Meaningful Metrics for Demonstrating Ethical Supervision of Unmanned Systems

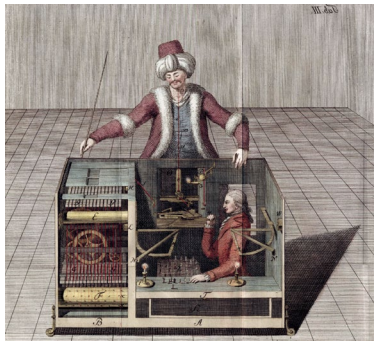
Workshop on Ethical Computing, AAAI Spring Symposium 2022
Metrics for Measuring AI's Proficiency and Competency for Ethical Reasoning

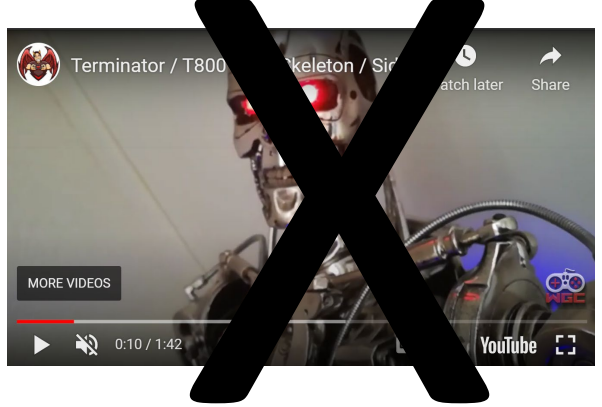
Don Brutzman and Curtis Blais
Naval Postgraduate School (NPS), Raytheon Corporation

23 March 2022

ICE BREAKER !!

Most egregious claims …





Many people (including big-name AI luminaries) seem to think that some kinda

AI Ethical Agent (perhaps a modern-day Homunculus)

Can sit as a monitoring process on top of any kind of robot software, ensuring that someone else's robot operates morally legally and ethically.

Such misconceptions have inhibited meaningful progress.

(Example: obligatory Terminator image with glowing red eyes.)

AAAI workshop on metrics is a great step in a necessary direction.

Abstract

Metrics for AI are important, as illustrated by the workshop topics of interest. We note that commonplace gaps in applied AI derive from “Here are the measurements we know how to take” which are too easily over-extrapolated into conclusions of interest. In other words, such precise metrics are necessary and appealing but may not broadly apply to general situations. We assert that necessary subsequent questions are “How do we define meaningful objectives and outcomes for a current unmanned system,” “How do we measure those characteristics that indicate expected success/failure,” and “Once we can measure meaningful results, how do we assemble exemplars into test suites that confirm successful completion across ongoing system life cycles?”

In our work, moral responsibility and authority for ethical behaviors by remote autonomous unmanned systems lies with the humans responsible for robot behavior. Lines of success or failure are clearly defined when delegating tasks to robots which have the capacity for life-saving or lethal force. Goals, constraints and metrics that are shared by humans and robots are formally verifiable as consistent and further testable in repeatable ways. This point paper explores potential design principles of broad value to ethical AI efforts.

First, some centering considerations

- Metrics are essential
- Precise metrics are necessary
- General metrics are elusive
- Humans own ethical responsibility and authority
  - not machines
- Can we evaluate whether we are getting better or worse?
  - system evolution is ongoing

How Do You Measure *That*?

Critical question that applies to all domains, all perspectives, all aspects

- Technical, philosophical, practical, developmental, operational
- This means us, i.e. everything in this workshop

If you can't measure a property or characteristic,

- Do you really understand the concept? Perhaps
- Have you defined it precisely, unambiguously, uniquely? Probably not

If you can measure something, then can analyze, assess, improve, etc.

- Conceptual clarity, representational crispness, repeatable
- If not, then people are stuck at talking about what they are talking about

Essential, necessary, elusive

1. **Metrics are essential**. Too many AI systems have ill-defined metrics that do not align with the ambitious goals being pursued.

2. **Precise metrics are necessary**. All claims are suspect if they are not built upon basis that clearly answers common-sense design question "How do you measure that?"
   - Carpenter's motto is good heuristic: measure twice, cut once…

3. **General metrics are elusive**. Metrics must inform the successful evaluations of objectives or else they are confusing and counterproductive.

Who's in charge? People, software, nothing?

Humans own ethical responsibility and authority.

