U.S. NAVAL RESEARCH LABORATORY

Assuring Autonomy for Defense Systems

Dr. Signe A. Redfield

Naval Research Laboratory

Washington, DC

Working notes

U.S. NAVAL RESEARCH LABORATORY

Why Should You Trust Me?

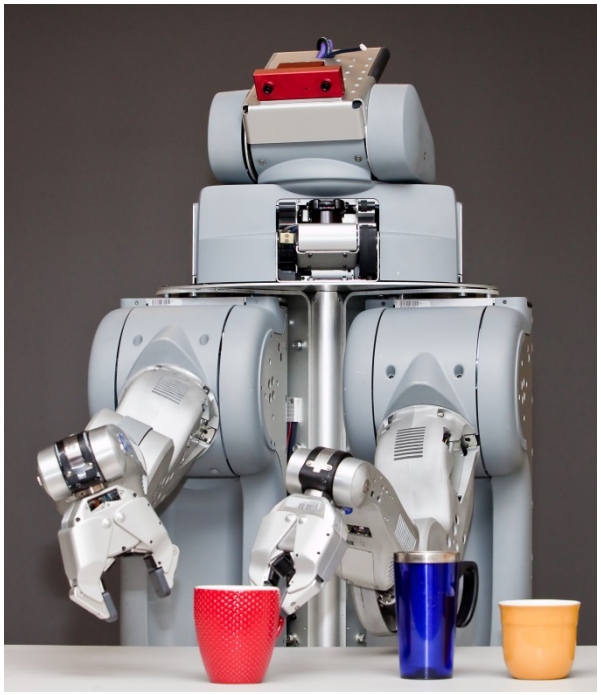
- Founding Co-Chair of the IEEE Robotics and Automation Society Technical Committee on Verification of Autonomous Systems
- Chair of the IEEE Guide for Verification of Autonomous Systems standard working group
  - Contributor to the approved IEEE Standard 1872-2015 “Core Ontologies for Robotics and Automation” and Secretary for the current IEEE Standard 1872.1 working group on Robot Task Representation
- 17 years of experience in autonomy and artificial intelligence for the Navy
  - Depth (underwater autonomy): 9 years at the Naval Surface Warfare Center in Panama City, FL doing basic research in autonomous behaviors for AUVs
  - Breadth (robotics in the international community): 3 years at ONR Global as the Associate Director for Autonomy and Unmanned Systems
  - Depth (space robots; verification of autonomous systems): 5 years at NRL
    - Designing the Payload Mission Manager (PMM) for a DARPA space robot program
    - Helping build formal models of the PMM, the fault management system, and autonomous operations
    - Helping to establish the verification of autonomous systems research community
  - Breadth (evaluation of artificial intelligence): 6 month detail as the acting Chief for the Test, Evaluation & Assessment group at the DoD Joint AI Center

U.S. NAVAL RESEARCH LABORATORY

Verification Summary (Functional Perspective)

- "Can it do the right thing?"
  - Not physically capable = don't need to evaluate further
  - Stage where individual behaviors and their integration are evaluated
- "Does it do the right thing?"
  - Decision logic is wrong = having the right components doesn't matter
  - Stage where we evaluate the system as a whole
- "What is the right thing, anyway?"
  - It does the wrong thing because we didn't understand what it needed to do
  - Particularly problematic for autonomous systems
    - Lack theoretical tools to answer whether it can or does do the right thing
    - Process is more expensive and time-consuming than for more mature disciplines

Can it pick up a cup?



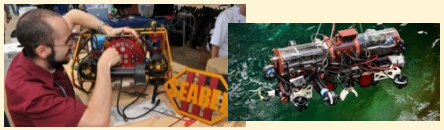
Which cup does it pick up?

Should it be picking up a cup?

