JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

APL AI

TECHNOLOGY ROADMAP

January 2020

APL Artificial Intelligence Technology Roadmap

Inventing the future of intelligent systems for our nation

Introduction

"Discovery consists of seeing what everyone has seen and thinking what no one has thought."
Albert Szent-György

Imagine a future in which a legion of autonomous systems collects and shares intelligence, senses the environment, continuously assimilates new data, learns, makes decisions, communicates, and acts cooperatively with humans and each other to carry out a mission while ensuring that the integrity and purpose of the system is not compromised or co-opted by cyber-attacks, insider threats, or misguided evolution.

Imagine a comprehensive home medical monitoring and diagnosis system that can be trusted to partner with a human on a daily basis to perceive and detect early-stage symptoms of all major chronic diseases, seasonal infections, and emerging health threats, and recommend courses of prevention and intervention.

Imagine a non-invasive brain-computer interface device that allows humans to communicate with machines at the speed of thought to solve complex problems, optimize mutual learning, develop research strategies, and create new works of art.

Imagine a fully autonomous mission to another planet in which robotic systems are trusted to develop their own courses of action to identify and explore interesting new phenomena, deploy sensors, collect data, and communicate results to researchers on Earth.

These are just a few of the many envisioned futures that are driving research initiatives in artificial intelligence across the Johns Hopkins University Applied Physics Laboratory (APL). As a nonprofit University-Affiliated Research Center, APL serves as a bridge between academia, industry and government to develop technologies that achieve mission impact in national security, space exploration and health. The enormous potential for artificial intelligence to dramatically transform most of the core technical activities of the Laboratory resulted in a decision by APL leadership to create a technology roadmap along with associated execution and engagement strategies. This report contains a description of APL's technology roadmap, which is the culmination of critical thinking by technical experts from across APL. In this introduction we provide context for the roadmap and an outline for this report.

As everyone engaged in science and technology knows, there has been an explosion of interest in artificial intelligence (AI) across the planet in recent years. The power of modern computing engines and the ability to store and access vast amounts of data have enabled the realization of a dream that started with the creation of the perceptron learning algorithm by Frank Rosenblatt in 1957. Today, the capability to develop deep neural networks and apply machine learning algorithms to complex systems has brought the power of AI to almost every discipline in science, technology, and the commercial world. The level of activity in the realm of AI theory and applications has literally reached the point of frenzy. Experienced researchers might be secretly worried about the danger of a technology trough of disillusionment followed by another "AI winter" given this extreme investment of human talent and computer resources. However, the applications of AI in image and speech recognition alone are so profound that this emerging field will endure even after we have reached the maximum on the hype cycle. To be clear, however, we have not reached that maximum yet!

In light of the intense interest in AI across the globe, the need for an APL roadmap to help guide internal research and development and external engagement was evident. Roadmaps are, of course, focused on the future. However, the roadmap that we will present is informed by the Laboratory's long history in space exploration, missile defense systems, undersea warfare, robotics, and autonomous systems. Building on this collective wealth of knowledge and experience, the roadmap development process began with envisioned futures in which AI is expected to play an essential role. Making exact predictions about the future is impossible, but this was not the purpose of our envisioned futures. The real goal here was to identify the common AI threads running through each story and to highlight these threads as the strategic vectors that will guide new research initiatives and future applications of AI. These strategic technology vectors form the foundation for the roadmap.

Amid a global landscape of institutions and researchers, APL has been deeply engaged in the theory and application of AI since its inception. The Laboratory's first defining innovation, the proximity fuze, was an intelligent system that could perceive its environment, decide whether the conditions for optimal impact were met, and then, at just the right moment, act. Every APL spacecraft ever built ventured into the extreme environment of our solar system with some level of autonomous reasoning to sense, navigate, and perform a mission. The Modular Prosthetic Limb with its neural interface demonstrated the amazing potential of human-machine teaming. More recently, APL has realized mission applications of AI in problems involving signal and image classification such as Deep Mine, the APL-developed system for autonomous classification of underwater mines, and a wide range of other problem domains. Applications of AI to health monitoring and medical diagnosis are expanding daily. APL has recently established a deeper partnership with Johns Hopkins Medicine resulting in new advancements such as the development of a deep convolutional neural network capable of classifying retinal images for age-related macular degeneration. APL's Intelligent Systems Center was established in 2015 to catalyze cross-displinary research and development in artificial intelligence, robotics and applied neuroscience and help accelerate the pace of breakthroughs like these.

At a high level, we have observed that there are three essential elements to successful application of AI to complex real-world problems: *domain expertise*, *data*, and *experimentation*. A brief discussion of these points will help frame the roadmap presentation that follows.

First of all, one might think that the term *domain expertise* refers to knowledge and expertise in using the tools of AI, such as machine learning and neural networks. On the contrary, here we are speaking of deep domain expertise in application areas such as undersea operations, space exploration, health systems, and numerous other domains of vital importance to APL and its sponsors. The tools and techniques of machine learning and neural networks that are essential for applications of AI can be straightforward to learn for technical staff with an advanced degree in physics, engineering, biology, computer science, or mathematics. In contrast, there is no substitute for the deep technical knowledge gained through years of applying foundational science and engineering skills to critical challenges. For this reason, an important component of our execution of the technology roadmap presented here is focused on workforce development and education, with the specific goal of providing a path for everyone at APL to augment their existing domain expertise with AI tools and techniques. Given the pace of technological advancement and the growing need for cross-disciplinary solutions, we seek to foster lifelong learning as an essential career pursuit.

