Trusted AI Challenge
Phase I: Architectures and Approach

Athul C. Dharmarajan, Zichong Yang, Bradley Feng, Ian Walter, Yupeng Zhou, Jitesh H. Panchal (panchal@purdue.edu)

School of Mechanical Engineering, Purdue University

Abstract

This whitepaper discusses the work done towards phase I of the Systems Engineering Research Center (SERC) Trusted Artificial Intelligence Systems Engineering Challenge. We describe our interpretation of the problem and the strategies we explored. We characterize the key assumptions required to propose a decision system for the task of navigating from the start and end points of the network. Then, we propose three different architectures for the decision system and measures to compare the performance of the proposed architectures. Finally, we list the variables and information required to compute the performance and additional information needed for Phases II & III.

Supported by:

SERC WRT-1085: Trusted Artificial Intelligence (AI) Systems Engineering (SE) Challenge: Seed Funding, Prime contract number: HQ003419D0003, DO HQ003423F0495, Subcontract Number 2103596-05

1 Problem Statement

The primary objective of the task is to ensure the safe passage from the start point to the endpoint along a defined network. An overview of the problem is shown in Figure *1*. The task involves identifying the path taken along the network. The network has links that might contain a mine/IED (Improvised explosive device). There is a central command and control center (C2) responsible for the operation along with a UAV (Unmanned Aerial Vehicle) and UGV (Unmanned Ground Vehicle), which travel along the network. UGV can remove a mine upon encounter, and the UAV can predict the likelihood of encountering a mine along a path using images fed to an AI-based model. C2 also has a human reviewer who can review the images and identify the likelihood of encountering a mine. We need to develop architectures of the decision system that integrate human operators with AI systems, ensuring that both entities can work together seamlessly. Another goal is to foster 'trust' between AI models and human operators by developing the notion of 'trust' between them. This involves identifying the specific roles that humans and AI should play within different architectural frameworks to optimize performance and identifying when a decision should be deferred to human reviewers. Finally, comparing various architectures across different performance metrics will help determine the most effective configurations for achieving these objectives.

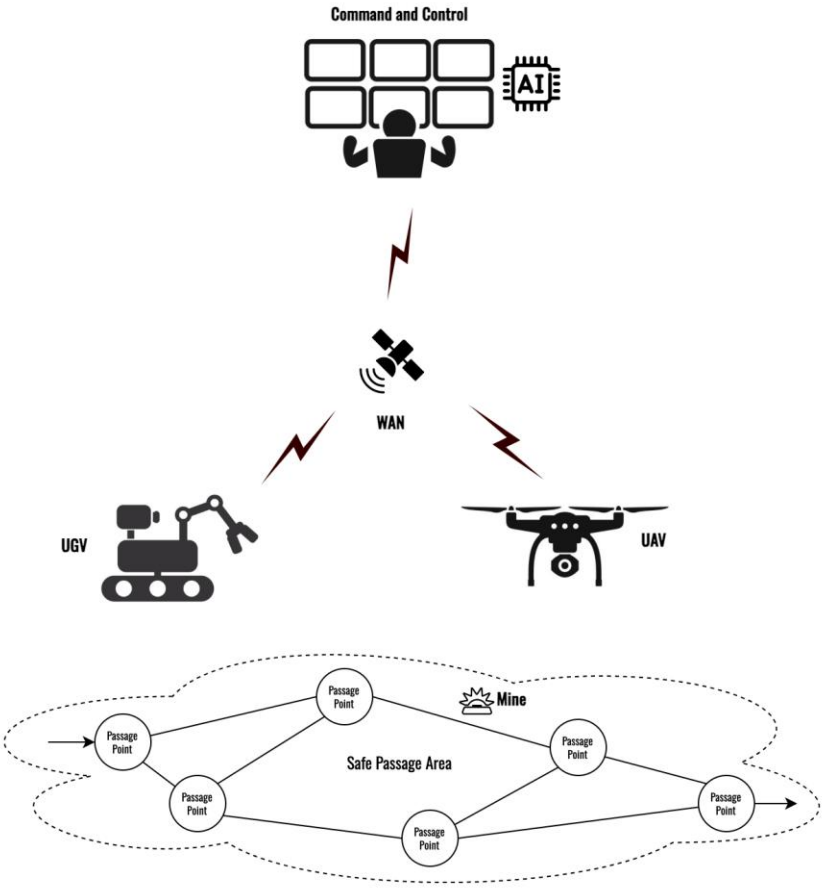

Figure 1: Overview of the problem

1.1 Assumptions

The main assumptions for each component are as follows:

- UGV
    - UGV can only receive instructions; no built-in processing or feedback mechanisms to send data to the Command and Control Center (C2)
    - No lethality/failure; UGV can clear the mines every time
- UAV
    - UAV has limited computational and communication capabilities; processing power onboard is weaker than C2 and can only run a lower fidelity version of the models, can't transfer all the high-resolution sensor data instantaneously (depends on the bandwidth)
    - On top of receiving instructions from C2, the UAV can share raw data collected from sensors (Video, images, LIDAR, etc.) and/or processed output (probability of encountering a mine)
- Dynamics of the environment
    - Initially, the conditions will be assumed static, i.e., the UAV needs only one pass to know the state of the network, and no changes are accounted for during operations. Later, this can be relaxed to account for changes in the network that happen during operation (fire breaking out, change in weather, etc.)
- Command and Control Center
    - The control center has enough human personnel available at all times to make decisions and analyze the inputs
- Coordination
    - The current strategy is for a single UAV and UGV. Later on, the architectures can be expanded to a swarm of UAVs and UGVs exploring the network in tandem

2 Architectures for the Decision-Making System

We describe the architecture of the decision-making system that determines the key decisions involved in the operation during a mission. An example of the decision can be which path to take at each step, depending on the likelihood of encountering a mine/IED. The architecture will define the flow of information, who is responsible for making key decisions, and how the decisions will be deferred to human reviewers due to lack of information.

2.1 Architecture A1: Centralized with Human Reviewer

In Architecture A1 (Figure *2*), the Human reviewer is responsible for all the key decisions. The human reviewer decides the path for the UAV to scan based on the images fed by the UAV. The reviewer also determines the path for the UGV to navigate based on the information from the UAV. Being a central architecture, this architecture depends on having enough bandwidth and low latency to ensure smooth communication between agents and the C2. This architecture

makes minimal use of artificial intelligence-based models, indicating a low amount of trust. Human review can be slow and comes at the cost of time.

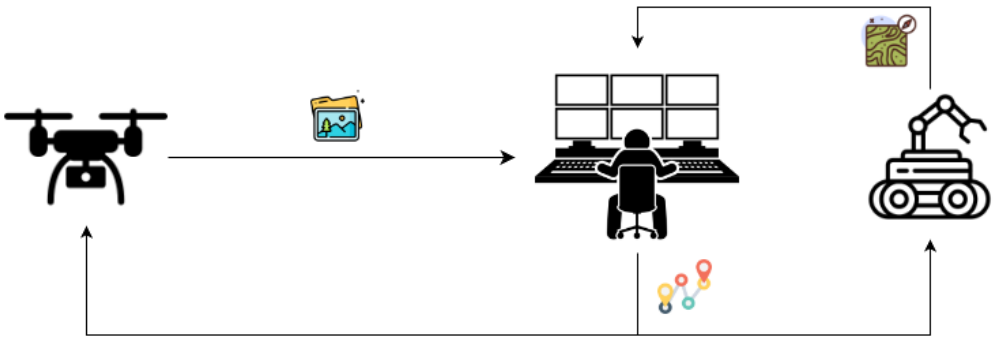

Figure 2. Architecture A1: Information Flow

2.2 Architecture A2: Centralized with AI

Compared to A1, AI assistance is introduced to Command and Control in Architecture A2 (Figure *3*). As using AI to detect mines is significantly faster, C2 can make faster decisions to return to the updated route and direct the UAV to capture pictures along the route faster, increasing overall efficiency. As for ensuring the accuracy of detecting the mines, C2 AI will hand over the task to the human reviewer if it has low uncertainty about whether there are mines in the captured terrain images. However, compared to A1 where the human reviewer is occupied, C2 AI in A2 will keep processing newly captured images and form the updated route.

