# The Johns Hopkins Institute for Assured Autonomy: Enabling a Future of Trust for Autonomous Systems

*Cara E. LaPointe, Anton T. Dahbura, David P. Silberberg, and Amber R. Mills*

## ABSTRACT

*Autonomous systems are becoming increasingly integrated into all aspects of our lives. To work toward ensuring these systems are safe, secure, and reliable and operate as designed, the Johns Hopkins University established the Johns Hopkins Institute for Assured Autonomy (IAA), run jointly by its Applied Physics Laboratory (APL) and the Whiting School of Engineering. The IAA takes a holistic approach to assuring autonomous systems by working across three pillars: increasing reliability of the technology, improving interactions within the integrated ecosystem, and engendering trust through policy and governance. This article discusses the need for the IAA, its goals and approach, and some of its initial research efforts.*

The Autonomous Future—"All the world is made of faith, and trust, and pixie dust."

J. M. Barrie, *Peter Pan*

It's an early spring day in 2035 and police officer Jake Lawton is awakened by his home service robot, Ratchet, announcing that it is 1 hour before he must depart and report for duty in Baltimore. Down the hall, Jake hears his kids shouting excitedly. He asks the autonomous personal assistant (APA) why they are shouting and is told that a commercial delivery drone (Figure 1) came overnight with items his family may want to purchase based on their shopping history and online personas. With the World Cup around the corner, it is likely this delivery will include a newly released "smart" soccer ball with sensors that transmit speed, rotation, and impact data to their 4K OLED gaming armbands. Ratchet tells the children to get ready for school and opens a video connection so Jake, still brushing his teeth, can reinforce Ratchet's instruction against wearing shorts to school today.

**Figure 1.** A delivery drone. Amazon announced plans to use drones to deliver packages more than 5 years ago. What seemed like a fantasy then is a reality now. (Bigstock image.)

Once ready, Jake comes downstairs to have breakfast with his children; before sitting down he quickly checks his front door drone-landing pad. Sure enough, there is a package. Selecting a biometric button on the side, Jake allows a multimodal facial/retinal scan and the package opens. After grabbing the new soccer ball and some other items, Jake closes the lid, and selects "Shopping Complete" on his device. He laughs to himself, as he knows the cat will go crazy when the drone comes midday to collect the box.

**Figure 2.** Smart cars operating autonomously with the help of a radar signal system and wireless communication. (Bigstock image.)

Sitting down at the table, Jake pours a cup of coffee and takes his smart pills. These pills simultaneously manage his blood pressure and transmit health status and drug regimen to a mobile device, which in turn updates his health plan and medical records. The APA compliments Jake on his recent weight loss and new commitment to exercise, which has lowered his cholesterol. With his blood pressure under control, Jake's medication can be reduced. Through an interface device, he then calls up the Intelligent Tutoring System to give the family an update on the kids' learning progress. Ratchet interrupts the conversation and informs everyone that the autonomous school bus will be out front in 3 minutes. Hurrying out the door, the kids grab their lunches and tablet bags. As they board the bus the kids are authenticated by the vehicle's identification and access management system.

Just then, Jake is alerted that an active shooter event is underway at Baltimore's Penn Station. He immediately jumps into his autonomous self-driving squad car (Figure 2), which shows the latest situational awareness data on his heads-up display. The car sets out on a route optimized for speed, which includes rerouting civilian vehicles to clear an emergency path to the scene. Security camera systems have already analyzed and identified two shooters and three possible victims using facial recognition technology. The Police Emergency Operations Center (EOC) instantly alerts the public about the danger. The autonomous locking system, using real-time artificial intelligence (AI) image and sensor data analytics, contains the shooters in a particular area of the train station, preventing their escape and ability to harm other passengers. As he gets closer to the station, Jake launches his vehicle's wingman drone to surveil the scene and provide the latest situational awareness. Arriving ahead of Jake is a Vertical Takeoff and Landing (VToL) emergency care vehicle, which deploys autonomous drones to seek out victims and assess their health and status. Jake helps establish an on-scene command hub that fuses the input from all drones, cameras, sensors, and citizen mobile device feeds to create a common operational picture of the situation. The shooters are rapidly cornered and apprehended by a human–machine team of officers, drones, and sensors that have trained repeatedly together for this kind of scenario. Lives were saved, the scene is cleared, and train operations resume.


