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

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Sami Saliba

Spring 2025

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

Advisor

Rider W. Foley , Department of Engineering and Society

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 sacrifices 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).

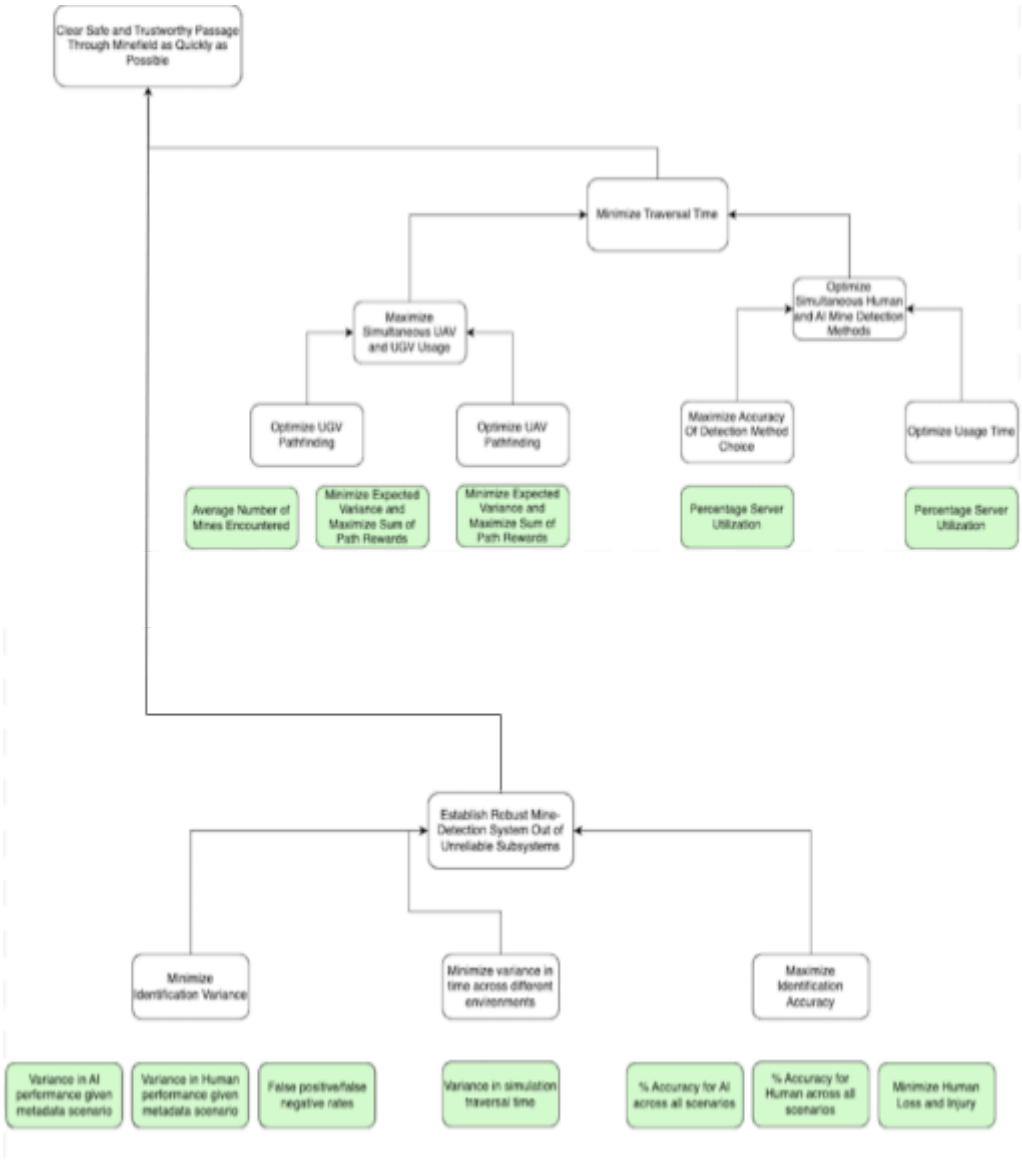
This project explores the intersection of trust in AI, specifically in ways to increase visibility, verifiability, and understanding in decision-making through the context of a life critical control problem. Trust is contextualized in this case through the exercise of troop movement and minefield traversal, 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



Hierarchy of system objectives

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).

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

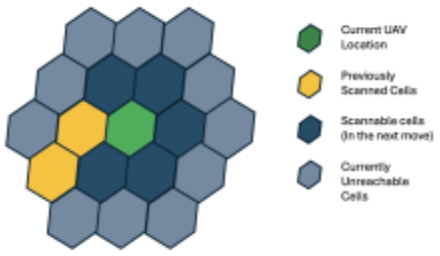
Process of converting metadata into prediction accuracy

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



Framework for UAV routing across cells

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

Framework for routing the UGV through prescanned cells

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.

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? This research examines how different stakeholders—military decision-makers, academic researchers, and engineering practitioners—evaluate trust in AI systems, using the Trusted AI Challenge as a structured case study. Given DEVCOM's direct involvement in shaping AI adoption, the competition provides a controlled environment to analyze how trust is assessed, quantified, and framed across professional perspectives.

This study employs a mixed-methods approach. Quantitatively, judges from government, academia, and industry scored each team across ten categories relevant to AI trustworthiness. These

scores were aggregated and compared across affiliations to identify trends in how different social groups prioritize aspects like risk management, explainability, and engineering rigor.

