Panel Summaries and Notes

First Workshop of Assured Autonomy

October 16-17, 2019
Sponsored by CCC

Summaries written by the panel leads:

- Laura Humphrey
- Howard Shrobe
- Lenore Zuck
- Nancy Cooke
- Nok Wongpiromsarn
- Missy Cummings
- Ashley Llorens

Notes taken by the CCC staff.

This note compiled by Ufuk Topcu.

Summary of the Safety, Verification, Certification Panel

Laura Humphrey, Phil Koopman, Darren Cofer, Julian Goldman

In terms of how assured autonomy is interpreted, several members of the panel noted that having an exact definition of autonomy is not necessarily the main issue; rather, it is recognizing that current verification and certification processes and practices do not apply well to the design methods and algorithms that underlie autonomous systems. For instance, a single test case can satisfy traditional software "code coverage" metrics (as in DO-178C) for an artificial neural network, yet it tells us almost nothing about whether the neural network is able to classify inputs correctly. In fact, it is difficult to know how much testing needs to be done to identify and eliminate faults in an autonomous system, since "edge cases" are hard to find yet can lead to catastrophic results if not found. The problem stems in part from the fact that autonomous systems are expected to act in dynamic, uncertain environments and also alongside humans, both of which are hard to model and interactions with which lead to huge system state spaces that are intractable to sufficiently test. A related point is that given the complexity of the operational environment, data-driven approaches are often used to train autonomous systems. However, it is difficult to know whether the amount, diversity, and quality of the training data is sufficient to cover all cases the autonomous system might encounter.

To address these issues, the panel expects that verification and certification processes and practices will need to change, which will also impact design approaches:

- System requirements will become more complex. They will have to precisely describe both the desired and undesired behaviors of the system while allowing enough degrees of freedom for the system to exercise autonomy.
- To help tackle complexity, we need to leverage more rigorous design methods that enable more automated analysis, including formal methods. We may also need more focus on system decomposition to enable compositional verification, both vertically (at different levels of abstraction) and horizontally (capabilities and system services interacting to form a system).
- However, rigorous design approaches such as formal methods need advancements. Such advancements include but are not limited to: an ability to analyze designs produced by data-driven approaches and machine learning (e.g. neural networks); to leverage machine learning and human intuition to guide and speed up analysis; to account for continuous dynamics of physical systems (e.g. reachability for hybrid systems); and to account for uncertainty (e.g. data-driven stochastic reachability).
- Methods and processes are needed to evaluate/verify/validate the quality and coverage of data used for training and to protect against rare events for which we have no current data.
- Autonomous systems will need to be outfitted with better monitoring and feedback mechanisms to detect invalid assumptions and gaps in coverage, as emphasized in UL 4600 and for medical applications in AAMI 2700-2-1 (under development).

- Also, autonomous systems or components of autonomous systems acting in the same domain should be designed to be interoperable both to share data on adverse events and to avoid unexpected, undesired emergent behaviors.
- Verification will need to be continuous across the lifecycle of the system to incorporate new data and address faults discovered as the system acts in real operational environments.
- Each autonomous system may be designed differently, precluding a standardized verification process. Instead, we may need to emphasize tailored safety cases and assurance cases.
- It may not be possible to verify that an autonomous system is sufficiently safe, in which case a run time assurance framework to safely bound behaviors of the system may be needed.
- Guidance on new approaches for safety, verification, and certification for autonomous systems is being generated (UL 4600, ASTM F3269-17), and there are defense-related programs exploring new approaches for assured autonomy (e.g. DARPA Assured Autonomy).
- There are also programs in the medical community that touch on issues related to safety, verification, and certification of autonomous systems (Federal Listening Session on Interoperability of Medical Devices, Data, and Platforms to Enhance Patient Care; Medical Autonomous Systems for Improving Healthcare Delivery; The Medical Interoperability Reference Architecture Research Collaboration), supporting research/tools (MD PnP Lab virtual hospital "sandbox"; Medical Device Interface Data Sheets; Clinical Scenario Repository; OpenICE open-source interoperability platform) and standards (IEC 60601-1-10, AAMI 2700-1, UL 2800, HL7).
- Finally, while the community is often focused on ensuring autonomous systems are not fielded unless there is ample evidence that they are safe, we should remember that *not* fielding autonomy also has safety implications. For example, it is estimated that there are 210,000-440,000 deaths due to preventable medical errors per year, which could potentially be greatly reduced if autonomous systems were more commonly used in the medical domain.

Summary of the Privacy and Security Panel

Howard Shrobe, John Launchbury, Todd Humphreys, Miroslav Pajic

Overall the panel dealt with the security challenges faced by autonomous systems. The panelists noted that autonomous systems are subject to the same security concerns as conventional IT systems plus several peculiar to systems that interact with the physical world.

Howard Shrobe began by providing a tentative definition of assured autonomy: “the expectation that a system entrusted to make decisions on its own will with high probability do what is expected and furthermore that it can explain its decisions. Secondly he defined a secure system to be one that will behave as designed and implemented even when attacked. As a consequence, system security is a necessary (but not sufficient) precondition for assured autonomy.

He went on to explain that system security can only be obtained as a “full stack” solution: vulnerabilities are exploited at many levels: processor, operating system, application software, language runtimes, etc. In addition, the manner in which components are integrated into the overall system has profound consequences. For example, in modern automobile, the “infotainment” system is connected to the control system; this made it easy to provide remote maintenance, but it also meant that security flaws in one part of the system can affect the behavior of a more critical component.

A second concern that Shrobe raised was the dependence of modern software systems on very deep supply chains. The development of a modern application involves the inclusion of dozens of libraries written by third parties; many of these are legacy systems whose provenance is uncertain. This means that developers need to worry not only about vulnerabilities in their own code but also about those in all the included software. As an example, he pointed out that several modern machine learning (ML) systems used to develop autonomous vehicles have vulnerabilities derived not from the code in the ML system but from image processing libraries loaded by the ML system.

Finally, Shrobe mentioned projects going on under the sponsorship of the DARPA SSITH project that are developing new hardware/software stacks that provide for inherent security against a broad range of attack categories. In particular he mentioned the HOPE project at Draper Laboratory and the CHERI project at Cambridge University.

John Launchbury presented a review of the NITRD workshop on AI and Cybersecurity, held in June 2019. The workshop’s purpose was to assess the key research challenges and opportunities in the interplay between cybersecurity and AI. Launchbury mentioned the DARPA Cyber Grand Challenge in which AI software autonomously crafted an exploit and then sent that exploit to a second machine in order to steal data resident on the second machine. A third machine saw that attack, reverse engineered it and applied a patch to protect itself from the attack. This all happened in 20 minutes, illustrating the power of AI techniques to operate at high tempo, both for the offense and the defense.

Launchbury pointed out the many traditional techniques for providing cyber security are difficult to apply in the context of cyber-physical systems. For example, traditional anomaly detection systems raise warnings when the system behaves outside of a pre-defined statistical profile; but in the physical world there is always a great deal of noise and uncertainty, making it difficult to distinguish unusual from malicious behavior. He the enumerated a number of challenges and potential new approaches:

- The use of cooperative and evolutionary game theory and multi-agent modeling. But he noted that in contrast to more traditional contexts, here the “game” is constantly changing.
- Machine Learning has provided a number of very powerful techniques but it is very fragile in adversarial settings and there are fundamental reasons for this that will be very difficult to overcome.
- Trust-worthiness in decision making is a fundamental requirement for autonomous systems; generally, we try to demonstrate this in test-beds. Even if the system can provide a rationale for its decisions, we still have the challenge of assuring that it would make similar decision in the real world.
- The use of out of bounds channels to get extra evidence (e.g. if GPS says you’re in the US but all radio stations are in Arabic, you’re probably not actually where GPS says you are).

- The absence of adequate testbeds, datasets, and tools.

Finally Launchbury described work in the DARPA High-Assurance Cyber Military Systems (HACMS) program which demonstrated that it was possible to do a cyber-retrofit on an unmanned Boeing aircraft.

Todd Humphreys talked about the challenges in developing self-driving cars. He began by pointing out that although almost all experts agree that self-driving cars must be more reliable than humans, they disagree on the size of the margin: Elon Musk - 2x; Ford - 100 x; Intel/Mobileye - 1000x. Humphreys' own estimate is somewhere between 100x and 1000x. The quandary is that to keep risk so low, automated vehicles must either drive uncomfortably cautiously (e.g. < 25 mph) or cooperate by sharing data to eliminate blind spots. He summarized this as: "cooperate, crawl, or crash". Humphreys offered several examples of where cooperation is beneficial: For example, suppose a car is at a green light but behind a bus. If the car, pulls out and passes the bus, it might hit an unseen pedestrian. But if another vehicle from across the street was sharing its data, its unobstructed view could have prevented the problem.