- Legitimacy cannot be fully delegated as ill-defined "autonomy" in AI systems.
- There is no omniscient Delphic-oracle homunculus agent.
- Human-machine combinations can (and perhaps must) be effective

Are we interested in deciding (making moral judgements) whether specific activities are ethical or not?

- Responsibility of human actors carrying out shared laws, governance.
- Machines not culpable, legally/ethically/morally, but still can be dangerous.
- In-between case: some human-led organizations avoid responsibility by deferring agency to "autonomous" actors, unsafe/unethical deployment

DoD Directive: Autonomy in Weapon Systems

- DoD Directive 3000.09, 21 NOV 2012 with change 1, 8 May 2017
- Original and update signed by DEPSECDEF Ashton Carter

CONTROLLING REFERENCE FOR U.S. DOD

1. PURPOSE

   a. Establishes DoD policy and assigns responsibilities for the development and use of autonomous and semi-autonomous functions in weapon systems, including manned and unmanned platforms.

   b. Establishes guidelines designed to minimize the probability and consequences of failures in autonomous and semi-autonomous weapon systems that could lead to unintended engagements.

4. POLICY (excerpted)

   a. Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.

   b. Persons who authorize the use of, direct the use of, or operate autonomous and semi-autonomous weapon systems must do so with appropriate care and in accordance with the law of war, applicable treaties, weapon system safety rules, and applicable rules of engagement (ROE). […]

6. RELEASABILITY. Cleared for public release. […]

REPEAT

People first – machines under human control

Only humans have authority, responsibility and accountability for actions with potential for lethal outcomes.

- Includes responsibility for unmanned systems under their command

Guiding heuristics for designing new AI-enabled mission capabilities:

a. How do humans accomplish such goal tasks today?
b. How might unmanned systems accomplish similar tasks in future?
c. How can human commanders safely direct and supervise such systems, retaining moral and legal authority over operations?

Building Ethical AI -

Moving from Theory to Practice

How do you measure that?

Trusting Software and Trusting Data

- Network Optional Warfare (NOW) Blog, January 2016
- In 1983, Dennis Ritchie and Ken Thompson jointly received the Turing Award for their development of generic operating systems theory, and specifically for the implementation of the UNIX operating system.
- Ken Thompson's lecture was Reflections on Trusting Trust, with the subtitle "*To what extent should one trust a statement that a program is free of Trojan horses? Perhaps it is more important to trust the people who wrote the software.*" This talk can still surprise: he describes source code that looks like it does one thing, but actually performs things that are quite different.
- So in effect, Ken Thompson chose his Turing Award moment to reveal to the world that he had superuser and user access for every Unix system and server on the planet. Further he revealed that, even with a great many people scrutinizing and rebuilding the source code, and even despite users banging on Unix daily everywhere, anyone else might use a super password for each and every account. Meanwhile no one else knew that the super password existed, much less that it quietly insisted on re-propagating itself in each fresh new copy of Unix. **No kidding.**
- How does the Navy get beyond software barriers to reach the next level of capability: **trust for shared data**?

