2021

Weapon Systems Safety When Deploying AI Technology

Berzins, Valdis A.

Monterey, California: Naval Postgraduate School

MONTEREY, CALIFORNIA

WEAPONS SYSTEMS SAFETY WHEN DEPLOYING AI TECHNOLOGY

EXECUTIVE SUMMARY

Principal Investigator (PI): Prof. Valdis Berzins, Graduate School of Operational and Information Sciences (GSOIS), Computer Science Department (CS)

Additional Researcher(s): Asst. Prof. Armon Barton, Graduate School of Operational and Information Sciences (GSOIS), Computer Science Department (CS); Asst. Prof. Joshua Kroll, Graduate School of Operational and Information Sciences (GSOIS), Computer Science Department (CS)

Student Participation: LT Sabrina Atchley, USN, Computer Science; LCDR Brandon Beckler, USN, Computer Science; Capt Sean Gilroy, USMC, Computer Science; CDR Edgar Jatho, USN, Computer Science; LT Cardavian Lowery, USN, Cyber Systems and Operations; Capt Benjamin Marsh, USMC, Computer Science; Capt Shane Wescott, USMC, Computer Science; Mr. Eugene Williams, Computer Science

Prepared for:
Topic Sponsor Lead Organization: ASN(RDA) - Research, Development, and Acquisition
Topic Sponsor Organization(s): Naval Air Warfare Center Weapons Division (NAWCWD)
Topic Sponsor Name(s): Senior Scientist and Systems Engineer, Mr. Bruce Nagy
Topic Sponsor Contact Information: Bruce.Nagy@navy.mil, 760-608-7285

This research is supported by funding from the
Naval Postgraduate School, Naval Research Program (PE 0605853N/2098).

Project Summary
We investigated how artificial intelligence (AI) technology that impacts system safety concerns should be developed. Guidance is needed for understanding the likelihood of unwanted events and making risk decisions on naval employment of systems that incorporate such technology. The project identified hazards associated with AI systems, analyzed differences between hazard analysis of such systems and those for explicit programmed systems, and explored possible mitigations for those hazards. We found that the current state of knowledge in the area does not yet provide effective mitigations for several of these hazards.

Keywords: *system safety, ordnance, artificial intelligence, AI, Naval Ordnance Safety and Security Activity, NOSSA, development processes, safety analysis*

Background
Safety concerns arise when deploying AI technology in weapon systems, autonomous systems carrying weapons/sensors/comms, and mission planning and decision aids. When AI technology is deployed in a weapon system, robot, or planning system, there is a likelihood that unwanted events occur. Naval Ordnance Safety and Security Activity (NOSSA) is responsible for understanding that likelihood and making risk decisions on naval employment.

Existing standards and guidance for assessing software system safety are geared towards explicitly programmed systems. These standards need to be augmented and extended to address systems that use machine learning to teach a computer to learn concepts using data without being explicitly designed and programmed.

Safety aspects of AI are complex and involve a wide range of concerns beyond the correctness of the algorithms used in deployed systems and in the tools used to train the models embedded in the AI. These models support decisions and recommendations made by AI subsystems, and the quality of the decisions and recommendations depends on how well those models match the real world, not just how well they match a finite set of available data. Relevant concerns include adequacy of the training data; validity of the training objectives; understanding implicit requirements derived from data; situational awareness for people teaming with the AI components; procedures for recognizing and recovering from failures; how AI system behavior interacts with doctrines, procedures and constraints of the larger systems and organizations in which the AI applications are embedded; and how well can AI behavior adapt when those doctrines, procedures, and constraints change.

Safety of AI systems is related to open fundamental issues in risk analysis. Almost all previous work on risk analysis defines risk as a function of (1) the probability of occurrence of an unwanted event, and (2) the severity of the consequences should that event occur. The assumption that the probability of occurrence does not change with time is implicit in this definition and current risk analysis practices. This assumption does not match the context of Navy operations. For example, the probability of hostile action

by adversaries is expected to be very different during active conflict and during peace time. We did not find formulations of risk analysis that account for this possibility.

The study was conducted by doing a literature search, analyzing previous publications in a research seminar with student and faculty participants, focused on previous work on identifying AI hazards and potential mitigations. Students explored case studies focused on effectiveness of possible mitigations. A PhD dissertation was initiated to further study problems related to this project.

