INSTITUTE FOR DEFENSE ANALYSES

AI + Autonomy T&E in DoD

SLIDES ONLY

NO SCRIPT PROVIDED

Rebecca M. Medlin, Project Leader

Brian D. Vickers
Matthew R. Avery
Rachel A. Haga
Mark R. Herrera
Daniel J. Porter
Stuart M. Rodgers

August 2023

IDA Document NS 3000083

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task C9082, "CRP Statistics WorkGroup." The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Dr. Shawn W. Whetstone from the Operational Evaluation Division.

For more information:
Dr. Rebecca Medlin, Project Leader
rmedlin@ida.org • (703) 845-6731

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS 3000083

AI + Autonomy T&E in DoD

Rebecca M. Medlin, Project Leader

Brian D. Vickers
Matthew R. Avery
Rachel A. Haga
Mark R. Herrera
Daniel J. Porter
Stuart M. Rodgers

Executive Summary

Interest in and technological progress in artificial intelligence (AI) are increasing rapidly. The technologies underlying AI include a broad range of algorithms, each with their own pros and cons regarding performance, reliability, robustness, test needs, and auditability, among others. There is not yet broad understanding of how AI systems work, how testing may change, or how AI is impacting DoD. This briefing is to help inform people on what they should know about AI at an introductory level.

The briefing takes a practical approach to the questions that DoD support staff are asking most commonly. These topics are organized into sections of the briefing, including:

1. What is AI? This section differentiates AI from autonomy, describes some of the algorithms that power AI systems, and introduces issues with the ambiguity of its use.
2. How does AI impact test and evaluation (T&E)? This section clarifies that we are in an early, experimental phase of T&E of AI where we know some of the common issues systems are running into but that we are still creating test methods and defining best practices.
3. What makes DoD AI unique? This section looks at various AI implementations in games, industry use cases, and DoD missions. It emphasizes that industry is having a hard time with complex AI use cases, and that DoD's missions are more complex and less constrained.
4. What is the warfighter's role? This section emphasizes that evaluating the role of humans working with AI systems becomes more important than historically because they are probabilistic and imperfect, and responsibility for system performance will still fall to the operator.
5. What is the state of AI T&E in DoD? This section provides an overview of how IDA is supporting AI across various DoD sponsors, pursuing efforts to educate and collaborate, and pushing the T&E community toward more rigorous evaluations.

IDA

AI + Autonomy T&E in DoD

Brian Vickers, Rachel Haga, Matt Avery,
Daniel Porter, Stu Rodgers, Mark Herrera

7-Aug-2023

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

Questions we're going to tackle

1. What is "Artificial Intelligence (AI)"?	The AI domain is filled with buzzwords. Focus on defining system features that require new test methods.
2. How does AI impact T&E?	AI isn't new, but systems with AI pose new challenges and require structural changes to how we T&E.
3. What makes DoD AI unique?	Industry AI applications often lack the task complexity and severe consequences of risk faced by DoD.
4. What is the warfighter's role?	T&E must assure warfighters have calibrated trust and an adequate understanding of system behavior.
5. What is the state of AI T&E in DoD?	

Chat GPT

THEWRAP

Chat GPT Proves That AI Could Be a Major Threat to Hollywood Creatives – and Not Just Below the Line | PRO Insight

AI Chatbot Wars: Google management on alert after seeing ChatGPT's potential

Could the AI bot one day replace the search engine?

DAILY BEAST

OpenAI's Impressive New Chatbot Isn't Immune to Racism

Tony Ho Tran

December 5, 2022 · 6 min read

The end of the world is near

The best thing since sliced bread

MOTHERBOARD
TECH BY VICE

OpenAI's New Chatbot Will Tell You How to Shoplift And Make Explosives

ChatGPT is yet another reminder that all AI systems are prone to bias and misuse.

By Janus Rose
NEW YORK, US

Let's Talk About ChatGPT and the Fall of Humanity

By Seth Freilich | Miscellaneous | December 22, 2022 | 38 Comments

ChatGPT is 80% effective at identifying Alzheimer's disease, study shows

The chatbot could help by picking up on clues in a spontaneous speech.


Test and
Evaluation is
the mirror
that reflects
back reality

1

What is “Artificial Intelligence”?

Why is it so hard to define?

"... the ability of machines to perform tasks that normally require human intelligence ..."

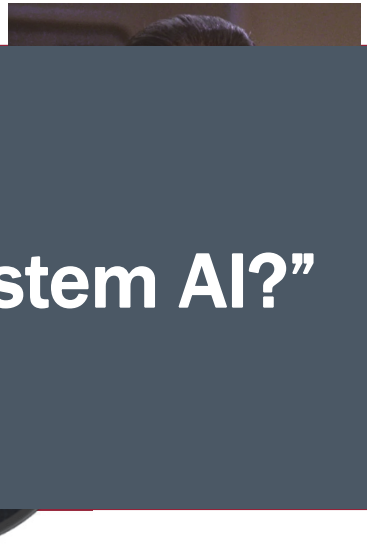
- DoD AI Strategy, 2018

What people think of

What meets the definition

All ~~models~~ *definitions of AI* are wrong, some are useful

"Is this system AI?"



What about AI needs to be tested differently?