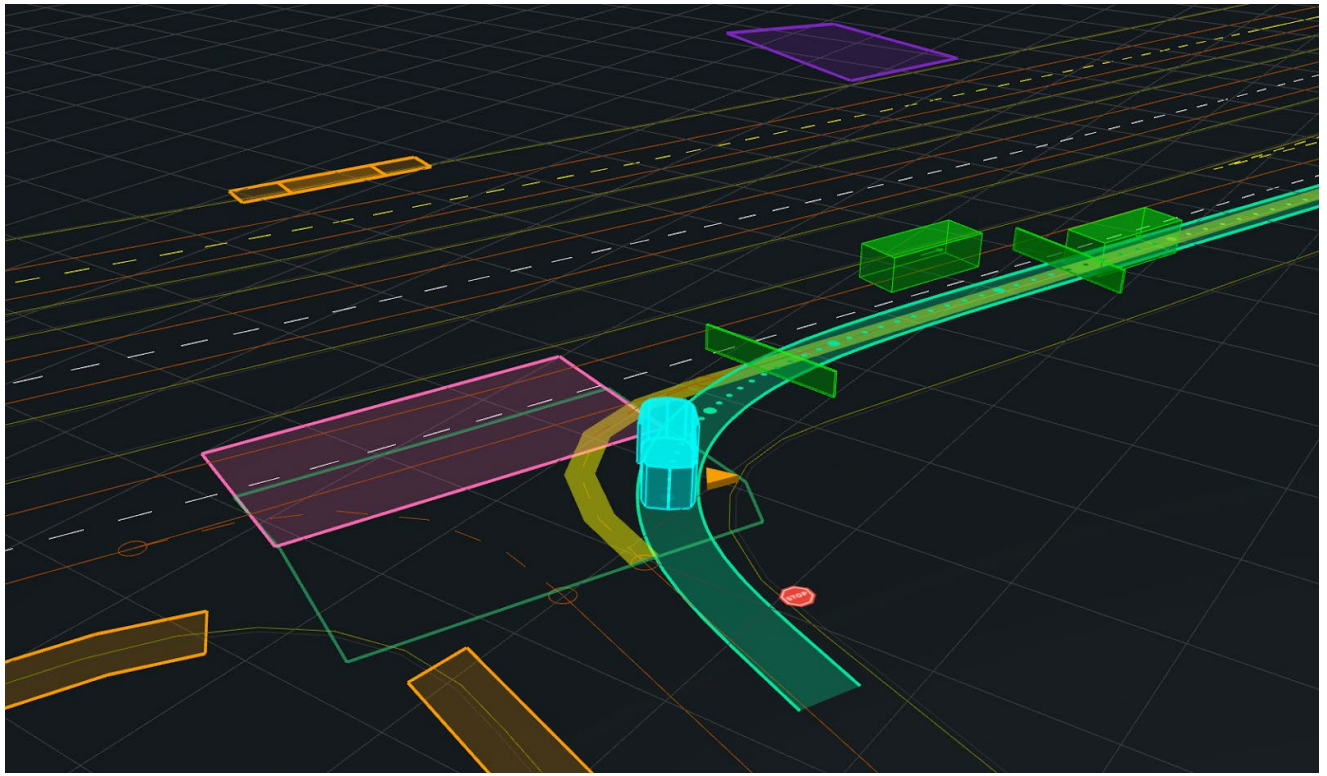
Coactive Design and Interdependency Analysis

- Matthew Johnson and Alonso Vera, "No AI is an Island: The Case for Teaming Intelligence", *AI Magazine*, vol. 40 no.1, Spring 2019
- *Abstract*. "The purpose of this article is to draw attention to an aspect of intelligence that has not yet received significant attention from the AI community, but that plays a crucial role in a technology's effectiveness in the world, namely teaming intelligence. We propose that AI will reach its full potential only if, as part of its intelligence, it also has enough teaming intelligence to work well with people. Although seemingly counterintuitive, the more intelligent the technological system, the greater the need for collaborative skills. This paper will argue why teaming intelligence is important to AI, provide a general structure for AI researchers to use in developing intelligent systems that team well, assess the current state of the art and, in doing so, suggest a path forward for future AI systems. This is not a call to develop a new capability, but rather, an approach to what AI capabilities should be built, and how, so as to imbue intelligent systems with teaming competence."
- Strong resonances exist with Ethical Control that deserve further exploration.

Best practice: “Off road, but not offline: How simulation helps advance our Waymo Driver”