Findings and Conclusions

We found that a major cause of accidents involving systems with embedded automatic control systems such as autopilots for airplanes and cars was lack of situational awareness when the systems recognized conditions they could not handle and handed control back to a human "safety driver" (Kroll & Berzins, 2021). We did not expect this finding, but after it was brought to our attention, students found Navy examples and identified a further implication: the skills and training of safety drivers tend to get rusty because they no longer get much practice in piloting their systems manually. Resulting recommendations included requiring safety drivers to do emergency recovery training while they were supervising autopilots operating under routine conditions, to focus that training on responding to a variety of failure conditions, and to add additional system design requirements for autopilots to provide "early warning" alerts when unusual conditions are detected. The early warnings could be used to cue the safety pilots to focus on the situation before it becomes a crisis and to bring the most experienced pilot available to the controls.

We believe this is a concern for the Navy because AI technology will not reach the state where it is sufficiently trustworthy for full autonomous control in the foreseeable future, at least for safety-critical applications such as weapons systems (Kroll & Berzins, 2022). This implies that man-machine teaming will be the expected mode of operation, and that these systems will have human "safety drivers" who will need mitigations for the hazard discussed above, such as additional training and additional design requirements for early warning of control handover.

Analysis indicates no defense in the literature currently stands as a viable answer to the problem of adversarial examples (Jatho et al., 2021). Adversarial examples are attacker-crafted inputs that reliably produce different outputs than developers of neural networks intended. Such networks take data as input, typically from sensors, and produce classifications as outputs, such as what kind of objects are present in an image. Adversarial examples is an active research area in AI. Fifty unique approaches to generating adversarial examples are represented in the Adversarial Robustness Toolbox alone (Nicolae et al., 2019), and new approaches generated regularly. A recent paper states that "The field has not advanced to the point where we know how to build solid defenses even against the attacks that we know already exist" (Barton et al., 2021). The project studied new approaches to improve these defenses and developed an approach that improves performance over all previously known methods (Barton et al., 2021).

Our overall assessment is that the state of the art in safety assessment of systems with AI components is not sufficient for Navy needs and that further development in this area is needed. Recommendations are provided below.

Recommendations for Further Research

The project identified hazards associated with AI systems, analyzed differences between hazard analysis of such systems and those for explicit programmed systems, and explored possible mitigations for those hazards. We found that the current state of knowledge in the area does not yet provide effective mitigations for several of these hazards. Further research is recommended to answer the questions below:

- How can AI systems provide early warnings when they encounter unusual conditions that are not well covered by their design or training data?
- How can the time-dependent nature of probability of hazard occurrence be incorporated into risk analysis?
- Can a game-theoretic approach to risk analysis be developed, and how would it apply to safety assessment with respect to contested environments?
- How should human safety operators responsible for overseeing AI systems be trained?
- How should doctrines, techniques, procedures and concepts of operations be reviewed and adjusted to ensure they do not pose safety hazards when interacting with AI systems?
- How to validate whether training data adequately covers possible real-world events that could impact operational safety?
- How to ensure that biases implicit in the way data sets are constructed do not degrade quality of decisions from AI components built using the data?
- How to assess risk exposure due to possible attacks on AI systems that use adversarial examples?
- How can risks due to adversarial examples be mitigated?
- How can risks due to faults in the optimization criteria used to derive AI decision rules from data be detected and mitigated?

References

Barton, A., Berzins, V., Jatho, E. (2021). Defending Against Adversarial Examples in Deep Neural Network Classifiers. NPS (Report No. NPS-CS-21-002). Naval Postgraduate School.

Kroll, J, &. Berzins, V. (2022). Understanding, Assessing, and Mitigating Safety Risks in Artificial Intelligence Systems. (Report No. NPS-CS-22-001). Naval Postgraduate School.

Jatho, E., Barton, A., Berzins, V. (Dec. 7-9, 2021). Defending deep neural networks with precise latent space declaration [Paper presentation]. MORS Emerging Techniques Forum, Arlington VA.

Kroll, J., & Berzins, V. (Dec. 7-9, 2021). Understanding and assessing safety in artificial intelligence systems [Paper presentation]. MORS Emerging Techniques Forum, Arlington, VA.

Nicolae, M., Sinn, M., Tran, M., Buesser, B., Rawat, A., Wistuba, M. … Edwards, B. (2019). Adversarial Robustness Toolbox (v1.0.0). arXiv: 1807.01069 [cs.LG].

Acronyms

AI	artificial intelligence
NOSSA	Naval Ordnance Safety and Security Activity