Qualitatively, written research papers and judge comments from the top-performing teams were thematically coded across four core dimensions: Best Practices, Novel Approaches, Systems Engineering Activities, and Trust Infrastructure. These categories reflect recurring themes in how teams justified their designs and how evaluators interpreted trustworthiness.

By combining these methods, this study explores both numerical trends and interpretive flexibility within SCOT—highlighting how professional backgrounds shape expectations of trusted AI and where consensus or divergence emerges between stakeholder groups.

Data Sources and Selection of Focus

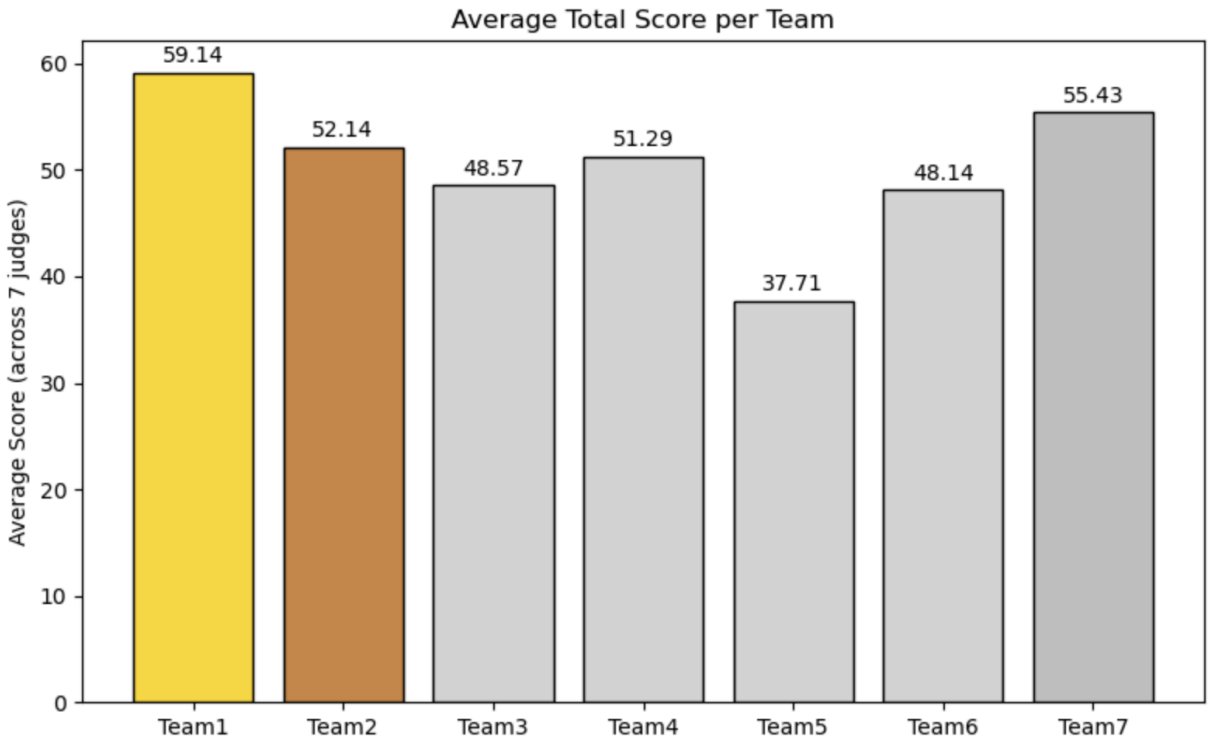
Quantitative Scoring

Seven judges rated teams on a 1–7 scale across ten categories relevant to trusted AI systems:

- **SE Activities** – How well systems engineering practices were applied.
- **Trust Infrastructure** – Measures ensuring explainability and oversight.
- **Key Workforce Skills** – Technical competencies necessary for AI trust.
- **Design Patterns** – Architectural decisions that promote reliability.
- **Risk-Based Monitoring** – Mechanisms for mitigating AI failures.
- **Quantitative Methods** – Use of data-driven validation techniques.
- **Best Practices** – Alignment with established industry and academic standards.
- **Novel Approaches** – Innovation in AI methodologies.
- **Future Plans** – Sustainability and adaptability of AI solutions.
- **Transition** – Feasibility of real-world implementation.

These judges, affiliated with industry, academia, or government, represent different social groups whose perspectives shape interpretations of what constitutes a "trusted AI solution." In keeping with SCOT's interpretive flexibility, each affiliation potentially ascribes distinct importance to certain categories (e.g., risk-based management vs. design patterns). From the seven participating teams, the top three (by highest mean overall score) were selected for deeper qualitative examination (Figure 5), as these submissions most successfully represent the principles of trusted AI according to the evaluators. Analyzing these high-scoring entries offers insight into which strategies resonated most across stakeholder groups and provides representative examples of how trust is constructed in practice.

Figure 5:



Average total score for each team across the seven judges. The maximum possible score is 70. Team 1 came in first with an average score of 59, followed by Team 7, and Team 2.

Quantitative Analysis

Score Aggregation and Ranking

The judges' ratings formed a matrix (Figure 6), where each row represented a team and each column represented a scoring category. Mean scores were calculated per category to assess trends, and overall team rankings were determined by averaging across all categories.

This ranking guided the selection of teams for deeper qualitative examination. An additional analysis explored whether specific social groups favored particular teams, highlighting potential biases introduced by professional backgrounds.