Cooperation, however, raises security (and other) concerns:

- Can the other party be trusted?
- Is the data exchange secure?
- Is the cost of the data exchange unaffordable?
- This approach requires precise pose information within common a common reference frame. Can this always be obtained?

A concern was raised from the audience that this approach seems to overlook privacy concerns. Humphreys responded that in current law, If you're taking data in public places, one doesn't have a reasonable expectation to privacy. However, these laws might change since when these laws where written we didn't have a panopticon. He conjectured that safety will trump privacy in such cases. He pointed out that one approach might be to transfer processed (and anonymized data, such as the location of a pedestrian but not the image of the pedestrian); but he pointed out that raw data is harder to fake. This is important since faked data can lead to dangerous misbehavior.

Miroslav Pajic addressed the problem of providing some level of quality of control even in these adversarial environments. He showed a video of an attack-resilient cruise control system. One key insight was that by using a variety of different sensors, even when some were under attack, they were still capable of providing the desired estimator. He also addressed the cost of this approach by pointing out that an adaptive cruise-control system can work effectively with only 10% of the messages. A second result was that if you authenticate only 20% of the messages, the attacker will either be detected or minimized.

The discussions centered around the following questions:

HACMS seemed to be very successful, what did we learn from it? Launchbury answered that *Cyber-retrofit* was a key concept that came out. Sensor security was one key. We developed some useful science but still need more. Last year's spending bill had language about HACMS and how to add them to similar projects/research.

The assumption that safety will trump privacy. Is it really a dichotomy? Can differential privacy offer anything? How about homomorphic encryption? Launchbury argued that you're always paying a cost for privacy. The question is to what extent are we willing to pay for it?

Homomorphic cryptography can provide strong guarantees but it only works if you find the exact right way to apply it. A general homomorphic encryption systems is orders of magnitude slower than normal computation. But, for example, Microsoft just announced a secure voting platform that uses homomorphic cryptography with affordable overhead by applying it to only a few specific computations. Shrobe argued that if you don't have the basics of a secure system, it can't provide privacy. If you need to be sharing info then you have a tension between what info to share and the privacy of that info. Humphreys argued that there is an inherent trade-off between safety and privacy. If you decide to blur someone's face before you share a video feed it reduces the receiver's ability to validate the video feed. Launchbury suggested that the privacy concern can be addressed by limiting the time that the data can only be retained to the short period needed to make a decision. He noted that there is some cryptographic research in this area.

*The problem of dealing with open-world environments****.*** Pajic noted that the first requirement is to have reasonable attack and threat models. But it was noted that validating these models is extremely difficult. Shrobe pointed out that usually the assumption is made that the attacker is only going after a single specific target; but a skilled attacker will attack multiple things at the same time. He also pointed out that it's very difficult to reason from first principles about very complex systems; instead we make abstractions to reduce the complexity, but these abstractions can hide critical issues. Launchbury pointed to the recent SPECTRE and MELTDOWN attacks; these are hidden by reasoning only about the macro-architecture abstraction. Finally Shrobe argued that you can never prove something is absolutely safe; but you can build something that is safe enough. Proving an entire complex system has no possible flaws is unlikely, but proving it would require a superhuman effort to find the flaw is probably good enough.

Summary of the Policy, Regulation and Ethics Panel
Lenore Zuck, Cara LaPointe, Nadya Bliss, Heather Roff

There is a common belief that security and privacy are conflicting terms. This is often not the case, since privacy can increase security.

Just like security, privacy is an afterthought. It's rare that systems are designed with security in mind, and less so with privacy in mind.

Let's refer to any data that may reveal information about an individual "privacy information." Note that this is not the strict definition of PII that, e.g., excludes biometric information.

In general, people don't want their privacy information to be leaked beyond their control. Yet, at times it is to the individual's benefit that private information can be transferred and easily access, most notably, in medical setup where the more information the treating team has about the patient, the better care the patient can get.
In contrast, companies are saving troves of private information on individuals, without a clear purpose in mind.

In between there's the government. For example, DOT has collected video tapes with driver's habits and behavior on the road. This data is only lightly anonymized and is, in principle, saved for a short duration.
Several points there:

- We don’t know how to anonymize data.
- We don’t know how to dispose of data, conceivably, what DOT can commit to is that they don’t know how to access the data, but not that it doesn’t exist after a certain point.
- Suppose there’re good solutions for (2), then should there be penalty for those who retain data beyond its necessary lifetime?

Yet another issue is that of privacy as an engineering, vs a policy, problem.

As to the “engineering” side:

- We should design mechanisms that will enforce restricted data collection
- We should design mechanisms that will enable using private information only for the purposes for each it is retained
- We need better mechanisms to prevent home devices (or other devices that are installed by choice) from recording private information. An example that was brought up was cameras in locker rooms.

As to the policy side:
Policies like HIPAA, in practice, prevent medical caregivers to obtain information they need while allowing leakage of information to where it is not intended. One of the issues is that HIPAA is not consistent (i.e., it may lead to different instructions/outcome for the same scenario) or complete (it doesn’t cover all cases). To scale, a policy has to be unambiguous and machine checkable. The bodies that usually determine such policy rare, if ever, include the “engineers” that can advise on what can be enforced.

While all the above issues already exist, the scale will be manifold larger when we deal with autonomous systems that can collect, and share, much more information. The main challenge is how to have these new systems with in-design privacy that will control and restrict the privacy information they collect and share, while allowing to “break the glass” when needed (medical emergency).

Summary of the Human-System Integration and Trust Panel

Nancy Cooke, Matt Johnson, Missy Cummings, Nisar Ahmed, Jessie Chen

How is assured autonomy interpreted in the context of human-system integration -- The term “autonomy” implies independence which may lead us to ignoring other agents that will be playing important roles in the system. However, as Matt Johnson articulated, “No autonomy is an island.” Autonomy is not independent; it is interdependent; and the need or interdependence (teamwork skills, communication, coordination skills) increases with increasing autonomy. Assured autonomy will need to team with humans and other autonomous agents. There is much more to teaming than putting agents together in the same room.

Key challenges --

- Assured autonomy requires human systems integration throughout the lifecycle of systems development. This means thinking about the role that autonomy will play and how it will interact with humans and other agents. Too often this is ignored. Human-autonomy interaction is considered as a last resort or a band aid.

- Autonomy often lacks transparency which negatively impacts human situation awareness.
- Algorithms that take in data can be very subjective, based on the choices that people make in the data that provide input.
- Standards for autonomy should not cost money and present financial barriers to good design.
- Computer science community seems to shy away from IRB.

The advances need to address these challenges –

- Need measures of team/system effectiveness.
- Machines need to be self-aware and know when they need to ask questions.
- Need models of reasoning under uncertainty.
- Need for interdisciplinary teams for assured autonomy. The [funding] agencies need to demand that experts on human-system integration be on the team.

The main take-aways of the panel –

- Assured autonomy needs to team with humans and human systems integration and team science are central to making this work. Design of autonomy is an interdisciplinary effort.
- Considerations of consequences for humans (i.e., bias) is important in selection of the data for machine learning.
- Autonomy can be a better team player with improved transparency, self-awareness (understanding of limits), and appreciation of teamwork.
- Measures and models are needed at the team/system level so that assured autonomy can be assessed in the intended ecosystem.

Summary of the Mobility Panel

Nok Wongpiromsarn, Natasha Neogi, Brian Sadler, Kevin Dopart

The panelists generally agrees that assured autonomy is a system-level problem that needs to take into account several considerations beyond science and technology. This includes public acceptance, regulation, ethics, balancing between rules and human behaviors, etc.

Kevin Dopart mentioned that the performance of motor vehicles depends on assured autonomy. He provided the United States Code for Motor Vehicle Safety (Title 49, Chapter 301), which defines motor vehicle safety as “the performance of a motor vehicle or motor vehicle equipment in a way that protects the public against unreasonable risk of accidents occurring because of the design, construction, or performance of a motor vehicle, and against unreasonable risk of death or injury in an accident.”

Natasha Neogi considered urban air mobility (UAM) and provided several types of barriers in airspace design. These barriers can be catagorized into

(1) UAM vehicle barrier: increasingly automated vehicle operation, vehicle design and integration, certification and operations approval, airworthiness standards and certification, manufacturing and supply chain, weather tolerant vehicles, and safe urban flight management),
(2) UAM airspace barrier: urban weather prediction, communication / navigation / surveillance / information and control facility infrastructure, safe, efficient, scalable and resilient airspace operations, fleet management, UAM port design, and operational rules, roles and procedures
(3) UAM community barrier: supporting infrastructure, local regulatory environment and liability, public acceptance, and operational integration, and
(4) UAM cross-cutting barrier: affordability, cybersecurity, safety, and physical security.

Moreover, UAM may be subject to several types of missions, including intra-metro air shuttle, public service vehicles, air medical transport, and package delivery. They also require infrastructure elements such as UAM traffic and weather monitoring sensors, recharging station, navigation tower, vehicle-to-vehicle communication, and UAM traffic monitoring station. As a result, this is a system-level problem, not a component level problem. We will limit ourselves if we think from a very narrow point of view.