**Figure 3.** Autonomy is ubiquitous. AI is being integrated into every sector and into every aspect of our lives—autonomous vehicles, AI-enabled personal assistants, military systems, delivery drones, interconnected infrastructure systems, and AI-powered health care. Many of these AI-enabled systems can act without human intervention, so it is critical that we assure their safe operation.

It's an interesting story. But what if it had gone another way?

Increasingly, AI is being integrated into every sector and aspect of our lives. Many of these AI-enabled systems are designed to take action without human intervention, autonomously. This particular scenario is filled with examples of autonomy, from delivery drones to AI-enabled personal assistants to smart pills. Autonomy is everywhere (Figure 3) and the stakes are high to get it right.

What if the commercial delivery drone had mistaken the window to the kids' room for the porch and smashed through it? Or if the box had been dropped on the cat sitting on the porch? What if the data from the smart-pill system was being auctioned off to organized criminals to steal Jake's personal identity information and to disrupt police activities? Or if Jake had forgotten that the blood pressure device was past its expiration date and providing inaccurate readings? What if the autonomous squad car was routed to the wrong location and the autonomous locking system trapped victims inside Penn Station with the active shooter, rather than allowing them to escape? Or, if the American public was so distrustful of autonomous technologies that none of these potentially life-enhancing (and life-saving) capabilities were launched in the first place?

## THE PROBLEM—THERE IS NO "PIXIE DUST"

Today, no one can vouch that autonomous systems are always safe, secure, ethical, and reliable, or that that they always do what they are supposed to do—in other words, that they are "trustworthy" or "assured" autonomy (Figure 4). The autonomous future has the potential to improve the human condition. But as many people and organizations race ahead to realize these benefits, there is a need for communities and societies to come up with ways to secure these systems against abuse and misuse, from criminal activity to innocent coding errors. There is no technological "pixie dust" that can be sprinkled over these systems to make them trustworthy. Rather, getting to a future with assured autonomy requires new knowledge and discoveries borne out by new techniques for engineering autonomous systems.

A holistic approach to assured autonomy requires many key elements.

First, autonomous systems must be **safe and reliable** when they operate in the unconstrained world. Achieving this is a significant challenge when underlying learning-based AI models cannot be explained and the boundaries are unclear between when algorithms will work well and when they will fail.

Second, these systems must be **secure and resilient** in the face of attempted attack and corruption by potential adversaries. These attacks could be physical attacks to alter critical sensor inputs, such as adding seemingly innocuous alterations to stop signs to cause visual-recognition algorithms to misread the signs, or virtual attacks such as denial of service or other cyberattacks.

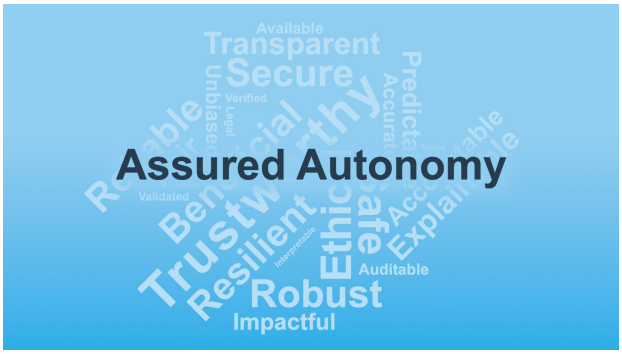



**Figure 4.** The many elements in a holistic approach to assured autonomy. Autonomous systems must operate safely, reliably, and predictably. They must be secure and resilient against attack and corruption. They must integrate seamlessly with humans and legacy technologies at scale. And, finally, autonomous systems must be ethical and beneficial to all members of society.

Third, autonomous systems must be **predictable** when operating within complex human ecosystems. The future world will include myriad interactions between humans and machines, spanning air, land, sea, and cyber domains. This complex ecosystem will require the **seamless and even synergistic integration** of autonomous technology with humanity and other legacy technologies at scale.

Finally, decisions in the design and integration of autonomous systems drive the social impact of those systems. It is important to ensure that autonomous systems are **ethical and beneficial** to all members of society, including the most marginalized and vulnerable.

