2021

Human-Machine Weapons Engagement Decisions: Systems Safety in Complex Decision Environments

Johnson, Bonnie W.; Miller, Scot A.; Green, John M.; Kendall, Walter A.; Godin, Arkady A.

Monterey, California: Naval Postgraduate School

NPS NRP Executive Summary
Human-Machine Weapons Engagement Decisions: Systems Safety in Complex Decision Environments
Period of Performance: 10/26/2020 – 10/23/2021
Report Date: 10/18/2020 | Project Number: NPS-21-N317-A
Naval Postgraduate School, Graduate School of Engineering and Applied Sciences (GSEAS)

HUMAN-MACHINE WEAPONS ENGAGEMENT DECISIONS: SYSTEMS SAFETY IN COMPLEX DECISION ENVIRONMENTS

EXECUTIVE SUMMARY

Principal Investigator (PI): Dr. Bonnie Johnson, Graduate School of Engineering and Applied Science (GSEAS), Systems Engineering (SE) and Mr. Scot Miller, Graduate School of Operational and Information Science (GSOIS), Information Sciences (IS)

Additional Researcher(s): Mr. John M. Green, Graduate School of Engineering and Applied Science (GSEAS), Systems Engineering (SE), Mr. Anthony Kendall, and Mr. Arkady Godin, Graduate School of Operational and Information Science (GSOIS), Information Sciences (IS)

Student Participation: Major Peh Ming Hui, Singapore Army, SE; Mr. Luis Cruz, SE; Ms. Ryane Pappa, SE; Ms. Savanna Shilt, SE; Ms. Angela Hoopes, SE; Major Samuel Wuornos, SE; USMC, Mr. J. Isaac Jones, SE; Mr. Russell Kress, SE; Mr. Adam Rahman, SE; Mr. William Newmeyer, SE; Mr. Wallace Fukumae, SE; Mr. Kheng Hun, SE; Mr. Robert French, SE; Mr. Obed Matuga, SE; and Ms. Caitlyn O'Shaugnessy, SE

Prepared for:
Topic Sponsor Lead Organization: N9 - Warfare Systems
Topic Sponsor Organization(s): Naval Air Warfare Center/Weapons Division
Topic Sponsor Name(s): Mr. Bruce Nagy
Topic Sponsor Contact Information: bruce.nagy@navy.mil, 703-939-1381

Project Summary

Advances in computational technologies and artificial intelligence (AI) methods present new opportunities for developing automated tactical decision aids to support warfighters making weapons engagement decisions. Tactical decisions become increasingly complex and can overwhelm human decision-making as threats increase in number, speed, diversity, and lethality. The deployment of such AI-enabled decision aids must consider system safety. This study explored the potential safety risks and failure modes that may arise as automation and AI technologies are introduced and implemented to support human-machine weapons engagement decisions. The study identified and evaluated safety risks and failure modes, root causes, and mitigation and engineering strategies to prevent, address, and recover from this new class of possible safety failures.

The study included five research initiatives. Dr. Bonnie Johnson conducted the first initiative to study the problem holistically and develop a taxonomy of safety risks, possible root causes, and mitigation strategies related to the introduction of AI into tactical decision aids. Naval Postgraduate School (NPS) students carried out the other four initiatives: (1) analysis of the decision risk involved in human-machine teaming for making weapons engagement decisions, (2) analysis of human-machine trust and weapons engagement decisions, (3) evaluation of the safety risks in implementing automated decision aids for air and missile defense, and (4) a study of data gathering and management, a critical enabler for AI safety, to support the Navy's development of AI systems.

This study found that developing and implementing AI systems for military applications, especially for applications involving weapon engagement decisions, introduces new and potentially dangerous safety risks that must be taken seriously. The root causes of these failure modes may be difficult to detect, predict, and prevent. The study recommends continued research the five focus areas and into systems engineering methods needed to ensure AI systems are designed, developed, implemented, and operated safely.

Keywords: *safety, artificial intelligence, AI, machine learning, ML, human-machine teaming, automated decision aids, weapons engagements, mission planning, metacognition, trust, risk, missile defense, failures, root causes, systems engineering, tactical warfare, battle management aids*

Background

The development of AI capabilities has the potential to transform the traditional battlespace. AI-enabled applications, such as automated decision aids for tactical missions (Johnson, 2019) and predictive analytics and game theory for mission planning (Johnson, 2020; Zhao & Nagy, 2020), offer huge gains in naval decision effectiveness and tactical superiority. The speed of warfare today often exceeds the cognitive abilities of humans to make decisions (Galdorisi, 2019). The Navy has acknowledged the need for AI and machine learning (ML) to support warfighters. Naval warfighters need real-time decision aids to support mission planning and battle decision aids.

