INSTITUTE FOR DEFENSE ANALYSES

DATAWorks 2024: Developing AI Trust: From Theory to Testing and the Myths in Between

John Haman, Project Leader

Yosef Razin

April 2024

IDA Product ID 3001946

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 BD-9-2299(90), "Methods Development," for the Office of the Director, Operational Test and Evaluation. 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. Breeana Anderson, Dr. Brian Vickers, Dr. John Haman, Dr. Kelly Avery, Dr. Keyla Pagan-Rivera, and Dr. Luis Aguirre from the Operational Evaluation Division.

For more information:
Dr. John Haman, Project Leader
jhaman@ida.org • (703) 845-2132

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 Product ID - 3001946

DATAWorks 2024:
Developing AI Trust: From Theory to Testing and the Myths in Between

John Haman, Project Leader

Yosef Razin

Executive Summary

The Director, Operational Test and Evaluation (DOT&E) and the Institute for Defense Analyses (IDA) are developing recommendations for how to account for trust and trustworthiness in AI-enabled systems during Department of Defense (DoD) Operational Testing (OT). Trust and trustworthiness have critical roles in system adoption, system use and misuse, and performance of human-machine teams. The goal, however, is not to maximize trust, but to calibrate the human's trust to the system's trustworthiness. Trusting more than a system warrants can result in shattered expectations, disillusionment, and remorse. Conversely, under trusting implies that humans are not making the most of available resources.

Executive Order 14110 requires "safe, secure, and trustworthy development and use" of AI. Furthermore, the desired end state of the Department of Defense Responsible AI Strategy is trust (DOD Adopts Ethical Principles for Artificial Intelligence 2020). Trust and trustworthiness are not well characterized and there is no standard, widely accepted model for understanding them; and methods for quantifying them in test and evaluation (T&E) are still evolving. This deficiency has resulted in trust and trust calibration rarely being assessed in T&E, in part due to the contextual and relational nature of trustworthiness. For instance, the developmental tester requires a different level of algorithmic transparency than the operational tester or the operator (Canellas et al. 2017); whereas the operator may need more understandability than transparency. This difference in needs means that to successfully operationally test AI-enabled systems, such testing must be done at the right level, with the actual operators and commanders, up-to-date CONOPS, and sufficient time for training and experience for trust to evolve.

Tests of AI-enabled systems should also be conducted sequentially through the life of a program, rather than all at once. This is because machine behaviors are no longer as predictable or static as traditional systems but may continue to be updated and adaptive (Yaxley et al. 2021). Thus, testing for trust and trustworthiness cannot be one and done.

As uncertainty increases with the complexity of AI-enabled systems, trust becomes increasingly important psychologically, as the means by which humans manage uncertainty while continuing to make decisions and operate.

Therefore, it is critical to ensure that those who work within AI – in its design, development, and testing – understand exactly what trust actually means, why it is important, and how to operationalize and measure it.

This presentation at DATAWorks 2024 is one of the products developed by DOT&E and IDA to discuss trust, and specifically the challenge of T&E for evaluating trust in AI during OT. We empower testers by:

- Establishing a common foundation for understanding what trust and trustworthiness are;
- Defining key terms related to trust, enabling testers to think about trust more effectively;
- Demonstrating the importance of trust calibration for system acceptance and use and the risks of poor calibration;
- Decomposing the factors within trust to better elucidate how trust functions and what factors and antecedents have been shown to effect trust in human-machine interaction;
- Introducing concepts on how to design AI-enabled systems for better trust calibration, assurance, and safety;
- Pointing to sources for finding validated and reliable measures for trust and trustworthiness;
- Discussing common cognitive biases implicated in trust and AI and both the positive and negative roles biases play; and
- Addressing common myths around trust in AI, including that trust or its measurement does not matter, or that trust in AI can be "solved" with ever more transparency, understandability, and fairness.

This presentation is based on the article *Developing AI Trust: From Theory to Testing and the Myths In Between*, published in the March 2024 issue of the International Test and Evaluation (ITEA) Journal.

IDA

DATAWorks 2024

Developing AI Trust:
From Theory to Testing and the Myths in Between

Yosef S. Razin,
IDA
yrazin@ida.org

Dr. Kristen Alexander,
DOT&E
kristen.l.alexander5.civ@mail.mil

February 28, 2024

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

If *trusted AI* is the DoD's goal, it is critical to improve the T&E community's understanding of what *trust* and *trustworthiness* mean.

AP Photo/Andrew Harnik

"Our operators must come to

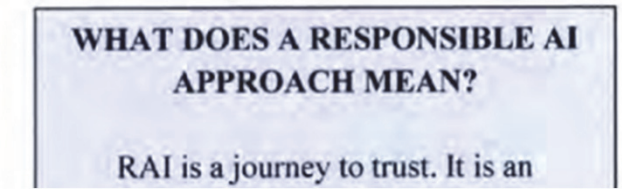
trust the outputs of AI systems,

our commanders must come to

trust the legal, ethical, and moral foundations of explainable AI,

and the American people must come to trust the values their Department of Defense has integrated into every application."