The second essential element for success in AI is *data*. Deep learning, currently the most prominent technique in AI, is typically utilized as a supervised learning approach that requires a large dataset of examples that have been properly classified and labeled by experts. The algorithm uses the labeled data to learn how to classify the examples in a training set and is then ideally able to properly classify previously unseen test examples. The training process for deep learning takes serious computing resources and often a very long time, but once the algorithm has been trained, it can classify new examples with relatively efficient computation. However, the real world is messy and presents many domain-unique challenges. For example, intelligent systems must often address rare events, such as actually encountering an underwater mine on an ocean floor that contains mostly rocks and garbage. In fact, for most real-world problems, a dataset, no matter how massive or well-curated, can never tell the whole story. Domain expertise is invaluable because it is necessary to fill in the gaps – inferring causal relationships that may not be present in the data – and using these insights to clean data, create synthetic data, develop simulations and otherwise find creative ways to train AI algorithms to perform the desired task.

General techniques like transfer learning have also proven to be valuable in data-limited situations. Computer simulations are especially important in the area of reinforcement learning. In this context, systems can learn to perform a task through thousands of randomized trials without wearing out a physical system in the process or making costly mistakes during real-world operation. Once the skill has been acquired in the simulation environment, the algorithms can be transferred to the physical system. For decades, APL has served in test and evaluation roles for various programs. The datasets and modeling and simulation capabilities developed in support of these programs may hold the key to enabling the development of AI-enabled real-world systems through emerging machine learning paradigms.

The third essential element of success in AI is *experimentation*. At present, a great deal of AI research is at the empirical stage and is not supported by a well-understood theoretical foundation. For example, the backpropagation algorithm that is used to train a neural network in deep learning gradually adjusts the parameters of each “neuron” in a large network of interconnected components. However, our understanding of just how the resulting trained network achieves the impressive performance that often follows this gradual tweaking of the parameters is still something of a mystery. AI researchers need the courage to experiment and the determination to persevere through noble failures. There is a human-machine team at work in every AI project: researcher plus computer. Success requires a lot of hard work by both members of the team. As a Laboratory, we aim to channel the results of these individual experiments into a larger body of knowledge that can be applied to current and future missions and, as appropriate, contribute new knowledge to our broader society.

One of the key characteristics of APL’s long-term vision for AI is to enable human designers to get more out of a system than they are capable of putting in. That is, we anticipate an increasing level of creativity and innovation in which a system, provided with high-level guidance, generates novel solutions to complex problems. Consequently, there will be an element of emergent behavior in the systems that we seek to create, presenting both exciting opportunities and significant challenges to overcome in realizing them.

This document is organized into the following two sections:

Section 1: What is an Intelligent System?
This opening section of the report offers our clearest and most relevant definitions of artificial intelligence, intelligent systems, machine learning, and related concepts. We discuss the common attributes of intelligent systems and provide a basic context for understanding the technology roadmap. These concepts are summarized in the **APL Intelligent Systems Framework**.

Section 2: The Technology Roadmap
The major technical elements of the roadmap are presented in this section. Based on envisioned futures formulated by experts from across APL, the technology roadmap is presented in the form of four **technology vectors**. Along each technology vector, a series of technical challenges of increasing complexity is identified. These challenges represent near, mid-range, and long-term technical goals and research focus areas that we believe will position APL among leading

institutions in the application of AI to critical challenges. In this context, we define near-term goals as objectives in the range of 0 – 3 years, mid-term goals as objectives in the range of 3 – 10 years, and long-term goals as objectives in the range of 10 to ∞ years.

1. What is an intelligent system?

Perceive, Decide, Act, Team, Trust

"Intelligence is the ability of an entity to achieve complex goals." Max Tegmark, *Life 3.0*

We define an **intelligent system** as an agent that has the ability to **perceive** its environment, **decide** upon a course of action, **act** within a framework of acceptable actions, and **team** with humans and other agents to accomplish human-specified goals. In addition to these qualities, we also require that the agent be able to achieve its objectives at the required level of **trust**. That is, we expect that humans and other agents will have an appropriately calibrated level of confidence in the ability of the system to perform as designed and in an ethical manner. Whether agents are embodied as robotic systems or exist only as computer software, to be regarded as intelligent systems they should have the attributes we have listed. Agents are usually implemented as systems in which the interactions among the components of the system give rise to intelligent behavior. Multi-agent systems may themselves be thought of as intelligent systems, where the intelligence arises from the ability of the agents to collectively accomplish a common goal. In all cases, trust is essential for effective teaming among agents as they perform complex tasks in challenging environments.