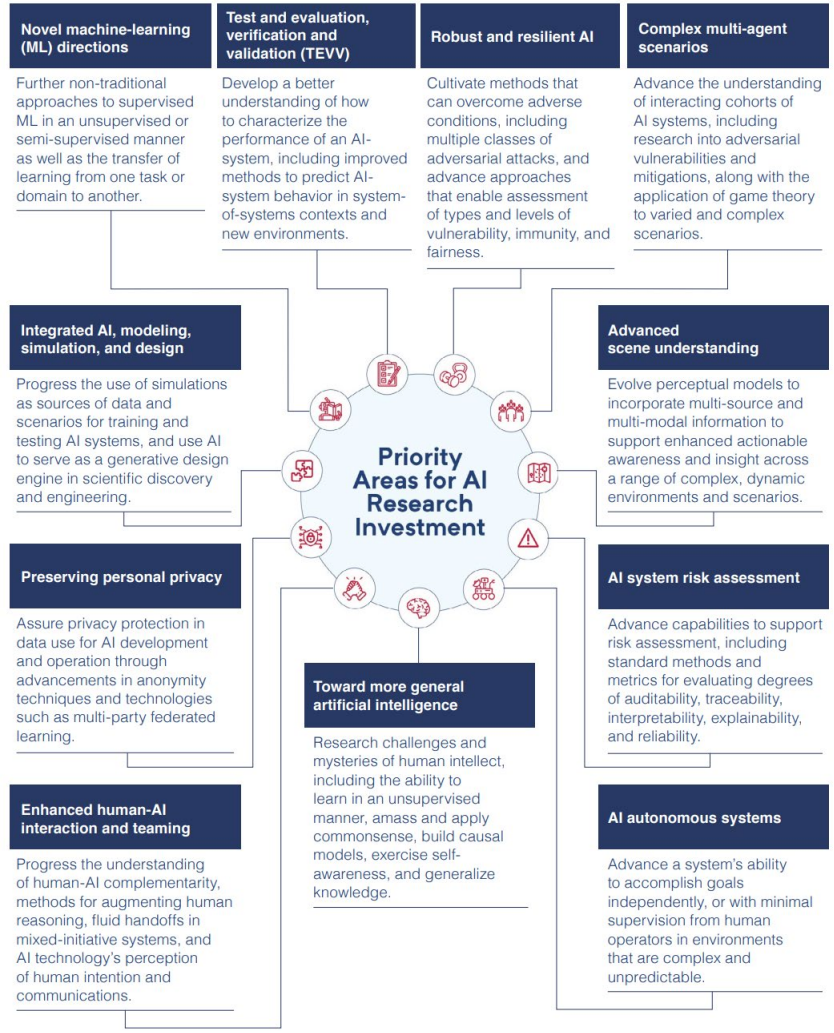
Industry is already operating at scale

- Google Waymo blog, 28 April 2020
- In simulation, "we drive around 20 million miles a day" and "over 15 billion miles“ total.
- LIDAR laser sensors looking in all directions
- Physically based modeling, simulation, reenactment
- Repeatable real/virtual replay of data and logic
- Large-scale regression tests
- Why not Navy and USMC?



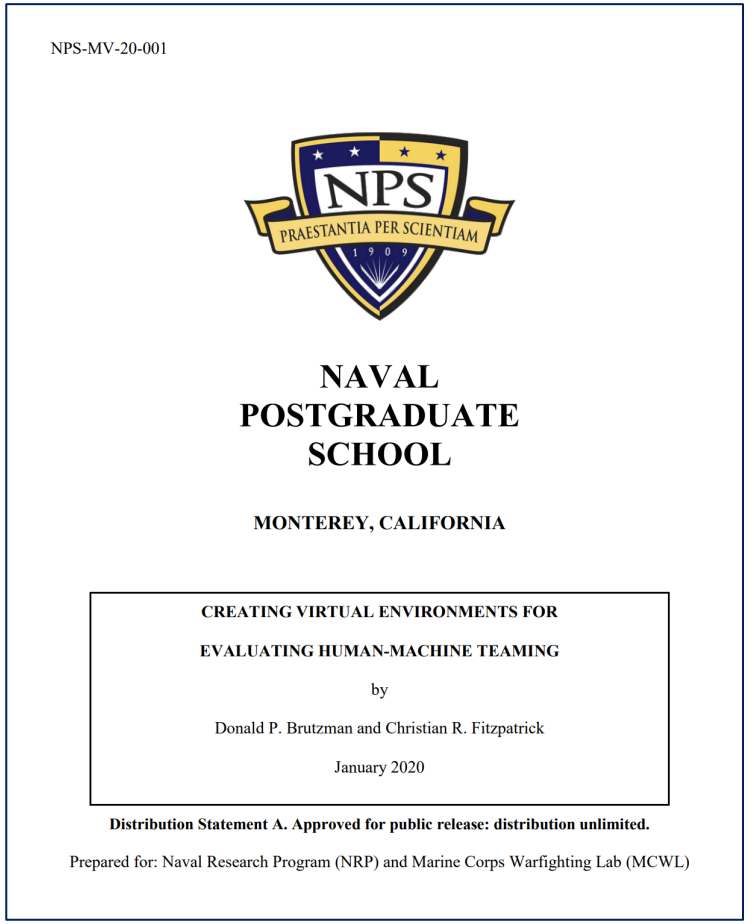
National Security Commission on Artificial Intelligence (NSCAI)

- The Final Report presents the NSCAI's strategy for winning the artificial intelligence era. The 16 chapters explain the steps the United States must take to responsibly use AI for national security and defense, defend against AI threats, and promote AI innovation. The accompanying Blueprints for Action provide detailed plans for the U.S. Government to implement the recommendations.
- https://www.nscai.gov
- NSCAI completed work 1 OCT 2021