AI & Autonomy

Autonomy

Behavior or outcome

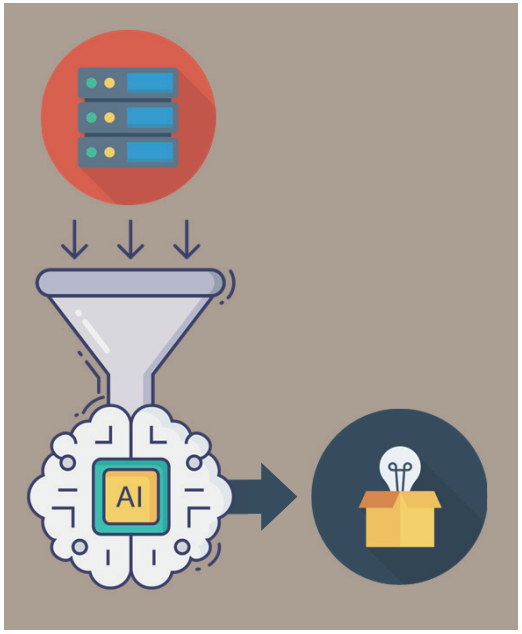
A system performing a task or making decisions independently

Artificial Intelligence (AI)

Underlying technology

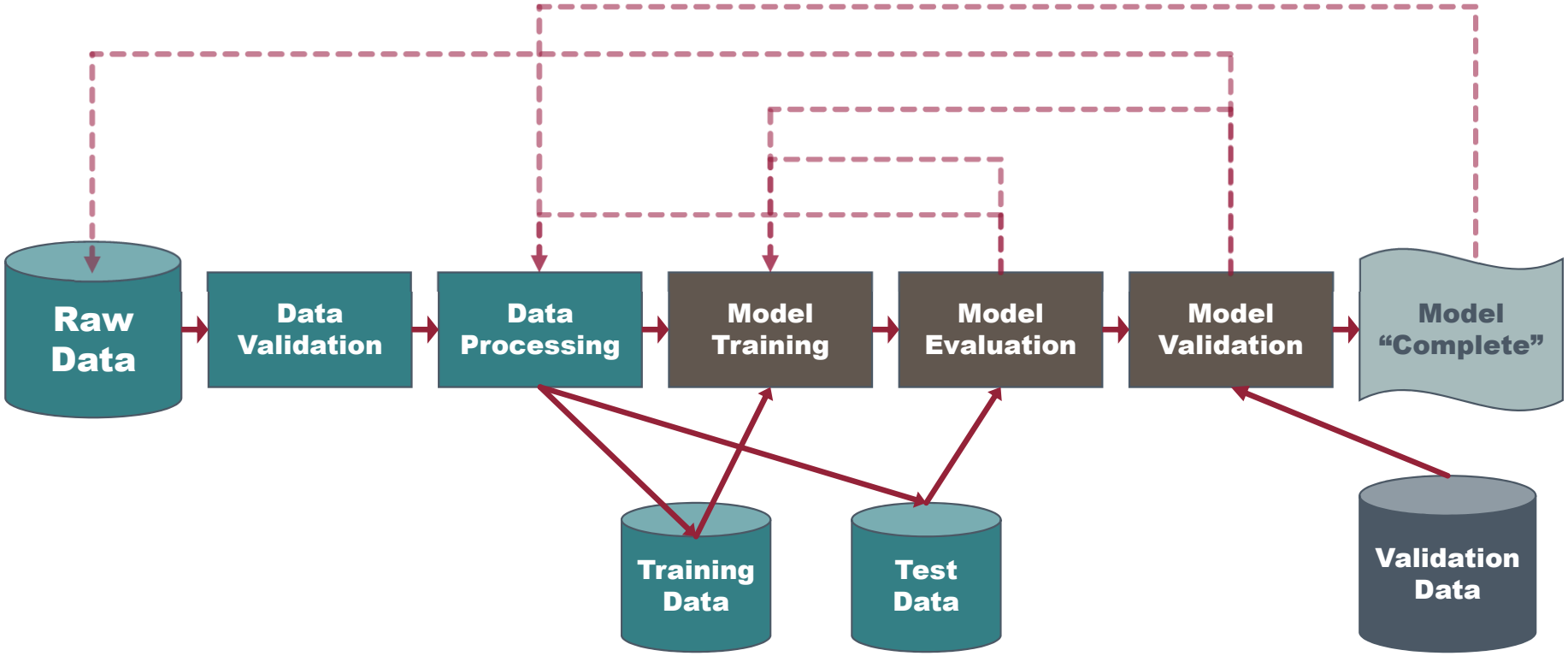
Technologies, algorithms, how task solutions are implemented

When people say “AI”, they usually mean “Machine Learning”