There is extensive research into data analytics, data fusion, AI, and ML; and automation taxonomies exist (Save, Feuerberg, & Avia, 2012). In parallel with the development of AI methods, studies are being conducted to understand the risks associated with AI capabilities and develop methods to ensure safety. Broad studies of the safety implications of AI systems are developing theories for achieving safe, intelligent systems (Kose & Vasant, 2017). Studies are developing taxonomies for the various pathways to dangerous AI (Yampolskiy, 2016). One concern is that adversaries may insert carefully crafted training sets with false information into ML models causing the ML systems to learn incorrectly (Chen et al., 2018). Some research is taking an opposing approach by taking the adversarial perspective and starting with the objective of how to create a malevolent AI system (Pistono & Yampolskiy, 2016). They hope to gain a deeper understanding of AI safety through this counter approach.

This study explored AI safety with a focus on the future application of AI methods for military applications. The study identified AI safety risks and developed high-level risk mitigation strategies. The study developed a framework for analyzing and engineering safety aspects of future AI systems for tactical decision aids and mission planning aids.

This study applied a systems analysis approach to understand the problem space and to develop engineered solution concepts. The study collected data and information through a literature review, participation in virtual conferences and workshops, and through discussions with subject matter experts. The study developed safety requirements for integrating automated decision aids into weapons engagement and mission planning decisions. The study explored the cognitive strengths of humans and machines to identify effective teaming arrangements in a variety of tactical and mission planning environments of increasing complexity. The study developed a set of complex threat scenarios to understand and evaluate human-machine weapons engagement teaming strategies. The study developed solution strategies throughout the systems engineering lifecycle of AI systems that need to be implemented to prevent, predict, mitigate, and recover from safety failures. The study involved a research team of faculty members at NPS and systems engineering student researchers. One thesis student and three capstone student teams contributed to the project.

Findings and Conclusions

The primary outcome of this study is the recognition that developing and implementing AI systems for military applications, especially for applications involving weapon engagement decisions, introduces new and potentially dangerous types of safety risks that must be taken seriously. The root causes of these failure modes may be difficult to detect, difficult to predict, and difficult to prevent.

This study discovered three fundamental reasons that the implementation of AI systems in the tactical military domain will lead to serious safety concerns.

The first is the nature of AI systems: they are non-deterministic, complex, and adaptive. AI systems learn while being trained and adapt to their operational environments as they receive data. AI systems often lead to emergent behavior and behavior that can be unexpected and unintended. As AI systems function and adapt to complex situations, they often contribute to the complexity. Future AI systems may even continue to learn in situ. For military applications involving weapon engagement decisions, it is critical to ensure that AI systems produce safe results.

A second reason is the role of data in the development and operation of AI systems. Developing AI systems requires a major effort in data gathering and management. Data must be representative of the operational scenario. Data is required to train, evaluate, validate, and operate AI systems. Data must be securely protected and evaluated, so it is free of bias and corruption.

The third factor is human-machine teaming. Human-machine teaming is a critical aspect of implementing AI systems effectively and safely. Appropriate trust must be established between human operators and AI systems. The appropriate level of automation (how automated or manual each decision needs to be) depends on how much decision risk is acceptable, and this depends on the complexity of the threat situation. For tactical decisions involving the use of weapons, the level of automation needs to adapt to the situation—if time is available, human operators will have time to weigh options; however, if time is very short, it may be necessary for engagement decisions to be made in a more automated mode to provide an effective defense.

These causal factors were studied by the NPS researchers and student teams using literature review, systems analysis, risk analysis, and through the study of operational use case analysis. The researchers identified a variety of tactical scenarios in which a future AI system could support situational awareness and tactical decision-making. The analyses revealed areas of safety failure modes, possible consequences, and potential root causes. The research led to the identification of system solution concepts for preventing, mitigating, or recovering from AI system failures. There are four categories of safety mitigation solution strategies that involve engineered capabilities as well as activities that must be performed during operations. Thus, the entire systems engineering lifecycle is affected. The research also led to the identification of systems engineering and program management practices that need to be implemented to support these AI system safety solution concepts.