She concluded that highly automated airspace and aircraft operations provide a path to address the UAM feasibility and scalability. However, enabling certifiable UML-4 operations requires fundamental changes to human and automation roles, responsibilities and authority.

Brian Sadler pointed out the challenge of unstructured environments and raised questions such as “Have we had prior access to the environment?”, “Do we know about the environment?”, and “Are we all able to navigate the road?” He further emphasized that risks must be known and that we need tools that enable us to physically reason about the environment.

Tichakorn Wongpiromsarn showed a video of autonomous driving in Singapore and pointed out that even during the quite hours, there were many instances where the autonomous vehicle needed to violate some traffic rules such as crossing the solid line to overtake an illegally parked vehicles. She emphasised that well-defined system specifications / requirements as well as validation methods are the key aspect of assure autonomy. The requirements need to have precise definitions and do not rely on the assumptions on human nature or capabilities. Furthermore, a hierarchy among the requirements needs to be defined for the case where not all the them can be satisfied. Finally, we need to establish validation methods that are objective and ensure sufficient coverage.

Several concerns were also raised by the audience, including
* The open world challenge leads to unreasonably large number of test scenarios and drives the question of edge cases. For example, from US data, we'll need 200 years mean time without disengagement for autonomous vehicles.
* The process of obtaining a driver license includes not only the written and driving test, but also the maturity of judgement with being 16. We may not be able to apply the assumption about human such as that of being 16 when writing a specification for autonomous vehicles. In fact, we only license, not certify, e.g., a pilot.
* Security is important and is not an after the fact question.
* We need infrastructure to help with human-autonomous vehicle communication. We have plan for vehicle-to-vehicle communication. But we have not quite considered vehicle-to-human or human-to-vehicle communication. This might require infrastructure support or change in behavior, e.g., slowing down to let potential jaywalkers know that the autonomous vehicle is about to stop for them.

Summary of the Space Panel

Missy Cummings, Danette Allen, Joel Mozer, Masahiro Ono

Dr. Allen discussed NASA's focus on boots on the Moon by 2024, with a permanent location on the moon a few years later. She emphasized performance and persistence in space and the need to do all autonomously, which will require the ability to fail in a way gracefully.

Dr. Mozer discussed the need for improved space situational awareness, as well as the difficulties in running concurrent operations, particularly in launching systems, that could greatly benefit from increased autonomy. He also emphasized the Air Force's need to develop an autonomy strategy in partnership with NASA, that could result in space trusted autonomy.

The last panelist, Dr. Ono, discussed the history of autonomy for NASA spacecraft, and posited that autonomous systems were a mainstay of NASA operations. He posited that the future needs of NASA in autonomy included trustworthy algorithms. His current concerns revolved around the difficulties in collecting sufficiently large datasets and restrictions and the culture surrounding third-party software use on-board spacecraft.

Dr. Cummings then asked a series of questions to the panelists that addressed their concerns and focus areas as they relate to assured autonomy, and then questions were solicited from the audience. The audience questions included an interest in hearing more about how the concept of heritage applied to system certification, and a discussion about how one can apply machine learning algorithms to Mars perceptions problems.

Main take-aways –

- Graceful degradation is going to be a critical element of assured autonomy
- Testing, validating and certifying autonomous systems are still areas of significant concern and they are going to be major chokepoints for assured autonomy, since they are critical for setting the boundaries of what it means to be assured.
- There are significant funding shortfalls in NASA (and other domains) that are obstacles to the development of the needed testing, validation, and certification protocols.
- Assured autonomy is a critical for the future of space operations but there are significant hurdles that must be surmounted to fielding such systems, not the least of which is cultural.

Summary of the Defense Panel

Ashley Llorens, Signe Redfield, Craig Lennon

Ashley Llorens began the panel by giving several examples of autonomous systems in defense applications including ground, air, space and maritime. Any autonomous system operating in a real-world environment faces an "open world problem" – that is, novel situations that systems encounter that were adequately represented or explicitly accounted for during the design, development and testing of the system. Defense applications often face an even tougher challenge in that operating environments tend to be even less structured and predictable than in commercial applications and may involve interactions with adversaries. Ashley gave an example of a marsupial robotic system – that is a nested team of ground and air robots – exploring an unknown indoor space. The system has the ability to navigate a cluttered environment through by deploying ground and air robots, but only has a limited ability to manipulate the environment. For example, the system can open some kinds of doors, but not others. Ultimately, there is a critical need to characterize and bound the performance of autonomous systems in their intended operating environments to enable the appropriate calibration of trust in the system by both institutions and human operators.

Signe Redfield spoke next and added further clarity to the unique challenges associated with of assuring autonomous systems, particularly in defense applications. She noted that verification of autonomy is not a traditional research domain in academia and that defense applications of autonomous systems present more difficult regimes with an even greater need for reliability. As compared with academic and industry applications, defense applications of autonomous systems:

- Can require operation durations of hours to months
- High variability environments
- Intermittent frequency of modification
- Low tolerance for failure
- Catastrophic consequences when mistakes are made

Ultimately, Signe identified the need to verify the safety, security and functionality of autonomous systems:

- Safety – includes safety of subject, safety of environment, safety of robot, safety of operator, safety of bystander – research focus typically on safety of subject (when "safety" explicitly considered) and safety of robot (during autonomy design)
- Security – critical need for defense-related systems.
- Functionality - critical problem not typically addressed by current verification tools; requires greater degree of certainty in some cases; presents unique challenges for learning systems.

Lastly, Craig Lennon gave a summary of efforts to create a roadmap for test, evaluation, verification and validation (TEVV) with the Department of Defense. The goals of this effort are to create:

- Verifiable requirements specifications:
    - Functional and traceable requirements language for describing desired behavior for autonomous systems
    - Algorithms for autonomous behaviors which facilitate verification
- Design and arguments for dependable autonomy:
    - Methods of designing systems for assurance
    - Arguments integrating safety, security, and software reliability
- Instrumented and measured autonomy:

    - Autonomy instrumented for evaluation and methods for measuring human-machine interaction
- Safe development:
    - Methods to protect life and property during experimentation, testing, and training with autonomous systems
- Infrastructure & tools for TEVV of autonomy:
    - Combine virtual and physical experimentation for effective and efficient testing of autonomous systems
- Dynamic assurance for adaptive systems:
    - Assurance methods for systems which are adaptive or being employed in novel environments

Assured Autonomy Workshop #1
October 16, 2019

Panel: Mobility

- How do you evaluate Assured Autonomy?
    - Kevin- The performance of motor vehicles depends on assured autonomy
    - Natasha- Think about assured autonomy in a system
        - There is entire airspace out there- safety, transportation, local regulatory environment
        - Uncertainty in the form of weather predictions
        - There are many cross cutting pictures, there are public usage factors
        - Infrastructure and certification
            - This is a system level problem NOT a component level problem
        - We will limit ourselves if we think from a very narrow point of view
    - Brian- Unstructured environments
        - Have we had prior access to the environment, do we know about the environment? Are we all able to navigate the road?
        - Constraints- We want to have constraining forces. My constraints might be time variant.
        - Risks must be known. Operate in different rules of engagements.
        - The way we are going to get around this is through diversity. This is also tied to the adversary question. Not everyone in the world is as ethically constrained as we are.
        - In order to do this, I need tools that allow me to access physical reasoning problems. We lack the ability to physically reason about the environment.
    - Tichakorn- Video of autonomous vehicles in Singapore
        - The debate about how best to handle various situations is ongoing
        - For Assured Autonomy
            - We need a well-defined system specification/ requirements
                - Precision definitions
                - No assumptions on human nature/capabilities
                - Rule hierarchy (not all the rules can be satisfied)
        - Validation methods
            - Objective measure
            - Sufficient coverage
    - Term Formalization
        - What is the right way to turn on the signal?
    - Behavior Specifications
- Discussion
    - Ashley- Open world challenge of taking systems. How much prior knowledge do you get from these situations.
        - Tichakorn- Various scenarios test different examples. Driving exam test.

        - Kevin- That drive the question of edge cases. Germany companies that look at 10,000 test scenarios. The current human vehicale system in the US, is 200 years between failed/error. We have a long way to go.
    - Matthew- Are we testing apples/apples here or are we testing something else?
        - The driving test
            - Driving test
            - Written test
            - Maturity of judgement with being 16
        - You license a pilot you don't certify a pilot
    - Brian- It is not an after the fact question. Security is important. Every aspect of the research/every sensor of the process. Another aspect of it.
    - Kevin- It is going to be limited driving abilities.
    - Frederick- How do you plan for an autonomous vehicle and the human communication in driving (nodding, waving... for example)
        - Kevin- We could plan, low latency, vehicle communication
        - Brian- I need a crossing guard advice. We need infrastructure like that. It is not a long term question it is an immediate question. We need to spend more on the infrastructure.
    - Tichakorn- How do we communicate our future? How future cars are going to predict what we are going to do? How do you communicate? This might go in conflict with the rule.
    - Danette- NASA
        - What kind of systems do we need to put into place to cap ground control?
        - Ground control radar.

