Phase I – White Paper

WRT-1085: Trusted Artificial Intelligence (AI) Systems Engineering (SE) Challenge: Seed Funding

Team: Afrooz Jalilzadeh, Alejandro Salado, Pratik Satam (The University of Arizona)

Document Number: UofA-DL-TR-002, v1

Date: August 13, 2024

Prime Contract Number: HQ003419D0003, DO HQ003423F0495

Subcontract Number: 2103596-04

Table of Contents

Problem statement from the sponsor and assumptions 3
Approach 3
Measures of effectiveness (MOEs) 4
Operational solution at the mission level (Soldier engagement) 4
Operational architecture at the system level 5
AI vs Human performance: Boundaries 9
Security 11
Model performance to system performance 12
Conclusions 14
Appendix. List of ancillary files 15

Problem statement from the sponsor and assumptions

The project consists of three phases:

- *Phase 1*. Explore performance of Artificial Intelligence (AI) models over variety of operational scenarios.
- *Phase 2*. Design of the decision system; human-machine teaming, resilience.
- *Phase 3*. Operational simulation of mission scenarios.

In essence, the teams must address throughout the three phases the following questions:

- Should a human operator assess imagery? Human review of a link takes 30 minutes vs 1 minute for the AI.
- What role should the human play?
- What notions of human trust in AI are important?
- Should the system be architected or operated to influence trust?

In Phase 1, our team has primarily focused on exploring the performance of the AI model and start the foundations to understand the potential design options of the decision system. We considered this necessary to properly contextualize the performance of the AI model.

Based on the documentation received from the sponsor and communications with them, our team made the following assumptions in our work:

- The Unmanned Ground Vehicle (UGV) is a mine clearing ground robot, which is a scarce resource for the mission.
- The Unmanned Aerial Vehicle (UAV) is a fast, multi-spectral video collection system.
- The time for the UGV to clear a path is:
    - 1 hour per link if a mine/IED is encountered.
    - 20 minutes per link if a mine/IED is not encountered.
- The AI performance data that were provided to the research team corresponds to the performance of the UAV.
- The human SME (operator) reviews video imagery from the UAV. During execution of the challenge, the SME will also get feedback from the UGV on whether a mine was present and will be able to correlate actual performance with predicted performance.

For Phase I, the sponsored required the following deliverables:

- Understanding of how AI performance varies with metadata.
- Initial ideas about relationships between model performance and system performance.

Both are provided in this white paper.

Approach

Most of our effort was focused towards understanding the problem and architecting some initial aspects of the solution. The work was driven by the information provided by the sponsor and our own knowledge of the problem domain. Because of our goal in this phase, we did not attempt to survey

existing literature for the specific details of the solution, which may become part of the next phases of the project.

Measures of effectiveness (MOEs)

First, we have refined the MOEs described in the documentation provided by the sponsor. We have both added precision to the definitions and incorporated additional measures that we believe necessary to adequately inform trade-offs later.

We recommend three overarching MOEs:

1) *Time to clear a path (MOE1)*; declaration. This MOE was identified by the sponsor. We interpret the MOE as the time that it takes the system to declare a path as cleared. By system we mean here the combination of the UAV, the UGV, the operator and any other element necessary for them to interoperate.
2) *Effectiveness (MOE2)*. This MOE was not identified by the sponsor, at least explicitly. However, this MOE is necessary to establish a coherent value model. This MOE refers to *how well* a path is cleared. This MOE is associated with the time to clear a path; clearly, declaring a path as cleared without being so would be inadequate.
3) *Trustworthiness (MOE3)*. This MOE was not explicitly identified by the sponsor but was implicit throughout the project documentation. This MOE appears in two ways:
    a. *As a factor that affects path clearing*. This refers to the extent to which information is trusted by the different actors. For example, low trust on an information set may trigger a call for additional information, which slows the clearance of a passage. This MOE is therefore an indirect measure to *Time to clear a path*, and not an MOE that is relevant on its own. However, because of the importance of trustworthiness in this project, we treat it explicitly.
    b. *As a factor that affects traversing the safe passage*. From a mission perspective, success in the soldiers traversing the path will depend on the extent to which they trust the path has been cleared. To what extent do I believe that the passage is actually safe? Passing is outside of the scope of modeling but is the key outcome of this system.