Creating virtual environments for evaluating human-machine teaming

- Donald P. Brutzman and Christian R. Fitzpatrick
- Technical Report, January 2020, Naval Postgraduate School
- *Abstract.* With the emergence of robots on the battlefield, it is critical for the Marine Corps to tactically integrate existing unmanned assets with manned systems during Marine Air Ground Task Force (MAGTF) operations. In parallel, the Marine Corps must also look forward to identify capability gaps that future unmanned systems might address. To do both requires extensive field testing, which is often unfeasible and always costly. This effort proposes the use of virtual environments (VE), virtual reality (VR) and agent-based modeling to conduct scenario-based assessments of Manned-Unmanned Teaming (MUM-T) during combat operations. To pursue such goals, the project examined a variety of relevant tactical scenarios where Marines and robots act in concert to achieve specific mission objectives. Such tactical scenarios are further assessed using deterministic combat simulations to create a valid methodology for behavior creation and assessment within each scenario-specific problem space.
- https://calhoun.nps.edu/handle/10945/64266


NPS-MV-20-001

NPS
PRAESTANTIA PER SCIENTIAM
1 9 0 9

NAVAL
POSTGRADUATE
SCHOOL

MONTEREY, CALIFORNIA

CREATING VIRTUAL ENVIRONMENTS FOR
EVALUATING HUMAN-MACHINE TEAMING
by
Donald P. Brutzman and Christian R. Fitzpatrick
January 2020

Distribution Statement A. Approved for public release: distribution unlimited.

Prepared for: Naval Research Program (NRP) and Marine Corps Warfighting Lab (MCWL)

Dimensions of Autonomous Decision-making

Dimensions of Autonomous Decision-making

A First Step in Transforming the Policies and Ethics Principles Regarding Autonomous Systems into Practical System Engineering Requirements

Michael F. Stumborg, Becky Roh, Mark Rosen

CNA, December 2021

This study identifies the dimensions of autonomous decision-making (DADs)—the categories and causes of potential risk that one should consider before transferring decision-making capabilities to an intelligent autonomous system (IAS).

Over 500 risk elements associated with real-world robotics usage distilled, correlated and grouped in semantic categories.

I would gladly agree with all the world to lay aside the use of arms, and settle matters by negotiations; but, unless the whole world wills, the matter ends, and I take up my musket, and thank heaven he has put it in my power. . . . We live not in a world of angels. The reign of Satan has not ended, neither can we expect to be defended by miracles.

-Thomas Paine
July, 1775

- https://www.cna.org/news/releases/2022-01-24
- https://www.cna.org/CNA_files/PDF/Dimensions-of-Autonomous-Decision-making.pdf

Qualification of unmanned systems: extend Verification Validation Accreditation (V V&A)

Proposed Future Work

SISO V V&A Study Group

How to test and certify robots will follow both orders and constraints? Humans confirm understanding and trust through qualification processes.

- Design, construct “qualification card” for testing unmanned systems…
- Comprehensive virtual environment, hardware/software in the loop.
- Carefully crafted scenario testing of key requirements and capabilities.
- Anti-pattern tests to provoke and confirm constraints are not violated.
- Record all unit-test decision trees, decision-branching traces, and results as a certification record for each hardware/software version of robots.
- Visualize realistic rehearsal, real-time and replay of robot operations repeatably using shared Web-based SPIDERS3D virtual environment.
- Humans assess mission logs and scenario outcomes for after-action analysis, lessons learned, and continuous improvement via suite of unit tests.

