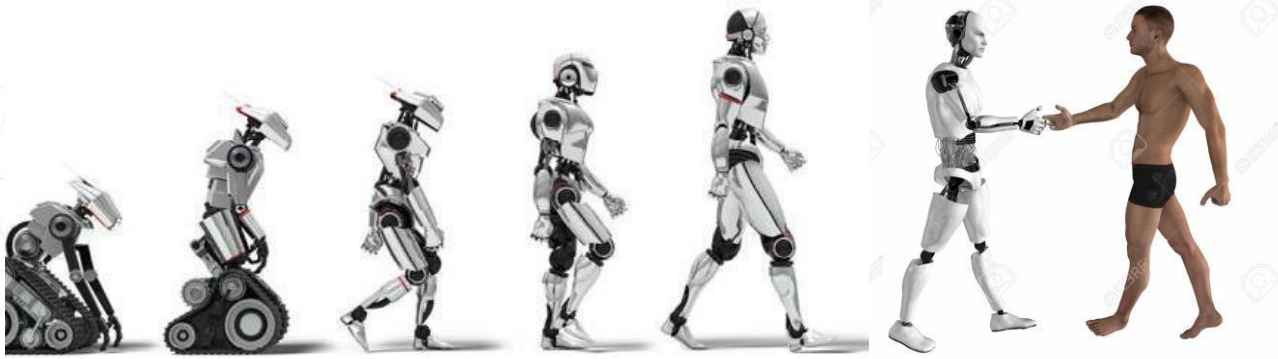
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