These three MOEs will form the basis of our value model: time to clear a path, effectiveness in clearing the path, and the trust that soldiers have in the clearance of the path. The performance of the AI model is anticipated to affect these three MOEs.

Operational solution at the mission level (Soldier engagement)

The operational solution at the mission level, particularly how soldiers plan to traverse the cleared path, may add considerations on the MOEs not yet stated. For example, will the soldiers wait for the path to be entirely cleared before traversing it, will they move one link behind the UGV, or directly behind the UGV? These different solutions pose different constraints on the MOEs stated earlier. For example, the operational solution will affect the importance of the time it takes the UGV to clear a path. If the soldiers wait to move until the path is cleared, there is likely a more stringent need for the system to clear the path. However, if the soldiers move right behind the UGV as it clears the path,

may relax such temporal need. In addition, there is a new dimension to trustworthiness that emerges in this case, since the soldiers' performance will not only be affected by the extent to which they trust the clearance of the path (as in the state of it being cleared) but also the clearing process, as their safety may be affected by the effectiveness of the UGV to clear a mine. (Note that clearing a path is different from clearing a mine in this case, as failing to clear a mine may only affect a soldier if it is erroneously detonated while the soldier is near the mine.)

We leave these aspects open at the moment and will address them in later phases of the project.

Operational architecture at the system level

We have identified the *level and distribution of autonomy* as the main design factor in this project. By level and distribution of autonomy, we mean the allocation of tasks, including decision ownership, between human operators, the UAV, and the UGV. For example, does the UGV decide on the clearing of a mine based on its own detection? Does the UGV move to designated areas by the UAV without human intervention?

To enable assessing these different design alternatives, we have developed in this Phase I a generic functional model of the system, which can be later instantiated into different architectures. Although the sponsor provided a model with the functions or activities performed by the different elements of the system, we believe the generic model was valuable to explore in depth the questions posed by the sponsor, as identified earlier in this paper. The functions (or activities) that have identified to complete the clearance of a path are the following:

- 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.

Note that there is no assumption about the temporal dependencies of the functions, other than they probably need to be completed in sequence from A1 to A8. Complete sequentiality may not be necessary, since some tasks might be operated in parallel. For example, A1 and A2 may be executed continuously while A4 through A8 are executed. Such decisions are outside of the scope of the work in Phase I and will be addressed in future phases.

Given the conditions of the problem presented by the sponsor, a minimal allocation of the functions to the different components can be performed, as shown in Figure 1. Note that the efficient survey of a large area can only be performed by the UAV and the safe survey of a zone and mine clearance

can only be performed by the UGV. All other functions or activities may be performed by any combination of the Operator, the UAV, and the UGV, including partitioning the activities or a more intricate allocation of subfunctions, depending, for example, on achieving certain thresholds on confidence, conditions of the terrain, expected performance, etc. Figure 2 and Figure 3 show examples of a human-heavy solution (where most tasks are allocated to the operator) and an autonomy-heavy solution (where most tasks are performed by AI), respectively. Figure 4 shows an example of a solution that is more intricate, where the specific execution of a task is performed by the operator or the AI depending on the performance of each on that task for the specific information at hand.

