DARPA Assured Autonomy Seeks to guarantee Safety of Autonomous Systems that leverage capabilities of machine learning

DARPA has launched a new research program called Assured Autonomy that aims to advance the ways computing systems can learn and evolve to better manage variations in the environment and enhance the predictability of autonomous systems like driverless vehicles and unmanned aerial vehicles (UAVs).

“Tremendous advances have been made in the last decade in constructing autonomy systems, as evidenced by the proliferation of a variety of unmanned vehicles. These advances have been driven by innovations in several areas, including sensing and actuation, computing, control theory, design methods, and modeling and simulation,” said Sandeep Neema, program manager at DARPA. “In spite of these advances, deployment and broader adoption of such systems in safety-critical DoD applications remains challenging and controversial.”

Autonomy refers to a system’s ability to accomplish goals independently, or with minimal supervision from human operators in environments that are complex and unpredictable. Autonomy delivers significant military value, including opportunities to reduce the number of warfighters in harm’s way, increase the quality and speed of decisions in time-

critical operations, and enable new missions that would otherwise be impossible

Autonomous systems are increasingly critical to several current and future Department of Defense (DoD) mission needs. For example, the U.S. Army Robotics and Autonomous Systems (RAS) strategy report for 2015-2040 identifies a range of capability objectives, including enhanced situational awareness, cognitive workload reduction, force protection, cyber defense, logistics, etc, that rely on autonomous systems and higher levels of autonomy.

Several factors impede the deployment and adoption of autonomous systems:

In the absence of an adequately high level of autonomy that can be relied upon, substantial operator involvement is required, which not only severely limits operational gains, but creates significant new challenges in the areas of human-machine interaction and mixed initiative control.

Achieving higher levels of autonomy in uncertain, unstructured, and dynamic environments, on the other hand, increasingly involves data-driven machine learning techniques with many open systems science and systems engineering challenges.

Machine learning techniques widely used today are inherently unpredictable and lack the necessary mathematical framework to provide guarantees on correctness, while DoD applications that depend on safe and correct operation for mission success require predictable behavior and strong assurance.

“Historically, assurance has been approached through design processes following rigorous safety standards in development, and demonstrated compliance through system testing,” said Neema. “However, these standards have been developed primarily for human-in-the-loop systems, and don’t extend to learning-enabled systems with advanced levels of autonomy where system behavior depends on its memory of received stimuli. The assurance approaches today are predicated on the assumption that the systems, once deployed, do not learn and evolve.”

One approach to assurance of autonomous systems that has recently garnered attention, particularly in the context of self-driving vehicles, is based on the idea of “equivalent levels of safety,” i.e., the autonomous system must be at least as safe as a comparable human-in-the-loop system that it replaces. The approach compares known rates of safety incidents of manned systems—number of accidents per thousands of miles driven—and conducting physical trials to determine the corresponding incident rate for autonomous systems. Studies and analyses indicate, however, that assuring safety of autonomous systems in this manner alone is prohibitive, requiring millions of physical trials, perhaps spanning decades. Simulation techniques have been advanced to reduce the needed number of physical trials, but offer very little confidence, particularly with respect to low-probability, high-consequence events.

Trustworthiness and Trust in Autonomous Systems

Most commercial applications of autonomous systems are designed for operation in largely benign environments,

performing well-understood, safe, and repetitive tasks, such as routing packages in a fulfillment center warehouse. Design for commercial systems rarely considers the possibility of high-regret outcomes in complex, unpredictable, and contested environments, says Report of the Defense Science Board Summer Study on Autonomy.

In military operations, these can include an adversary whose goal is to neutralize the use and effectiveness of such systems, either through deception, direct force, or increased potential for collateral damage or fratricide. Although commercial applications are gradually expanding beyond these controlled environments, e.g., self-driving cars, delivery drones, and medical advisory systems, fielded autonomous systems do not yet face a motivated adversary attempting to defeat normal operations. Trust is complex and multidimensional.