Figure 6 (Example Judge Scoring):

	Team 1						
Factor	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7
SE activities	6	6	4	6	5	6	7
Trust infrastructure	6	6	7	6	6	5	2
Key workforce skills/abilities	6	7	7	6	5	7	7
Design patterns	6	4	7	6	6	6	6
Risk-based monitoring/mgmt	6	4	6	7	6	7	6
Quantitative methods	6	7	7	7	7	7	6
Best practices	5	7	4	6	5	6	6
Novel approaches	5	6	4	7	7	6	5
Future plans	6	6	6	6	5	6	6
Transition	6	6	5	6	6	6	6
Total	58	59	57	63	58	62	57

Judge color labeling indicates their background, with green representing government, orange academia, and yellow being the sole industry judge.

Affiliation-Based Cross-Check

To investigate interpretive flexibility, scoring patterns were compared across the three social groups. Industry, academia, and government judges each have distinct concerns when evaluating AI:

- Industry tends to emphasize efficiency, practicality, and scalability, favoring approaches that are immediately implementable (Peres et al., 2020).
- Academia values novelty, methodological rigor, and quantitative validation, prioritizing innovative techniques even if they require additional testing (Uddin, 2024).
- Government places the highest weight on risk-based monitoring, explainability, and oversight mechanisms, ensuring that AI systems align with regulatory and security standards (Ahn & Chen, 2022).

By analyzing how each group scored different categories, patterns of divergence and consensus were identified.

Qualitative Analysis

Qualitative Research Papers and Coding Approach

Each of the top three teams submitted a research paper detailing their approach to building trusted AI systems. Within the SCOT framework, these documents serve as a lens into how trust is interpreted and prioritized by different social groups. To systematically analyze these strategies, the papers were thematically coded across four dimensions central to AI adoption: Best Practices, Novel Approaches, Systems Engineering (SE) Activities, and Trust Infrastructure.

These categories capture a range of design priorities, from alignment with industry and government standards, to the use of innovative methodologies, structured engineering practices, and mechanisms for reliability and oversight. This structure allows for a direct comparison of how teams balanced innovation with feasibility and how those choices resonated with evaluators from different backgrounds.

To deepen this analysis, statistical testing using the Kruskal–Wallis test was conducted across all categories. The results showed statistically significant differences in how affiliations scored Best Practices, SE Activities, and Trust Infrastructure, providing further insight into the interpretive flexibility among social groups and the priorities they emphasize in trusted AI.

Results

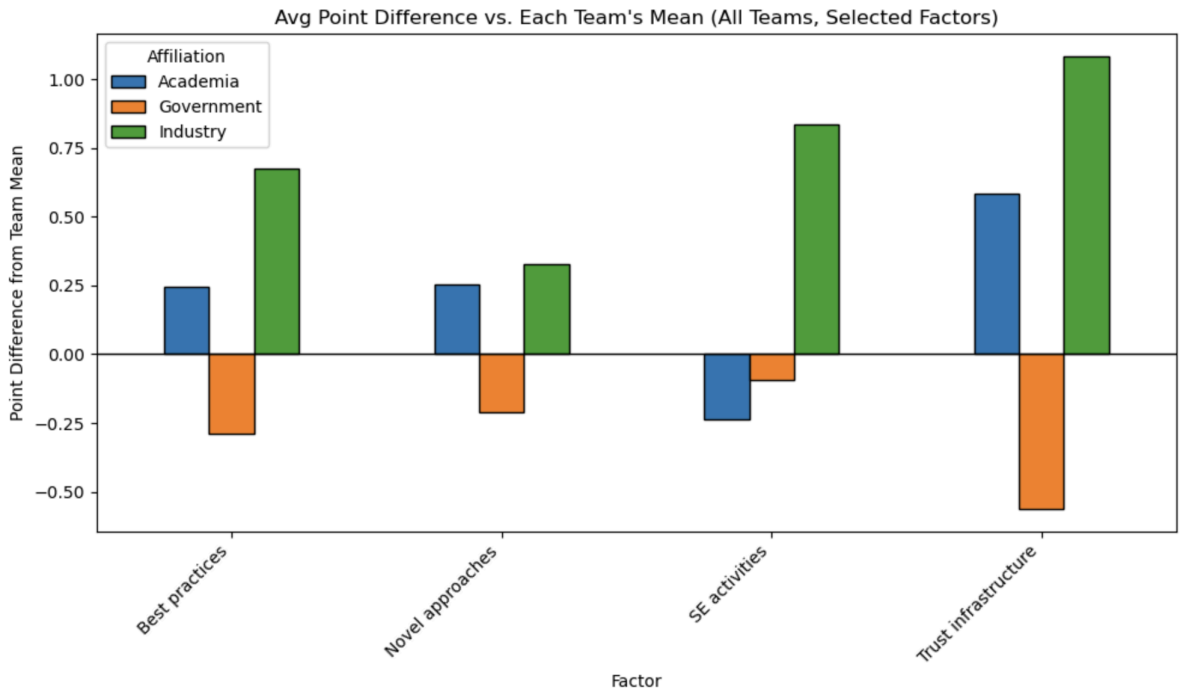
Overview of Findings

This section examines how the different social groups, industry, academia, and government, assessed AI trustworthiness through both quantitative rankings and qualitative analysis. The results highlight the interplay between structured engineering practices, novel methodologies, and risk management in high-stakes AI applications.