Panel: Privacy and Security

- Howie Shrobe: this panel is representing challenges rather than needs
    - Assured autonomy - the expectation that a system entrusted to make decisions on its own will with high probability and can explain its decisions
    - A secure system is what that will behave as designed and implemented even when attacked. Can't have A without B
    - example of hacked autonomous car from 60 minutes
    - this is an across the stack solution
    - vulnerabilities are exploited at many levels: processor, software, integration (e.g. entertainment system was connected to the control system - made it easy to maintain but it was a larger system in which the security can affect one part of it)
    - how many lines of code are yours vs. somebody else's? Even if your code is good need to worry about the others
    - In a vehicle you have 100s of processors and lots of hardware
        - vulnerabilities in language libraries (e.g. LibC, LibPython)
        - supporting libraries (e.g. for image processing)
    - vulnerabilities scale with lines of code
    - ex. tensorflow vulnerability list

    - project going on at CSAIL, Draper, and MIT-LL to address this problem. Design hardware that tags everything, bring in metadata and policies for that metadata. Come up with new ways of building the stack. Rather than a single kernel, use distributed setup
    - computer hardware and many language runtimes don't represent or properly manage important semantic distinctions
- John Launchbury
    - briefing on NITRD workshop held in June. approx 70 people. brought together cybersecurity and AI experts
    - workshop purpose: asses the key research challenges and opportunities in the interplay between cybersecurity and AI
    - when we think about cybersecurity traditionally we think about hardening/closing things
    - ex. AI looked at code and crafted an exploit, then sent that code to another machine which stole that data. A third machine saw that attack, reverse engineered it and applied a patch to protect itself from the attacker. This all happened in 20 minutes
    - You might have an AI developer sitting beside you, noticing the things that you are doing
    - responding to (semi-)autonomous cyber action. Might be attacked by bot systems. Need to be able to respond at speed
    - detecting "real" anomalies is difficult since there are many "anomalies"
    - system could be watching the news to see where the new attacks are coming from
    - game theory tie in: cooperative and evolutionary game theory and multi-agent modeling. The "game" in this domain is changing all the time
    - ML is very fragile in adversarial settings. There are fundamental reasons for this
    - Notion of trust-worthiness in decision making, might know how it's making those decisions but won't know how those decisions will apply in the real world
    - might want to compartmentalize parts of the world (e.g. languages, driving, etc.)
    - we demonstrated that we could do a cyber-retrofit on the unmanned Boeing aircraft
    - using out of bounds channels to get extra evidence (e.g. if GPS says your in the US but all radio stations are in Arabic, you're probably not actually where GPS says you are)
    - need testbeds, datasets, tools
- Todd Humphreys
    - how reliable must self-driving cars be?
        - Elon Musk - 2x human reliability
        - Ford - 100 x
        - Intel/Mobileye - 1000x
        - my guess is somewhere between 100 and 1000 x
    - to keep risk within tight allocation, automated vehicles must either:

        - drive uncomfortably cautiously (e.g. < 25 mph)
        - cooperate: share data to eliminate blind spots. This is where security concerns come in
    - autonomous vehicle trilemma: cooperate, crawl, or crash
    - 2 “sensorias”
    - e.x. you have a green light but you’re slightly behind a bus. If you speed past the bus you might hit an unseen pedestrian
        - if another vehicle from across the street was sharing it’s data it could let you know that a pedestrian is trying to cross because it’s view was unencumbered
    - what is cooperation undesirable:
        - cost of data exchange
        - difficulty of ensuring trustworthy data
        - requires precise pose within common frame
    - Question: you seem to be eliminating privacy?
        - yes, I’m just focused on the security aspect. If you’re taking data in public places I’m assuming you don’t have a reasonable expectation to privacy. However, these laws might change since when these laws where written we didn’t have a panopticon
        - privacy concerns are greater than this. Do pedestrians have a right to not be recorded by cars everywhere they go? My conjecture is that safety will trump privacy in this case
    - structure emerges in different sensing modalities
    - insights gained so far
        - cooperation is essential to extremely low risk
        - data integrity will be a key challenge
        - raw data easier to validate harder to fake
        - managed data exchange
    - Q: Why is fake data a concern?
        - it’s for safety primarily. Imagine I’m doctoring my video feed that enables you to see in your blindspot
        - Launchbury: I could send out fake data to make cars move for me for example
- Miroslav Pajic
    - how are we providing some level of quality of control even in these adversarial environments
    - attack-resilient cruise control demo
        - we were able to play with the robot for a while
    - Next thing was on an american built car. Here we didn’t have the car we used models
        - use a lot of different sensors, even when some were under attack still had desired estimator
    - how are you validating trust in a cost-effective manner?

    - I will be able to do adaptive cruise-control with 10% of data messages (?)
    - non-secure vehicle platooning. If you authenticate on 20% of the data the attacker will either be detected or minimized
    - how are building to improve context and have a higher level of confidence you are not under attack
    - security-aware human-on-the-loop planning
- Todd- 3 of you were involved with hackems and I was struck that if they could meet these goals we would solve all these problems. What was the outcome? What did it reveal?
    - Launchbury: I think when hackems were successful at DARPA. Cyber-retrofit was a key concept that came out. Ex. Earthquake protection, it's not enough just to put a lock on the door, there is seismic retro-fiting. We did the cyber version of that with the Boeing UAV
        - sensor security was also key. Developed some useful science but need more
        - Last years spending bill had language about hackems and how to add them to similar projects/research
- Heather: I want to question the assumption that safety will trump privacy. Why do we have to make it a dichotomy? Anything with regards to differential privacy etc? Designing for instead of trading off
    - Launchbury - you're always paying off some kind of cost for privacy. The question is to what extent are we willing to pay for privacy. Homomorphic cryptography only work if you find the exact right way to apply it. We explored designing a general systems but it was a million times slower than normal computation
        - Microsoft just announced a secure voting platform that uses homomorphic cyptography on only a few specific computations so that's promising
        - There will need to be a jump in engineering
    - Howie - if you don't have the basics of a secure system it can't provide privacy. If you need to be sharing info then you have a tension between what info to share and the privacy of that info
    - Launchbury - it will be critical to have legal frameworks in place for privacy
    - security by obscurity means privacy
        - Todd - I'm talking about safety vs. security. If you decide to blur someone's face before you share a video feed it reduces their ability to validate the video feed
    - Miroslav - what is the level of privacy that we can tolerate such that we can provide these safety critical services
    - Launchbury - how long will I allow my data to be available. Say it exist for 10 seconds and then must be deleted
        - Are there ways to do that?
        - There is some cryptographic research in these areas
    - Alwyn - things become tricky in open world environments

        - Miroslav - you have to have some kind of attack model and threat model
        - Alwyn - How do we validate that?
        - Miroslav - need to look at some kind of cross-layer
        - Howie - if the attacker is attacking multiple things at the same time there are more challenges. Can you reason from first principles about very complex systems
        - you give up some privacy to interact in the social setting
        - Launchbury - nobody worried about SPECTRE and Meltdown because it was theoretical...until it wasn't
        - Howie - you can never prove something is safe, you build something that is safe enough. Proving an entire complex system has no possible flaws is unlikely, but proving it would require a superhuman effort to find the flaw is probably good enough
        -

Breakout Session 1 (Dan Lopresti, Matthew Johnson, Joel Mozer, Howard Shrobe, Heather Roff, David Kuehn, Erik Blasch, Meeko Oishi, 3 other people)

What are the goals of "assured autonomy" along the following dimensions?

