Trusted Artificial Intelligence for Warfighter Routing in High Risk Environments

Analyzing the Requirements of Trust in the Adoption of Artificial Intelligence in Military Operation

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Systems Engineering

By
Sami Saliba

November 8, 2024

Technical Team Members:
Stephen Durham, Hannah Palmer, Justin Abel, Andrew Edwards

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS
Rider Foley, Department of Engineering and Society

Hunter Moore, Department of Systems Engineering
William Scherer, Department of Systems Engineering

Introduction

The U.S. Army Combat Capabilities Development Command (DEVCOM) group sponsors a nationwide competition to identify systems engineering artifacts that build trust in Artificial Intelligence (AI)-enabled systems. As AI advances, it offers significant advantages over traditional methods in military logistics and planning (Szabadföldi, 2021). The superiority of high-complexity AI systems, however, often sacrifice human readability and understandability (Dwivedi et al., 2023). Without validation and transparency built into the design, allowing users and operators to audit accuracy and verify performance at all times, AI systems will face significant challenges to adoption (Svenmarck et al., 2018). Trust—ensuring that users, operators, and decision-makers can rely on AI systems—is crucial, even if the entire decision-making process is not fully understood (Leike et al., 2017). Trust is a factor that affects all autonomous systems, but in most control environments, statistical modeling and proven techniques are able to provide support to decision making (Matt et al., 2014). AI has no such statistics underpinning performance, and with lives potentially at stake, a lack of trust has delayed the integration of AI into current warfare tactics (Castelvecchi, 2016).

This project explores the intersection of trust in AI, specifically in ways to increase visibility, verifiability, and understanding in decision-making through the context of a life critical control problem. Trust is contextualized in this case through the exercise of troop movement and minefield traversal, a problem characterized by uncertainty and high risk. In this problem, soldiers must navigate a simulated minefield with unreliable mine detection methods. My technical topic seeks to improve operational robustness and user confidence, or trust, in AI enabled systems through the integration of explainable statistical models, data, and decision methods into opaque AI architecture. Next, I will focus on the requirements of the adoption and usage of AI in military operations through the Social Construction of Technology framework, focusing on the social and ethical dynamics shaping its acceptance.

Developing Strategies for Safe and Trusted Minefield Navigation

To explore methods of improving trust in AI pipelines, the goal of this work is to create a system that can efficiently route mine-defusing Unmanned Ground Vehicles (UGVs) and troops through simulated mine-laden terrain under various environmental conditions, as quickly as possible. The complexity of this problem stems from varying accuracy of mine detection methods. In this work, two systems are employed: a human observer and an AI. These methods have different accuracies depending on environmental factors such as visibility, time of day, and precipitation. Additionally, the processing times differ significantly with the AI able to evaluate a cell in one minute, whereas the human takes 30 minutes to evaluate the same cell. To enable the mine detection methods, a routable Unmanned Aerial Vehicle (UAV) is utilized to provide aerial reconnaissance of each possible traversal location. The overall problem can be visualized in the following objective tree (Figure 1).

Figure 1: Technical Objective Tree

With the scale of the United States military, the complexity of moving troops, supplies, and other goods from point A to B becomes a logistical challenge that is compounded by potentially hazardous terrain, inaccurate and delayed evaluation systems, and the countless environmental conditions encountered across all seven continents (Siegel, 2002). Ultimate decisions regarding current traversal methods are primarily human based using statistical prediction models, and heuristic planning to determine the safest path available (Serrano et al., 2023). Additionally, there are limitations of current technologies (e.g., mine detection systems) where operational efficiency, resource utilization, and

accuracy is highly variable (McCormack, 2014). Although there has been some exploration in incorporating AI based hazard detection and routing methods, the dependency of an autonomous system that must prioritize the preservation of life requires ethical consideration in the design of the AI model (Sarker, 2024). AI can help optimize paths, predict risks dynamically, and adapt to rapidly changing conditions, such as evolving enemy tactics or environmental hazards using a variety of inputs (Bistron & Piotrowski, 2021).