The analysis begins by exploring affiliation-based scoring trends (Figure 7), demonstrating how social group priorities shaped category rankings. Next, it delves into the top three teams' research papers, identifying best practices, innovation, and gaps in trust-building efforts.

Affiliation-Based Scoring Analysis

Figure 7:



Deviation was calculated based on the groups compared to the mean of all judges. Compared across all teams.

(Figure 7) Illustrates the average point difference from each team's mean across selected factors, categorized by judge affiliation.

- **Best Practices:** Industry judges rated this category the highest, while government judges rated it the lowest, suggesting that practical implementation was valued more by industry than government evaluators.
- **Novel Approaches:** Academia scored this category the highest, aligning with their preference for innovative methodologies, while government judges rated it below the mean, likely due to concerns over unproven techniques.

- **SE Activities:** Industry and academia rated this category positively, reinforcing the importance of structured engineering processes, while government judges scored it slightly below the mean, possibly due to different prioritization of risk-based frameworks.
- **Trust Infrastructure:** Industry and academia rated this category significantly above the mean, whereas government judges rated it well below, confirming that government evaluators prioritize explainability and oversight more critically than other groups.

These trends align with the statistically significant differences found in the Kruskal-Wallis test, reinforcing how different groups prioritize AI trustworthiness.

Comparison of Top Three Teams

Performance Breakdown

For each team, several excerpts were selected and coded based on their relevance to these categories. The following table (Table 1) summarizes how the top three teams performed in key aspects, emphasizing the distinction in their strategies. The full coded analysis is available in (Appendix B).

Table 1

Category	Team 1	Team 7	Team 2
Best Practices	Explainable AI, adaptable frameworks **Score: 5.57**	Multi-agent coordination, structured data processing **Score: 6**	Path-planning algorithms, redundancy mechanisms **Score: 5.42**
Novel Approaches	Hybrid RL + statistical models for decision-making **Score: 5.71**	Trust-dependent hierarchy, UAV-UGV cooperation **Score: 6**	Multi-arm bandit for AI-human review **Score: 5.28**
SE Activities	Modular design, structured reward function **Score: 5.71**	Extended V-Model, layered validation **Score: 6.27**	Large-scale simulation, iterative testing **Score: 5.14**

Table 1

Trust Infrastructure	Uncertainty quantification, real-time model interrogation **Score: 5.42**	AI confidence reporting, behavior auditing **Score: 4.21**	Interpretability dashboard, trust-weighted heuristics **Score: 4.85**

Emerging Trends and Patterns

Trust Evaluation

Of the three winning teams, Team 7 excelled in multiple categories, achieving a near-perfect score in Systems Engineering Activities. However, their overall ranking was impacted by a lower score in Trust Infrastructure, where reliance on automated auditing and hierarchical oversight resulted in lower scores from government evaluators. Team 1, the overall winner, scored the highest in Trust Infrastructure with an average of 5.42, incorporating uncertainty quantification models, trust calibration thresholds, and real time model interrogation to enhance AI reliability and transparency. This approach, which emphasizes statistical modeling to characterize uncertainty, was highly rated across all judging categories for its auditability and explainability. However, a single government judge assigned them a score of two, lowering their average from six. Government evaluators consistently favored explainability and human-in-the-loop oversight, making Team 1's approach particularly well-received. In contrast, Team 2, which employed a trust-weighted heuristic and AI-human negotiation model, was rated lower due to its less explicit focus on real-time trust calibration, further emphasizing the importance of transparency and risk mitigation in military AI adoption.

Emerging Themes and Analysis

SE Activities: The Most Universally Valued Category

Across all three top-performing teams, a strong emphasis on structured systems engineering methodologies was evident. Each team implemented architectures to support human-AI collaboration, along with iterative validation cycles to refine AI decision-making over time. These practices were

consistently recognized and rewarded by judges regardless of affiliation, suggesting a consensus that strong engineering design is foundational to building trust in AI systems.

Through the lens of SCOT, this alignment represents a shared interpretive closure around engineering discipline as a baseline requirement for trustworthiness. Regardless of whether evaluators prioritized innovation or oversight, the presence of structured systems engineering activities provided a common ground that satisfied differing expectations. This highlights the stabilizing role that established engineering norms play in facilitating the adoption of new technologies, particularly in high-stakes domains like military operations.

Disparities in Trust Infrastructure Priorities

While the SE Activities of the top performing teams garnered widespread support, significant divergence emerged in the evaluation of Trust Infrastructure. Government-affiliated judges placed the greatest emphasis on explainability, accountability, and human-in-the-loop oversight. Their preference for real time model interrogation, uncertainty quantification, and confidence calibration thresholds reflects a strong concern for risk mitigation and operational control, key themes within the government's role as a regulatory and safety focused stakeholder.

In contrast, industry and academic judges scored Trust Infrastructure more moderately, often prioritizing efficiency or algorithmic novelty over exhaustive oversight mechanisms. This difference illustrates SCOT's notion of interpretive flexibility: trust is not a fixed attribute but is socially constructed and shaped by the unique concerns of each group. For government evaluators, trust is grounded in transparency and verifiability; for others, it may be contingent on performance or theoretical soundness.

Innovation vs. Practicality: A Key Trade Off

A recurring tension observed across submissions was the trade off between innovation and practicality. Teams that pursued novel, complex AI approaches, such as hybrid reinforcement learning models or hierarchical trust based architectures, were well received by academic evaluators, who value cutting edge research and theoretical advancement. Team 1 and Team 7 represented this with their use of deep learning combined with explainability frameworks, which were praised for pushing the boundaries of trusted AI design.