The current approach—dealing with these concerns on an incremental and largely piecemeal basis—will not enable the bright future promised by autonomy. The full potential of autonomy will only be realized by understanding the complex and dynamic interactions across the IAA's three pillars of focus. We see both an opportunity and an imperative for Johns Hopkins University (JHU) to become a leader in the field of autonomous system assurance—one that moves the world toward a brighter future through these technologies while avoiding a darker one. And, as APL founding director Merle Tuve once wrote when urging the government to staff the effort to perfect the proximity fuze during World War II, "the time [to act] is shorter than we think."

## LAUNCHING THE JOHNS HOPKINS INSTITUTE FOR ASSURED AUTONOMY

Building on a $30 million internal investment from the JHU president's office, APL and the Whiting School of Engineering (WSE) have joined to establish the IAA. Led in its formative stage by co-directors Dr. Cara LaPointe of APL and Dr. Tony Dahbura of WSE as they

have prepared for the appointment of an executive director, the IAA fills a critical gap by helping to assure our autonomous future.

The IAA's vision is to drive a future where autonomous systems are trustworthy contributors to society (Figure 5). To realize that vision, the IAA aims to bring together the entirety of the university and other key players across the country and beyond; develop the right partnerships across government, industry, and academia; and attract top academic minds and industry thought leaders with the intent of assuring the autonomous world. The IAA achieves its vision by covering the full spectrum of research and application, thought leadership, partnerships and collaboration, education and workforce development, and translation and entrepreneurship.


**Figure 5.** IAA vision.

The IAA takes a holistic approach to address the key elements of assured autonomy by working across three pillars (Figure 6): increasing reliability of the technology, improving interactions within the integrated ecosystem, and engendering trust through policy and governance. The IAA seeks to understand negative consequences of autonomous systems and to find ways to prevent or mitigate them. Under the oversight of research director Dr. David Silberberg, the IAA conducts a robust portfolio of internally funded research to create tools and methods to drive assurance into the design, development, operation, and protection of autonomous systems. Although the projects are arranged into specific categories, assuring autonomous systems requires solutions that may overlap and cross domains (Figure 7).

**Technology:** Autonomous technologies employ AI to simulate human cognition, intelligence, and creativity. AI-enabled autonomous systems can often learn from their experiences, modify their own logic to achieve better results, and improve their capabilities. These systems are increasingly incorporated into infrastructure that is critical to the nation's security, health, quality of life, and commerce. To lead advancements in technology, the IAA is funding relevant projects that include:

- **Identifying Factors to Explain the Behavior of Deep Learning Systems**, with co-principal investigators Anna Buczak (APL) and Mark Dredze (WSE): A unique blending of three advanced AI, machine learning, and human language technology methods to support human-understandable explanations of AI-system behavior

- **Regression Analysis for Autonomy Performance Comparison**, with co-principal investigators Marin Kobilarov (WSE) and Paul Stankiewicz (APL): A holistic performance testing approach for successive versions of software to ensure that new versions do not break old versions and introduce new failures

- **Verified Assured Learning for Unmanned Embedded Systems (VALUES)**, with co-principal investigators Greg Hager (WSE), Marin Kobilarov (WSE), and Aurora Schmidt (APL): A pioneering methodology to unify physics-based modeling, continuous learning neural network methods, and formal approaches for verifying safety and performance

New techniques are also needed to ensure autonomous systems are **secure and resilient** to malicious deception and spoofing. Malicious adversarial attacks can be made on an autonomous system itself such as via cyber hacking or they can be focused on physically or digitally disrupting the sensor and data inputs. Cybersecurity is critical to

