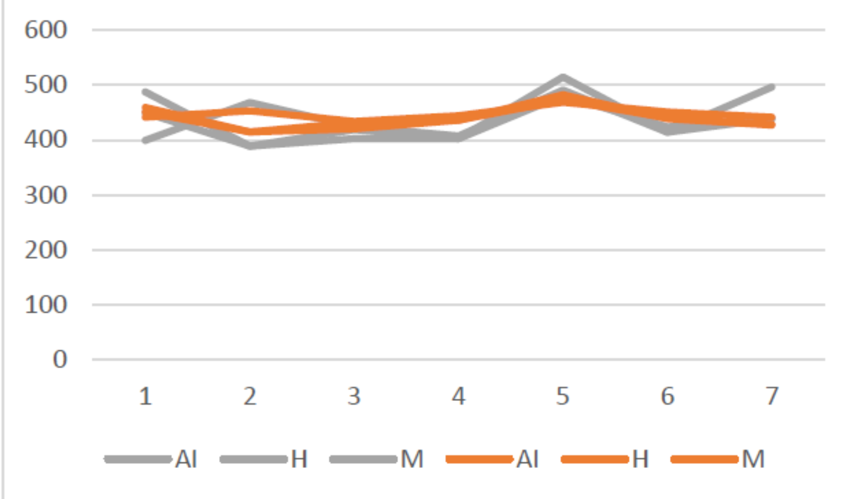


Target
Proposed Path of the UGV at t=0
Proposed Path of the UAV at t=0





Target
Proposed Path of the UGV at t=0
Path taken
Proposed Path of the UGV at t=n



Lessons Learned:

- Deeply study the situation at hand, **identifying all sources of uncertainty, operational variations,** etc.
- **Disaggregate the metrics that characterize the effectiveness of the system** to be implemented considering real-world situations.
- **Decouple action taking from predictions**/inference.
- Let decision making **algorithms incorporate uncertainty**.
- **Design systems** such that their functional flows **react to different uncertainty levels**.
- **Study architectural variations** that can **respond to communications breakage**.

THANK YOU

samuelcornejo@arizona.edu
alejandrosalado@arizona.edu