However, these same innovations drew more cautious responses from government judges, who favored systems with clear operational feasibility and lower risk. Team 2, which relied on the well-established A* algorithm and a conservative trust weighted heuristic, was rated more favorably by

both government and industry, suggesting a preference for proven, stable methodologies that prioritize control and accountability over novelty.

Within the SCOT framework, this divergence underscores how stakeholder priorities shape technological interpretation. Academia constructs trustworthiness through innovation and exploration, while government frames it through regulatory alignment and risk aversion. These competing interpretations create friction but also guide the iterative refinement of trusted AI systems as developers respond to varied expectations.

Discussion

This research situates the evaluation of trusted AI in military applications within the SCOT framework, demonstrating that AI trustworthiness is not solely a technical concern but also a social negotiation. By analyzing how industry, academia, and government assess AI-enabled systems, this study highlights how social groups shape technological meaning through their unique priorities, efficiency and scalability for industry, methodological innovation for academia, and risk mitigation for government.

The findings of this study align with broader research on explainable AI (XAI) (Gunning & Aha, 2019), which emphasizes the need for transparent decision-making frameworks to enhance adoption in safety-critical environments. Similar studies on AI trustworthiness have found that stakeholders in high-risk applications prioritize clear interpretability over raw performance (Doshi-Velez & Kim, 2017) (Glikson & Woolley, 2020). This research reinforces those findings by demonstrating that government judges consistently rated AI systems lower in Trust Infrastructure unless clear explainability mechanisms were embedded. Unlike prior studies that primarily focus on user acceptance in commercial AI applications, this work extends the discussion to military AI evaluation, where stakes are significantly higher and oversight is a fundamental requirement for deployment.

The results also connect to literature on human-AI teaming (ZhangRui et al., 2021), where trust is often contingent on predictability, transparency, and control mechanisms. Prior research has shown that trust-building in human-AI collaboration depends on a balance between autonomy and human oversight (Hancock et al., 2024). The findings from this study confirm this, as government evaluators favored systems that explicitly retained human-in-the-loop frameworks, while industry judges were more comfortable with higher levels of automation. This adds to the existing body of knowledge by

demonstrating that trust evaluations in military AI applications are significantly influenced by evaluator affiliation and domain priorities.

Finally, this research contributes to multi-agent systems research (van der Hoek & Wooldridge, 2008), particularly regarding trust-weighted AI decision hierarchies like those employed by Team 7. The variability in how stakeholders rated novel AI approaches underscores a challenge in emerging AI technologies: the trade-off between innovation and operational feasibility.

From a SCOT perspective, the current state of trusted AI in military applications suggests that we are still in an *open interpretive phase*. While structured engineering practices show signs of interpretive closure, being consistently valued across all stakeholder groups, other dimensions, such as Trust Infrastructure and Novel Approaches, remain contested. The divergence in how different social groups assess transparency, explainability, and autonomy indicates that the field has not yet reached full stabilization. Competing interpretations persist, particularly around how much human oversight should be retained and how novel methods should be evaluated for safety and reliability. Until consensus emerges around these core issues, trusted AI in military contexts will continue to evolve through negotiation, iteration, and refinement.

Limitations and Caveats

While this study provides valuable insights into AI trust-building, several limitations must be acknowledged. First, the sample size was relatively small (seven teams and seven judges), limiting the generalizability of findings. A larger-scale study with a broader panel of evaluators could reveal more nuanced trends in how social groups prioritize AI attributes. Additionally, the competition setting may not fully replicate real-world AI deployment, where factors such as organizational culture, regulatory constraints, and long-term performance evaluation also shape trust.

Another limitation is this study focused primarily on high-level evaluation categories, meaning that deeper sub-category analysis (breaking down Trust Infrastructure into explainability, redundancy, and failure tolerance) could provide even richer insights. As a final caveat, there was a limited number of judges, across categories, with industry having one representative, academia only two, and four from government backgrounds. With a greater sample and response rate, a better understanding can be determined

Future Improvements and Next Steps

If conducting this research again, several modifications would be beneficial. First, incorporating direct end-user feedback (from soldiers, AI engineers, and command decision-makers) would provide a more holistic perspective on how AI trustworthiness is assessed beyond competition judges. Expanding the quantitative component by applying factor analysis or machine learning clustering on scoring patterns could also reveal hidden correlations between scoring categories and social group priorities.

Another key improvement would be testing AI decision frameworks in simulated or live operational environments. While this study analyzes perceptions of AI trust, actual AI trustworthiness should be measured through performance benchmarking, stress testing, and real-time explainability demonstrations. Additionally, more rigorous interviews or surveys with judges could clarify why certain scoring patterns emerged and whether specific AI design choices influenced their evaluations.

Conclusion

This research provides critical insights into how industry, academia, and government evaluate trusted AI in military applications, reinforcing that AI adoption is not just a technical challenge but a social negotiation. By applying the SCOT framework, this study demonstrates that different stakeholder groups prioritize distinct aspects of AI trustworthiness, with industry valuing practical implementation, academia favoring innovation, and government emphasizing risk mitigation and oversight. These findings have broader significance for AI development in high-risk, mission-critical environments, where ensuring stakeholder alignment is crucial for adoption.