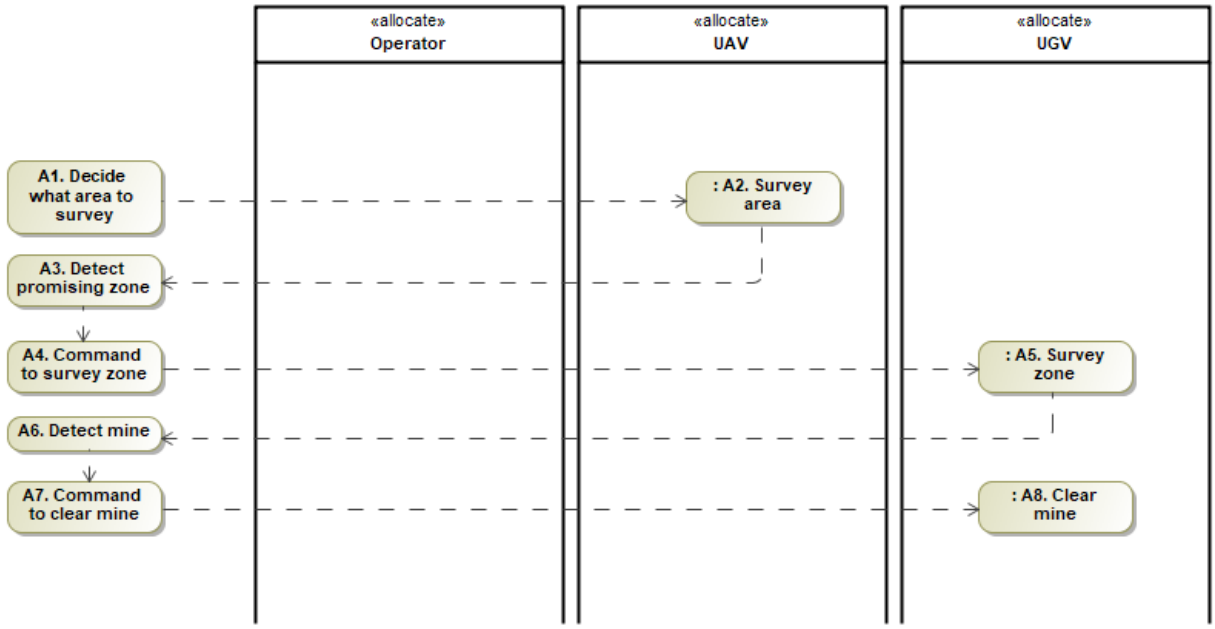

Figure 1. Minimal functional allocation

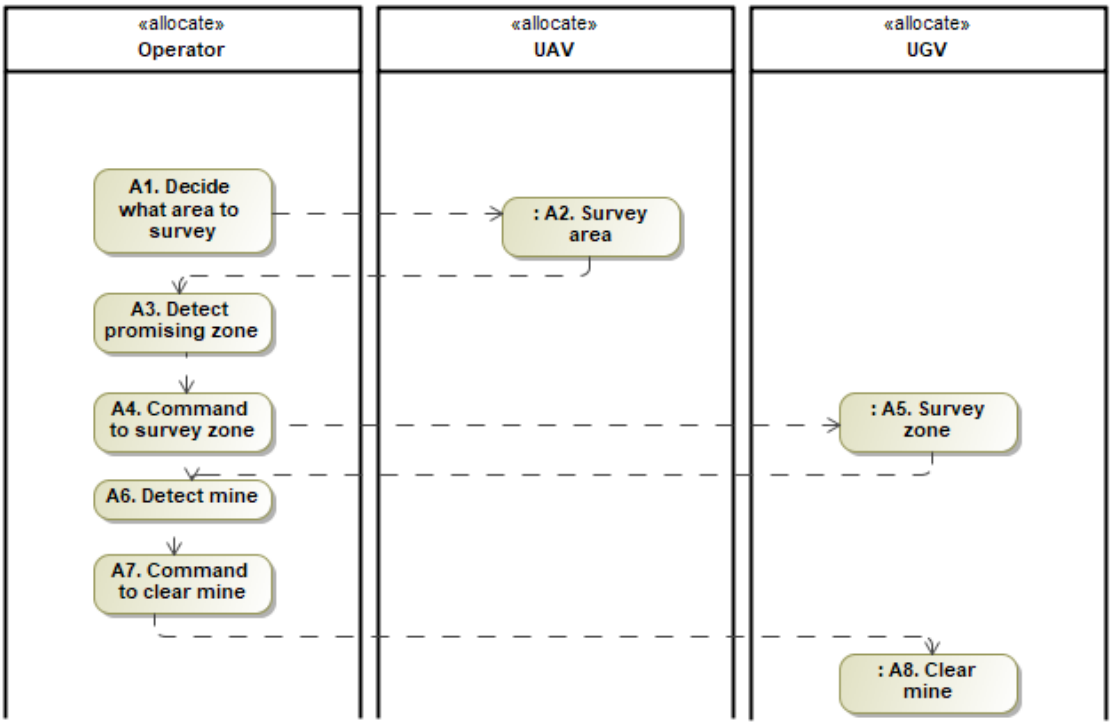

Figure 2. Human-heavy functional allocation

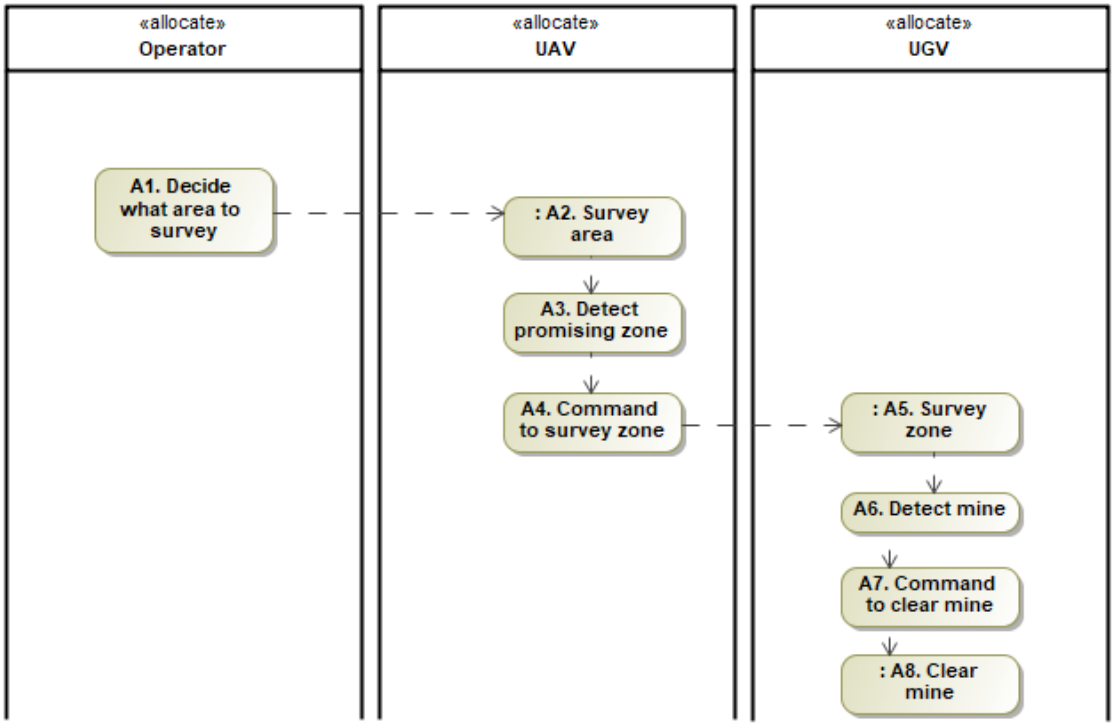

Figure 3. AI-heavy functional allocation

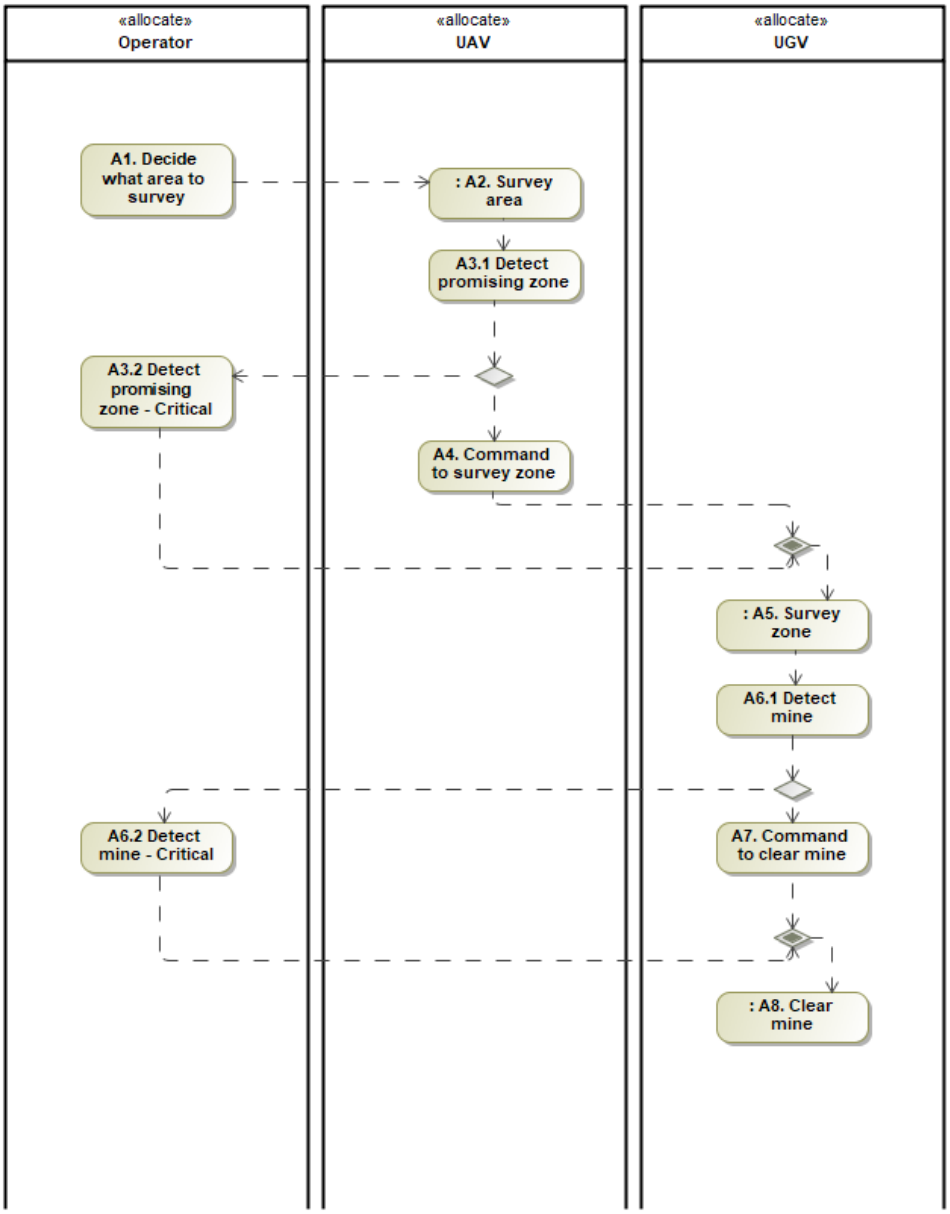

Figure 4. Intricate functional allocation, depending on AI vs human performance

From this model, we can also derive the fundamental factors of trustworthiness (the MOE):

- The confidence derived from the performance of the AI models and/or the human. That is:
  - Does the operator trust the performance of the AI models.
  - Does the UGV trust the information provided by the UAV, if applicable?
  - Do the UAV and UGV trust the information provided by the operator?
- The vulnerability of each function. That is, is the output generated by the function of high integrity or has the function been compromised?

- The vulnerability of each information interface (i.e., between the operator, the UAV, and the UGV). That is, is the input I receive of high integrity, or has it been compromised?

These factors contribute to the system level performance, which will be elaborated later in this white paper.

AI vs Human performance: Boundaries

We developed a predictive model using a neural network to assess AI and human performance based on environmental metadata. The features considered in the model include Surface type, Time of day (Day or Night), Temperature, Wind speed, and Visibility. By training the neural network on the available data, using 70% for training and 30% for testing, we enabled it to predict outcomes under various environmental conditions and determine the accuracy of both AI and human performance. The neural network model was evaluated using the Mean Squared Error (MSE) on the test data. The MSE was found to be 8.8532e-05 and 7.3178e-4 for AI and Human, respectively, indicating the average squared difference between the predicted and actual values. This metric demonstrates the predictive model's ability to predict accurately. The neural network consists of a single hidden layer with 5 neurons. The activation function used was ReLU (Rectified Linear Unit), and the output layer consisted of a single neuron with a linear activation function. The network was trained using the Levenberg-Marquardt algorithm.

