Attack Surfaces of Autonomy

David Tate
Institute for Defense Analyses

Work sponsored by OASD(R&E)

March 2019

We know TEVV* for autonomous systems will be hard

Can't test exhaustively – state space is too big

Can't statistically sample – response isn't smooth

New learning can invalidate past results

Rich human-machine teaming (HMT) makes humans part of the system under test

*Test, Evaluation, Verification and Validation

Evidence-Based Licensure might be a solution

Define **dependability cases**

Accumulate **evidence**

Construct explicit **dependability arguments**

Establish **third -party confidence**

License for use within **defined limits**

Form 25

6390

Department of Public Safety - Bureau of Building Inspection

CITY OF PHILADELPHIA

Elevator Operator's License

Olga Villari

having been duly examined and having qualified in accordance with the Ordinance of Council, Ordinance Revised, June 11, 1940, is hereby licensed to operate a Passenger Elevator for one year from date of issue.

Geo. U. Siegrist

CHIEF

DATE OF ISSUE Oct. 23, 1944

“Dependability” means everything we care about

Mission performance

Safety

Security

Reliability

Each aspect of dependability generates attack surface

If your system can be made unsafe, you lose

If your system can be made unreliable, you lose

If your system can be made to fail the mission, you lose

If adversaries can make your system
do what they want…

Is “cybersecurity” really a distinct thing for autonomy?

Are cyberattacks importantly different from other threats to dependability?

Is robustness to potential adversarial actions importantly different from robustness to environmental complexity?

Claim:

Cyber assurance is just one dimension of dependability, and is best treated as part of a holistic dependability case

The capabilities that enable autonomy are…

Perception

Reasoning

Planning / Deciding

Learning

Self-organizing behavior

Human-Machine Teaming (HMT)

The technologies that enable these include…

Machine learning

Computer vision

Sensor fusion

Knowledge representation

Inference engines

Path planning

Optimization

Expert systems

HMT CONOPS

How do these generate attack surfaces?

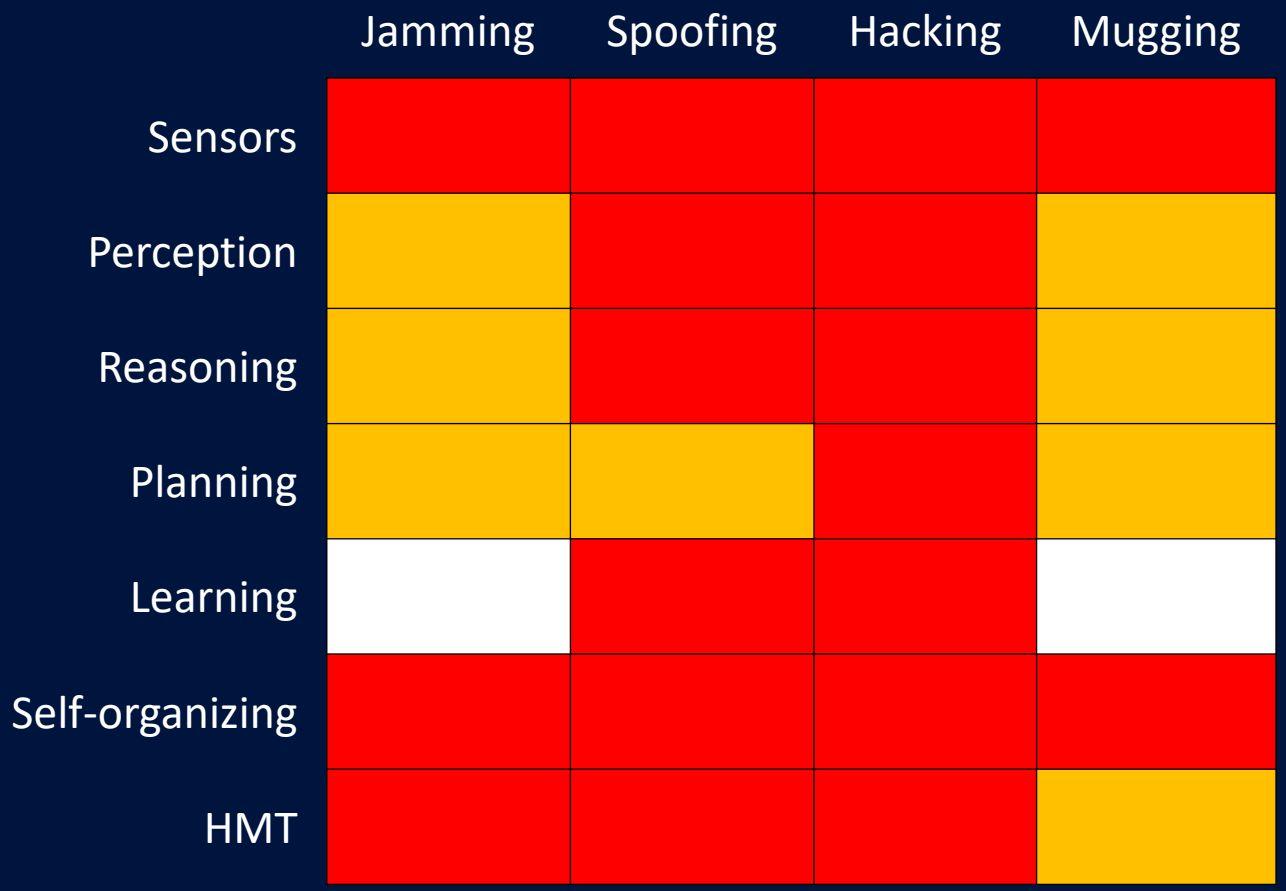
Attack surfaces arise from the **inputs** to these technologies and capabilities…

Perception : sensors, algorithms, stored data, training
Reasoning : world model, ontology, algorithms
Planning : world model, stored data, algorithms
Learning : [inputs], training, algorithms, architecture
Self-organizing : world model, algorithms, sensors/ comms
HMT: messages sent/ received, world model, CONOPS

An informal taxonomy of mischief

1. "Jamming" – information denial (to or from)
2. "Spoofing" – input impersonation or confusion
3. "Hacking" – unauthorized access / control
4. "Mugging" – threat of physical harm

How can adversaries make autonomy undependable?