Recommendations for Further Research

This study recommends continued research in five focus areas: artificial intelligence (AI) system safety solution concepts, AI system applications for the weapon engagement kill chain, human-machine teaming for this application, risk management, and data management.

The first focus area that requires continued research is AI system safety solution concepts. This study identified four categories of solution strategies that span the systems engineering lifecycle: inherently safe

designs, building in safety reserves, developing mechanisms to allow for safe fails, and implementing procedural safeguards during operations. Each of these types of safety solution strategies needs to be carefully studied for any application that will be enabled by AI system capabilities. Additionally, these solution strategies must be tailored to the specific application domain.

The second focus area for future research is on the design and development of specific AI methods for implementation in future tactical kill chains. Initial findings from this study show that different AI methods will be needed for the different functions in the kill chain, and a more complex mapping will be required rather than a simple one-to-one mapping. It is likely that a federated learning approach will be required that orchestrates a heterogeneous set of machine learning algorithms and AI methods that can handle the highly complex spatial-temporal dynamics of a tactical battle scenario. A significant level of research is required to identify and evaluate different and novel AI algorithms and methods for this application.

The third future research area is human-machine teaming. This study focused on the trust relationship between human operators and future AI-enabled tactical decision aids and risk levels associated with different levels of automation. Similar research is required to study many other aspects of human-machine teaming including: explainability, useability, human factors, human-machine interdependency, cognitive loading, and adaptive and agile levels of autonomy.

A fourth area of future research is risk management. This study identified a set of operational use cases representing the use of future AI-enabled decision aids for air and missile defense missions. The study conducted a risk analysis to identify and analyze potential failures, consequences and root causes. Additional risk analysis is required for these threat scenarios as well as for many other mission domains. Risk analysis needs to be performed continuously during the systems engineering lifecycle of the design, development, and implementation of future AI systems for military applications—especially for weapon engagement decisions.

A fifth area that requires future research is data management in support of AI development. This study revealed the importance of data management for effective AI system design, training, evaluation, and operations. Additionally, the study found that acquiring, curating, formatting, validating, and using data in support of AI system development is a major undertaking and systems engineering task in its own right. It is crucial that the Navy recognizes the importance of appropriate data acquisition and management to support AI development and ensures that funding, acquisition, and program management supports it—for weapons engagement applications and other mission domains.

References

Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I., & Sruivastava, B. (2018). Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv: 1811.03728*.

Galdorisi, G. (2019, May). The Navy needs AI, it's just not certain why. In *USNI Proceedings* (Vol. 145, No. 5, pp. 1-395).

Johnson, B. (2019). Artificial intelligence—an enabler of naval tactical decision superiority. *AI Magazine*, *40*(1), 63-78.

Johnson, B. (2020). Predictive analytics in the naval maritime domain. In *Proceedings of the AAAI Symposium on the 2nd Workshop on Deep Models and Artificial Intelligence for Defense Applications: Potentials, Theories, Practices, Tools, and Risks.*

Kose, U., & Vasant, P. (2017, September). Fading intelligence theory: a theory on keeping artificial intelligence safety for the future. In 2017 *International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1-5). IEEE. doi: 10.1109/IDAP.2017.8090235.

Pistono, F., & Yampolskiy, R. (2016). Unethical research: how to create a malevolent artificial intelligence. *25th International Joint Conference on Artificial Intelligence (IJCAI-16)*. https://arxiv.org/ftp/arxiv/papers/1605/1605.02817.pdf

Save, L., Feuerberg, B., & Avia, E. (2012). Designing human-automation interaction: a new level of automation taxonomy. *Proc. Human Factors of Systems and Technology*, *2012*.

Yampolskiy, R. V. (2013). Artificial intelligence safety engineering: why machine ethics is a wrong approach. In V. Müller (Ed.), *Philosophy and theory of artificial intelligence* (Vol. 5, pp. 389-396). Springer. https://doi.org/10.1007/978-3-642-31674-6_29.

Zhao, Y., & Nagy, B. (2020, May). Modeling a multi-segment war game leveraging machine intelligence with EVE structures. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II* (Vol. 11413, p. 114131V). International Society for Optics and Photonics.

Acronyms

AI artificial intelligence
ML machine learning
NPS Naval Postgraduate School