Must… have…. data…… more…… data……. mo

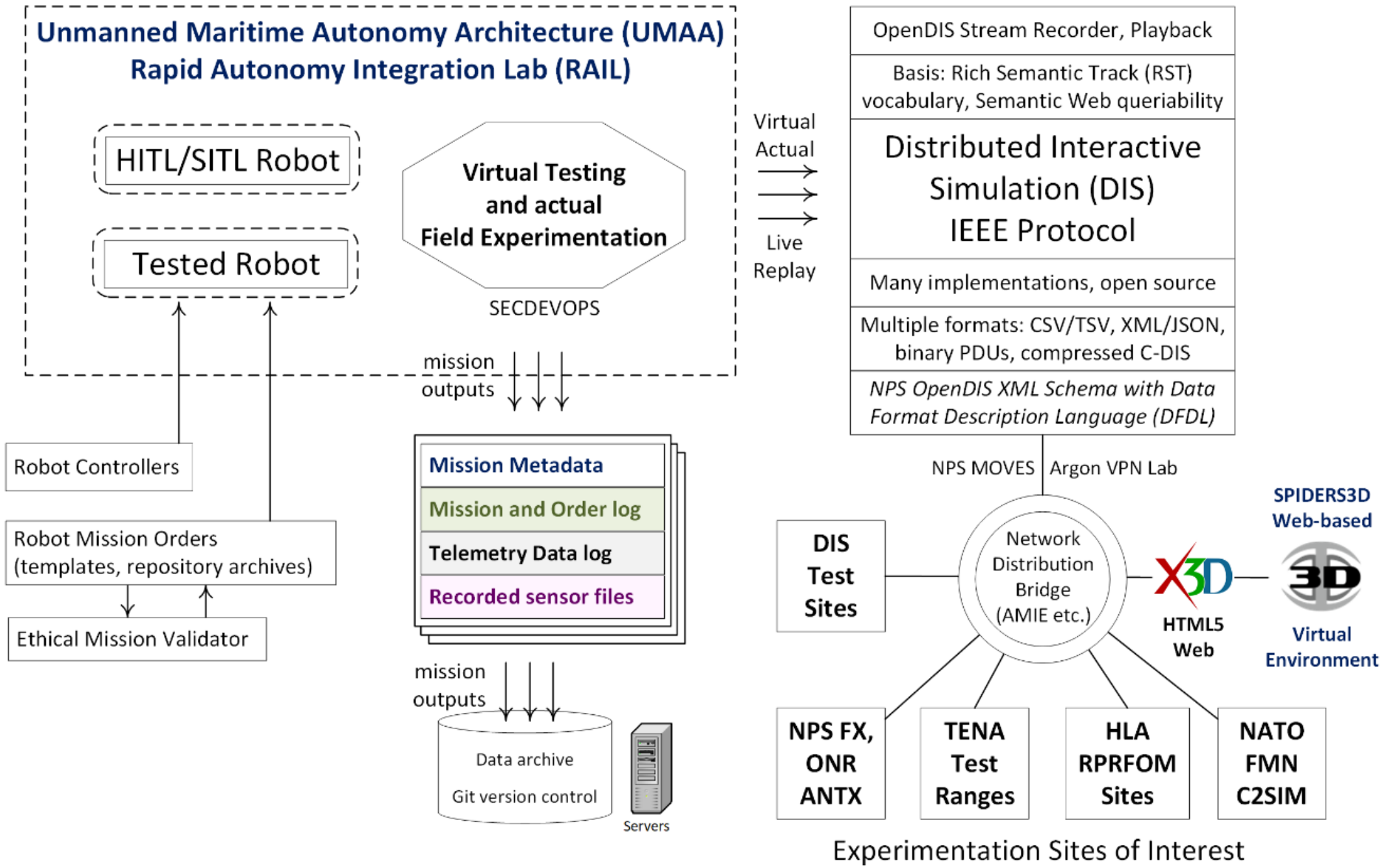
Data Strategy for Unmanned Systems

Field Experimentation (FX), Simulation and Analysis

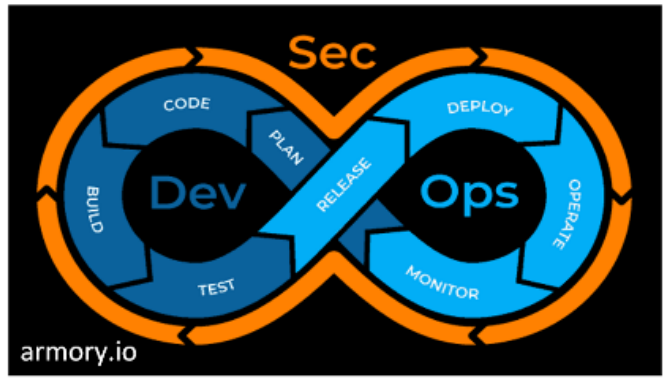
Abstract. Data collection and analysis techniques for robot experiments are haphazard and incomplete. Building best-practice workflows for data and metadata from unmanned systems can leverage both field experimentation (FX) and simulation to support archival data re-use and repeatable analysis. Reusable end-to-end data workflows are needed.

https://wiki.nps.edu/display/NOW/Data+Strategy+for+Unmanned+Systems

Live-Virtual-Constructive (LVC) Data Capture for Unmanned Systems Mission Analysis



Can we **Accredit** with automated simulation testing as part of DevSecOps and Digital Engineering?



- Incremental design and continuous testing for rapid revisions, upgrades
- Agile methods for code quality
- Unit test suite to validate full set of functional, security requirements
- Automatic confirmation of all design and security requirements before new release
- Result: automatic, more thorough V V+A

Existing VV+A procedures, with proposed standard

- Define Goal Outcomes
- Measurable System Requirements
- Test cases to demonstrate required result
- Conduct simulation "episode"
- Measure test outcomes
- Record results (pass/fail)
- Repeat steps for accreditation

Simulation stream

Accredited tests

- Record simulation "episode"
- **Verification** outcome measurement
- Confirm correct reference test results

Simulation episode test repository

Goal: create repeatable suite of tests for V V+A ?