The individual making the decision to deploy a system on a given mission must trust the system; the same is true for all stakeholders that affect many other decision processes. Establishing trustworthiness of the system at design time and providing adequate indicator capabilities so that inevitable context-based variations in operational trustworthiness can be assessed and dealt with at run-time is essential, not only for the operator and the Commander, but also for designers, testers, policy and lawmakers, and the American public.

Establishing trust is difficult in Systems that learn. Machines are being developed with experience that change their capabilities and limitations and adapt to their use and environment. Such systems will outgrow their initial verification and validation and will require more

dynamic methods to perform effectively throughout their lifecycle.

Assured Autonomy program

The goal of the Assured Autonomy program is to create technology for continual assurance of Learning-Enabled, Cyber Physical Systems (LE-CPSs). Continual assurance is defined as an assurance of the safety and functional correctness of the system provided provisionally at design time, and continually monitored, updated, and evaluated at operation-time as the system and its environment evolves.

An LE-CPS is defined as a system composed of one or more Learning-enabled Components (LECs). A LEC is a component whose behavior is driven by “background knowledge” acquired and updated through a “learning process,” while operating in a dynamic and unstructured environment. This definition generalizes and admits a variety of popular machine learning approaches and algorithms (e.g., supervisory learning for training classifiers, reinforcement learning for developing control policies, algorithms for learning system dynamics). The generalization is intentional to promote abstractions and tools that can be applied to different types and applications of data-driven machine learning algorithms in Cyber Physical Systems (CPSs) to enhance their autonomy.

In order to ground the Assured Autonomy research objectives, the program will prioritize challenge problems in the militarily relevant autonomous vehicle space. However, it is anticipated that the tools, toolchains, and algorithms created will be relevant to other LE-CPSs. The resulting technology from the program will be in the form of a set of publicly

available tools integrated into LE-CPS design toolchains that will be made widely available for use in commercial and defense sectors.

In contrast to prescriptive, process-oriented standards for safety and assurance, a goal-oriented approach, such as the one espoused by Neema, is arguably more suitable for systems that learn, evolve, and encounter operational variations. In the course of Assured Autonomy program, researchers will aim to develop tools that provide foundational evidence that a system can satisfy explicitly stated functional and safety goals, resulting in a measure of assurance that can also evolve with the system.

Paul Brubaker, the CEO and president of Alliance for Transportation Innovation, said the formation of this DARPA program was good news. “This activity will go a long way to assuaging public concern regarding the safety of autonomous systems and self-driving vehicles,” Brubaker said in an email. “Kudos to DARPA for launching this project – it’s the long pole in the tent to increasing the comfort to self-driving and turning the possibilities into reality.”

Program objectives

DARPA Goal – Develop rigorous design and analysis technologies for continual assurance of learning-enabled autonomous systems, in order to guarantee safety properties in adversarial environment

• Increase scalability of design-time assurance
• What is the baseline capability of the proposed methods, in terms of the hybrid state-space and number and complexity of

learning-enabled components
• How do you plan to scale up by an order of magnitude?
• How will you characterize the tradeoffs between fidelity of your modeling abstractions and scalability of the verification approach.
• Reduce overhead of operation-time assurance
• What is the baseline overhead of the operation-time assurance monitoring techniques?
• How do you plan to minimize it to be below 10% of the nominal system resource utilization?
• Scale up dynamic assurance
• What is the size and scale of dynamic assurance case that can be developed and dynamically evaluated with your tools?
• Reduce trials to assurance
• How will your approach quantifiably reduce the need for statistical testing?

The program has 3 phases and all three phases should complete within 4 years.

References and Resources also include:

https://www.darpa.mil/program/assured-autonomy

https://www.darpa.mil/news-events/2017-08-16

https://www.nextbigfuture.com/2017/08/darpa-seeks-scalable-and-dynamic-testing-for-learning-robots.html

https://fas.org/irp/agency/dod/dsb/autonomy-ss.pdf