- **Societal**
- **Programmatic**
- **Infrastructure building**
- Any other dimensions we should be thinking about?
- Dan - I think programmatic are the kinds of research investments that need to be made to push forward the technology
- Heather - I see programmatic and infrastructure not being very different
    - everything is part of society so…
    - how is this different from breakout 3
- David - Health and safety would be one area
- Erik - if you go to the locus point there is no infrastructure so it's really simple. If we go all the way out on infrastructure then everything is interconnected and is very complicated
- Meeko - a lot of the hullabaloo about autonomy is that it will provide new function that don't exist today. Is that part of our charge to address?
- Erik - the error domain has allowed us to have more efficient air travel. The goal is that autonomy would enable some service in society. What will it actually do for us? Not just about building more efficient autonomous systems
- Heather - job scaling in the workforce due to increased autonomous/AI systems. Would it be bad then to create these systems?
    - Hard to address because autonomy can mean different things and apply to different domains
    - Heather - true, washing machines freed women from home labor for example
    - anti-lock brakes were safer structurally but lead to people driving more

        recklessly, more people driving off the road.
            - If you have a car with 360 vision you might choose to drive 100 mph because I can see the road is clear for example

- Heather - I don't know if talking about the end state as the goal is particularly correct
- Howard - there are different readings of what these 3 things mean
- Some say that if you have an autonomous car it will reduce fuel consumption, while others argue rather than pay for parking they will just send their car to circle the block until they come out
    - Heather - there are also tertiary effects. Less organ transfers and therefore more deaths from organ failures. More crime because people can more effectively do criminal activity
    - Erik - Could you prevent that from happening
        - once you provide a backdoor your system is not secure any more
- Air taxi lands on someone who's not traveling
- Heather - autonomous shipping of transport goods
- Erik - some research goals aren't necessarily technology goals. What the funding agencies want is the fundamental math/science to explore these areas
- David - in some ways I think our models capture assured autonomy better than they capture real autonomy. Not good at capturing failure cases
- Heather - we have a pipeline problem. Need interdisciplinary background. Create more programs that intersect
- David - there are some job classes where taking people out of situations is the goal. We want miners (or minors?) operating the equipment from the office rather than on the floor
    - Erik - there are things this country has cleaned up that other places haven't. Bring up the quality of life, not necessarily income but in other ways
- Howard - we're not on an incremental improvement curve. We need to rethink the whole way of doing things
    - also every job may not be able to be made autonomous
- Heather - thinking only about the singular task and ignoring the external impacts will lead to more of these problems
- David - if we don't work toward short terms autonomy we have adversaries that are, so it may be worth the risk in the broader goal of national security
    - Heather - They are looking at autonomy, I don't know about assured autonomy per se
- David - harms from the internet tend to not immediately kill people unlike these systems
- two levels of goals: assurance it will do what you want and assurance it won't do something malicious
    - David - I would say you can assure it does X, but it might not do good things (due to individual use cases)
    - social integration piece: benefits of injecting that system into society isn't more costly than the benefits provided by the system

- sometimes you can safety through operational restrictions
- blue sky of robotic surgery. Before you automate it, might want to make sure it's the right choice to begin with
- Meeko - some sort of structure for asking questions at that interface about societal implications
    - David - needs to be interdisciplinary at the design level but also monitoring it in practice
- ultimately there are systems we cannot assure
- how much of these discussions is about assurance vs. autonomy
- hard to metricize something and say this definitely made things safer
- we live in a different era for risk. Much stricter regulation in domains like aviation. This is an open societal questions but there are these very high levels of assurance present
    - David - what about a domain that doesn't have a standard. Say commercial space travel
    - current space travel is very unsafe, but there is a perception in the public that this a relatively safe thing to. Public perception that any commercial space travel will be held to a very high standard
- Erik - I know in Canada they don't have flood insurance because they don't authorize it. After certain floods the government just tells people you can't rebuild here, it's too flood prone
- Heather - we know consent based models fail all the time. No one reads the OS update info when they need to update their phone
-

Breakout Session 2

(Albert Wavering, Sushil Birla, Laura Humphrey, Hiro Ono, Nok, Danette, Nancy, Nadya Bliss, Ufuk, Signe)

Topic: Why is assured autonomy becoming critical now?

1. What has happened the last 5-10 years that makes "assured autonomy" an issue?
2. What are you expecting in the next 5-10 years that will make "assured autonomy" an issue?
3. What will happen if we don't care about assured autonomy now?

Discussion

Question 1: What has happened the last 5-10 years that makes "assured autonomy" an issue?

    - Brief answer: Autonomy is mature enough to field but hard to V&V. Elaboration of the answer:
        - Economic incentives for autonomy:
            - Cost of enablers (e.g., elements for computation, storage,

communication, sensing) has come down.

- Economics of retrofit market: Seems much cheaper to replace human with automation (when risk of inadequate assurance is not considered).
- Economics of new applications, e.g.: delivery by unmanned aerial systems (drones); driverless hailed rides (Uber; Lyft; etc.); other kinds of autonomous vehicles.

- Lowered barriers to adoption of Autonomy: Comfort zone; willingness to accept or try …:
    - Ubiquity across society (smartphones, home robots, hobby drones) faulty, classification algorithms, agriculture)
- Recent awareness of the risks (adverse consequences such as fatalities in automated vehicles) when Autonomy is not ASSURED adequately.
    - Realization that it is in the best interest of the concerned parties.
- Technological enablers for better Assurance are emerging:
    - Capability to understand risks (and associated hazards).
    - Technical services available to analyze risks (and associated hazards).
    - Capability to control (i.e., eliminate, avoid, mitigate) hazards rooted in engineering deficiencies.
    - Other technological enablers are emerging from defense applications.

Question 2: What are you expecting in the next 5-10 years that will make “assured autonomy” an issue?

- Rapid change.
- Demonstrations of more than one system at a time.

2. Increasing complexity.
3. Lack of theoretical foundation and analytical methods for assurance-> empirical expensive.
4. Empiricism bites- we can’t do this for an autonomous system that doesn’t have any machine learning.
    - Yes, there are additional problems but we haven’t dealt with the fundamental ones yet.
    - The learning piece changes things.
5. Learning from data exacerbates the problem.
6. From the robot standpoint-things are different.
7. Computational limitations.
8. Human- Automation Interaction.
9. The availability of sensors/data has now opened the door at a lot of data. We have an explosion of data.
    - Issue: Inadequate validity of the data.
    - Issue: Limited information value without the context of the data.

- Need to understand the context of the data.

Question 3: What will happen if we don't care about assured autonomy now?

- Could lose the business case.
- Could lead to "Autonomy winter" (an analog of the "AI winter").
- Catastrophic accident

Breakout session 2 items not aligned with the three assigned questions (CCC might find other use for this information):

1. We are doing autonomy with real data and it is showing. Our data itself is not always at it needs.
2. We lack certifying agencies.
    - FCC.
    - It is not just the algorithms.
    - You can start to build in lessons learned.
    - I don't have answers for you, but I do have questions that you should be asking along the way.
3. 5-10 years?
    - IEEE guide.
4. Signe- How much work is done that we can't understand or reach?
    - Policy and regulation provides guardrails.
    - What is going to get people so alarmed that will demand some kind of insurance?
        - If you are not smart enough about doing something a little more pro-activity, then perhaps we do it.
    - The certification agency could address all of these systems.
5. Danette- It is going to take a catatrasphone
    - Something bad is going to happen in 5-10 years, then in 5-10 years after that we will have regulation.
6. Government or other certification/regulation
    - May have to be application/specific domain
    - May be driven by lawsuits & courts
    - How do you evolve you legal and ethical frameworks?
    - Policy is your strategy
7. NASA- will have to get ahead of this. If you want to go from the moon to mars. We need to convince ourselves. Maybe in 5 years there will be a change.
    - Hiro- Autonomous driving.
8. Defining safe boundary of system operations
    - Nature is an adversary for a lot of problems.
9. Humans are jerks.
10. Hiro- if there is any essential difference in AI, then perhaps we can have to switch.

Breakout Session 3

(Julian Goldman, Lenore Zuck, Kevin Dopart, Miroslav Pajic, Philip Koopman, Nisar Ahmed, Frederick Leve, Sean Phillips, Todd Humphreys, Missy Cummings, Ann Schwartz, Ashley Llorens, Jessie Chen, Craig Lennon, Alwyn Goodloe)

What are the different dimensions for assured autonomy, i.i. What factors make achieving assured autonomy a challenge? What are the dependencies, if any, between dimensions?

Dimensions:

Todd - safety, security, integrity, privacy

Lenore - how do you explain, what does it mean that this type of system is safe to a person who has no understanding of systems?

Julian - assured autonomy requires safety, security, integrity, privacy; is this necessary and sufficient?

Todd - these are the additional requirements on top of the operational requirements

Alwyn - functionality has to deliver these 4 things

Ashley - performance, than attributes of how you achieve performance

Sean - if you apply these, but doesn't do what it's supposed to?

Lenore - performance is how well does it do the job?

Functional - it does what it does

Performance - how well does it to do what it does?

Julian - a system that supports assured autonomy must also provide properties in the following areas:

Frederick - there needs to be domain dependence, it separates general intelligence from autonomous systems

Domain dependence should be understood / incorporated

Ashley - list topics, then add "within an operational domain"

Jessie - And within stakeholders

Alwyn - different level of expectation of reliability (parachutes); operational context with functional requirements; what levels of safety and requirements

Todd - running danger of attacking a problem that's too broad

Ashley - value to get to the dimensions; we've articulated one of the major problems: specifying a domain

Todd - pick a domain and use it as emblematic: anesthesia

Julian - Assured autonomy as a system that supports assured autonomy must provide properties in these areas (safety, security, integrity, privacy); but perhaps must be able to provide **guarantees** in these areas. Those guarantees are driven by the domain itself. May be probabilistic guarantees.