The key takeaway is that AI trust-building must be approached holistically, integrating structured engineering methodologies, transparent decision-making, and trust calibration mechanisms to satisfy diverse stakeholder expectations. Future research should build on these findings by expanding the sample size of evaluators, incorporating direct operational testing, and examining how trust in AI evolves over time. Additionally, policymakers, engineers, and AI developers should collaborate to establish standardized frameworks that balance innovation with real-world feasibility, ensuring that AI solutions are both technically sophisticated and socially accepted.

Moving forward, researchers should explore how trust metrics can be operationalized to create AI systems that adapt dynamically to user confidence levels, bridging the gap between autonomy and human

oversight. By prioritizing transparency, adaptability, and risk-awareness, the AI community can advance deployable, trusted AI technologies that meet the demands of military, industry, and regulatory stakeholders alike.

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

Ahn, M. J., & Chen, Y.-C. (2022). Digital transformation toward AI-augmented public administration: The perception of government employees and the willingness to use AI in government. *Government Information Quarterly*, *39*(2), 101664. https://doi.org/10.1016/j.giq.2021.101664

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

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning* (No. arXiv:1702.08608). arXiv. https://doi.org/10.48550/arXiv.1702.08608

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

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), Article 2. https://doi.org/10.1609/aimag.v40i2.2850

Hancock, P., Billings, D. Schaefer, E., Chen, C., Parasuraman R. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. (2024). *ResearchGate*. https://doi.org/10.1177/0018720811417254

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

Peres, R. S., Jia, X., Lee, J., Sun, K., Colombo, A. W., & Barata, J. (2020). Industrial Artificial Intelligence in Industry 4.0—Systematic Review, Challenges and Outlook. *IEEE Access*, *8*, 220121–220139. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3042874

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

Uddin, M. M. (2024). Rejection or integration of AI in academia: Determining the best choice through the Opportunity Cost theoretical formula. *Discover Education*, *3*(1), 249. https://doi.org/10.1007/s44217-024-00349-7

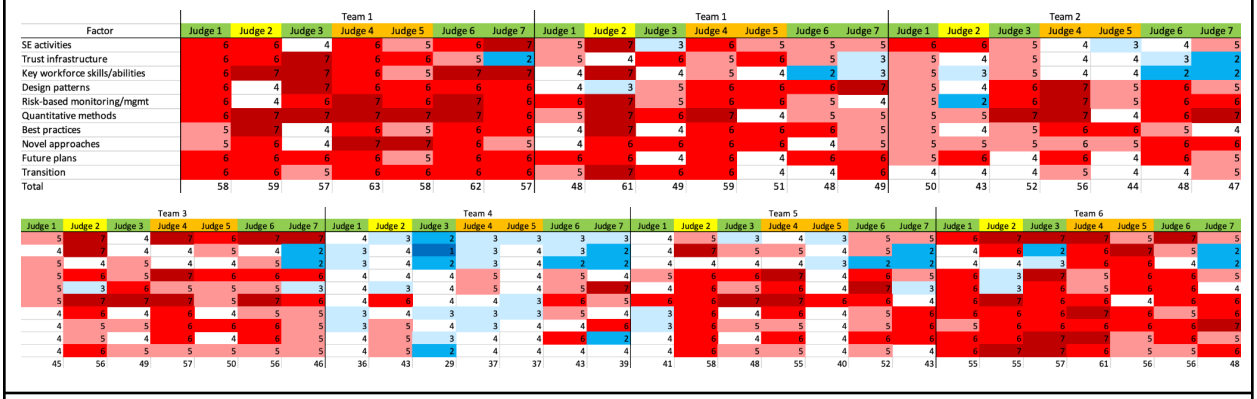
Umbrello, S., Torres, P., & De Bellis, A. F. (2020). The future of war: Could lethal autonomous weapons make conflict more ethical? *AI & SOCIETY*, *35*(1), 273–282. https://doi.org/10.1007/s00146-019-00879-x

van der Hoek, W., & Wooldridge, M. (2008). Chapter 24 Multi-Agent Systems. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Foundations of Artificial Intelligence* (Vol. 3, pp. 887–928). Elsevier. https://doi.org/10.1016/S1574-6526(07)03024-6

ZhangRui, J, M., FreemanGuo, & MusickGeoff. (2021). An Ideal Human. *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3432945

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

Appendix A (Team Scoring Evaluation):



Factor	Team 1 Judge 1	Team 1 Judge 2	Team 1 Judge 3	Team 1 Judge 4	Team 1 Judge 5	Team 1 Judge 6	Team 1 Judge 7	Team 1 Judge 1	Team 1 Judge 2	Team 1 Judge 3	Team 1 Judge 4	Team 1 Judge 5	Team 1 Judge 6	Team 1 Judge 7	Team 2 Judge 1	Team 2 Judge 2	Team 2 Judge 3	Team 2 Judge 4	Team 2 Judge 5	Team 2 Judge 6	Team 2 Judge 7
SE activities	6	6	4	6	5	6	7	5	7	3	6	5	5	5	6	6	5	4	3	4	5
Trust infrastructure	6	6	7	6	6	5	2	5	4	6	5	6	5	3	5	4	5	4	4	3	2
Key workforce skills/abilities	6	7	7	6	5	7	7	4	7	4	5	4	2	3	5	3	5	4	4	2	2
Design patterns	6	4	7	6	6	6	6	4	3	5	6	6	6	7	5	4	6	7	5	6	5
Risk-based monitoring/mgmt	6	4	6	7	6	7	6	6	7	5	6	6	5	4	5	2	6	7	5	6	6
Quantitative methods	6	7	7	7	7	7	6	5	7	6	7	4	5	5	5	5	7	7	4	6	7
Best practices	5	7	4	6	5	6	6	4	7	4	6	6	6	5	5	4	5	6	6	5	4
Novel approaches	5	6	4	7	7	6	5	4	6	6	6	6	4	5	5	5	5	6	5	6	6
Future plans	6	6	6	6	5	6	6	6	6	4	6	4	6	6	5	6	4	6	4	6	5
Transition	6	6	5	6	6	6	6	5	7	6	6	4	4	6	4	4	4	5	4	4	5
Total	58	59	57	63	58	62	57	48	61	49	59	51	48	49	50	43	52	56	44	48	47