Simulation test pass/fail ?

Validation set of test results

Developmental tests

- Record simulation "episode"
- **Verification** outcome measurement
- Measure *latest* test results

Simulation stream replay repository

Simulation stream

Conduct simulation "episodes"

Automating Verification Validation Acreditation V V+A

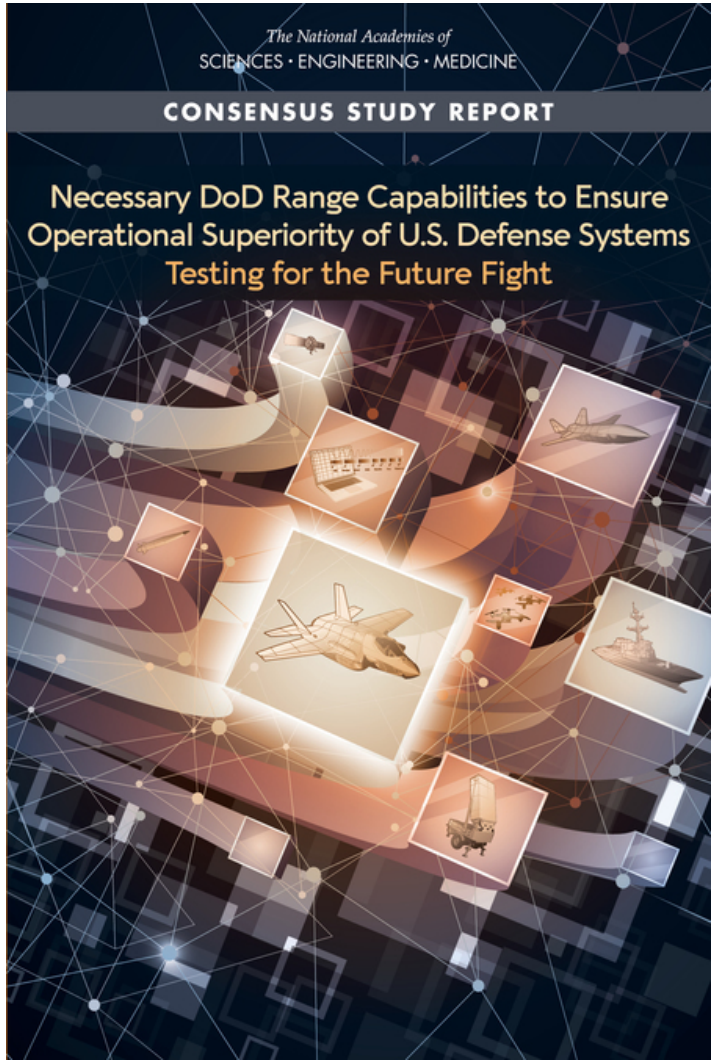
February 2021
SISO SIW Workshop
V V+A meetings

Imperatives and possibilities in military context

Assessing Physical and Technical Suitability of DoD Test and Evaluation Ranges, Infrastructure

Rigorous operational testing (OT) of weapon systems procured by the U.S. Department of Defense (DoD) is fundamental to ensuring that these sophisticated systems not only meet their stated requirements, but also perform under realistic operational conditions when faced by determined adversaries employing their own highly capable offensive and defensive weaponry. DoD's test and training range enterprise provides the geography, infrastructure, technology, expertise, processes, and management that make safe, secure, and comprehensive OT possible.

The challenges facing the nation's range infrastructure are both increasing and accelerating. Limited test capacity in physical resources and workforce, the age of test infrastructure, the capability to test advanced technologies, and encroachment impact the ability to inform system performance, integrated system performance and the overall pace of testing.