Methodology and Technical Evaluation:

To simplify the overall approach, the overall problem is considered a system of subsystems in three parts: Evaluation of which method (human or AI) should scan the potential location, Routing of the UAV, and Routing the UGV and troops.

Modeling Detection Reliability

The first step is modeling the inherent unreliability of AI and human detection methods in a way that subsequent decision making can be validated and audited through statistical techniques. Bayesian estimation can be employed to update the probability of mine presence per cell as additional data is provided (Zyphur & Oswald, 2015) (Figure 2). The accuracy, or inaccuracy, of prediction are constantly updating to maximize performance.

Figure 2: Accuracy Prediction Flow

Optimizing UAV Routing

Optimizing UAV routing (Figure 3) is essential in reducing mine encounters. Incorporating Baysesian estimation into Deep Reinforcement Learning (RL) (Li, 2018), we seek to create adaptive UAV pathfinding. Modeled as a Markov Decision Process (MDP) (Puterman, 1990), with states including UAV

position and scanned data, while actions involve choosing cells to scan. Bayesian estimates provide an accuracy and performance metric, demystifying some of the black box aspects of RL.

Figure 3: UAV Navigation

Routing UGV and Troops

Finally in routing the UGV and soldiers, a method is needed to minimize traversal time and avoid mines using the UAV data. Pathfinding algorithms are able to calculate the cost for each move (Foead et al., 2021), and find the shortest possible path based on the likelihood of a mine. For instance, if a mine adds 40 minutes to a cell traversal, a cell with a 50% mine probability adds 20 minutes to the base time. This ensures the UGV selects the safest and most efficient route, updating paths in real-time to align with mission goals and enhance operational efficiency.

Figure 4: UGV Navigation

Evaluation Criteria

To evaluate the success of the overall system, several criteria must be addressed to ensure optimal performance and mission success. The primary criterion is *trust and reliability*. This encompasses verifying that the system functions as intended and inspires confidence in soldiers to rely on AI. Evaluating trust involves analyzing metrics such as accuracy, false positive/negative rates, and the

system's consistency across varied environmental conditions. Although false positives would require rerouting but maintain safety, a false negative is unacceptable. The system would be trained to minimize this feature at all costs so that troops are not unknowingly directed to a mine. *Traversal time* is another essential criteria, focusing on minimizing the duration of missions to reduce exposure to potential threats. This metric includes the expected time under normal conditions and variance to capture delays caused by obstacles like mines or shifting environments. Performance variability under different environmental conditions must also be assessed. This ensures the system's resilience and reliability when faced with diverse, unpredictable situations. Key indicators include how environmental changes affect detection accuracy and the system's adaptability to unforeseen conditions and thus, *environmental resilience* is an important outcome. Lastly, *resource utilization* is critical for operational efficiency, involving concurrent processing by both AI and human systems. Effective parallel processing ensures real-time data analysis and decision-making. Metrics such as the average number of concurrent processes and server utilization rates help identify how well resources are managed throughout identification, routing, and mine-clearing operations.

A successful implementation of this system must be able to effectively balance and optimize for each sub-objective. Thus, the primary exploration of the technical aspect in this work is: *How can a balance of these sub-objectives be reached to maximize trust and minimize total traversal time?*

Analyzing the Adoption of AI in Military Operations

For AI to be trusted in control of critical, life-sensitive scenarios, methods must be developed to ensure performance with or without human involvement. The Social Construction of Technology (SCOT) framework (Pinch & Bijker, 1984) provides a lens to examine how relevant social groups - users, decision-makers, and other stakeholders - shape the design and adoption of technology. SCOT highlights the process of interpretive flexibility, where these groups ascribe different meanings, uses, and priorities to the technology. Over time, as negotiations between social groups resolve conflicting interpretations,

closure and stabilization occur, solidifying the technology's form and function. In the context of this project, these stages will be analyzed in the remainder of this section.