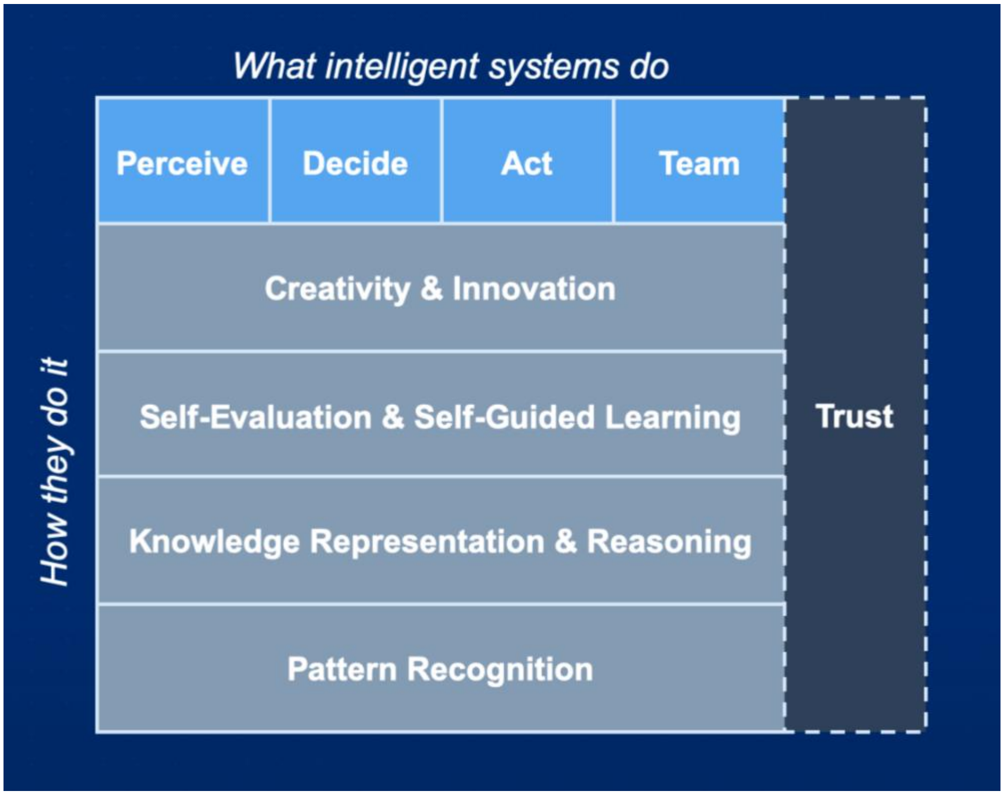
An **autonomous system** is a system that has been delegated the authority to act within specific bounds. A driverless car is a good example of a system that is both intelligent and autonomous. Another example is the automatic landing system on an aircraft. The term **machine learning** generally encompasses a wide range of algorithms that improve their performance through data and experience and which can be generally categorized as supervised learning, unsupervised learning, or reinforcement learning.

In general, the term **artificial intelligence** refers to an agent that has one or more of the attributes of an intelligent system — that is, some ability to perceive, decide, act, or team in a trusted manner. A deep neural network that classifies images as either being a photograph of a cat or not being a photograph of a cat perceives the environment and makes a decision, but it probably does not take any actions based on that decision. We would like to include this as an example of AI, but we might stop short of calling this neural network an intelligent system. At the same time it is possible that a neural network of this sort could be a component of an intelligent system.

The experts sometimes refer to **narrow AI** as distinct from **general AI**. A thermostat would qualify as an example of narrow AI, but a fully functioning iRobot as envisioned by Isaac Asimov would be regarded as general AI. One also hears the term **artificial general intelligence** as a description of a machine system capable of human-level thought and reasoning. There is considerable debate in the scientific community on the topic of whether artificial general intelligence is even possible.

Along with describing *what* intelligent systems do, we have also tried to identify *how* intelligent systems achieve their goals. This combination of attributes and means forms the **APL Intelligent Systems Framework**. There are four main elements that capture how intelligent systems achieve their goals and we list them in order of increasing sophistication: **pattern recognition**, **knowledge representation and reasoning**, **self-evaluation and self-guided learning**, and **creativity and innovation**. The Intelligent Systems Framework can be visualized as follows.

The APL Intelligent Systems Framework


What intelligent systems do
Perceive
Decide
Act
Team
Creativity & Innovation
Self-Evaluation & Self-Guided Learning
Knowledge Representation & Reasoning
Pattern Recognition
Trust
How they do it

In the applications of AI that are most important for APL and our sponsors, trust is a key requirement. For this reason, our visualization shows that every aspect of *what* intelligent systems do and *how* they do it must be grounded in trust. APL's Intelligent Systems Framework provides a basic context for the discussions that follow. By separating the fundamental AI-enabled capabilities and attributes from the tools we use to enable them (i.e., deep learning algorithms, knowledge graphs, etc.), this framework has provided us with a basis for long-term planning even as the landscape of enabling tools continues to evolve over time.

2. The Technology Roadmap

The strategic technology vectors driving our AI future

"I don't have any magical ability. Before I work out any details, I work on the strategy. Once you have a strategy, a very complicated problem can split into a lot of mini-problems."

Terrance Tao, Fields Medalist

Based on envisioned futures spanning national security, space exploration and health, we have formulated four **technology vectors** to guide APL research and development for AI over the next decade. Each technology vector is presented in the form of an aspirational goal. Along with each goal we provide near-term, mid-term, and long-term technological advances that we believe will enable us to reach the goal.