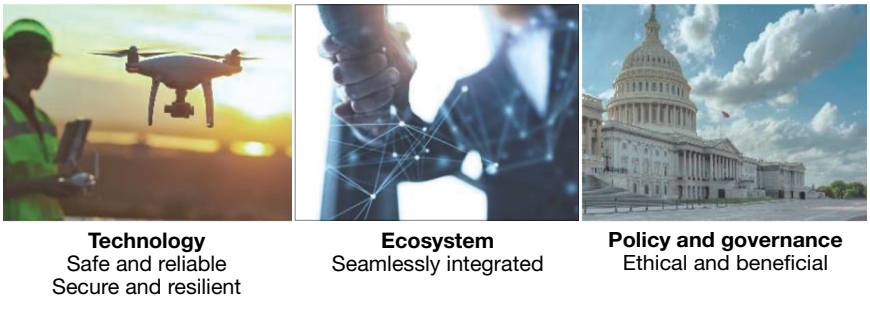



**Figure 6.** IAA approach. Following a holistic approach, the IAA will address the key elements of assured autonomy by working across three pillars: identifying the complex interactions between **technology** design and development, integrating the **ecosystem**, and the developing tools of **policy and governance**.

ensure privacy, integrity, and operability of autonomous systems. The sheer complexity and opacity of machine learning, plus the degree of connectivity and the distributed nature of emerging autonomous systems, will present even more potential vulnerabilities than today's most complex systems. Relevant IAA-funded projects include:

- **Physical Domain Adversarial Machine Learning for Visual Object Recognition**, with co-principal investigators Yinzhi Cao (WSE), Philippe Burlina (APL), and Alan Yuille (WSE): A new technique for increasing the resilience of deep learning systems to physical attacks that present as patch-based and occlusion-based attacks

- **Risk-Sensitive Adversarial Learning for Autonomous Systems**, with co-principal investigators Raman Arora (WSE) and Ryan Gardner (APL): A novel learning framework that incorporates risk-sensitivity factors for applying deep reinforcement learning to real-world autonomous systems applications

**Ecosystem:** Any number of things can go incredibly wrong in a highly connected and complex ecosystem populated with both people and autonomous systems—from crowded streets to networked smart cities to state-of-the-art medical facilities. New algorithmic and systems engineering approaches must be developed to deal with the unprecedented level of dynamic interactions between autonomous platforms in an assured ecosystem. Autonomous systems must predictably and **seamlessly integrate** with other autonomous systems, legacy technologies, and individuals. Human–system interaction must provide people with an understanding of autonomous systems' decisions and actions, the ability to interact at appropriate levels of abstraction, and the ability to override the system's actions. New human-system engineering techniques are needed to ensure autonomous systems will be smoothly and readily adopted into society. Relevant IAA-funded projects include:

- **RADICS: Runtime Assurance of Distributed Intelligent Control Systems**, with co-principal investigators Yair Amir (WSE) and Tamim Sookoor (APL): A novel combination of monitors and governors over reinforcement learning algorithms and traditional algorithms to ensure the safety of city-scale critical infrastructure systems

- **Assuring Autonomous Airspace Operations**, with co-principal investigators Lanier Watkins (APL) and Louis Whitcomb (WSE): A prototype autonomous traffic management system to safely manage the increasing density and ubiquity of low-altitude autonomous unmanned aerial systems

**Policy and Governance:** It is difficult to overestimate the ramifications that autonomous systems will have on society and the importance that these systems are **beneficial and ethical**. People are concerned about the potential dangers of these systems and want to be assured they will behave legally, ethically, fairly, and transparently while preserving privacy and fairness. Effective policy and governance are critical to ensuring a thriving environment that is hospitable to emerging autonomous technologies and ecosystems while providing important guardrails to prevent and mitigate any potential negative consequences. Effective policy and governance provide the tools to codify societal norms and incorporate community priorities into the requirements for autonomous systems. Relevant IAA-funded projects include:

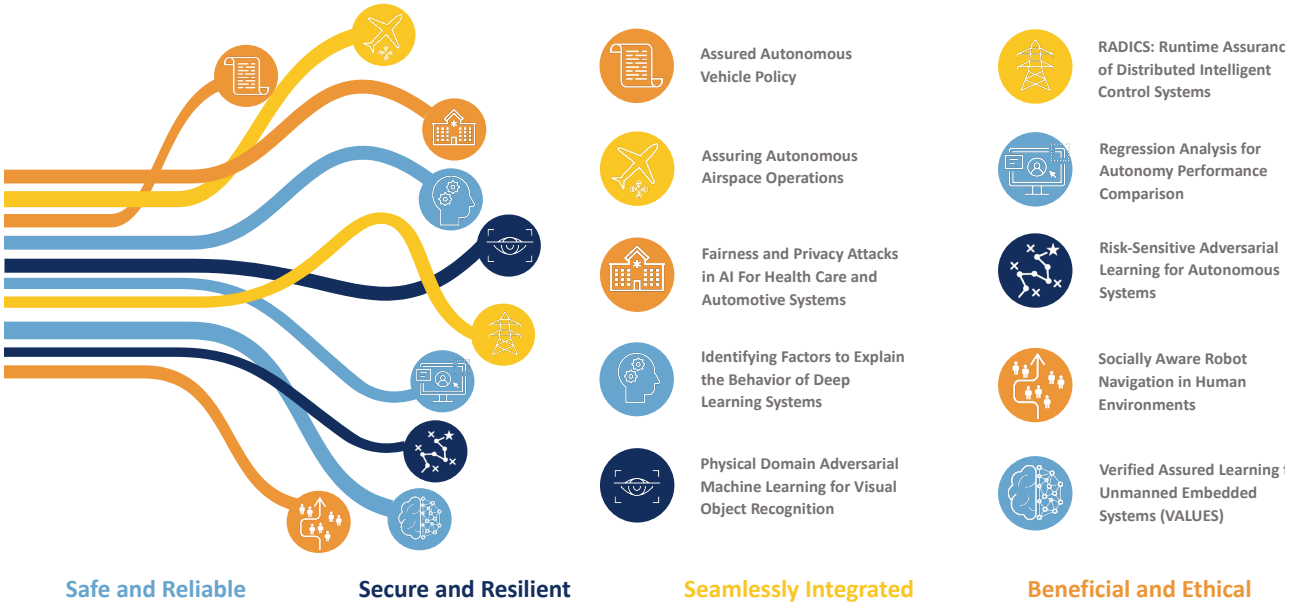



**Figure 7.** IAA seed-funded research projects in 2020–2022. The IAA has funded projects in each of its three pillars.

- **Assured Autonomous Vehicle Policy**, with co-principal investigators Johnathon Ehsani (BSPH), Tak Igusa (WSE), Jeff Michael (Bloomberg School of Public Health, BSPH), Joshua Mueller (APL): Research to create a novel policy-making framework verified by simulation that enables simultaneous and symbiotic development of both technology and policy for autonomous vehicles
- **Fairness and Privacy Attacks in AI for Health Care and Automotive Systems**, with co-principal investigators Philippe Burlina (APL) and Yinzhi Cao (WSE): Novel algorithm development to address fairness and effectively defend against privacy attacks
- **Socially Aware Robot Navigation in Human Environments**, with co-principal investigators Chien-Ming Huang (WSE) and I-Jeng Wang (APL): Research to develop autonomous agents, such as robots or vehicles, that navigate an indoor environment considering both social boundaries and physical boundaries

## TESTING THE TOOLS AND METHODS OF ASSURANCE

To move from research to application, the IAA will sponsor flagship projects as test bed demonstrations of assurance tools and methodologies, testing research and development projects for their effectiveness in applied real-world settings. These demonstrations will also measure research and development for progress toward the IAA's objectives of advancing the safety, security, reliability, and trustworthiness of autonomous systems.

The IAA anticipates conducting multiple demonstrations in core domains including autonomous vehicles, safe cities, and health care. For each, the IAA will engage stakeholders from across the JHU enterprise and partner organizations to identify compelling, effective use cases for testing and demonstration.

## COMMUNICATING THOUGHT LEADERSHIP

Our industry seeks to define, understand, and reach some agreement on the questions, What is successful autonomy and assured autonomy? In a setting or community where AI-enabled systems are operating independently without human intervention, what does it mean for them to be assured and trustworthy?

To help build public understanding and trust, the IAA is developing thought leadership and bringing together leaders from academia, industry, and government for critical conversations about what assured autonomy is and how we achieve it successfully.

The IAA will gather leaders to share and advance thought leadership, Intellectual influence, and pioneering thinking, in a range of forums and channels such as

- Research, trade, and mainstream publications
- Literature reviews
- Research analyses
- Seminars

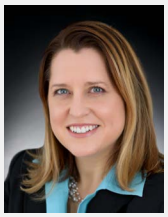

**Figure 8.** Building a community. The IAA is growing a broad community, working across APL and JHU and leveraging APL's relationships with academia, government, and industry.