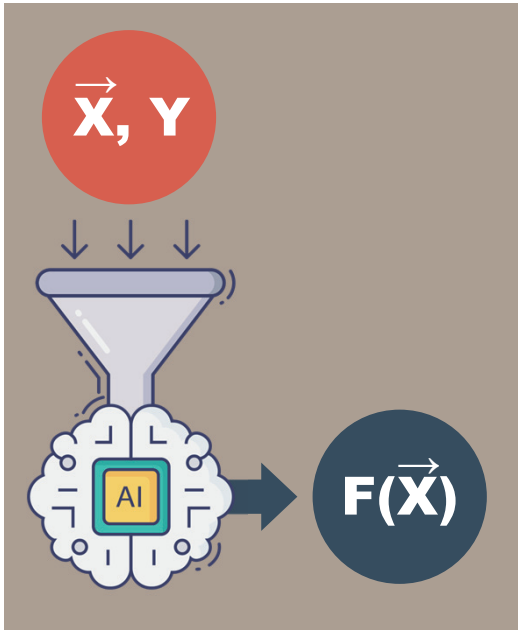


- ***Learning*** is the process of drawing inferences from data
- ***Machine Learning*** uses algorithms defined in software to build “machines” that “learn” – consume data and return inferences

Machine Learning Ops (ML Ops)



Supervised Learning

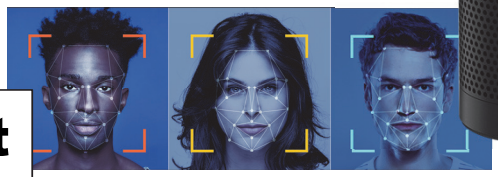


$\vec{X}_{new} \rightarrow F(\vec{X}) \rightarrow \hat{Y}$

Uses **data** or ***features*** mapped to an **outcome** to guide the learning process and produce a model that can predict the outcome

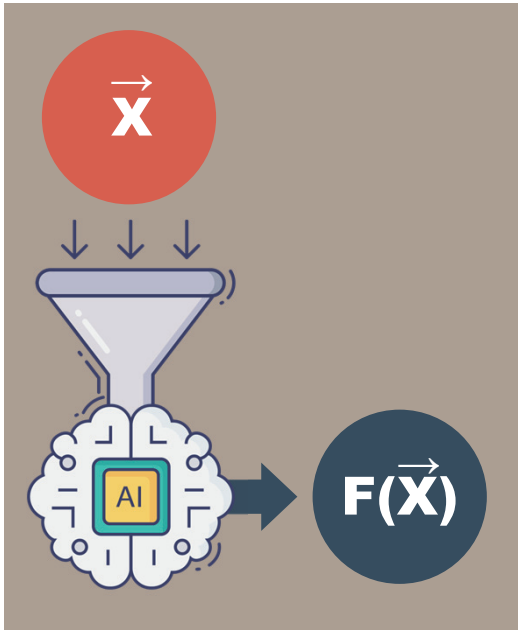
Algorithms: Regressions (e.g. Linear, Logistic), Classifications, Neural Nets

Applications: Forecasting Modelling, Computer Vision, Optical Character Recognition (OCR)



Project Maven

Unsupervised Learning



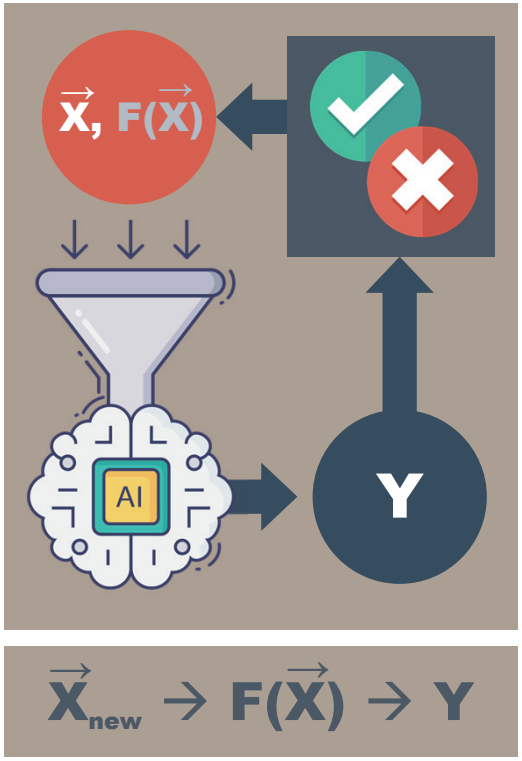
$\vec{X}_{new} \rightarrow F(\vec{X}) \rightarrow Y$

Based on data alone with no measurements of outcome

Algorithms: Clustering, Topic Modeling, Latent Semantic Indexing

Applications: Dimension Reduction, Recommender Systems

Reinforcement Learning



Incorporates feedback over time in response to its outputs

Algorithms: Model-free vs. Model-based, Reinforcement Learning with Human Feedback

Applications: Games (Chess, Go, Doom, StarCraft), Chat Bots, Shield AI

OpenAI

AlphaGo

What is "Artificial Intelligence"?
Why is it so hard to define?

The "AI" domain is filled with buzzwords, ambiguity, and disagreement.

Proposed definitions often do *not* lend themselves to practical applications.

Different types of Machine Learning (ML) present different challenges.

- **Not all AI/ML will require additional T&E considerations.**
- **We must identify what characteristics of AI/ML warrant extra T&E attention.**

2

How does AI impact T&E?

How is it different from traditional software?

Testing something you do not understand

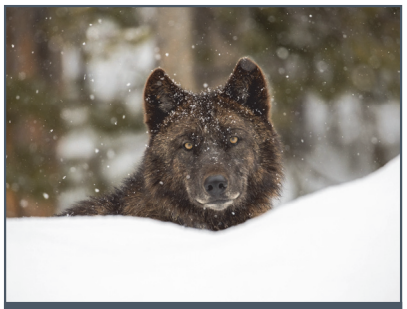
Do we understand AI enough to adequately test and evaluate it?