U.S. NAVAL RESEARCH LABORATORY

Academic Drivers

- Academia: verification of autonomy is **not a traditional research domain**
- Community is open to doing the research, but needs guidance about the problem
- **Industry guidance ≠ DoD guidance**: DoD has more difficult regimes and needs better reliability



Research

Operational Time: ~2 hours

Variability of operations environment: Moderate

Frequency of modification: Frequent

Acceptable frequency of failures: Frequent

Mistake consequences: Negligible

Risk Tolerance: High

Industry

Operational Time: Days to months

Variability of operations environment: Low

Frequency of modification: Never

Acceptable frequency of failures: Low

Mistake consequences: Expensive

Risk Tolerance: Moderate

DoD

Operational Time: Hours to months

Variability of operations environment: High

Frequency of modification: Intermittent

Acceptable frequency of failures: Very low

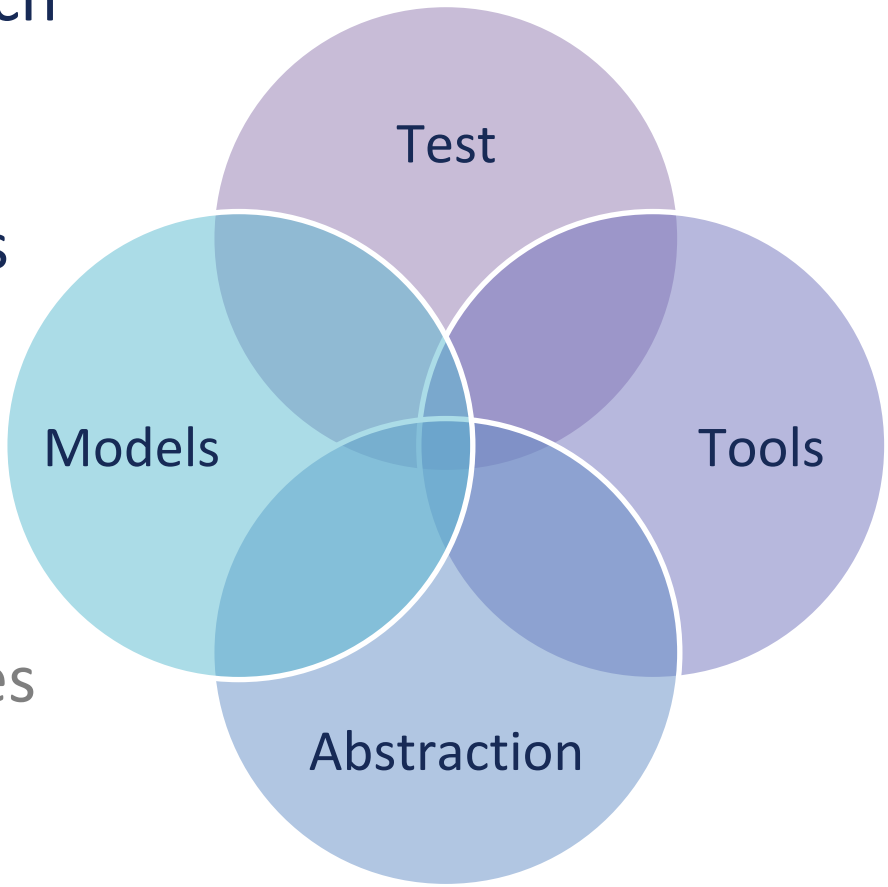
Mistake consequences: Catastrophic

Risk Tolerance: Low

U.S. NAVAL RESEARCH LABORATORY

Generic Challenges

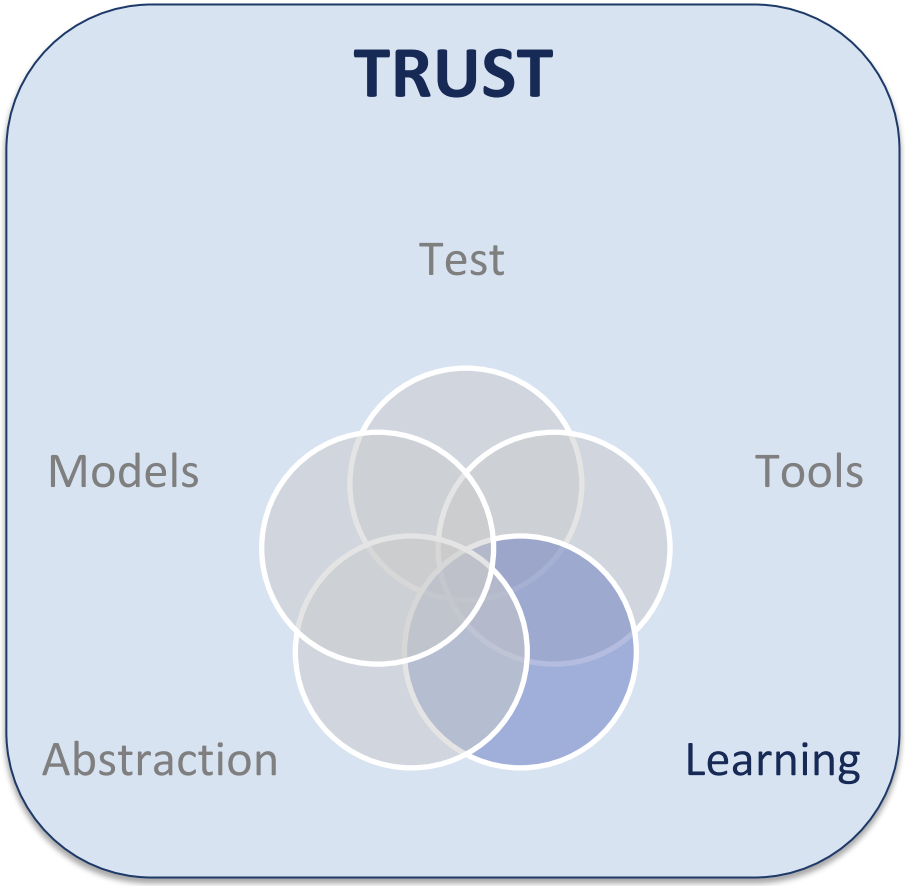
- In 2014, identified 26 research challenges in verification
- Four main categories (for now); some overlap in topics
  - Abstraction
  - Models
  - Test
  - Tools
- Work ongoing, but challenges being added faster than they're being solved
  - Deep Learning



U.S. NAVAL RESEARCH LABORATORY

New Challenge Areas

- In 2014, identified 26 research challenges in verification
- Four main categories (for now); some overlap in topics
  - Abstraction
  - Models
  - Test
  - Tools
- Work ongoing, but challenges being added faster than they're being solved
  - Deep Learning
- The purpose of verification is to build trust
  - Active research area within the robotics community



U.S. NAVAL RESEARCH LABORATORY

Defense-Centric Challenges

- Need to verify:
  - Safety
    - Known process in the formal verification community for some aspects of problem
    - Autonomous Systems: 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)
    - DoD: More dangerous robot combined with more stringent guidance re: safety of environment, asset, bystander
  - Security
    - Known process in the formal verification community
    - Autonomous Systems: security typically near the bottom of the list of needs, well below "does it work"
    - DoD: Critical need
  - Functionality
    - Autonomous System: Critical problem not typically addressed by current verification tools
    - DoD: requires greater degree of certainty in some cases, can accept less in others.
    - Especially critical in the context of learning systems
    - Interesting to autonomous system designers in academic community

- Approval from users and stakeholders
  - **Boils down to trust**

- Academic Timescale
  - Verification Working Group has been active for 5 years
    - Have still only addressed a fraction of the challenges identified