Judge	Affiliation
Judge 1	Gov't
Judge 2	Industry
Judge 3	Gov't
Judge 4	Academia
Judge 5	Academia
Judge 6	Gov't
Judge 7	Gov't

Appendix B (Coded Research Analysis):

Team 1	Excerpt	Annotation
Best Practices	"Our approach adapts quickly to new scenarios, provides explainable statistical outputs and behavior, and effectively manages the variable accuracy of the two prediction subsystems."	The team emphasizes explainable AI and adaptability, which align with industry standards for AI transparency and robustness.
Average Score: 5.57	"We introduce trust into high-complexity AI systems by simplifying data inputs and providing a structured behavioral framework,	References best practices in AI modeling by reducing complexity and aligning AI with established human

Appendix B (Coded Research Analysis):		
	similar to human decision-making."	decision-making frameworks.
	"Our framework incorporates real-time model interrogation, allowing users to query AI predictions and receive explainable outputs."	The inclusion of real-time interrogation aligns with best practices in AI transparency, ensuring verifiability and reducing the black-box nature of deep learning models.
	"By integrating statistical methods with reinforcement learning, we create a system that enhances mine detection reliability while maintaining efficiency."	This hybrid approach follows best practices by combining statistical verification with machine learning, reducing the unpredictability of AI-based decision-making.
Novel Approaches	"Our approach uses a detailed minefield simulation to train explainable statistical models and a reinforcement learning (RL) algorithm that guides the UAV in selecting scan locations and methods."	The use of reinforcement learning for UAV path selection is positioned as a novel application in military mine detection.
Average Score: 5.71	"We rely on explainable statistical models in the form of linear regressors to infer and verify accuracy based on observable environmental metadata and prediction estimates."	This introduces a novel hybrid method of integrating statistical inference with reinforcement learning.
	The simulation environment we developed includes a configurable network with dynamically generated terrain, metadata-driven mine placement, and adaptive AI scanning behavior."	The team proposes an advanced simulation environment that integrates AI adaptability and metadata-driven risk assessment, which is novel in high-risk military contexts.
SE Activities		
Average Score: 5.71	We developed a modular simulation environment to train an intelligent RL agent using statistical models that estimate the accuracy of mine detection methods."	The structured lifecycle approach of simulation training aligns with SE methodologies.
	Our mission wrapper is designed in a modular way to extend on functionality within the base mission class, providing helper functions and streamlining certain processes."	References structured software engineering methodologies, ensuring modularity and extendibility.
	"We use reinforcement learning with a structured reward function that penalizes inefficient routing and unsafe traversal while prioritizing accuracy."	This structured reward function follows SE methodologies by ensuring systematic, quantifiable optimization criteria.

Appendix B (Coded Research Analysis):		
Trust Infrastructure	“We ensure system reliability by integrating uncertainty quantification models that provide confidence intervals for mine detection probabilities.”	This promotes trust by offering explicit uncertainty measures, which allow users to assess the reliability of AI-driven decisions.
Average Score: 5.42	“Trust calibration is built into the system through a reinforcement mechanism where AI predictions are flagged for human review when confidence levels fall below a defined threshold.”	This explicit confidence thresholding mechanism ensures that trust is dynamically adjusted based on AI reliability, reducing risks in high-stakes scenarios.
	“By maintaining human oversight, human control is retained and the AI is prevented from overriding human decisions.”	Ensuring human-in-the-loop oversight reinforces trust in AI autonomy.
	“Monitoring these metrics provides a quantification of trust. Our approach prioritizes robustness and reliability, which are critical in high-risk tasks like minefield traversal.”	Explicitly addresses quantifiable trust metrics, a key factor in AI-enabled military operations.
Team 7	Excerpt	Annotation
Best Practices	“The system implements a structured workflow for mine detection and clearance operations… The AI system processes data significantly faster, completing analyses in approximately one minute, though its accuracy varies based on environmental conditions.”	Emphasizes efficiency and structured data processing, which follows best practices in AI reliability.
Average Score: 6	“The mine-clearing operation employs a multi-agent architecture comprising human operators, AI, UAV, and UGV working in coordinated roles.”	This structured multi-agent approach aligns with established SE frameworks.
	“The system prioritizes explainability by implementing an interface that visually represents AI confidence scores alongside human annotations.”	This practice aligns with AI transparency guidelines, ensuring that human operators can visually assess AI predictions.
	“Our multi-agent system follows a structured command hierarchy where AI provides initial assessments, but human operators have final decision authority.”	This hybrid AI-human structure follows best practices in decision assurance, ensuring that AI complements rather than overrides human judgment.

