DEPARTMENT OF THE AIR FORCE
HEADQUARTERS AIR FORCE
WASHINGTON DC

Understanding and Avoiding Unintended Behaviors in Autonomous Systems

Abstract

The U.S. Air Force (USAF) Scientific Advisory Board (SAB) study on Understanding and Avoiding Unintended Behaviors in Autonomous Systems addresses potential approaches to avoiding, reducing, and mitigating unintended behaviors across the entire autonomy lifecycles from system design through deployment. The study includes theoretical frameworks, as well as approaches to software architecture and design, and approaches that are employed at run-time.

The Study Panel examined past cases of unintended behaviors in autonomous systems across the defense organizations and the commercial sector. Additionally, the study reviewed theoretical frameworks that allow probabilistic or deterministic assessments of near-infinite state software systems and evaluated software architectures and design approaches that support verification of autonomous systems. The Panel assessed approaches for detecting, alerting, and avoiding unintended behaviors during operational run-time. Finally, the study provided a roadmap for near, mid and far-term science and technology efforts to guide USAF research and investments as well as organizational requirements for development and fielding. The Panel used evidence gathered from multiple briefings from sources across the Department of Defense, private industry and academia to determine several recommendations. The SAB recommends the USAF:

- Advance new processes for agile spiral development of autonomy software while minimizing unintended behaviors, including continuous development, integration, testing, evaluation, verification, and validation. The process should allow rapid updating of autonomous systems during operations, including adjustments based on new training data to eliminate unintended behaviors.

- Develop the infrastructure and processes required to enable robust run-time assurance. Continual monitoring of safety-critical functionality and development of new assurance methods for updating software on-line are key to avoiding dangerous unintended behaviors.

- Develop a strategy that ensures future autonomous systems will be continually tested via data-driven simulation. This strategy should include the ability to continually plan and implement adversary simulations, automatic processes to collect and leverage limited experimental data, and the use of Advanced Framework for Simulation, Integration, and Modeling to simulate autonomy under attack.

- Co-design, co-validate and co-train future human-autonomy teams to enable resilient performance. Approaches should be developed to define system-level requirements for human-autonomy teams, and to enable mutual understanding of both explicit and implicit intent via communication and training.

- Develop processes and research for scaling resilient autonomy in teams of many humans and autonomous systems. Extending system-level safety and mission assurance to heterogeneous teams as well as developing new approaches for agile autonomy-at-the-edge software will ensure safe operational baseline and high-performance for future missions at scale.

SAF PA Approval: 2021-0043