U.S. NAVAL RESEARCH LABORATORY

DoD Needs

- Higher required certainty coupled with complex system interactions, increased flexibility, and extended mission duration
  - Make verification of modifications cheaper and faster
  - Establish doctrine/tactics/principles of operation to support well-defined operational performance bounds
- Safety, security AND functional verification
  - Functional verification growing in academia, especially for learning and HRI communities
- Reduced economies of scale
  - Compensate with investment in process and tools

BACKUP SLIDES: CHALLENGES

U.S. NAVAL RESEARCH LABORATORY

Challenges, 1

- How is an adequate model of the system created?
- Common models and frameworks need to describe autonomous systems broadly enough so they can be used to standardize evaluation efforts and interfaces to the system.
- How should models of black box autonomous systems be developed and debugged? How is a mathematical and/or logical model suitable for formal analysis produced from empirical observations?
- How should one identify and model components that are not captured yet (and what are their properties)?
- What determines the level of simulator fidelity to extract the information of interest?
- How is the level of abstraction determined for the robot model, its behaviors, and the simulation that tests the model? How many environmental characteristics need to be specified? What are the aspects of the environment, the robot, and the autonomy algorithms that cannot be abstracted away without undermining the verification?
- Where is the transition from specifying system requirements to designing the system and how are principled requirements developed so they do not devolve into designing the solution?
- How is it ensured that the implicit and the explicit goals of the system are captured? How is a model of the system goals from a human understanding of the task goals, the system, and the environment created?
- How are performance, safety, and security considerations integrated?
- At what point is there enough evidence to determine that an autonomous system or behavior has been verified?
- How does one ensure it is possible, in the physical world, to test simulated situations that result in boundary cases?

U.S. NAVAL RESEARCH LABORATORY

Challenges, 2

- How would tests be designed so that passing them indicates a more general capability?
- How are challenging design reference missions selected so that performing well against them indicates a more general capability for the system rather than for specific system components?
- How can test scenarios be produced to yield the data required to generate mathematical / logical models or to find the boundary locations and fault locations in the robot state space?
- Once an adequate model is created how is it determined whether all resulting emergent behaviors were captured and what are appropriate performance measurement tools for this?
- Measurement and evaluation are generally poorly understood – operators can describe the tasks for the robot but lack tools to quantitatively evaluate them. How should autonomous behaviors be measured so they consistently and accurately describe the capability embodied by a robot?
- How is a metric defined for comparing solutions?
- How is the optimal defined against which the verification is performed? ? How is the solution shown to be in fact, optimal? How is the performance of the system measured?
- How is the performance from finite samples of the performance space generalized across several variables and parameters?
- Autonomy frameworks are unable to determine whether all the resulting emergent behaviors have been captured or to supply performance measurement tools.
- What new tools or techniques need to be developed?

U.S. NAVAL RESEARCH LABORATORY

Challenges, 3

- In general, how do we verify the fitness of a given physical robot structure for a given task or environment (obviously, a robot that cannot sense color or is operating in the dark with an infrared sensor is unfit to sort objects on the basis of color)?
- Descriptive frameworks are either too specific and constrain the developer to specific tools when designing the autonomous elements of the system or, too broad and difficult to apply to specific cases. Tools are needed to analyze systems at both the specific and the broad levels.
- How is a structured process that allows feedback between the physical/ground truth layer and the formal methods verification tools developed?
- How to disambiguate between cases where the specification was incorrect (task description abstraction failed to capture some required system action) and those where the environmental model was incorrect (environmental abstraction failed to capture some critical system-environment interaction)? How to identify not just individual situations but classes of situations where the vehicle fails to be safe or to achieve safe operation (e.g. a front wheel often falls off the cliff but the back wheels never do). How should unanticipated unknowns be accommodated?
- If an algorithm, or patch to an existing algorithm, was replaced can it be proven that no new failure modes were introduced without re-doing the entire verification process?
- How do we verify systems that incorporate deep learning?
- In what ways should the training data used by learning components integrated into the verification testing and analysis processes?