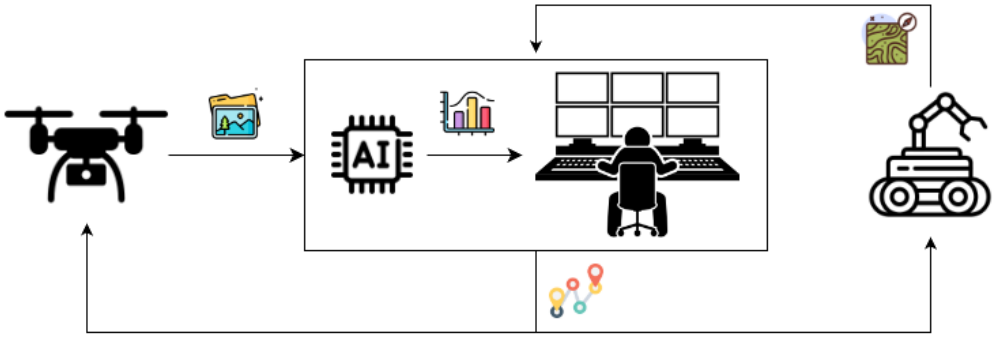

Figure 3. Architecture A2: Information Flow

2.3 Architecture A3: Decentralized with AI

With the emergence of on-device machine learning technology, we can also consider a situation where a lightweight on-device AI on UAV can complete most of the detection tasks, as shown in Architecture A3 (Figure *4*). In this situation, the AI-equipped UAV can directly talk to UGV to update the routes if it has high confidence about the detection results. In this situation, C2 AI and C2 Human will only get involved when UAV AI is uncertain. This will make the route update even faster, potentially shortening the total time to pass the region as the UGV will traverse fewer terrains with mines.

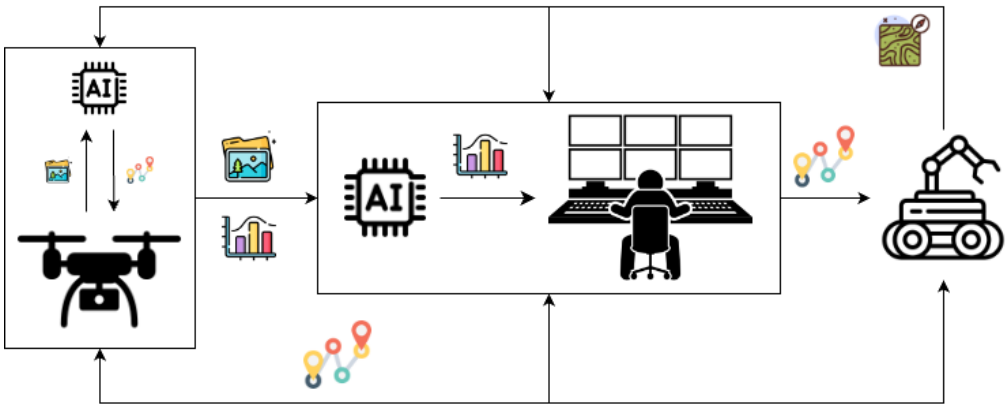

Figure 4. Architecture A3: Information Flow

3 Comparison of the Architectures

To effectively compare and evaluate the architectures of autonomous systems involving UAVs and UGVs, we come up with four core metrics, including expected time, lethality, suitability across different network types, and computational load on UAVs.

Expected time refers to the anticipated time it takes for a UAV or UGV to complete a mission from start to finish. The expected time for safe passage can be computed given the probability for each decision and state of the network. Shorter expected times generally indicate a more efficient system. By comparing the expected times of different architectures, we can evaluate which system achieves objectives more quickly, thus maximizing operational efficiency.

Lethality refers to the potential risk of failure that could result in damage or destruction (e.g., how often a mine is not successfully cleared). Although it has not been modeled in the current phase, comparing lethality rates in later phases can help determine which architecture is safer. Lower lethality rates would be preferable, indicating a higher reliability and safety of the system in clearing hazards without incident.