Technology Vector 1: Autonomous Perception

As we have seen in the Intelligent Systems Framework, the first attribute of an intelligent system is the ability to perceive its environment. APL has been in the business of creating sensors throughout its entire history, and our experience in sensor development includes radar systems, hyperspectral imaging systems, brain-computer interface systems, quantum sensors, geolocation systems, and computer vision systems, as well as a wide range of mission-specific sensors. Our first technology vector builds on this wealth of experience. ***The vision of autonomous perception is to develop sensing systems with the capability and authority to reason about their environment, focus on the mission-critical aspects of a scene, understand the intent of humans and other machines, learn through strategic exploration, and demonstrate curiosity-driven perception strategies.*** Systems with these capabilities will deliver more accurate and insightftul information about the state of the world and ultimately contribute to the discovery of new forms of fundamental knowledge. The steps along the roadmap toward realizing autonomous perception are envisioned as follows.

Near-Term: In the near-term we expect research and development to focus on systems that can reason over spatial and temporal representations of objects and scenes, and also learn strategies for focusing their attention on the salient information in a scene. Representations of objects and entities with integrated models of how they relate and interact form an agent's world model. For example, in mobile robotics applications, a world model might contain information on the geometric properties of objects in a scene. A computer vision algorithm with the ability to reason spatially and temporally could form hypotheses about the possible locations of an object in a scene after it is concealed or becomes occluded. This is a form of what is referred to as object permanence reasoning in human psychology, an ability that children learn during their first year of life. This level of logical inference, broadly speaking, is beyond the ability of current deep learning algorithms. In information retrieval applications, a world model might be constructed as a knowledge graph representing relationships among people in a social network.

Mid-Term: The next step in autonomous perception will involve machines that can reason about agent intent and also learn to predict behavior. While models of the intent, strategies and behaviors of other agents can be considered part of an overall world model, we see reasoning about other agents as presenting unique challenges with respect to other components of world models such as objects and locations. Next-generation computer vision algorithms will make real-time predictions about the future actions and locations of people in a scene based on observations of past behavior. For example, a person passing by a camera's field of view carrying an empty water bottle might tend to walk in the direction of a water fountain. In defense applications, surveillance systems with this capability will be able to infer adversary tactics, and eventually high-level strategies, through partial observations of troop movements and logistics. In general, the ability to reason about the intent and capabilities of other agents will help to make intelligent systems more capable teammates in real-world environments.

Long-Term: Our long-term vision for autonomous perception is to create systems that can form hypotheses about the world through causal and counterfactual reasoning as the basis of learning through exploration. We envision that in the future, systems will build and refine their world models in real time as they explore complex ecosystems. For example, if a robotic system is searching a building, we would like it to recognize a door and understand that there is another part of the building to search. Space systems like New Horizons have demonstrated an amazing ability to achieve onboard fault tolerance to realize highly autonomous yet highly scripted fly-bys during long-duration missions. We envision future space exploration systems with the ability to select creative ways of using a suite of instruments to pursue scientific goals with only high-level direction from scientists and engineers. Future perception systems will build and expand world models on the fly, discovering and reporting new insights about the nature of the universe, the object and agents therein, and even the physical laws that govern it.

Technology Vector 2: Superhuman Decision-Making and Autonomous Action

As we have discussed above, the first attribute of an intelligent system is the ability to perceive the environment. The next two key attributes of an intelligent system are the abilities to decide and act. Effective decision-making requires the aptitude to search, evaluate, and select a course of action among a vast space of possible actions towards accomplishing high-level goals. This past decade has seen impressive advancements in the facility of robotic systems and platforms to walk, run, swim, fly, and even do backflips. APL has conducted leading research in swarming systems, such as autonomous sea-surface vehicles and marsupial robot teams for accessing physically constrained spaces. We expect that this enhancement of robotic capabilities for both commercial systems and national security applications will continue. However, what is needed now is to reduce the heavy reliance of these systems on humans to decide how and when to perform actions in dynamic scenarios. This reliance can be especially unfortunate given the limitations of human intelligence in considering large numbers of complex alternative strategies, coupled with limitations on the speed and effectiveness with which we can carry them out. **The vision of superhuman decision-making and autonomous action is to create systems that blend human and artificial intelligence to identify, evaluate, select, and execute effective courses of action with superhuman speed and accuracy for real-world challenges.** We envision that future systems will radically enhance the ability of human decision-makers to coordinate actions and effects across large-scale systems of systems to achieve strategic goals for the nation and for humanity. Progress along this technology vector will benefit any application that involves autonomous action, from distributed search and rescue operations to cyber operations and automated drug discovery. With respect to this technology vector we also note that there is, of course, considerable controversy over delegating authority for autonomous action to certain intelligent systems, including weapons systems, for example. Therefore, it will be essential that the available courses of action to a system are grounded in a framework of values, ethics, and mission objectives.

The steps along the roadmap towards achieving superhuman decision-making and autonomous action are envisioned as follows.