- Limited understanding of compressible flow
- Required “proof of concept” testing in operational setting
- Accepted a lot of risk
- Most instrumented aircraft – collecting lots of data

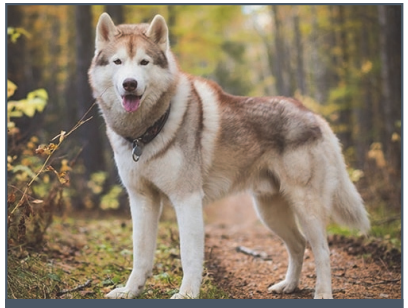
Is it a tank?

Wolf

Husky



Wolf



Husky

Wolf

Husky

*Images notional

Is it a husky or a wolf?

Is current domain expertise enough?

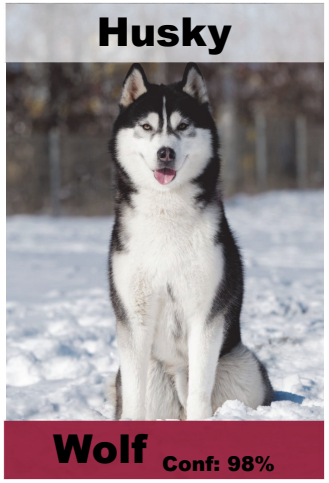
*Images notional

T&E of AI will require new expertise.

And in many cases, this expertise may not exist.

Image Quality SME

- Target in motion / blurry
- Target contrast with background
- Orientation
- Abnormal configurations



Dog vs Wolf SME

- Body Size
- Head Size
- Teeth length
- Muzzle
- Eyes
- Ears

Assurance starts with data and model access

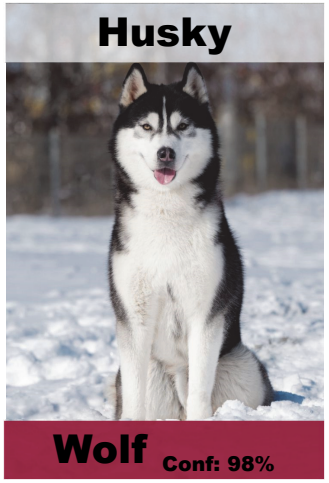
*Images notional

- **Data is a key driver of AIES performance**
- **AI exacerbates "Tech Debt" and requires operational realism earlier in the lifecycle**
- **Beyond data quality, understanding data informs test designs**

Test design of AIES will be informed by the data and model.

Image Quality SME

- Target in motion / blurry
- Target contrast with background
- Orientation
- Abnormal configurations



Dog vs Wolf SME

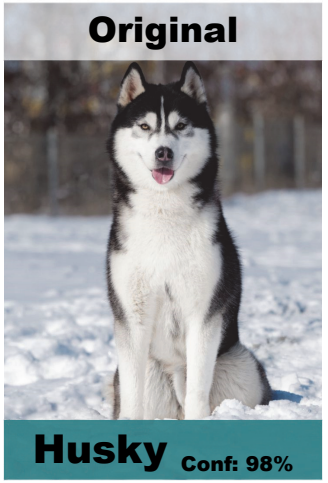
- Body Size
- Head Size
- Teeth length
- Muzzle
- Eyes
- Ears

New Adversarial Attack Vectors

*Images notional

AI creates new cyber survivability challenges.

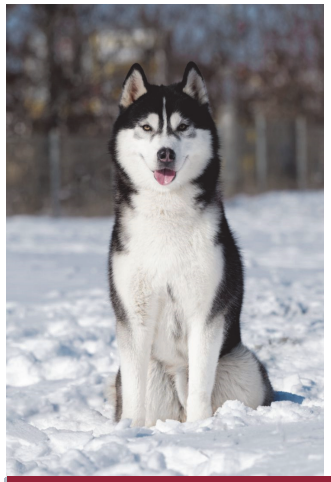
Specialized expertise will be required to test for cyber attacks targeted at AI algorithms!



What can I infer from my test?

- ‘Black box’ systems are not transparent across stakeholders
- Even with ‘complete’ transparency, many model behaviors are difficult to ‘explain’
- Model *not* inherently tied to physics and/or geometry, resulting in unanticipated results

*Images notional



How do we design tests (with limited resources!) to mitigate these issues?

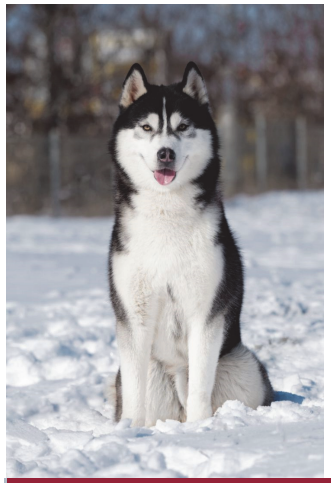
How much can I generalize my test results?

- The non-linear, discontinuous nature of many AIESs can make behavior difficult to predict
- The state space of possible behaviors and outcomes is vast
- AI is commonly “brittle” and overfit to its training data

Is my test operationally realistic?

