*PROPOSAL in response to FOA ARO #W911NF-19-S-0008, FY 2020 DoD MURI*
*Topic 11: (ARO) Adaptive and Adversarial Machine Learning*
*Research Topic Chief: Purush Iyer, (919) 549-4204, s.p.iyer.civ@mail.mil*

# Robust Concept Learning and Lifelong Adaptation Against Adversarial Attacks

Insup Lee (PI), Osbert Bastani, Kostas Daniilidis, Eric Eaton, Dan Roth, James Weimer (University of Pennsylvania)
Julia Parish-Morris (Children's Hospital of Philadelphia)

This proposal addresses the design and analysis of robust and adaptive machine learning based on how children learn. Autonomous assets that utilize learning systems and interact with humans can revolutionize the battlefield, but also present new sources of uncertainty and vulnerabilities. Learning systems are vulnerable to adversarial inputs that are malicious (e.g., image tampering) and non-malicious (e.g., dynamic environments). Moreover, these adversarial inputs can affect the learning systems at design-time (i.e., in the training data) and runtime (i.e., in the test data). Providing predictive analytics that deliver intelligent battlefield services in an adversarial setting requires cross-thrust research addressing fundamental adversarial learning challenges. First, modern machine learning generally lack adversarial robustness, both at design-time and run-time, where state-of-the-art approaches tend to perform substantially worse in the presence of adversarial examples such as sensor noise (e.g., dust on the camera lens), image transformations (e.g., translations, rotations, and scaling), and distribution shifts (e.g., daytime images to nighttime images). Second, concepts learned by both children and learning systems must be robust to adversarial examples in real, large-scale dynamic environments, requiring new concepts built upon previous knowledge. However, achieving this improved robustness requires increased reliance on prior concepts (e.g., relational models, causal models, and shapes) that may also be susceptible to changes in the environment. Third, safety critical autonomous systems require verifying and monitoring the robustness of concept-based learning, which is challenging in dynamically evolving systems based on the sensed environment.

In stark contrast to current supervised approaches to machine learning, young children are not passive learners that ingest millions of pre-labeled images; rather, they are concept learners that are constantly generating and testing concept-based hypotheses about the world around them. Their learning is broadly hierarchical, connectionist, and sensitive to statistical regularities and invariances in the environment. Perhaps most importantly, young children are motivated learners. They seek novelty and salience, they attend to experts (parents, siblings) for knowledge, and they are sensitive to social context when learning. Their knowledge is never static; it responds to their own curiosity, mixed in with developing “common sense” and a strong sense of whimsy. Young children play never-ending games of prediction, expectation, problem-solving, and learning through multi-sensory hands-on experience with the real world. Although young children are exquisitely attuned to any opportunity to learn, they also quickly decide who is “trustworthy” and refuse to learn from sources who are “untrustworthy”.

In this project, we aim to develop the foundations for robust and adaptive learning based on childhood development consisting of three research thrusts: (Thrust I) Concept-based Learning Robust to Adversarial Examples; (Thrust II) Adaptive Learning in Dynamic Environments; and (Thrust III) Verification and Monitoring of Learning. Our approach will enable adaptive childlike learning that is robust to adversarial examples by utilizing concepts (e.g., prior models and shapes) inherent in the physical world (Thrust I), while simultaneously detecting and adapting to changes in the environment and concepts (Thrust II), such that robustness claims can be validated through a combination of offline verification and runtime monitoring (Thrust III). Evaluations of the techniques will be performed on an interactive robotic platform as a surrogate for future military applications involving cooperative robotic system with learning in a battlefield environment.