Near-term: In the near term, we expect that research and development will focus on enabling systems to autonomously select appropriate sequences of actions as a function of mission state, given well-defined yet potentially competing objectives. Current systems rely on scripted behaviors and rulesets, sometimes implemented as finite state machines, that spell out how and when behaviors should be performed. These rulesets can be brittle and unresponsive to complex and rapidly evolving mission scenarios. A near-term goal is to build on recent developments in machine learning to create effective decision-making architectures that increasingly incorporate learned, data-driven behavior selection. For example, a small team of aerial vehicles with effective sequential decision-making capabilities could accomplish competing objectives such as protecting teammates and protecting themselves while fluidly shifting tactics as the mission evolves. An unmanned aerial vehicle (UAV) would be capable of autonomously selecting combinations of actions that accomplish competing objectives such as the need to play defensive and offensive roles. A key challenge here is determining the most effective representation of possible actions. An AI system controlling a UAV, for example, may be given the ability to choose among a coarse set of highly-scripted action sequences. Limiting the AI in this way limits the potential performance improvement. However, training is easier and there will be a greater trust factor in the system as well. At the other extreme, the AI may be given the direct control of the UAV's low-level control system, offering the opportunity for novelty and greater performance improvement while increasing the difficulty of learning and the realization of trust. In general, advances in machine decision-making will enable agents to decide and act at machine speed across a broad range of applications.

Mid-term: The next level of research in decision-making and action will involve multi-agent systems that generate collective courses of action guided by human intent. Multi-agent systems of this type offer the opportunity to realize emergent collective behaviors with the potential to solve complex problems in unexpected ways. Systems in this generation will coordinate actions across large numbers of agents spanning land, air, sea, space, and cyber domains while mitigating the challenges of position, navigation, timing, and communication. A superhuman aspect of decision-making for these systems will be the ability to react and act faster than humanly possible. For example, distributed groups of maritime platforms will autonomously coordinate the scheduling of resources to defend themselves against missile raids while defending the homeland against incoming threats such as ballistic missiles. Multi-agent medical systems will seamlessly coordinate the flow of patients and resources across institutions and practitioners to optimize health outcomes.

Long-term: In the far-term research and development plan for superhuman decision-making and autonomous action, we envision systems that are capable of reasoning and shaping our world across extended periods of time spanning years or even decades. This will require much more abstract reasoning capabilities. Portfolio optimization systems with these attributes could aid decision-makers in the Pentagon in selecting the optimal fighting force to build or help guide long-term investments at the National Science Foundation in addressing critical challenges like global warming.

Technology Vector 3: Human-Machine Teaming at the Speed of Thought

All intelligent systems team with humans to some extent, either directly or through acting in accordance with human-specified goals. **The vision of human-machine teaming at the speed of thought is to create systems that can be trusted to understand human intent while collaborating to perform tasks that are difficult, dangerous or impossible for humans to carry out with speed and accuracy.** We envision a progression from humans utilizing machines as tools to humans interacting with machines as trusted teammates, and ultimately fusing with machines as seamlessly integrated extensions of our bodies and minds. In addition to the ambitious technological advancements we outline along this vector, creating effective human-machine teams will require appropriate calibration of trust across people and institutions. This will require new human-centered methodologies for test and evaluation, along with clear policy guidance for determining which decisions and actions are appropriate for an intelligent system to perform within a framework of societal values and ethics, and in accordance with the system capabilities.

Near-term: As a foundational element of human-machine teaming at the speed of thought, we expect that near-term research will focus on establishing common world models for machines and humans. Common world models will enable developers to effectively bootstrap learning systems with human knowledge while enabling operators to quickly establish shared situational awareness. While deep learning algorithms provide powerful pattern recognition capabilities, developers primarily interact with the system by providing labeled training data. Richer architectures for machine perception are needed to enable a higher level of intuitive collaboration between humans and systems. Enabling research along this vector will require development environments with human-machine interfaces that allow agents to select actions as a function of the intent and state of their human teammates.

Mid-term: The next step after building a foundation of common world models is to create systems with the ability to learn physical and cognitive tasks by observing and imitating human teachers and to self-improve through simulation. Continued advancements in natural language processing and human-centered analytics from industry and the academic world will assist in reaching these goals. However, targeted research and development will be necessary to bridge the gaps between generalized language understanding and mission-specific workflows and interaction paradigms.

Long-term: The intelligent systems of the future will serve as seamlessly integrated extensions of our bodies and minds, enabled by shared cognitive models. Achieving this human-machine fusion will require significant advances in the ability of machines to essentially read our minds — that is, to infer our intent through observation and interaction, enhanced by measurements of our physiological and cognitive states. Towards realizing this future, APL aims to continue advancing research in brain-computer interfaces, neurally-integrated prosthetics, and functional analysis of the brain, together with a focus on privacy and ethical issues.

Technology Vector 4: Safe and Assured Operation

Achieving assured operation of intelligent systems refers to the goal of developing systems that are robust, resilient and reliable across the full range of situations that may be encountered during the course of their intended use. It will become increasingly vital to advance the science and technology of assuring intelligent systems as we seek to employ them in safety-critical military and civilian applications. **The vision of safe and assured operation is to develop intelligent systems that are robust to the perturbations of real-world environments, resilient to adversarial attacks, capable of ethical reasoning and guaranteed to pursue goals that remain aligned with human intent.** In addition to the challenges of operating in the real world, we anticipate that resilience to adversarial attack will become increasingly important over time. Cyber attacks, in particular, have been in a constant state of evolution for decades and there is no reason to expect that this trend will slow down any time soon. So, we can expect that new ways of subverting the intended operation of intelligent systems through cyber attacks will also be forthcoming in the years ahead. At the present time, we see at least three approaches to altering the intended operation of a system. *Adversarial input attacks* are designed to evade the perception capability of an intelligent system. For example, at APL we have demonstrated that by placing a small patch on a person, it is possible to convince a computer vision system that the person is a teddy bear. Similar experiments show that using masking tape to slightly alter STOP signs can fool the perception system of a driverless car. *Data poisoning attacks* refer to attempts to alter the data on which a perception system is trained so that incorrect conclusions are reached. *Model stealing attacks* are aimed at understanding enough about how a perception system works to enable one to disguise objects and evade detection. Clearly, all three of these approaches are related and we can expect more sophisticated attacks in the years ahead.