We will need to test more to establish current statistical confidence in tests

**Images notional*



How do we design tests (with limited resources!) to mitigate these issues?

How much can I generalize my test results?

- **There is no standard definition of "M&S" or "Digital Twin"**
- **Using M&S to validate AI is immature**
- **Unclear how the AI brittleness will impact M&S requirements**
- **Real-world tests will continue to be crucial to OT&E**

How does AI impact T&E?
How is it different from traditional software?

- **Predictability + Transparency**
- **Impact of Data on Performance**
- **New Adversarial Attack Vectors**
- **Overfitting / Brittleness**

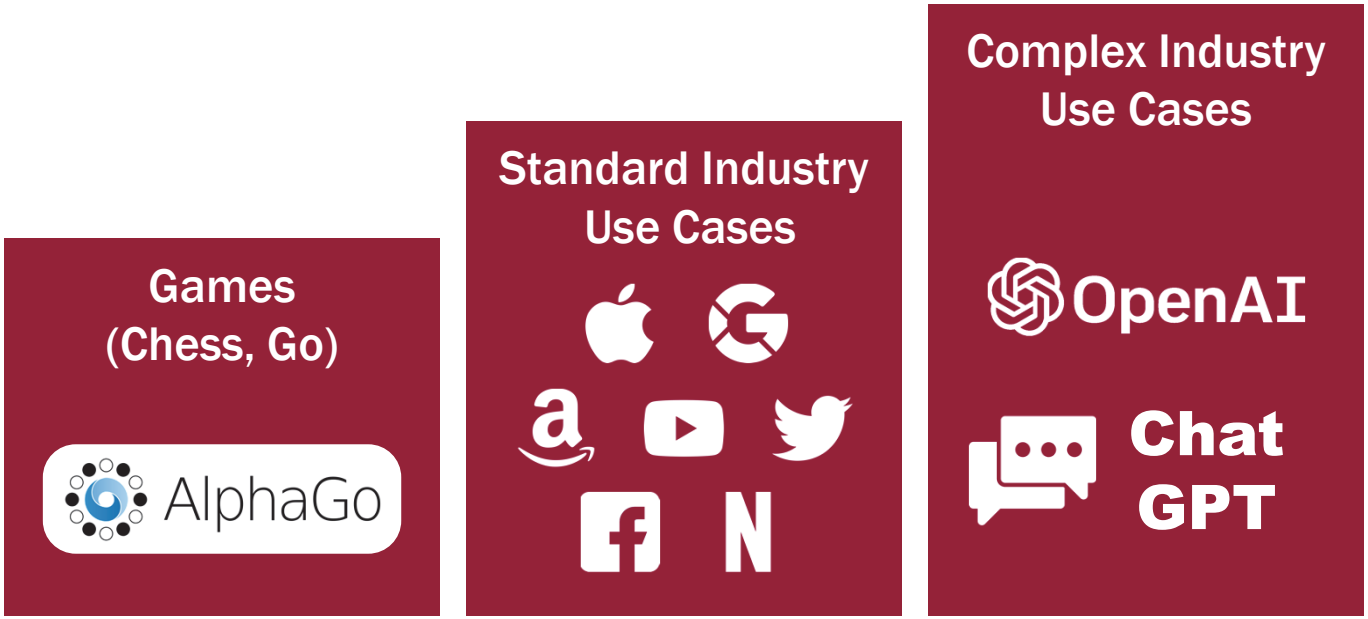
Failing to expand and adapt current T&E practices will result in increased risk when fielding DoD AI&A systems

3

What makes DoD AI unique?

And why should you care?

What makes DoD AI&A unique?



What makes DoD

July 18 (Reuters) - The National Highway Traffic Safety Administration (NHTSA) is opening a special investigation into the crash of a 2021 Tesla Model Y vehicle that killed a motorcyclist in California, it said on Monday.

Since 2016, NHTSA has opened 37 special investigations of crashes involving Tesla vehicles and where advanced driver assistance systems such as Autopilot were suspected of being used. A total of 18 crash deaths were reported in those Tesla-related investigations, including the most recent fatal California crash.

- Interactions with other (unknowable) drivers and pedestrians
- Behavior governed by regulation and signage that change with locality
- Construction that must be detected and requires 'breaking' simple traffic rules
- Environmental cues and sensors can be obstructed with road debris

Complex Industry Use Cases

TESLA

What makes DoD AI&A unique?

Games (Chess, Go)

AlphaGo

- Interactions with adversaries actively trying to undermine the mission
- Operating with limited regulatory guidance and unclear constraints
- Deviations from nominal may be common place and higher risk
- Harsh environments will require robust systems and well-maintained sensors

DoD Use Cases

DIU Ground Vehicle Autonomous Pathways project

What makes DoD AI&A unique?

Games
(Chess, Go)

AlphaGo

- Testing demonstrated that ocean environment impacts performance
  - E.g. Depth, Bottom Sediment, Acoustic Noise
- Knifefish OT occurred in similar regions where the developer trained the system; Issues with test data independence
  - New clutter/false alarms and ocean environments must be tested

DoD
Use Cases

Knifefish

What makes DoD AI unique?
And why should you care?

AI is more than just math.

- Data and computational power
- Ability to crowd source
- Task complexity and ambiguity
- Harsh Environments

AI T&E is not "solved."