Ashley - notions of graceful degradation and failsafe

Frederick - runtime assurance

Lenore - with a grain of salt (probability)

Todd - Nok - there's going to be a priority of goals, and you will break some of your rules, just hope to break highly prioritized rules last.

Ashley - shared situational awareness between the human and the machine

Philip - this is the solution space and not the problem space. You don't have to have graceful degradation, you may want it.

Frederick - on specific missions, there are some very stringent constraints → constrained problem

Philip - is that autonomy? Autonomy generally implies opening in an open world

Ashley - operating in uncertainty

Philip - an open world, there are shades of gray - are assuming from square one that you will not know all that could happen

Frederick - anything that you specify is wrong; you have context about the domain (submarine, don't need to worry about air dynamics but do need to worry about fluid dynamics); can have specifications on things you know, but there will always be unknown unknowns.

Challenge: Definition of autonomy, uncertainty of scope

Julian - autonomous medical systems if we have the right drivers

Lenore - add compositionality

Julian - anesthesia systems are autonomous within a closed system

Miroslav - what is the vision vs. state of the art?

Julian - state of the art is very limited, but next part of the complexity is running

- How to help blood pressure while keeping patient asleep and fluids at right level...

Missy - navy is looking at it (change in respiration, etc.) / change in need for prolonged care (72 hours) in the field - automated critical care unit

Todd - act of intubation is extremely difficult; who would do it?

Julian - one of the hardest problems is patient assessment; can be made easier by smart clothing, etc. - it's all contextual information.

Todd - goals (based on specification and requirements documents), models, tools, impact

Frederick - domain is important!!!

Lenore - can't do goals and tools separately

Todd - important to decouple them

Ashley - assurance (Oxford): a positive declaration intended to give confidence

- Human beings relation to the system
- An assured system is one that can be trusted to satisfy agreed upon guarantees

Julian - why do we need assured

Ashley - now using just autonomy

- Unique challenges - open world, impossible to fully specify the operating domain, large number of possible actions to the agent, …, effective communication to the human workflow
- (check with him for his notes)

Julian - no funding agency has ownership of the medical space

Alwyn - from NITRD standpoint: hoping we can have academics look at safety and security levels that have to be met (assurance may not be defined the same way by industry or regulatory, but there are issues which must be met, and academics can be working on)

Julian - do we want to see applied research with deliverables that can be adopted OR theoretical, forward looking

Frederick - by bringing up the domains, we're sometimes able to find thing to abstract away

Breakout Report Back

- Breakout 1:
  - need for assured autonomy to be interdisciplinary
  - need for continued monitoring or traceability of assumptions and system bounds everyone had considered at time of time, through its lifecycle
- Breakout 2:
  - interest now because the technologies are established enough to extract economic value
  - human-automation interaction, transfer of control back and forth, quality of data, inadequate validation
  - need ability to better define safe boundaries

    - more and worse catastrophes if we don’t start caring about assured autonomy now
- Breakout 3:
    - system degradation outside of system specifications
    - impact on society
    - appropriate scoping

Panel: Safety, Verification, Certification

- Darren Coffer:
    - who would put AI in a safety-critical application
    - how is “assured autonomy” interpreted in your field (ex. aerospace: set objectives target, avoid other aircraft, route planning, etc.)
    - perception problem is the most difficult to deal with. How do I take that sensory input and deal with it
    - assurance is compliance with air regulations using accepted means of compliance. Legally we have to comply with this. Different parts for different kinds of aircraft
    - how do you know you’ve done enough testing? We have some rigorous structural coverage metrics to evaluate inadequacies
    - how do I achieve the equivalent assurance for those underlying assumptions
    - recently worked with NASA on a certification consideration for adaptive systems.
    - DARPA program called assured autonomy: new test methods/test generation and metrics, and applying formal methods to the analysis of certain methods
    - ASTM F3269-17 standard is floating around now for UAVs. Bounded autonomy approach
- Julian Goldman:
    - preventable medical errors are the cause of 210 - 440k deaths per year (45 per hour)
        - another study needs to be done prospectively, the current data is very soft but even if the number is off by 100k or so it is still one of the top causes of death in the US. #3 currently
    - clinical environments are not engineered
    - scene in an OR has not changed since I was trained (except for the colored monitor, used to be monochome)
    - what if integrating clinical environments can enable apps to rapidly and safely implement solutions based on the intended use of the system
    - patient controlled analgesia setup: currently nurse at monitoring station monitors your blood oxygen level from a desk. Sees a spike and runs down the hallway to see that it is a false alarm because you are gripping bed, phone etc.
        - family member pushes PCA button while patient is asleep. Since there were a number of false alarms the nurse takes her time, but now it’s too late and the patient has overdosed and died

        - you could instead monitor 2 signals at the same time. When both go off the PCA stops delivering analgesia and sends the nurse.
        - too much noise pollution in hospital
        - difficult to get accurate measurements
            - only 5% of medical devices have the right clock time
            - taking measurements is very error prone, e.g. blood pressure can be affected based on arm positioning
            - pulse Ox averaging time affects the reading
    - Gaps/needs:
        - adverse medical events are not shared and analyzed at a a national level. FDA oversight limited to medical devices
        - How can emergent hazards of heterogeneous point-of-care composable systems be identified
        - NITRD RFI released last year and a workshop was held in the summer.
    - built a virtual hospital sandbox. clinical scenario repository
- Philip Koopman
    - brute force road testing - good for identifying easy cases (but 1 billion ++ miles). expensive and potentially dangerous
    - event if you go to simulation (removes danger) you can't get an accurate open world
    - novel objects are "triggering events" (e.g. furry conference in pittsburgh every year)
    - Mask-R CNN hides systemic problems
        - can't detect "camouflage" (similar clothes to background object), children next to adults, bare legs (I think because they did the training data in the winter) and khaki pants (vertical brown things are trees), sun glare, red objects in certain context (e.g. 5 signal traffic light with red on top is person), people next to columns, construction workers
    - UL 4600 key ideas
        - goal: structured way to argue that AV is sufficiently safe
        - driver is a fault absorber. Drivers do more than drive (e.g. get out to clear snow)
        - system level safety for autonomous operation and lifecycle
    - what does "unreasonable risk" mean?
        - better than human drivers.
            - 94% is human bad choice is a big mis-quote. Which human drives do you mean? all drivers include drunks, 16-year old etc
        - insurable
            - cost is not much worse than human drivers. Cost of settling with victims < cost of the fix (this actually happens, economically that's what you're incentivized to do)
        - satisfies a combination of moral and economic imperatives
            - "stop bothering us with assurance, we're too busy saving lives"

- Meeko Oishi
    - What is assured autonomy?
        - Probabilistics assurances in dynamic, uncertain environments
            - Interactions between
                - Dynamical components
                - Complex mode-logic
                - Learning elements
                - Humans in the loop
        - Challenges
            - How can existing methods accommodate systems?
            - What do we need to address some of these?
            - We need to be able to integrated informal methods?
        - Stochastic reachability analysis
            - Based on numerical solutions
                - Convex optimization
                - Scenario-based optimization
                - Chance constrained optimization
        - Data-driven stochastic reachability
            - Data-driven approach or implications
            - Quick and dirty system ID
            - Doing something that is model based.
- Laura Humphrey
    - We don't need formal methods for everything. It depends on what the consequences are.
        - We are still going to be testing.
        - It would be good to use
    - Autonomy stack
        - There are a functional correctness in the software. So we are looking at various tools. There are some really great case studies.
    - It is getting harder to separate varitification.
        - Richer elements are harder to test. How do you write the requirements for something? So that they systems have a way to reuse?
            - Do you have to start over and redo everything?
    - We need to enable different processes. Interest education. There is not going to be one group that can solve all these problems.

Discussion

- Philp- Standard updates- safety case that will be changed by the change. You only have to evaluate the parts of the case that were affected.
- Heather- We don't have data in this area. There seems like there is going to be a real problem with getting access to data. How are you thinking about getting access to these kids of data.

    - Julian- It is good to think about the different kinds of data. We lack the data on the analysis on those cases. There is a dataset that would help us. If we were to reinvent the preventative streams we would have a different stream.
        - The ways we are doing things today is putting patients at risk.
- Data Centric tradition-We know how to evaluate systems based on physical systems. Repeatable on how it is going to operate. Once again, even though you are not going to get an exact answer. It might not have the same affect.
- Natasha- what is the right artifact?
    - Near terms plan- continue to find ways to blame the human.
    - Every single autonomous system is not going to cut it.
    - Every system is going to have its own “what does it mean to be safe”
        - Going to need its own road test to line up.
    - Philp- Autonomy fundamentally breaks how we do safety.
        - Breaks how we think humans are doing.
        - There are so many assumptions of what drives. That goes unstated.
    - If you had requirements you would need machine learning.
- Julian- physiological control systems
    - Mimic what human operators do/systems do
    - Systems are more vigulate that what the nurses are.
    - The outcomes are hard to prove.