Even without any interference from human adversaries, developing competent intelligent systems for real-world applications remains a fundmental challenge – even simple goals can be difficult to achieve in a complex world. For these resaons, we fully expect that safe and assured operation will be as much an ongoing pursuit as a strategic goal to be achieved.

Near-term: In the near term we expect that work will focus on identifying system designs that optimize the tradeoffs between modular architectures that emphasize interoperability and assurance and data-driven learning architectures that emphasize system performance. Bridging the gap between simulation and real-world testing will be key to enabling impact in complex, safety-critical applications.

Mid-term: Building on the foundation of assured single-agent architectures, the mid-term goal will be to develop multi-agent architectures that are robust against real-world perturbation and adversarial influence. We can expect that the adversaries will attempt to disrupt the operation of multi-agent systems by degrading collective perception and decision-making capabilities, as well as disrupting the ability of the components to team with other agents and humans. Research will also focus on characterizing and bounding the extent to which emergent behaviors arising from multi-agent interactions lead to unintended outcomes.

Long-term: The long-term goal of research efforts towards safe and assured operation is to produce systems that can achieve goals in challenging environments through creative problem-solving while producing outcomes that are consistent with human intent. Ensuring that systems with this degree of autonomy remain aligned with human-specified goals is a grand challenge that will involve decades of research. The emerging academic field of AI Safety has largely focused on existential risks that AI might pose to humanity. APL aims to engage with this community by helping to align AI Safety research with the practical realities and challenges of developing intelligent systems for real-world operation. Ultimately, we must work together across disciplines and institutions to ensure that advances in AI are beneficial to humanity's future.

The Four Technology Vectors

Technology Vector	**Near-Term** (1-3 years)	**Mid-Term** (3-10 years)	**Long-Term** (10-∞ years)
Autonomous Perception	Machines that learn to reason over world models and also learn strategies for focusing their attention	Machines that learn to reason about humans and other machines and learn to predict behavior	Machines that learn to perform causal and counterfactual reasoning and also learn by strategic exploration
Superhuman Decision-making and Autonomous Action	Machines that perform and select appropriate behaviors using data-driven architectures	Machines that generate and execute courses of action across multi-agent networks	Machines that reason and act strategically and learn across levels of abstraction and time
Human-Machine Teaming at the Speed of Thought	Machines that depend on human support, common world models, and labeled data	Machines that interact with humans through natural language and demonstration	Machines that primarily rely on human intent inferred through shared cognition
Safe and Assured Operations	Machines with assured system components and architectures	Machines with assured multi-agent architectures and resilience to adversary attacks	Machines with goal alignment, risk sensitivity, and ethical and moral reasoning

Note that, with respect to this chart, any particular project aimed at developing an intelligent system would occupy a vertical slice. In other words, a project is developed at a specific point in time and would make use of the most advanced family of capabilities spanning all four strategic vectors at that time.

Strategic Research Areas

Advancements across a broad range of technology areas will be necessary to realize the long-term goals laid out in the roadmap. Among many related pursuits, the following represent areas of increasing emphasis for AI-related R&D at APL.

Fundamental Advances in Machine Learning Algorithms: The promise of machine learning is to automate the development of complex functions that cannot be explicitly encoded. Leveraging the current landscape of machine learning algorithms is necessary but not sufficient to drive progress along our technology vectors. While there will undoubtedly be continuing research on new algorithms in both academia and industry, it will be critical to pursue new ideas in machine learning that are tailored to the unique aspects of key applications.

Supervised machine learning techniques, such as deep neural networks, are generally more mature than experiential learning techniques like deep reinforcement learning for real-world problems. Deep reinforcement learning has recently shown great success for decision-making when applied to board games and video games (like Go and chess), but applications of this general family of algorithms to APL problems will require robust world models as discussed above. Ensuring adequate representation of real-world conditions in a dataset or simulation is an open problem and active area of study.

Another aspect of fundamental research in machine learning algorithms concerns what APL technical experts call “enormously small” datasets that are clutter-rich and target-poor. Training an algorithm through supervised learning often requires millions of labeled images, for example. However, in some defense and intelligence challenges there are only a few hundred labeled images available. Major advances in this area are needed, such as learning from synthetic data, learning with fewer labels, and developing techniques for automatically labeling data.

Online machine learning, or “learning on the fly,” refers to machine learning from incoming data in real time. In this case, the algorithm must dynamically adapt to new patterns of data without forgetting previously learned information. This challenge applies to any real-time monitoring system, including ubiquitous checkpoint and cyber monitoring systems. In some domains, online learning will be especially difficult due to austere computing environments where size, weight, and power are limited. A danger associated with online learning, in general, is the lack of robustness to real-world conditions coupled with the potential for adversaries to subvert system goals.