We can learn from industry, but we must recognize that our applications of AI will have different T&E requirements

4

What is the warfighter's role?

How does this impact Human-Systems Integration (HSI) T&E?

Offloading warfighter work to AI doesn't make HSI less important.

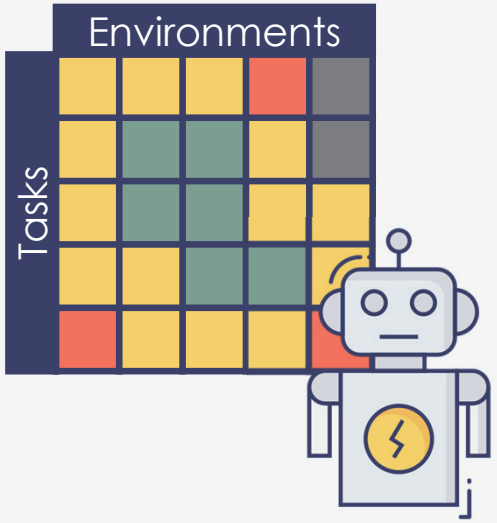
HSI is more important than ever.

Calibrating Trust

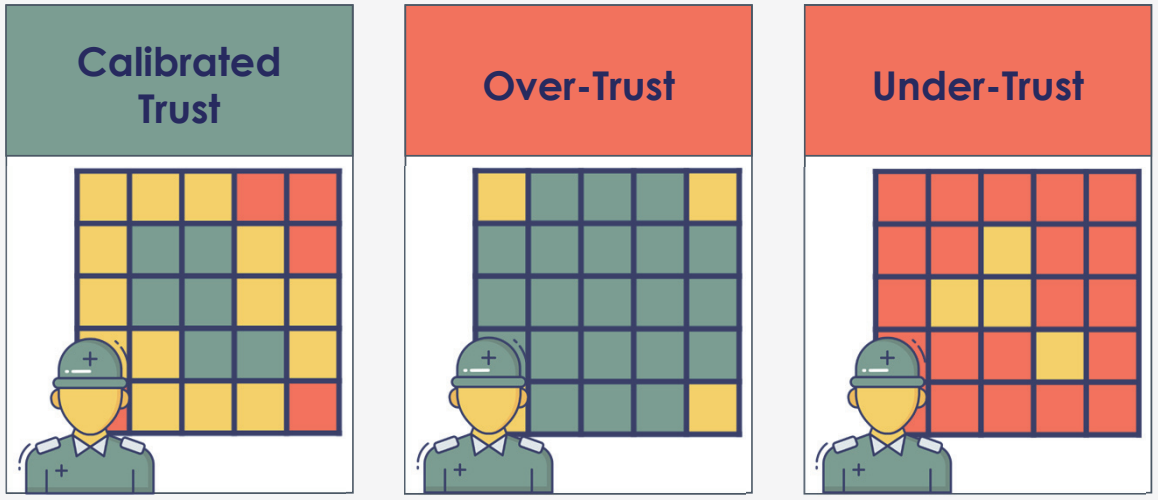
Trust is a person's belief that something can be depended on in vulnerable or uncertain situations. The critical outcome of trust is reliance, which is the behavioral, continued use of a system.

Calibrating Trust

Calibrated Trust occurs when the warfighter's operational reliance aligns with the system performance for a given context.



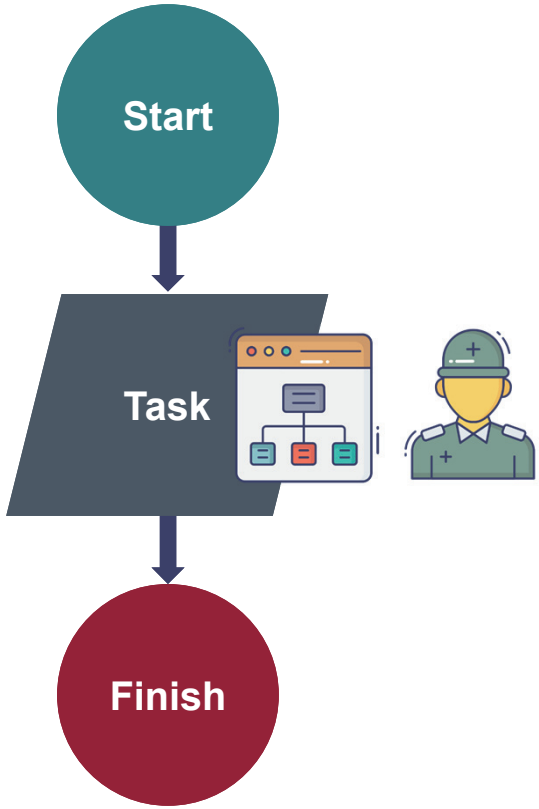
System performance



Warfighter's trust and subsequent reliance on the system

Assessing Usability

Usability is the fitness of a tool for a task. Composed of **utility** and **ease of use.**