APL JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

JHU Institute for Assured Autonomy

Assuring the Future Autonomous World

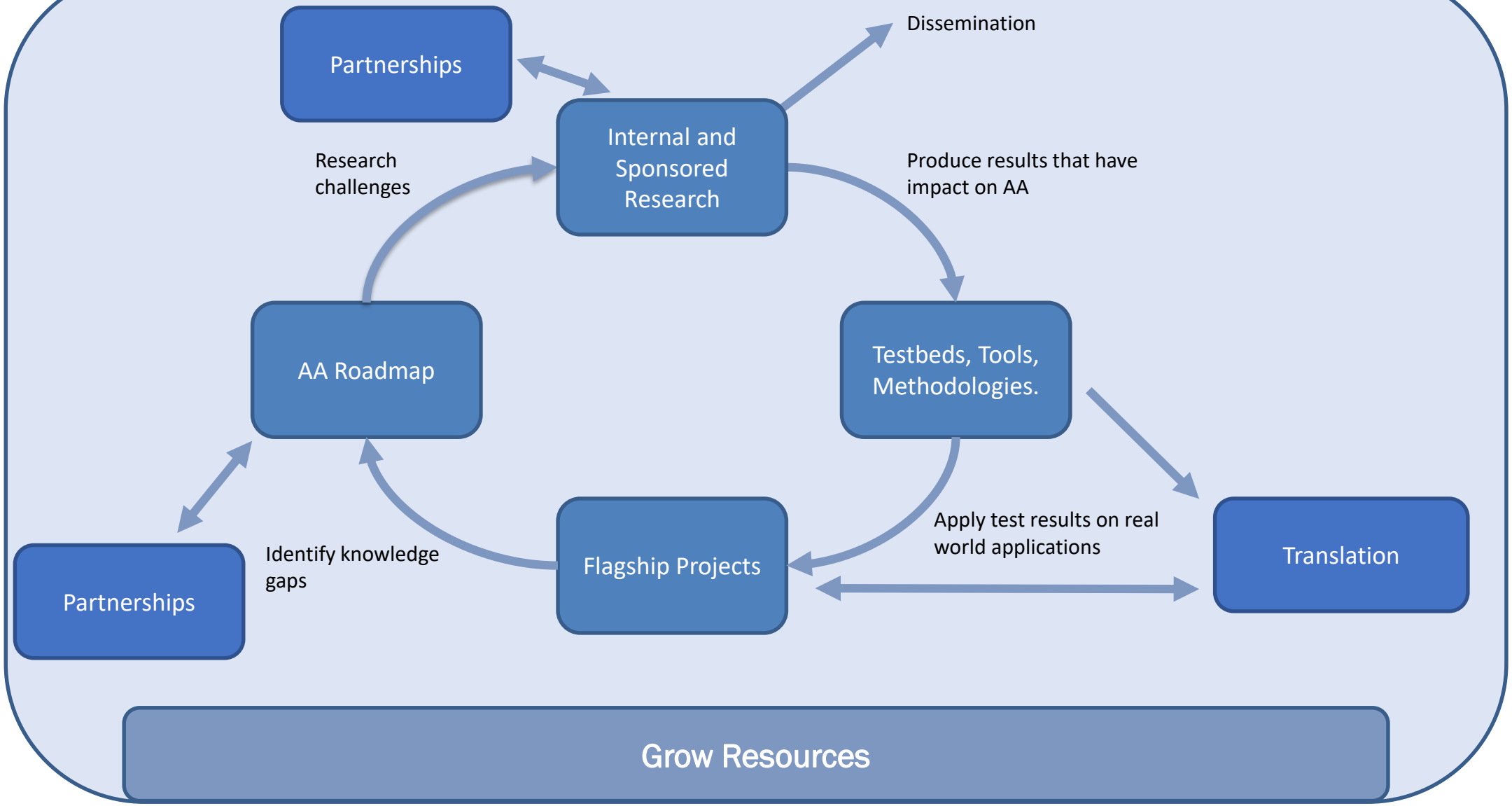


February 2020

Tony Dahbura (AntonDahbura@jhu.edu)

Cara LaPointe (Cara.LaPointe@jhuapl.edu)

IAA Strategic Approach



IAA Status and Current Activities

- JHU President Ron Daniels has committed $25M over the next five years for research, faculty slots and facilities- largest commitment of its kind;
- Spring Assured Autonomy workshop drew 150 participants from WSE and APL;
- First round of internal funding has resulted in 48 pre-proposals and 23 proposals of which 10 are being funded for $3.2M (two years);
- BDP Executive Director search is underway;
- Candidates for faculty slots are being considered;
- Space reserved in the Stieff Silver Building for IAA headquarters.

Autonomous System Lifecycle

8- Regression Analysis for Autonomous Performance Improvement

12- Methods to Avoid Bias and Data Leaks in DL Systems

13- Safety and Performance Verification for ML Systems

21- Explainable AI

7-White Box/Black Box Monitoring of Autonomous System Operation

9- Assured Resource Managers for Assured Autonomy (Airspace Ops)

3- Human-System Interactions (human intent)

11- Adversarial ML for Visual Object Recognition

23- Adversarial Learning Using Learning Agents

Policy & Governance

5- Assured Autonomy Policy Development Formal Methodology

BIAS AND PRIVACY ATTACKS IN AI FOR HEALTHCARE AND AUTOMOTIVE SYSTEMS (Burlina, Cao)

- Deep learning algorithms can leak private information and may be gender/race/disease biased.

- The proposed work develops algorithms to address bias as well as approaches to assess possible risks in existing algorithms for privacy / membership attacks, and proposes ways to effectively defend against such privacy attacks.

- The investigators use:
  1. **directed data augmentation using synthetic data produced from deep generative models** to address both bias and privacy challenges; and
  2. **identity-obfuscation pre-processing** to reduce the risk of membership and related attacks on privacy while maintaining the performance of diagnostic models.

PHYSICAL DOMAIN ADVERSARIAL MACHINE LEARNING FOR VISUAL OBJECT RECOGNITION (Yuille, Cao, Burlina)

- Addresses adversarial attack and defense techniques for machine learning and deep learning applied to visual object recognition, specifically including methods that implement patch and occlusion attacks.

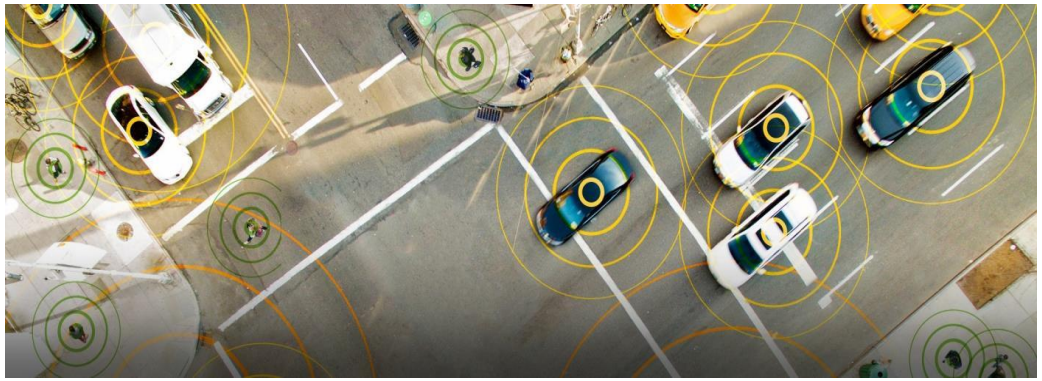
- The objective is to establish an ecosystem composed of adversarial machine learning (AML) attack/defense algorithms as well as a testbed specifically for evaluating non-differentiable patch- and occlusion-based physical AML algorithms.