Multi-modal and contextual learning is another important area of basic research for APL. Information used to develop and maintain situational awareness of adversary activities often comes from disparate sources, including, but not limited to, images, text, hyperspectral data, acoustical data, and radar. Effective integration of information from diverse sources requires contextual awareness of each modality for accurate reasoning. This challenge is also complicated by the fact that the data from some sources will be unstructured, noisy, and conflicting. Machine learning algorithms that can develop hypotheses by leveraging contextual information from multiple domains will be especially important in applications that involve maneuvering in cyber space, closed-loop intelligence, surveillance, and reconnaissance, and intelligent control of spectrum resources.

Human-Machine Fusion: Realizing the long-term vision for Human-Machine Teaming at the Speed of Thought will require a seamless connection between AI systems and the human brain. Developing a non-invasive brain-computer interface system is a promising path towards shared cognition and an active area of research at APL. Two important challenges in this area are physically accessing brain signals in a non-invasive manner and decoding these signals in a reliable way to understand human intent. Past APL research using invasive implants was focused on the motor regions of the brain. Future research will focus on accessing and analyzing signals carrying more complex thought patterns. Another essential challenge for brain-computer interface research is to send signals into the brain. A bi-directional non-invasive device could hold the key to dramatic advancements in education, health, scientific discovery, and military operations.

Enhanced by measurement and interpretation of neural signals, humans and machines will ultimately collaborate to make inferences and decisions based on incomplete and uncertain information. New algorithms and teaming paradigms will be needed under these regimes to ensure that decisions made and actions taken by human-machine teams are both maximally effective and aligned with human values. Therefore, in addition to continuing our research on the actual interface system between humans and machines, effective operational use of AI will require research on: integrating uncertainty into machine decision processes and communicating it to humans; conveying machine recommendations that consider information unseen to humans; “explainability” of the often opaque machine learning process; and ensuring that machine recommendations are consistent with human intent.

Novel AI Substrates and Neuro-Inspired Computing: A fundamental long-term research objective for APL is to use the multidisciplinary power of neuroscientists, computer scientists, and mathematicians to discover new computational paradigms modeled on the human brain. In pursuing this goal we will continue to explore the use of emerging databases of reconstructed neural circuits that has been created as part of programs like the Machine Intelligence from Cortical Networks (MICrONS) program. As we noted above, current approaches to deep learning tend to rely on heuristically constructed networks rather than exploiting a principled design strategy. Similarly, prototype neuromorphic processor platforms have the capacity to simulate neurons statistically, but draw little architectural inspiration from biological brains. Of the core principles believed to fuel the computational power and efficiency of the biological brain, arguably the least explored is the concept of motifs – repeated structural and functional computational units that can be observed at different scales and modalities. We believe that motifs account for some key gaps between today’s artificial neural networks and future computational architectures that truly approximate the human brain. Recent advances in brain imaging and subsequent image processing are making novel, large-scale structural and functional brain information available, creating the first significant opportunity for motif and computational architecture discovery. Connectomics refers to the emerging science of the interplay between brain structure, function, and connectivity across multiple scales. APL is a leader in this field and we are actively developing novel, scalable mathematical tools to extract knowledge from brain-mapping datasets. This work will provide critical insights into the highly efficient processing power of the brain and serve as the foundation for a new generation of architectural designs for truly neuromorphic processors. Combined with advances in low size, weight, and power hardware, connectomics will allow us to push machine perception to the tactical edge. Applications for this work will include detecting, tracking, and reacting to time-critical threats by carrying out intelligence processing at the sensor.

AI Safety Research**:** To establish trust in intelligent systems for critical applications, we must enhance our understanding of machine learning algorithms in both theory and practice. This will require an understanding of the extent to which AI systems are robust to real-world conditions, vulnerable to adversarial attack and ultimately, the extent to which they can be trusted to perform critical tasks autonomously. In particular, as the Department of Defense and the Intelligence Community seek to leverage AI to improve decision-making, automate defense systems, and reduce the burden on analysts, the requirements for security and robustness of AI systems will be increasingly important. Finding new techniques for identifying and controlling bias in AI systems will be key to facilitating near-term progress. Further, we must continue to work towards new methods that take advantage of the potential for intelligent systems to learn on the fly while mitigating the many inherent risks associated with doing so. Ultimately, we seek fundamental advancements towards developing self-aware systems that can identify and address their own vulnerabilities, explain their decisions, correctly interpret and safely execute human goals, and ask for guidance when necessary.

To help accelerate research in AI Safety and trusted autonomy more generally, Johns Hopkins has recently established the Institute for Assured Autonomy (IAA). IAA is an emerging national center of excellence ensuring the safe, secure, reliable, and predictable integration of autonomous systems into society by covering the full spectrum of research across the three pillars of technology, ecosystem, and policy and governance.