Panel: Space

- Danette- NASA know how to build complex software systems
    - Boots on the moon by 2024, permanent location on the moon a few years later
    - Then moon to mars, then thinking about deep space observations as well.
    - Performance and Persistence in space- need to do it all autonomously.
    - Need to fail in a way that we can retreat and restart.
- Joel- end user of assured autonomy
    - What are our needs. What are some of our challenges? We provide the GPS that got you here today. We also do space situational awareness. Eastern/western range we support launches.
    - We need to reduce our reaction times and respond. We need to free up our precious humans. We spend a lot of time/operating dozens of satellites.
    - We can only launch one rocket at a time. Then we developed a system that can control and regulate self destruction- so now we can launch more than 1 rocket at a time.
    - Autonomy strategy- in partnership with NASA, we have an effort in space trusted autonomy.
- Hiro- Spacecraft Autonomy
    - At first we were 100% autonomous
        - That dropped but there is now a growing need for more autonomy.
    - Automated scientific scene interpretation
        - search/priorities the data- false positives

        - Europa Lander Mission Concept
            - Can only survive 20 days on the surface.
        - Creating “fake mars”
        - Outstanding challenges in spacecraft autonomy
            - Lack of capable on-board computers
            - V&V (validation and verification)
        - Difficulty in collecting sufficiently large dataset
        - Restrictions/culture in using third-party software on-board
    - Discussion-What keeps you up most at night?
        - Missy- Batteries that propelled space propulsion.
            - Huge problem in robotic surgery is that it hasn’t been certificated. The idea of heritage.
        - Danette- We need to be more agile and test more state-of-the-art
            - I worry we won’t be able to establish an equivalence. How do we certify these systems? My fear is if we continue to look at the problem, all we need are new methods. Not fully automated or retractable that they might be.
        - Joel- Just as autonomy is on the rise. We are also on the verge of a new space age. A lot of new costs and space age. Finding new space and commercializing space. I think of great power competition.
        - Hiro- There is a piece of code which works on the links and systems. The algorithm that works on the systems.
        - Missy- I think this country underfunds autonomy.
        - Nancy- When they talk about autonomy do they talk about heritage?
            - Missy- Yes.
        - Ashley- You are starting to apply machine learning to perception problems. Machine learning models wouldn’t be entirely accurate.
            - Hiro- Machine learning is not different from the visual.
        - Missy- most important things in the next year
            - Danette- test and environments for modeling and simulation
                - Different than how we have traditionally done it.
                - Environment that allows me to create an agent
            - Joel- Change the culture/how we operate/we are trying to come up with a strategy. Find out the use cases. Building the culture and use cases.
            - Hiro-Short term, break the systems and autonomy will apply and get everyone on board. We need some examples. My goal to apply machine learning to mars. Make everyone filed confined and safe.

Panel: Policy, Regulation, Ethics

- Lenore- Challenge 1
    - Digital Assistants
    - Deep Learning
        - We don’t have the system in place for this

- Cara LaPonite- Ecosystem design approach.
    - If you can't integrate it into the design approach it. We have three pillars that you can look at. We have a pillar around policy and government.
        - Big role of the road map is to build a comprehensive view.
        - Policy is critical. Not domain expert.
    - Can up with a list of policies that had to be updated.
    - This takes real work and real effort.
- Nadya Bliss- Significant vulnerabilities
    - CS has seen a lot of vulnerabilities.
    - Tremendous advancement in the research but not in the algorithm space.
    - There is a lot more to be done in this space.
- Heather Roff- Policy happens at the international level (as well as at the white house)
    - If you are a small delegation- you might not have support. Could have issues when coming up with ideas and not having the necessary technical support.
    - There is a potential for misinformation
    - Swimming in your lane- "why are you doing this and why are you doing that, stay in your lane" Most of the people are doing this. Best way to do this is to do interdisciplinary work. The more that that culture can change, the better we can be.
    - Science of Induction- You might fundamentally miss things. If you expect to test one thing and get something else, you might miss thing.
    - Fundamental core problems of science might be a fundamental core problem of policy.
- Cara- Technology has ethics embedded into it. Right vs. right problem.
    - In assured autonomy- There are always trade offs you make for security and privacy. You build in those trade-offs. Everytime you make a decision for how you are trading those off, you are making a decision for how to trade those off.
    - Coming up with a design philosophy. Design philosophy got us there.
    - Every time you make a trade off- you are creating policy

Discussion

- Danette- There are often no AI experts on panels either.
    - Heather- Interdisciplarinity is key! Need to get the right people around the table.
- Marc S.- What about educational options to education policy people in this space? MOOCs?
    - Nadya- Interesting we see staffers being hired with expertist. It is key to have the ability to call on people who know.
- Kevin- Handful of committee staff have some expertise.
    - Trying to figure out what is trustworthy. How do we help those regulators and inform these new technologies?
- Cara- You have to inform those policies.
- Signe- Structure to learning.
- Induction arguments- inductive training of our systems.
    - Heather- I think it is a huge problem.

Panel: Defense

- Ashley Llorens
    - I’m with JHU/APL
    - from our perspective an intelligent system will act with some degree of autonomy
    - every intelligent system teams with humans in some way, so it must retain that ability
    - could be physical systems (e.g. spacecraft, maritime, robotic prosthetics), national security analysis, cyber operations, etc.
    - video of fully autonomous swarming boats. Commanded by a single operator who is on shore
        - self organize based on the commands of the operator
        - still need people on the boats even in this very controlled environment (Navy test waters)
    - robotic system that needed to open a door
- Signe Redfield
    - no one was even asking can it do the right thing? Does it need a color camera or is a black and white camera enough?
    - Does it do the right thing?
    - What is the right thing anyway?
    - academic drivers:
        - research: operational time approx. 2 hours, frequently modified, acceptable frequence of failures frequent, little mistake consequence
        - industry: operational time days to months, frequency of modification: never, acceptable frequency of failures: low, mistake consequences: expensive
        - DoD: operational time: hours to months, frequency of modification: intermittent, acceptable frequency of failures: very low, mistake consequences: catastrophic
    - in 2014 identified 26 research challenges in verification. four main categories: abstraction, models, test, tools
        - work ongoing but challenges being added faster than their being solved
        - How will we prove anything with deep learning if we don’t know what’s going on inside it?
        - the purpose of verification is to build trust
    - defense-centric challenges: need to verify safety, security, and functionality
    - critical mass of researchers beginning to look at this as a problem
    - how much trouble am I going to get in if I put this out and it fails vs. I need to get this job done?
        - need to ID times and places it shouldn’t be used
    - academic timescale: we’ve been working on this for 5 years and am only just chipping at this
    - DoD has reduced economies of scale
- Craig Lennon

    - Part of the DoD Test & Evaluation, Verification & Validation working group
    - Dod acquisition lifestyle structure. Goals: verifiable requirement specs, arguments for dependable autonomy, instrumented and measured autonomy, convince people it's safe to test, tools for testing, dynamic assurance for adaptive systems
    - love to have your info if you're interested
- Q&A:
    - Cara - talked about all these big challenges, have you ID'd impact points and if so what are they?
        - Signe - the working group who developed the challenges had no funding. At our last meeting we started working through a roadmap
            - funding priorities have dictated which things get worked on. Some people have been developing solutions to specific pieces of challenges. Working on a book that addresses those, but we haven't had a big funding program to address these challenges
        - Craig - not a giant pot of money that pays for assurance challenges. Tried to break it down by who pays for what. The funders have to prioritize it
        - Ashley - join AI center for the DoD chooses what to fund. Once we start to put out our roadmaps we can start to choose challenges for funding to push things forward
    - Besides dedicated research for assurance, do you see this making its way into other projects
        - SIgne - i've been working with NRI on a formal model for this. This project from a systems engineering project is basically backwards because they already had hardware and an operational method before they had a customer. Seems to be working but not sure it could be exported as a project anybody else could use
        - Ashley - blooper reel behind every success
    - Hiro - autonomy is broad and diverse. Can we find a spec for certain types of algorithms or diverse algorithms?
        - Craig - I don't know that I have that wisdom but some program that's leading, like and airforce program will figure out how to do it for their program and we can modify it
    - Phillip - I have seen some specs like you're talking about. Autonomous systems have a safety shutdown box that was really reliable. Not so good for mission completeness but didn't run over the bystander
        - Ashely - we've also got an onboard watchdog, simple verifiable
    - Lenore - when I grew up safety was a property that once lost cannot be regained. A system that does nothing satisfies the safety problem
        - safety engineers have a different, slightly broader view then CS people
        - Phillip - 2 terms used: safety (no loss event) and permissiveness (scope of the control space you are allowed to operate in that still guarantees nothing bad will happen). Want permissive space to be as big as possible

        to complete mission, but the bigger it is the more likely you are to have loss events.
    - Signe - haven't pulled safety out as a specific thing because so much of what we are concerned with is the safety of the system itself. Don't want the robot to fall down the stairs and break. Bystander safety: don't want to hit someone. Don't know how to bound the list of all possible things in the world that the robot can interact with