Suitability assesses how well the given architecture adapts to different network environments, which could vary in terms of size, complexity, or threat level. The system will potentially be deployed in varied or unpredictable environments or in other different networks. Thus, suitability is a crucial parameter to evaluate the robustness of the system.

Hardware requirements and limitations refer to the amount of data processing work that the UAV must perform during a mission. This helps us understand how much processing is offloaded to the command center and how much is handled on-board. Because the limited space on UAVs restricts the limitation of computational resources on UAVs. Too much computational load, which is beyond the capability of UAVs, will not only lower the efficiency but also lower the correctness. Architectures that manage to balance this load efficiently while maintaining system performance are generally more scalable and resilient.

4 Variables Needed to Compare the Architectures

4.1 Variables for Time Calculation

The expected time for the mission completion is the most important index for the effectiveness of the architectures. Here's a breakdown of the variables listed and their implications for system design and evaluation:

1. **Time of Decision Making:**
   - *t_ai_decision* = 1 minute: Time taken by the AI to make a decision based on imagery captured by the UAV. This reflects the processing speed and efficiency of the AI system.
   - t_human_decision = 30 minutes: Time taken by a human operator to make a decision using the same imagery. This highlights the difference in decision-making speeds between humans and AI, emphasizing the potential time savings offered by AI.
2. **Time of Moving Physical Entities:**
   - t_uav = 1 minute: Time for the UAV to capture and transmit imagery to Command Control. This includes the operational speed of the UAV and its communication efficiency.
   - t_ugv_block_no_mine = 20 minutes: Time for the UGV to traverse terrain where no mine is detected. This gives an idea of the UGV's speed in a non-threat environment.
   - t_ugv_clear_mine = 40 minutes: Time for the UGV to clear a detected mine, indicating the efficiency of the UGV in handling threats.
   - t_ugv_block_with_mine = 60 minutes: Total time for the UGV to traverse and clear a block with a mine. This is a combination of movement and operational time, showing how threat detection and clearance significantly increase operational time.

By analyzing these time variables, we can calculate some parameters such as decision-making efficiency and operational efficiency. So that we can assess the operational effectiveness of different system architectures and make informed decisions on system improvements, scalability, and deployment strategies.

4.2 Variables for Entity Trustworthiness

Although we assume 100% effectiveness, unlimited battery life, etc., there are still many places we didn't assume that it would function perfectly. It is essential for us to come up with a method to evaluate system effectiveness to maintain high standards of safety and efficiency in autonomous operations. We can characterize how much trust we can put in each entity by using the following metrics.

- Accuracy
  - Probability that the Human makes the correct decision
  - Probability that the AI makes the correct decision

    - Will accuracy increase if more pictures (from different angles) of a block is taken?
- Degradation of signals
    - Probability of frame loss from the UAV
    - Probability that directions from Command and Control to UAV and UGV is lost
- System failure
    - Probability that AI system is down
    - Probability that UAV/UGV did not respond to commands

Evaluating these variables provides a quantitative basis for trust in the system's human and machine components. It helps in understanding the system's robustness against failures and inaccuracies, ensuring that operational plans are both realistic and appropriately cautious.

5 Next Steps

In Phase II, the focus will be on the software implementation of the methods and the simulation of the architectures proposed in Phase I. During this phase, the measures of performance identified in Phase I will be computed to evaluate the effectiveness of the proposed architectures. These architectures will then be compared based on their performance. Additionally, it is crucial to account for any changes in the operational environment that may occur during the task. This includes considering factors such as limited bandwidth and potential errors or loss of information during communication, which could impact the overall performance and reliability of the system.

In Phase III, the architecture will be tested on previously unseen network structures to assess its robustness and adaptability. This phase will involve using new mission scenarios to evaluate the performance of the different architectures under varied conditions. Another aspect of this phase will be considering lethality in decision-making processes to ensure the effectiveness and safety of operations. Additionally, the strategy can be expanded to leverage the capabilities of multiple UAVs and UGVs, enhancing the overall operational efficiency. Human movement can also be planned in conjunction with UGVs, leveraging seamless coordination and integration between human operators and autonomous systems.