Team

Angelina Boampong	Christine Morris
Taylor Buck	Hitesh Patel
Kyle Casterline	Nathan Parrish
Bob Chalmers	David Patrone
Michelle Chen	John Pino
Matt Dinmore	**John Piorkowski****
Nathan Drenkow	Chris Ratto
Neil Fendley	**Matthew Rich****
Dean Fisher	Corban Rivera
Matt Giarra	**Pedro Rodriguez****
Kris Gibson	Heather Roff
Chad Hawes	**Jennifer Sample****
Jay Huang	**Jim Schatz****
Karl Hibbitts	Blake Schreurs
Kapil Katyal	Ted Staley
Marty Keutel	Lee Stearns
Nathan Kuo	Sarah Stevenson
Ian MacLeod	Bruce Swett
Dave Nobles	Mimi Szeto
Cara LaPointe	Michael Tabernero
Kevin Ligozio**	Anthony Tripoli
Ashley Llorens*	**Matthew Wagner****
Jared Markowitz	**I-Jeng Wang****
Jimmie McEver	Eddie White
Andrew Merkle	Mike Wolmetz

Additional ideas and contributions from numerous experts across APL

* Team lead
** Sub-team lead

References

Transforming National Security in the Era of Artificial Intelligence – APL White Paper prepared for General Shanahan.

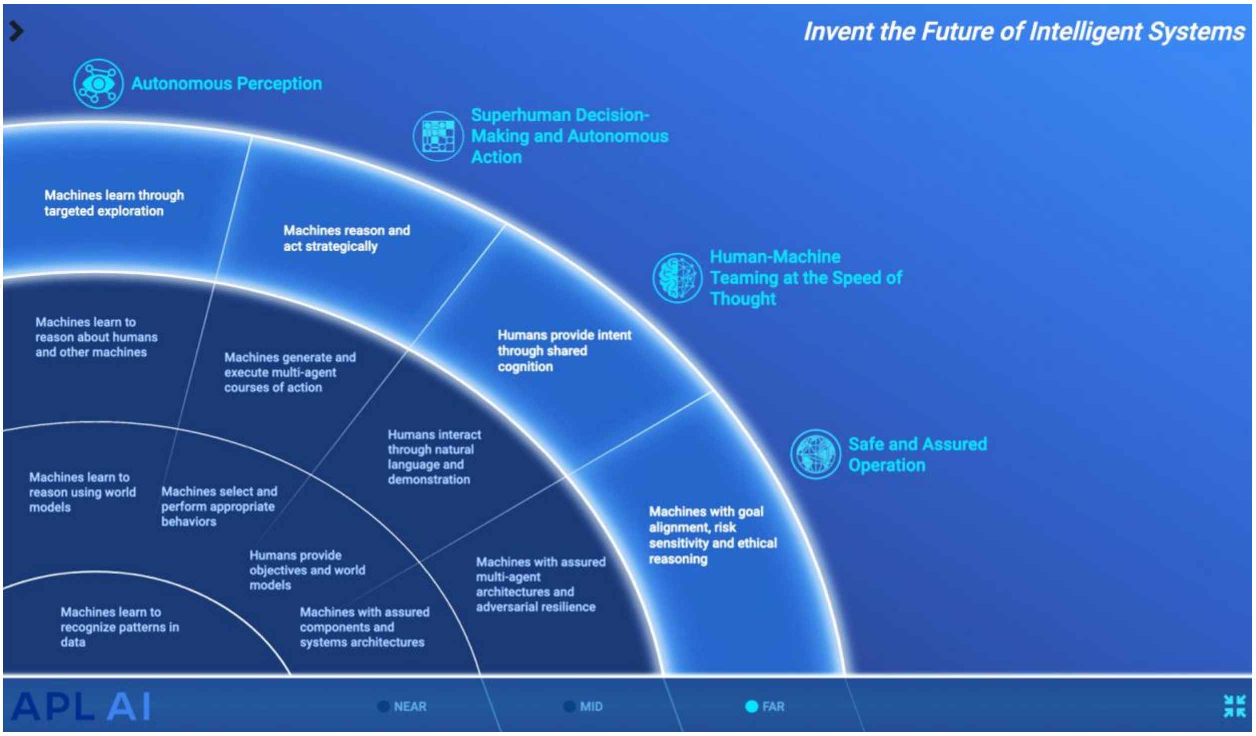
Rush to Military AI Raises Cyber Threats – *Defense Industry News* article by Theresa Hitchens, 4/26/2019 (quotes Fred Chang, former Director of Research, NSA)

Summary of the 2018 Department of Defense Artificial Intelligence Strategy – *Harnessing AI to Advance our Security and Prosperity*

A 20-Year Community Roadmap for Artificial Intelligence Research in the US – Community Computing Consortium and the Association for the Advancement of Artificial Intelligence

The One Hundred Year Study on AI – 2016 Report - Stanford University

National Security Commission on AI – Initial Report, July 2019



Invent the Future of Intelligent Systems

Autonomous Perception

Superhuman Decision-Making and Autonomous Action

Human-Machine Teaming at the Speed of Thought

Safe and Assured Operation

Machines learn through targeted exploration

Machines reason and act strategically

Machines learn to reason about humans and other machines

Machines generate and execute multi-agent courses of action

Humans provide intent through shared cognition

Humans interact through natural language and demonstration

Machines learn to reason using world models

Machines select and perform appropriate behaviors

Machines with goal alignment, risk sensitivity and ethical reasoning

Humans provide objectives and world models

Machines with assured multi-agent architectures and adversarial resilience

Machines learn to recognize patterns in data

Machines with assured components and systems architectures