- The new models contain **explicit representations of object parts** and detect the objects if a significant number of these parts have been detected in plausible spatial configurations.

RISK-SENSITIVE ADVERSARIAL LEARNING FOR AUTONOMOUS SYSTEMS (Llorens, Arora)

- Deep reinforcement learning (DRL) formulations have several shortcomings, including a general lack of robustness when employed in dynamic and uncertain environments.
- The researchers develop an adversarial learning framework for developing robust, risk-sensitive DRL agents.
- DRL typically relies heavily on simulation due to the impracticality of replicating large numbers of diverse trials in the real world; however, simulated environments invariably differ from their real-world counterparts.
- The researchers pose the learning problem as a competition between an agent that seeks to avoid undesirable outcomes and a parameterized environment that dynamically reconfigures itself in order to cause them.
- Experimentation using application-inspired agents, simulation environments and a proof-of-concept hardware demonstration.

IAA Research in Assured Transportation



IAA Research in Assured Public Safety and Security

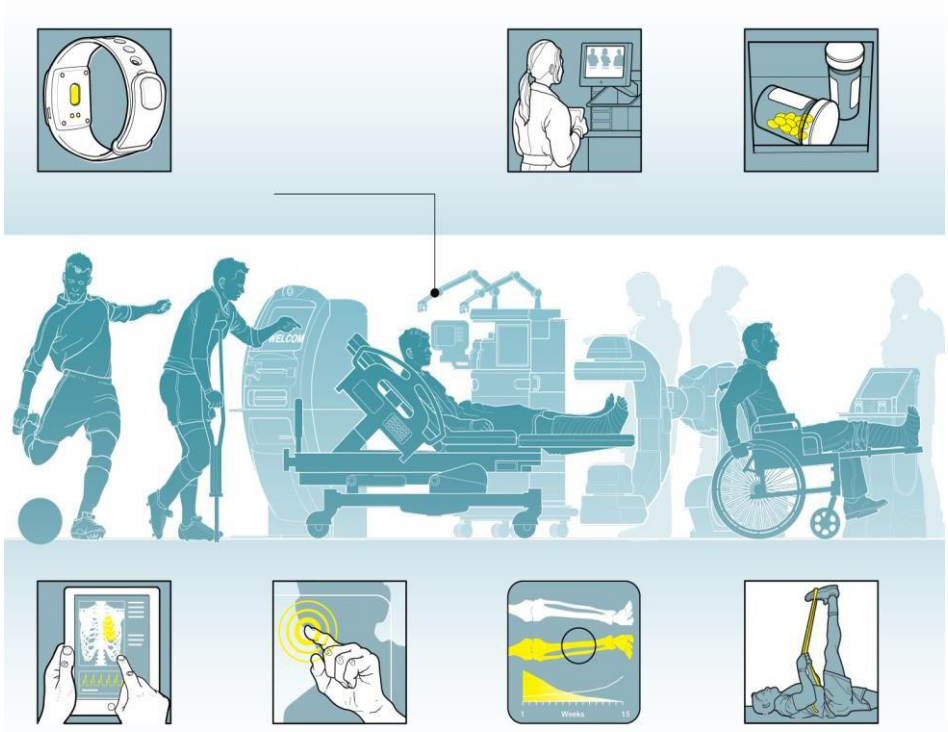
- Top-level goals include:
  - Create a cyber-secure central system for monitoring and managing a large presence of campus IoT devices
  - Partner with government, industry and academia to facilitate the development of assured autonomous devices for augmenting the IoT network and building additional safety services onto the network to enhance campus-based smart functions & services
  - Increase the level of trustworthiness in individual technologies and integration of these technologies into systems to facilitate the deployment of research prototypes without the fear that the technology is likely to be misused or be unavailable when truly needed

https://goo.gl/images/kdNQWV https://goo.gl/images/MxGk3w

IAA Research in Assured Health Systems

- Top-level goals include:

  - Partner across JHU to evaluate existing and emerging health technologies that provide trustworthy autonomy and eliminate cyber-related harm to patients and other healthcare stakeholders
  - Develop standards of practice as well as implementation frameworks and be seen as the world's preeminent trustworthy autonomy and cyber-safe healthcare institution
  - Partner with government, industry, and academia for the development of assured autonomous medical systems, leveraging JHUs medical research labs for the assurance of legacy and next-gen intelligent medical systems
  - Explore human-machine teaming for advanced, efficient, reliable, and trusted medical and health services and share knowledge of how to achieve this with the world



https://goo.gl/images/uiLe77

Questions?

Thank you!