Give Initial Orders

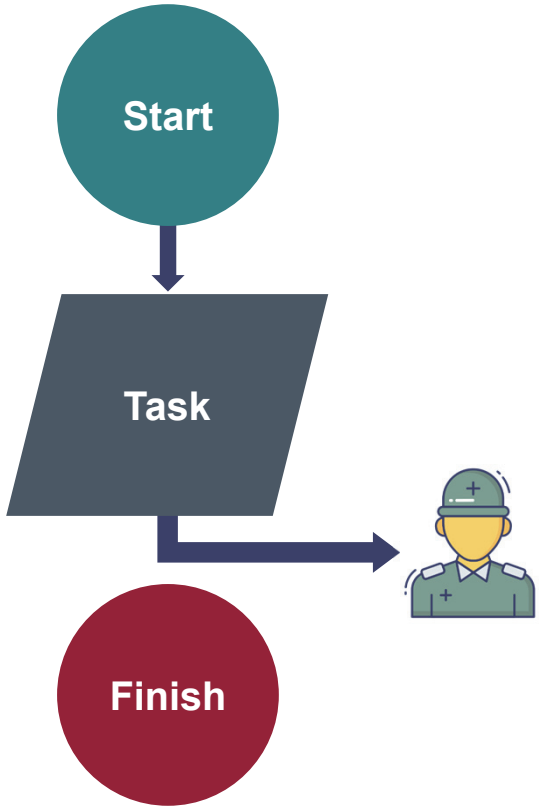
Must assess whether the system achieved the warfighter's intent

Extract Information

Must confirm warfighter has information needed to be situationally aware

Assessing Usability

Usability is the fitness of a tool for a task. Composed of **utility** and **ease of use.**



Give Initial Orders

Must assess whether the system achieved the warfighter's intent

Extract Information

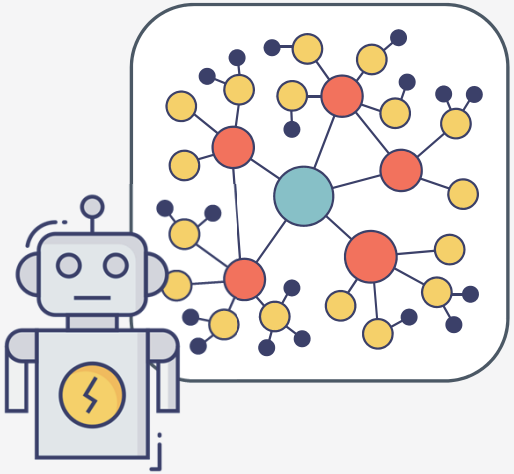
Must confirm warfighter has information needed to be situationally aware

Intervene & Take Over

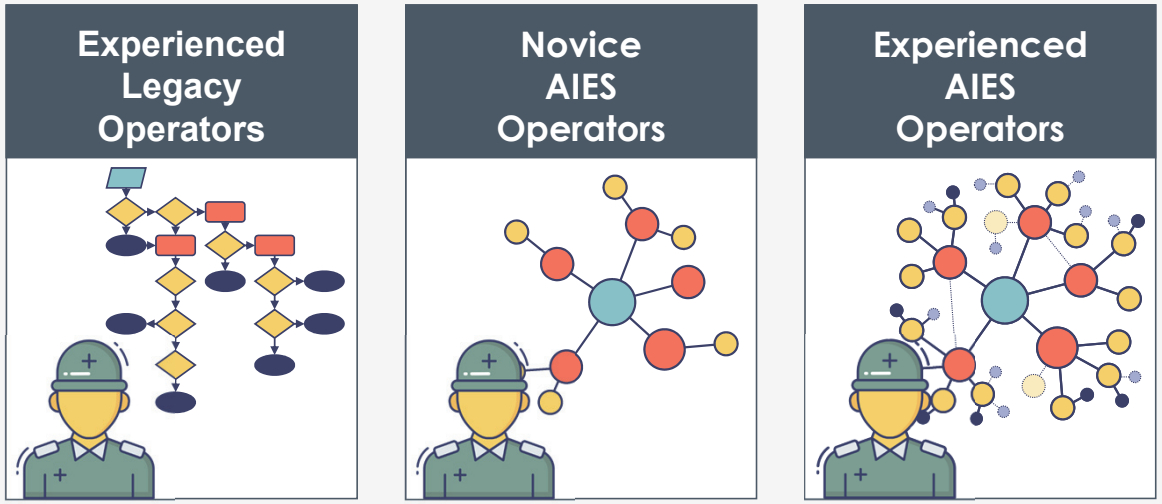
Must evaluate off-nominal situations where tasks are handed off to warfighter

Building Mental Models

Humans form mental models of automation that allow them to infer the current state of a system from incomplete information and make predictions about future states based on the current one.