The adoption of human-out-of-the-loop systems is shaped by the expectations and needs of its relevant social groups. Factors such as human impact, ethical considerations, and international law are critical in influencing these groups' interpretations and acceptance (Amoroso & Tamburrini, 2020). For example, autonomous systems that cannot be fully explained require mechanisms to ensure their functionality and accuracy (Umbrello et al., 2020). Without sufficient trust in these systems, relevant social groups may resist their adoption, regardless of technical improvements. Military leaders, as a key social group, must balance innovation with strategic, ethical, and safety considerations, prioritizing systems that allow human oversight and maintain consistent performance (Nuechterlein, 1976). This drives developers to incorporate trust mechanisms, such as real-time anomaly detection and explainable outputs, to meet the standards set by these influential groups.

The trust gap between humans and autonomous systems presents a significant barrier. Soldiers, as another relevant social group, may resist adopting systems perceived as "black boxes" due to their lack of interpretability, even if these systems demonstrate superior performance. Leaders face the challenge of reconciling the potential benefits of these systems with the need for rigorous validation and ethical deployment. These conflicts underscore the importance of developing hybrid control systems and explainable AI methods that allow human oversight without compromising efficiency (Bao et al., 2021).

Stabilization in the context of autonomous systems for military operations can only occur when both soldiers and leaders reach a consensus on the systems' trustworthiness and reliability. This requires technologies that incorporate explainable outputs and verifiable decision-making models. When these systems satisfy the needs and expectations of relevant social groups, the technology can transition from contested adoption to widespread use, achieving closure.

The interpretive flexibility inherent in SCOT is evident in how different social groups engage with autonomous systems. For soldiers, these systems must instill confidence and facilitate decision-making, while for leaders, they must align with broader strategic objectives and uphold

accountability. These differing interpretations shape the trajectory of the technology, guiding it toward designs that incorporate transparency and explainability to gain the full trust and acceptance of all relevant social groups.

Research Question and Methods:

Through the SCOT framework, I seek to answer: *What factors are influencing social groups in adopting AI technology in military environments?* For this question, I will utilize the competition reports and submissions for the DEVCOM competition from the nine competing schools across the country. With the direct involvement of DEVCOM in shaping the adoption, operationalization, and integration of emerging technologies into military settings, the competition results and submissions will be valuable in creating a complete picture of the requirements, concerns, and social factors influencing the adoption and trust of AI enabled systems.

The results of the competition provide a quantitative and qualitative assessment of various approaches in building trust in AI systems. DEVCOM's scoring framework ranks submissions based on various criteria, such as accuracy, explainability, and operational feasibility, which reflect the priorities and expectations of military stakeholders. These rankings not only reveal technical strengths but also highlight the relative importance of social factors like usability, transparency, and trustworthiness. By analyzing how these metrics were weighted and how the winning teams addressed stakeholder concerns, I can identify patterns in what is deemed essential for building trust in AI systems.

As an additional component, the written methodologies and design rationales submitted by the competing teams will be analyzed. These documents provide insight into how technical teams interpret and address the expectations of relevant social groups, such as military leaders and operators. Through SCOT's interpretive flexibility lens, I will examine how teams framed their solutions to align with different social group priorities, uncovering competing narratives about what constitutes a trustworthy AI system. By utilizing both methods, the SCOT analysis will reveal how relevant social groups interpret and

influence AI adoption, the competing interpretations that arise, and the conditions necessary for achieving closure and stabilization in military applications of AI.

Conclusion

This project examines both the technical and social dimensions of trust in autonomous systems for military applications. The technical focus is on enhancing the operational robustness of AI-enabled systems for minefield navigation through explainable models and verifiable decision-making frameworks. The social dimension employs the Social Construction of Technology (SCOT) framework to analyze how stakeholders, such as soldiers and military leaders, influence the adoption and stabilization of these technologies. Together, these efforts aim to bridge the gap between technological capability and social acceptance, providing solutions that ensure safer, more effective military operations while addressing the ethical and operational concerns that shape trust and adoption.