Appendix B (Coded Research Analysis):		
Novel Approaches	“The AI’s decision-making is guided by a hierarchical model where higher trust in AI reduces human intervention, while lower trust increases manual validation.”	The dynamic trust-dependent decision hierarchy is an innovative way to balance efficiency and human oversight.
Average Score: 6	“Instead of a static traversal plan, we employ real-time environmental adaptation where AI adjusts UAV and UGV movement based on newly observed terrain conditions.”	This approach introduces dynamic adaptation, making AI decisions more context-aware and reducing the risks of outdated planning.
	“The optimization algorithm incorporates a bi-level planning approach to enable the independence of the UAV and UGV.”	Introduces a hierarchical approach to autonomous system coordination.
	“To configure the UGV for real-time path optimization under uncertainty, the D* Lite Algorithm will be employed.”	D* Lite is an advanced path-planning method adapted for real-time AI decision-making.
SE Activities	“Applying the Extended V-Model to Foster Trust... By treating human operators and their interactions with AI as integral elements of the system rather than peripheral concerns, the extended V-model encourages human-centered AI considerations.”	Clear evidence of SE framework integration, particularly for human-AI interaction modeling.
Average Score: 6.28	“The architecture follows a networked structure where information flows bidirectionally between human operators, AI, and unmanned vehicles, enabling real-time adaptation to changing conditions while maintaining operational coherence.”	Systems thinking approach ensuring flexibility in AI-human interactions.
	“We implement an iterative development cycle with regular validation checkpoints to ensure that AI behavior aligns with human expectations.”	This aligns with SE best practices, ensuring that system performance is continuously evaluated and improved.
	“Our architecture follows a layered approach, separating perception, decision-making, and execution layers to improve system maintainability.”	This structured design methodology is a common SE strategy that improves fault tolerance and system scalability.
Trust Infrastructure	“Operators receive automated reports on AI decision confidence, allowing them to review potential anomalies before execution.”	Automated reporting enhances transparency and gives human operators insight into AI reliability, reinforcing trust.
Average Score: 4.71	“We introduce an AI behavior auditing mechanism where historical decisions are logged and compared against human	This auditing feature enhances accountability and provides an additional layer of verification to ensure

Appendix B (Coded Research Analysis):		
	assessments for alignment.”	AI trustworthiness.
	“Trust becomes essential as operators weigh AI-driven rapid assessments against their own analyses.”	Focuses on the trade-off between AI speed and human accuracy as a trust-building mechanism.
	“The trust-weighted heuristic prioritizes paths where AI predictions align with human assessments.”	Directly incorporates trust calibration into AI decision-making.
Team 2	Excerpt	Annotation
Best Practices	“The UAV follows a path planned by the A* algorithm… If a mine is detected, the C2 will plan a new path, treating the node with a mine as an obstacle.”	References an industry-standard path-planning algorithm used in robotics.
Average Score: 5.42	“Human review can be slow and comes at the cost of time. However, this ensures high accuracy in high-risk decision-making.”	Balancing speed and accuracy aligns with best practices in AI-human teaming.
	“The decision system is built using a fail-safe redundancy model, ensuring that in the event of AI uncertainty, human operators assume control.”	Redundancy is a critical best practice in safety-critical AI applications, ensuring continued functionality even when AI confidence is low.
	“All system decisions are validated against real-world military traversal data, ensuring alignment with historical mission outcomes.”	Grounding AI decisions in empirical data ensures that recommendations align with real-world operational standards.
Novel Approaches	“Multi-arm bandit agents determine whether to use a human reviewer or AI for reviewing footage based on environmental factors.”	The use of multi-arm bandits for dynamic task allocation is a novel approach to human-AI interaction.
Average Score: 5.28	“We introduce a human-AI negotiation model where AI recommendations are presented with alternative paths, allowing humans to override decisions dynamically.”	The ability to negotiate AI decisions in real-time is a novel approach that blends automation with human control.
	“A reinforcement learning-based trust model continuously updates AI behavior based on human feedback, ensuring that AI learns from operator preferences.”	This feedback-driven trust adaptation mechanism introduces an interactive, evolving model of AI-human trust calibration.

Appendix B (Coded Research Analysis):		
SE Activities	“We simulate 10,000 environments with 1,000 episodes, testing human-AI collaboration strategies under different reward functions.”	Large-scale simulation with structured experimentation aligns with SE validation methodologies.
Average Score: 5.14	“The system is tested under multiple environmental conditions, including extreme weather simulations, to evaluate robustness.”	Stress testing across variable conditions follows SE validation methodologies, ensuring AI reliability under diverse operational scenarios.
	“We employ a structured testing framework that includes unit testing, integration testing, and field testing before deployment.”	This staged testing approach aligns with SE standards for verifying system reliability before real-world implementation.
Trust Infrastructure	“The trust-weighted heuristic prioritizes paths where AI predictions align with human assessments.”	The explicit weighting of AI-human agreement fosters trust.
Average Score: 4.85	“If priority is only accuracy, the human reviewer is overwhelmingly preferred. When we prioritize both cost and accuracy, AI is preferred in some cases, but human validation remains critical.”	Balances trust in AI based on situational trade-offs.
	“The AI system includes an interpretability dashboard that visually breaks down decision factors, allowing operators to verify logic before execution.”	Providing operators with visual explanations of AI decision-making fosters transparency and trust.
	“To mitigate concerns about AI bias, the system uses an ensemble decision-making approach where multiple AI models independently assess the same data.”	Using ensemble models reduces individual model biases and increases confidence in AI recommendations.