To optimize the performance of the neural network, we experimented with different numbers of hidden neurons. We found that using a small number of hidden neurons, such as one or two, resulted in an underfitting model that could not capture the complexity of the data. On the other hand, using a large number of hidden neurons led to overfitting, where the model was too complex for the relatively small and straightforward dataset, resulting in poor generalization to new data. After careful consideration, we settled on five hidden neurons, as this provided a good balance—allowing the model to learn the necessary patterns without becoming overly complex.

Figure 5 represents the training processes of two distinct neural network models—one focused on AI and the other on Human performance. Both plots depict the Mean Squared Error (MSE) across training epochs, comparing the model's behavior on training, validation, and test datasets. In both plots, the MSE for the training, validation, and test datasets shows a consistent downward trend, with all three curves closely aligned. This indicates that the model is well-trained, with minimal overfitting, and generalizes effectively to unseen data.

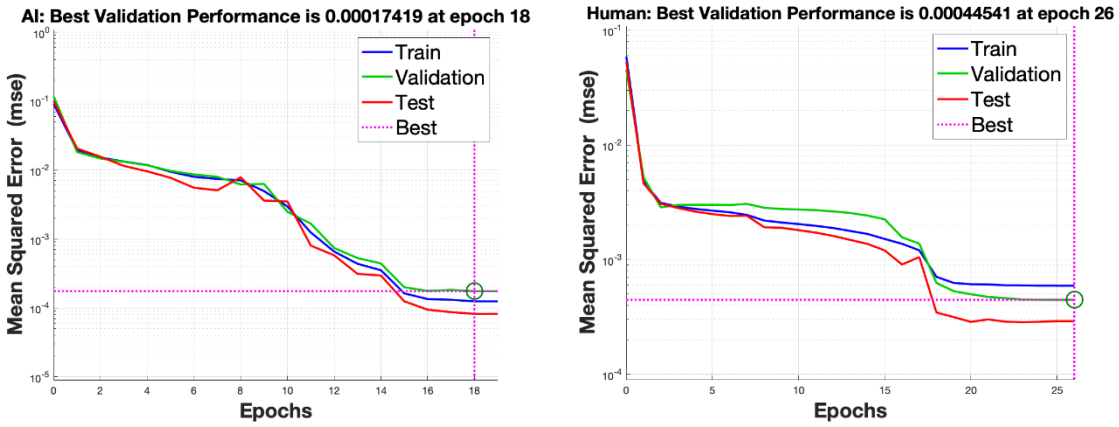

Figure 5. Training process of neural network models (left: AI performance; right: human performance)

The model now allows users to input specific environmental conditions—such as road type, temperature, visibility, and other relevant factors—and receive a predictive output on whether AI or human decision-making would be more effective under those circumstances. This interactive capability ensures that decision-makers can tailor their strategies to real-time conditions, optimizing the performance of the system in diverse environments.

It is important to note that the model is designed with flexibility in mind and can be updated or retrained when new data becomes available. This capability ensures that the predictive model remains accurate and relevant as additional information is gathered, allowing for continuous improvement in performance predictions for both AI and human assessments.

Considering the extensive data on surface conditions but limited information on other features like time of day, temperature, wind speed, and visibility, we now prioritize road condition as the primary variable to derive meaningful insights into when AI outperforms human decision-making. By analyzing the data, we calculated the mean and 95% confidence intervals for the predicted outcomes across different road types. This analysis allows us to clearly identify the scenarios where AI has an advantage over human decision-making and where it might fall short, enabling more informed decisions on which approach to rely on in specific situations. The analysis is visualized in Figure 6. The first plot shows the mean accuracy and the 95% confidence interval for AI across different road conditions, while the second plot presents the same metrics for human decision-making. Both AI and Human models demonstrate lower accuracy in wooded and rocky conditions, with the Human model exhibiting slightly higher accuracy in these challenging environments. However, the confidence interval for the AI model is tighter, indicating less variability in its predictions.