Panel: Human System Integration, Trust

- Nancy- Assured Autonomy requires human systems Integration throughout the lifecycle of systems development
    - Robocop
    - MQ-1/9 Operator Control System
- Human Autonomy Teaming Considerations
    - Team members have different roles and responsibilities-do not replicate humans
    - Effective team have team members who are interdependent and need to interact/communicate
    - Interpersonal trust is important to human teams
    - To get reliable, verificable Human- autonomy team we need to measure team/system effectiveness.
        - Need to be able to measure the team coordination and effectiveness.
- Matthew- No Autonomy is an island
    - Accomplish skills independently.
    - If you are going to accomplish skills, you need to address them.
    - This should change your view of how you are building autonomous systems
    - Misconceptions and Myths when it comes to "autonomy"
        - We need to make sure we chose the right words when we are looking at them.
    - As you increase autonomy does it create more or less independent?
        - More or less teaming capability
- Missy- Working on a paper on data curation
    - Analysis of crash predictors
    - The outcomes depends on the choices people make makes it very subjective.
    - Independent certification of driverless car/modulated cars
    - I think it is really wrong for DOT standards to cost money
        - We have to appreciate this.
- Nisar- Technical problems
    - Human-autonomy interaction considered afterthought or band aid
        - Safety and assurance vi. Layers of defense
    - Theorist/programmer/system designer
        - Uncertainty- context, task, environment
        - Models of reasoning under certinatily
    - Intelligent competent machine=loner know it all

        - Smart competent people ask questions of themselves and others
        - We don’t build smart autonomous machines to do this well.
    - Using self trust to adjust user trust.
- Jessie Chen- We design human machine interfaces to be more transparent.
    - We have come up with a framework that is situational awareness based. It is based on there levels of situational awareness.
        - Detection
        - Comprehension
            - Reasoning process
            - Constraints
        - Protection
            - Projected outcomes. We are trying to look at the human factors aspect. We have conducted a series of factors and we found out consistently that there is an interface. Humans working with more transparent agent.
    - What is the information requirement that we need to address?

Discussion

- Mark- I might argue that some attitudes have improved a lot in the past 10 years
    - What do you actually need to do to incorporate change.
    - Structured analysis and experimentation.
    - Matthew- People talk about experimentation. People don’t know what is meant by teaming or how important it is. As people are working in the field they need to be clear about what they are doing and what they are doing differently.
        - That paradigm exists very very strongly.
        - Big gap between what we think is important.
- Missy- This is a backlash behind some of that group. Need agencies to promote interdisciplinary teams!
    - I think every proposal should say “someone who does human in the loop” needs to be on it
- Matthew- DARPA has a proposal that just came out that says that.
    - Nancy- there is a collaboration on that as well, so it isn’t exactly true.
    - We should be on high collaboration that connects these groups and ties the cognitive science departments.
- Julian- Boundaries and scope with our autonomous systems. When it comes to Human-in-the-loop, has to do with the discomfort with the scope. Defining the scope. Even reviewers, when you need that subject matter expert brought in you need that scope.
    - Is there a way to better scope it? So you can engage?
    - Nancy- I do think we need more methods to understand the whole system that includes the humans. Perhaps then we can address Mark’s problem and get people involved in the system.

    - Heather- There are people out there who could model the process and institutions for us
- We don’t really have a culture like the hospital culture does
    - Ashley- We inherited IRB process from medical community.
    - Matthew- DOD end users

Breakouts

Sushil- Challenge Problems (Heather Roff, Hiro Ono, Matthew Johnson, Sushil Birla)

- What properties to assure? Need to isolate properties of interest.
    - [Darren] from natural environment.
- Is this a case that is actually tractable?
- Dealing with the Unknown-Unknown problem.
- The level of assurance needed implies that a mishap would be a rare event. The treatment of outliers in current ML techniques might not be suitable for a rare event application environment.
- [Darren; Heather] Perception (and modeling the open world) with Assurance.
- Heather- we don’t have to solve the problem, or expect it to be solved right away.
    - How do we know that we are not going to miss something?
    - The reasoning behind uncertainty is that it is huge.
- [Hiro] Adversarial challenge.
- Ability to determine when the system is at the boundary of its design operating environment/envelope (e.g., when it is at the boundary of the envelope for which it is trained).
- [Heather; Sushil] Ability to reason about UNCERTAINTY.
- [Heather; Sushil] Ontology of “what can go wrong” (mishaps; misbehaviors; failures; unexpected adverse conditions) over which the system should survive.
- (Corresponding to “what can go wrong”), adaptation of goals dynamically, so that all is not lost.
- Traditional process has been determined.]
- Assumptions that don’t necessarily hold.
- Heather- This is a lot more human machine interaction. You have to think through a lot of the information in ways that you might not see before.
    - Nisar- There has to be some kind of change or growth for the humans.
    - Heather- What are the competency boundaries?
        - That would be part of the challenge itself.
    - Heather- There is a need for a theory building part of this as well.
- Kevin- You have to think about different properties. It could be put in something so crazy that it won’t be achievable.
    - Matthew- The evaluation question that is raised is very important too. The challenge that is separated into not.
        - Independent verification
        - How do I constrain and verify?
    - Only caring about single environmental failures.

        - Heather- perhaps we define these failures.
        - Given the limitations of my vehicle there will edits and changes to take on
- Matthew- Given the domain that was put in front of me (NAVIGATION), there are changes that will be needed for higher precision in the space-time trajectory.
    - Sushil- Generalize the navigation challenge “higher precision in the space-time trajectory” to “... solution trajectory” (trajectory of the control variables in general).
    - Hiro- There is a right and wrong answer. And still the question is can I run the test, yes or no.
        - You can say no.
    - Heather- planners and autonomous system need to adapt and see if something is there. It is going to be in the bounds of something that you can do.
    - [Heather] (Context: Military applications) Emergent properties -> need for adaptation.
    - [Darren] Need new methods of verification which expose and “test” for the corner cases or edge cases.

- Ashley- Challenge problems (Ashley Lennor, Missy Cummings, Laura Humphries, Signe, Ufuk, Nancy Cooke, Kevin Dopart, Jessie Chen, Todd Humphreys, Craig Lennon, Tichakorn Wongpiromsarn)
    - Laura - in order to require new methods of assurance you need a high fidelity simulator. When you get in the real world to test, things break down
    - Ashley - driverless car simulations
    - Tich - understanding what is the good driving behavior through observation. How to get companies to share the data that they have?
        - Ashley - maybe provide real world examples of where the system would fail. Turn the testing method on its head
        - Signe - we don’t necessarily know what failure looks like
    - Ashley - always easy to find a corner case that breaks. If the adversary is unbounded the adversary always wins
        - DARPA grand challenge but focused on assurance
    - Missy - I think the elephant is the room is assured autonomy is really a process not a think
    - Ashley - testing violations of some sort of specialized bounds
        - Laura - a lot of formal method success are proprietary and people won’t reveal the problems they find
        - they don’t want to hear I found 10 bugs they want to hear I found a small probability of failure or whatever.
    - Signe - UAVs?
        - Ashely - challenges?
        - would it make sense to have the challenge structure such that the challenge organizers provide the structure and say find the problems or they develop the system?

      - Ashley- I think you could go either way
    - Signe - blimps wall climb. When they get close to the wall they stack to go over. Quad-copters don’t do that. IDing those kinds of emergent behavior would be worth assuring
    - Ashley - if it’s outside of your specification you can’t win
      - Signe - I don’t think any of us could write a bullet point specification for this. I can’t do it for something as basic as object avoidance
  - might be ways to find things in the environment that aren’t obviously part of the cost analysis
    - couldn’t this be done in simulation though?
    - yes, sure
    - Signe - it’s the challenge of not knowing what level of abstraction of the environment in the simulation or what in the environment is important
      - I don’t think we can build tractable simulators that consider everything, but we don’t know what to include. Our current method is just to iterate and add more stuff as you find it effects performance
    - I like you image net parallel.
  - Nancy - This is what we’re doing in DARPA assist. They’re not talking about the assurance of their agents, but if this could be added it would be good
  - what are our metrics for assurance?
  - Signe - what we need is the creation of a program that will involve solving those problems
    - Ashley - we would need to make a proposal for a seedling effort

Report Back 2:

- Group 1:
  - want something challenging, not open world, but a lot of factors
- Group 2:
  - hypothesis was yes and we were exploring that angle
  - blue team/red team kind of approach. Blue team creates autonomous system with performance balance. Red teams responsible for manipulating violations to reduce safety of performance environment
- Group 3:
  - government is pretty good at privacy. Even if it collects it doesn’t stay for a long time. We know companies are not
  - medical industry - the more data they collect, the better they can treat you
  - how you would design systems technically based on what data you can or can’t collect from the sensor forward and the societal controls around the use of data