Direct vs. Indirect effects

New tools for new kinds of dependability case

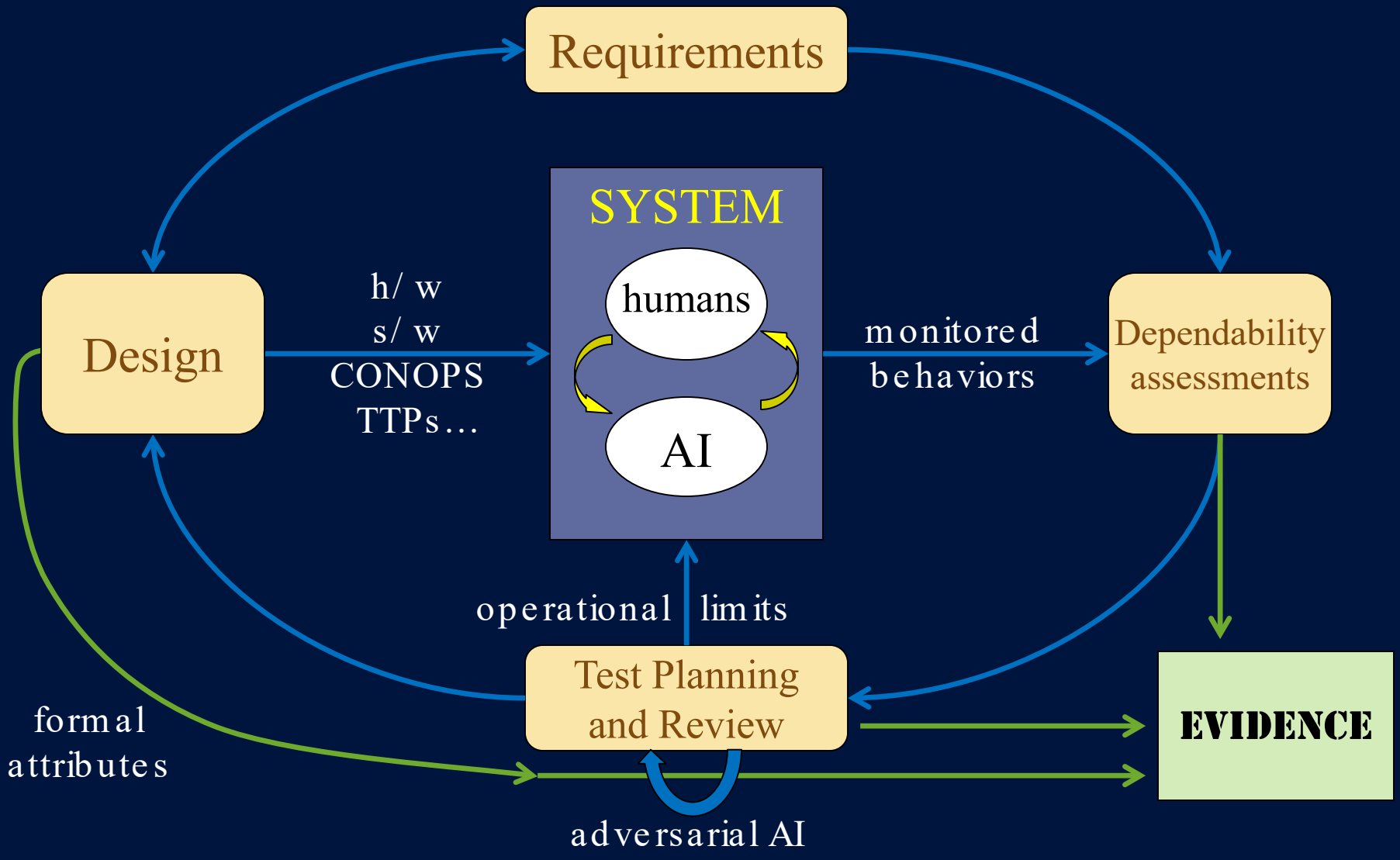
Formal methods for software verification

Cognitive instrumentation / explainable AI

Greatly expanded use of M&S, including LVC experimentation and intelligent adversarial testing

Rapid automated sequential test design

Here's the skeleton of an iterative test framework



What does success look like?

The **time series of performance** against the requirements, in the face of adversarial testing intended to induce dependability failure, provides a **richer body of evidence** toward licensure than simple pass/ fail testing

This approach **quantifies robustness**, based on the history of behavior and resilience under stress

Supports **partial licensure** by identifying **operational bounds** within which performance is most dependable, **evolving over time** pre- and post-fielding

Summary

Autonomous systems face adversarial threats (including environmental complexity) to cognitive abilities

Explicit dependability arguments for autonomous systems will need more compelling evidence than pass/ fail testing can provide

Intelligent adversarial testing will be key to developing the evidence that supports confident fielding

Questions?

Backup

Human-machine teaming makes everything harder

Nontrivial teaming of humans with machines that have some autonomous capabilities causes rapid expansion of the state space, the set of failure modes, and the unpredictability of the human-machine system

"Level of autonomy" is not the relevant metric for how hard TEV&V will be – less autonomy can mean more complicated human-machine teaming, and thus more failure modes and broader attack surfaces