National Academies of Sciences, Engineering, and Medicine. 2021. Necessary DoD Range Capabilities to Ensure Operational Superiority of U.S. Defense Systems: Testing for the Future Fight. Washington, DC: The National Academies Press. https://doi.org/10.17226/26181 (includes presentation)

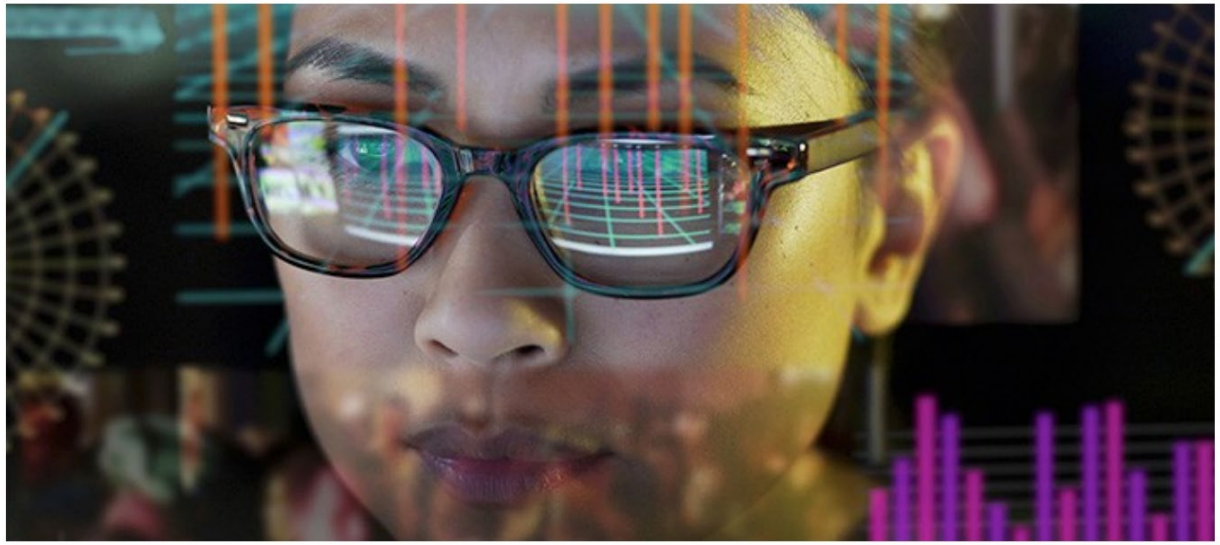
UNESCO member states adopt the first ever global agreement on the Ethics of Artificial Intelligence

AI is pervasive, and enables many of our daily routines - booking flights, steering driverless cars, and personalising our morning news feeds. AI also supports the decision-making of governments and the private sector.

AI technologies are delivering remarkable results in highly specialized fields such as cancer screening and building inclusive environments for people with disabilities. They also help combat global problems like climate change and world hunger, and help reduce poverty by optimizing economic aid.

But the technology is also bringing new unprecedented challenges. We see increased gender and ethnic bias, significant threats to privacy, dignity and agency, dangers of mass surveillance, and increased use of unreliable AI technologies in law enforcement, to name a few. Until now, there were no universal standards to provide an answer to these issues.

https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence



Audrey Azoulay, Director-General of UNESCO presented Thursday the first ever global standard on the ethics of artificial intelligence adopted by the member states of UNESCO at the General Conference.

This historical text defines the common values and principles which will guide the construction of the necessary legal infrastructure to ensure the healthy development of AI.

Recommendations: Protecting data, Banning social scoring and mass surveillance, Helping to monitor and evaluate, Protecting the environment

Analytic questions of interest – to ethical AI practitioners

Repeatability

- What happened? What changed? What's next? What will happen?
- What can we learn from all data + all operations? … Will we prevail?

Understandability

- Do I entrust this robot with authority for lethal or lifesaving force?
- Do these combined human-machine actions make sense to team?

Trust and Verification, Validation, Accreditation (V V+A)

- Is a robot system (hardware, software, sensors) qualified to deploy?
- What is Modeling + Simulation basis for analysis-development loop?

Conclusions and Recommendations

- Metrics are essential for ethical AI repeatability
- Metrics actually drive conceptualization, operationalization, evolution
  - You Get What You Measure
- Testing brings theory into practice
- Live-Virtual-Constructive (LVC) approaches are possible
- Building testbed frameworks for measurable performance establishes basis for verification validation + accreditation (VV+A), robot certification, even actionable conventions, professional practices, and laws

- Is it time for AI ethics to “get real” and “walk the walk” yet?

Contact

Don Brutzman

brutzman@nps.edu
http://faculty.nps.edu/brutzman

Code USW/Br, Naval Postgraduate School
Monterey California 93943-5000 USA
1.831.656.2149 work
1.831.402.4809 cell

Contact

Curt Blais

clblais@nps.edu
https://nps.edu/faculty-profiles/-/cv/clblais

Code MV/Bl, Naval Postgraduate School
Monterey California 93943-5000 USA
1.831.656.3215 work