- Conferences
- Workshops
- Speaking engagements
- Opinion and commentary essays
- Research and development forums
- Internet/digital/social media

## BUILDING A COMMUNITY

No one institution or sector will be able to independently realize a future in which autonomous systems are trustworthy contributors to society. Therefore, the IAA is working across APL and JHU, as well as leveraging APL's natural role as a bridge between academia, government, and industry, to motivate the broader community to collectively bring to bear their skills, creativity, innovation, and ingenuity in assuring our autonomous future (Figure 8). Partnerships across sectors and throughout our communities are critical to achieving a future where autonomous systems are seamlessly integrated into human ecosystems.

There are a number of opportunities for those who are interested in learning more or contributing to the mission of the IAA. To learn more about the IAA, visit the website at https://iaa.jhu.edu.

**Cara E. LaPointe,** Asymmetric Operations Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

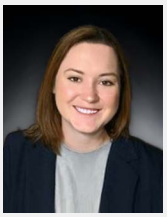
Cara E. LaPointe is a futurist who works at the intersection of technology, policy, leadership, and ethics. She earned a BS in ocean engineering from the US Naval Academy, an MPhil in international development studies from Oxford University, an MS in ocean systems management from the Massachusetts Institute of Technology (MIT), an Eng degree in naval engineering from MIT, and a PhD in mechanical and oceanographic engineering awarded jointly from MIT and the Woods Hole Oceanographic Institution. She is the co-director of the Johns Hopkins Institute for Assured Autonomy, focused on ensuring that autonomous systems throughout society are safe, secure, reliable and ethical. Cara has served as an advisor to numerous global emerging technology initiatives at the National Academy of Medicine, the United Nations, and the Organization for Economic Cooperation and Development. She is the author of *The Blockchain Ethical Design Framework*, a patented engineer, and a White House Fellow. Her email address is cara.lapointe@jhuapl.edu.

**David P. Silberberg,** Asymmetric Operations Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

David P. Silberberg is the IAA research director. He received master's and bachelor's degrees in computer science from the Massachusetts Institute of Technology (MIT) and a PhD in computer science from the University of Maryland, College Park. Dr. Silberberg has conducted extensive research and development in the areas of leading-edge AI and machine learning algorithms, including graph analytics, distributed and large-scale architectures, intelligent access to distributed and heterogeneous database systems, and semantic graph query languages. He led APL's Large-Scale Analytic Systems Group, which applies machine learning and AI-base algorithms to perform descriptive, predictive, and prescriptive analytics on large and complex data. Dr. Silberberg also served as chief architect for the deep archive of NASA mission data and for the Hubble Space Telescope data archive. He teaches the Large-Scale Database Systems course in the Johns Hopkins University Engineering for Professionals program. His email address is david.silberberg@jhuapl.edu.

**Anton T. Dahbura,** Department of Computer Science, Johns Hopkins University, Baltimore, MD

Anton T. Dahbura is a computer scientist with research interests in information security, fault-tolerant computing, testing, optimization, and applied analytics. He received a BSEE, an MSEE, and a PhD in electrical engineering and computer science from Johns Hopkins University. He is the co-director of the Johns Hopkins University Institute for Assured Autonomy. Anton has served as a researcher at AT&T Bell Laboratories, as an invited lecturer in the Department of Computer Science at Princeton University, and as research director of the Motorola Cambridge Research Center. He serves as executive director of the Johns Hopkins University Information Security Institute. He also holds an appointment in the Johns Hopkins University Malone Center for Engineering in Healthcare. He is a fellow of the IEEE. His email address is antondahbur@jhu.edu.



**Amber R. Mills,** Asymmetric Operations Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

Amber R. Mills is an engineer by trade but focuses on the intersection of technologies and their user. She received a bachelor's degree in civil engineering from The Citadel, The Military College of South Carolina. She is currently pursuing a master's in Computer science from the Johns Hopkins University Whiting School of Engineering. At APL, she has served as a systems engineer focusing on operational test and evaluation and integration of solutions for the Department of Defense and the Department of Homeland Security. She has extensive knowledge in the integration of technologies into complex, heterogeneous ecosystems. Her email address is amber.mills@jhuapl.edu.