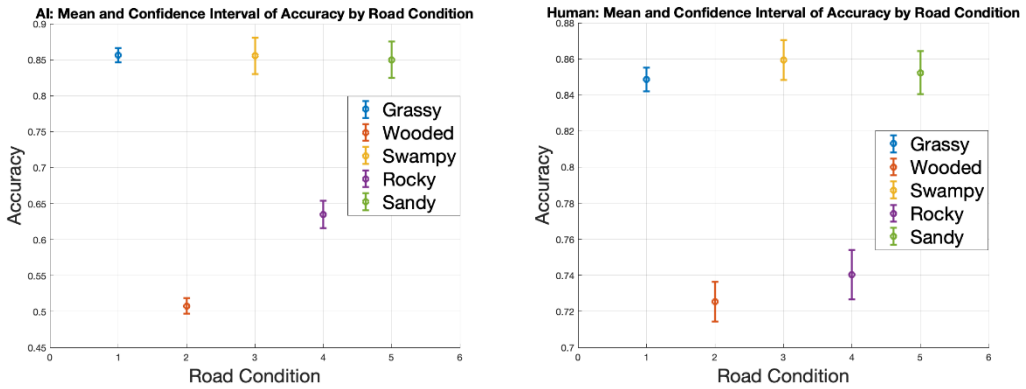

Figure 6. Comparative analysis AI vs human performance

Additionally, Table 1 summarizes the mean accuracy for both AI and human performance under each road condition. This table complements the plots by offering a quick reference to the average performance in each scenario, facilitating a more straightforward comparison between AI and human decision-making capabilities across the different terrains.

Table 1. Mean accuracy summary for AI and human performance under each road condition

	Grassy	Wooded	Swampy	Rocky	Sandy
AI	0.8561	0.5076	0.8555	0.6352	0.8500
Human	0.8485	0.7254	0.8594	0.7403	0.8524

The ultimate goal is to solve the shortest path problem within a mine-clearing mission. The approach will begin by determining whether to rely on AI or human evaluation based on the predicted accuracy for each link in the network. This decision-making process is critical and presents substantial challenges that require a carefully optimized model. The accuracy of predictions not only dictates whether AI or human evaluation is chosen but also significantly impacts the overall time required for demining operations. An incorrect prediction could lead to delays—either by directing the UGV along a path where mines are missed or by wasting time clearing paths that are already safe. Therefore, the next phase of this project will focus on developing an optimization model that carefully balances prediction accuracy with operational efficiency, minimizing the risks associated with incorrect predictions and ensuring the mission is completed as quickly and safely as possible.

Security

As previously discussed, we suggest that the existence of vulnerabilities is a major contributor to trustworthiness. Moreover, the critical nature of the Silverfish operation makes it necessary to characterize it as a Zero Trust Cybersecurity Problem. The Zero Trust Cybersecurity Problem deviates from the traditional defense in depth framework by focusing on the philosophy- “Trust nothing, and verify everything”. This section highlights a threat model, evaluated with the Zero Trust Cybersecurity in mind, analyzing the Silverfish system as a 3-Way Security Problem.

The Silverfish system has the potential for targets from different cyber-attack vectors. Here, we highlight some of the potential cyber-attack vectors that will be evaluated in future phases of the project, as part of architecting the potential solutions to the given problem:

1. *UGV/UAV*: UGVs and UAVs are complex systems that depend on various sensors, actuators, and communication networks for their normal operations, opening up opportunities for attackers to target and comprise them. These systems can be targeted with Injection attacks, trojans, side channel attacks, and reverse engineering attacks. These attacks can be used to target the UGV/UAV's control system, sensing systems, and communication systems. The mode of attack execution could be different avenues including networking-based attacks, attacks on the supply chain, or software/firmware attacks. It is critical to secure the UGV/UAV from such attacks, for example through the use of Intrusion Detection Systems and Intrusion Prevention Systems, that are able to detect and stop such attacks when they happen.
2. *AI*: AI models are notoriously vulnerable to adversarial attacks, wherein an attacker injects perturbations into the AI data to cause the AI system to misclassify their answers. Such adversarial attacks are broadly of the following types: a) Evasion Attacks: The attacker modifies the AI Input slightly so that the model makes an incorrect prediction or classification; b) Poisoning Attacks: The attacker injects misleading or malicious data into the training dataset; and c) Model Extraction and Inversion Attacks: an attacker aims to replicate an AI model by querying it with inputs and using the outputs to reverse-engineer the model. An attacker can perform such attacks through many means, including targeting the supply chain and innovative use of deception and camouflage techniques to cause AI misclassifications. The system can be secured against such attacks through the use of adversarial training and detection systems, for example.
3. *Insiders*: The Silverfish system is vulnerable to insider attacks who can comprise different aspects of its operations, especially the decision-making process. Such insider attacks can be mitigated through strict access control, monitoring, security awareness training, behavioral analytics, and incident response planning.