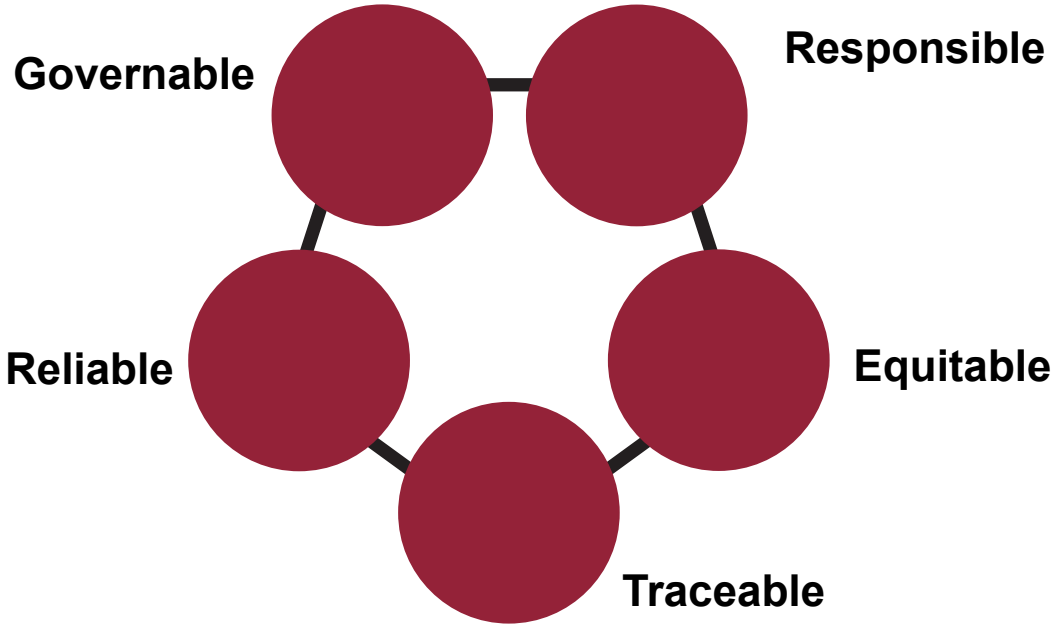


System's underlying model



Warfighter's mental model of system behavior

In 2020, DoD adopted 5 principles for the ethical use and development of AI



The DoD has started using the umbrella term

"Responsible AI (RAI)"

Work that predates the principles used the term

"Legal, Moral, Ethical (LME) AI"

What is the warfighter's role?
How does this impact HSI T&E?

Increased autonomy will result in fewer opportunities for human intervention—making limited warfighter touchpoints critical

T&E must assure warfighters have calibrated trust and an adequate understanding of system behavior and failure modes

5

What is the state of AI T&E in DoD?

Where do we go from here?

How is IDA involved with DoD AI&A?

IDA AI&A work is cross-divisional

USD P&R

National AI T&E Infrastructure Capability Gap

CDAO

1. What is the **demand** of AI&A programs?
2. What is the **supply** of relevant AI&A T&E infrastructure?
3. Where are **gaps** between AI&A demand and T&E infrastructure supply?

Addressing AI T&E Challenges

Learn and Educate

- Educate current workforce on AI T&E challenges (**in progress**)
  - AI & Autonomy DODM, AO Course AI Briefs
- Create an AIES test concept (**in progress**)
  - Identify features of AIES that warrant additional T&E
  - Identify artifacts that would be required to build ‘assurance'
- Collaborate with other DoD stakeholders. **Bring OT perspective!** (**in progress**)
  - E.g. DAU AI Training; NAITIC Study

Update OT Requirements

- AI and Autonomy T&E DODM (**in progress**)
- Require access to data and models
- Require programs to report the usage of ML algorithms
- Consider constraints on “continuous learning” systems
- Test for calibrated Trust, Adversarial AI attack vector, and “responsible” employment

Questions we [hopefully?] tackled

1. What is "Artificial Intelligence (AI)"?	The AI domain is filled with buzzwords. Focus on defining system features that require new test methods.
2. How does AI impact T&E?	AI isn't new, but systems with AI pose new challenges and require structural changes to how we T&E.
3. What makes DoD AI unique?	Industry AI applications often lack the task complexity and severe consequences of risk faced by DoD.
4. What is the warfighter's role?	T&E must assure warfighters have calibrated trust and an adequate understanding of system behavior.
5. What is the state of DoD AI T&E?	

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-08-2023	Final	

4. TITLE AND SUBTITLE

AI + Autonomy T&E in DoD

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Haga, Rachel A.; Vickers, Brian D.; Avery, Matthew R.; Porter, Daniel J.; Rodgers, Stuart M.; Herrera, Mark R.

5d. PROJECT NUMBER

C9082

5e. TASK NUMBER

C9082

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER

NS - 3000083

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Central Research Project

10. SPONSOR/MONITOR'S ACRONYM(S)

CRP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

This OED Draft has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

13. SUPPLEMENTARY NOTES

Project Leader: Medlin, Rebecca M.

14. ABSTRACT

Test and evaluation (T&E) of AI-enabled systems (AIES) often emphasizes algorithm accuracy over robust, holistic system performance. While this narrow focus may be adequate for some applications of AI, for many complex uses, T&E paradigms removed from operational realism are insufficient. However, leveraging traditional operational testing (OT) methods for to evaluate AIESs can fail to capture novel sources of risk. This brief establishes a common AI vocabulary and highlights OT challenges posed by AIESs by answering the following questions:
1. What is "Artificial Intelligence (AI)"?
a. A brief "AI Primer" defines some common terms, highlights words that are used inconsistently, and discusses where definitions are insufficient for identifying systems that require additional T&E considerations.

15. SUBJECT TERMS

artificial intellgience; test and evaluation; human-system integration

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Rebecca M. Medlin
Unclassifed	Unclassified	Unclassified	SAR	48	19b. TELEPHONE NUMBER *(Include area code)* (703) 845-6731