References:

Amoroso, D., & Tamburrini, G. (2020). Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues. *Current Robotics Reports*, *1*(4), 187–194. https://doi.org/10.1007/s43154-020-00024-3

Bao, Y., Cheng, X., Vreede, T. D., & Vreede, G.-J. D. (2021). Investigating the relationship between AI and trust in human-AI collaboration. *Hawaii International Conference on System Sciences 2021 (HICSS-54)*. https://aisel.aisnet.org/hicss-54/cl/it_enabled_collaboration/3

Bistron, M., & Piotrowski, Z. (2021). Artificial Intelligence Applications in Military Systems and Their Influence on Sense of Security of Citizens. *Electronics*, *10*(7), Article 7. https://doi.org/10.3390/electronics10070871

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, *538*(7623), 20. https://doi.org/10.1038/538020a

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G.,

& Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9), 1–33. https://doi.org/10.1145/3561048

Foead, D., Ghifari, A., Kusuma, M. B., Hanafiah, N., & Gunawan, E. (2021). A Systematic Literature Review of A* Pathfinding. *Procedia Computer Science*, *179*, 507–514. https://doi.org/10.1016/j.procs.2021.01.034

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2017). *AI Safety Gridworlds* (No. arXiv:1711.09883). arXiv. https://doi.org/10.48550/arXiv.1711.09883

Li, Y. (2018). *Deep Reinforcement Learning: An Overview* (No. arXiv:1701.07274). arXiv. https://doi.org/10.48550/arXiv.1701.07274

Matt, P.-A., Morge, M., & Toni, F. (2014). *Combining statistics and arguments to compute trust*.

McCormack, I. (2014). The Military Inventory Routing Problem with Direct Delivery. *Theses and Dissertations*. https://scholar.afit.edu/etd/684

Nuechterlein, D. E. (1976). National interests and foreign policy: A conceptual framework for analysis and decision-making. *Review of International Studies*, *2*(3), 246–266. https://doi.org/10.1017/S0260210500116729

Pinch, T. J., & Bijker, W. E. (1984). The Social Construction of Facts and Artefacts: Or How the Sociology of Science and the Sociology of Technology might Benefit Each Other. *Social Studies of Science*, *14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Puterman, M. L. (1990). Chapter 8 Markov decision processes. In *Handbooks in Operations Research and Management Science* (Vol. 2, pp. 331–434). Elsevier. https://doi.org/10.1016/S0927-0507(05)80172-0

Rigby, J. C., McWilliams, J., & Johnson, J. (Eds.). (2018). Generational Shift: How technology is shaping a step change in the future of mine counter-measures. *Conference Proceedings of INEC*. https://doi.org/10.24868/issn.2515-818X.2018.067

Sarker, I. H. (2024). AI for Critical Infrastructure Protection and Resilience. In I. H. Sarker (Ed.), *AI-Driven Cybersecurity and Threat Intelligence: Cyber Automation, Intelligent Decision-Making*

and Explainability (pp. 153–172). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-54497-2_9

Serrano, A., Kalenatic, D., López, C., & Montoya-Torres, J. R. (2023). Evolution of Military Logistics. *Logistics*, *7*(2), Article 2. https://doi.org/10.3390/logistics7020022

Siegel, R. (2002). Land mine detection. *IEEE Instrumentation & Measurement Magazine*, *5*(4), 22–28. IEEE Instrumentation & Measurement Magazine. https://doi.org/10.1109/MIM.2002.1048979

Svenmarck, P., Luotsinen, L., Nilsson, M., & Schubert, J. (2018). *Possibilities and Challenges for Artificial Intelligence in Military Applications*.

Szabadföldi, I. (2021). Artificial Intelligence in Military Application – Opportunities and Challenges. *Land Forces Academy Review*, *26*(2), 157–165. https://doi.org/10.2478/raft-2021-0022

Umbrello, S., Torres, P., & De Bellis, A. F. (2020). The future of war: Could lethal autonomous weapons make conflict more ethical? *AI & SOCIETY*, *35*(1), 273–282. https://doi.org/10.1007/s00146-019-00879-x

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian Estimation and Inference: A User's Guide. *Journal of Management*, *41*(2), 390–420. https://doi.org/10.1177/0149206313501200