Model performance to system performance

A general model to link AI model performance to system model performance is not possible in this phase because of heterogeneity of the possible solutions at this time, both of the operations at the mission level and of the system architecture, as presented earlier.

For example, the Silverfish mission could be modeled as a tree traversal and an optimization problem. Here, all the available routes could be modeled as a spanning tree, with edges being set to weights characterizing the terrain type, traversal time, and the probability of encountering a mine (or not). By framing this solution then as an optimization problem, different conditions and algorithms to find the best generic algorithm could be explored. For instance, one example framework that could be explored is to structure the problem as a deep spanning tree, wherein the soldiers (moving with the UGV) traverse the first few nodes in a greedy manner incurring the cost of demining. While the UGV-team travels the path incurring the worst-case costs, the drone flies a few nodes ahead, and analyzes the future-path. The aim in this case would be to achieve a more accurate analysis

closer to the destination by buying in analysis time required by the human, while accepting the cost of greedy travel closer to the UGV-team's starting point.

Before proceeding with such a detailed model, higher level trade-offs would be necessary to determine the specific mission-level operations and overarching system architecture. Such an effort is planned for future phases of this project.

Meanwhile, in this phase, we have identified variables that we consider critical to feed the analytical models necessary to assess system performance from the AI model performance. The most relevant ones are discussed below.

Likely, the most impactful performances, besides direct measures, are related to the zone identification error and the mine detection error. Zone identification and mine detection may individually and/or jointly produce two types of error:

- Type 1. A mine exists but is not detected by the UAV.
    - Type 1.1. If the UGV detects the mine, this error leads to a higher time for clearance (MOE1).
    - Type 1.2. If it does not, this error will likely lead to a defective mission, as the UGV may be severely damaged (MOE2).
- Type 2. A mine does not exist but is detected. The impact of this error on timelines may be twofold (MOE1). If the selected path is through the "error," the time will be shorter. If the error has led to choosing a different path, the overall time will be suboptimal but no penalization on planned time will exist. That is, while the path is not optimal, it may still be considered acceptable, as it was chosen.

The knowledge by the soldiers and/or operations team of the different errors also impacts trustworthiness (MOE3).

Therefore, the system-level performance model will likely end up being fed by:

- A baseline time value for each activity to accomplish the mission (i.e., those identified earlier), which may be stochastic and will depend on the actor that executes the activity (e.g., a human takes 30 minutes to review imagery vs AI doing it in 1 minute).
- An activity error for each activity (e.g., detection error), which may lead to latency and/or overall system effectiveness.
- A trust matrix, which captures the trust that each actor has on the effectiveness of each activity.

The overall time to clear a path, depending on the final architecture, will likely depend on the baseline time values, the repetition time due to errors, and the potential delay in processing due to trust issues. Similarly, the overall system effectiveness will likely depend on the activity errors and activity allocation decisions that may depend on trust (e.g., an operator may allocate an activity to actor A because of lack of trust even though the effectiveness of actor B is better). As discussed, the specifics of the model will depend on the selected architecture.

Analytical and computer models will be developed in future phases.

Conclusions

We have presented some initial ideas related to the performance of the AI model, the relationship between AI model performance and system level performance, and the different factors that should be considered when architecting the Silverfish model.

Our findings indicate that trustworthiness will play a critical role in both the selection of the architecture and the performance of the system within the mission. Security will likely play a major role in determining trustworthiness.

Finally, we believe that the characterization data of the performance of both the human and the AI models should be leveraged to develop fluid architectures that maximize the performance of each system component at the mission level. Dynamically updating such characterization during field operations might further improve such a performance.

Appendix. List of ancillary files

The following table lists the files that were generated in completing Phase I. They have been submitted with this white paper as a .zip attachment.

File	File name
Codes used to assess the bounds and performance of the AI and Human models	codes-selected.zip