Deputy Sec. of Defense Kathleen Hicks



DoD Responsible AI (RAI) Strategy and Implementation Pathway

Executive Order 14110

If *trusted AI* is the DoD's goal, it is critical to improve the T&E community's understanding of what *trust* and *trustworthiness* mean.

AP Photo/Andrew Harnik

"Our operators must come to

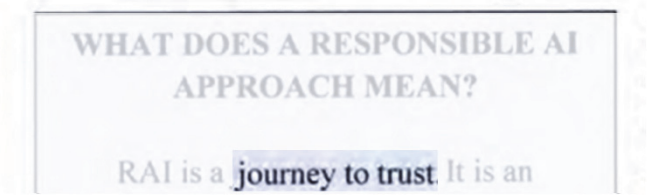
trust the outputs of AI systems,

our commanders must come to

trust the legal, ethical, and moral foundations of explainable AI,

and the American people must come to trust the values their Department of Defense has integrated into every application."

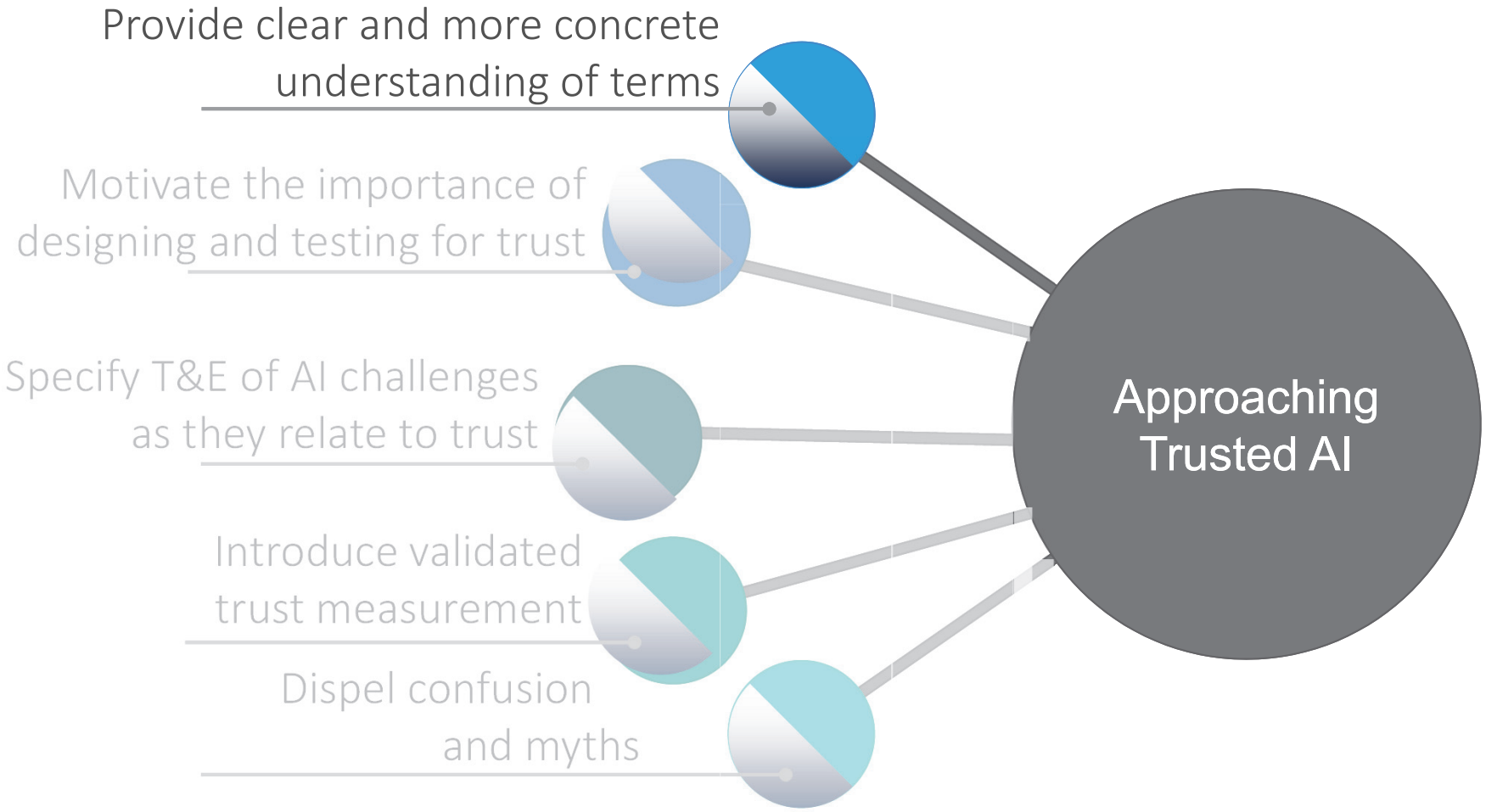
Deputy Sec. of Defense Kathleen Hicks



DoD Responsible AI (RAI) Strategy and Implementation Pathway

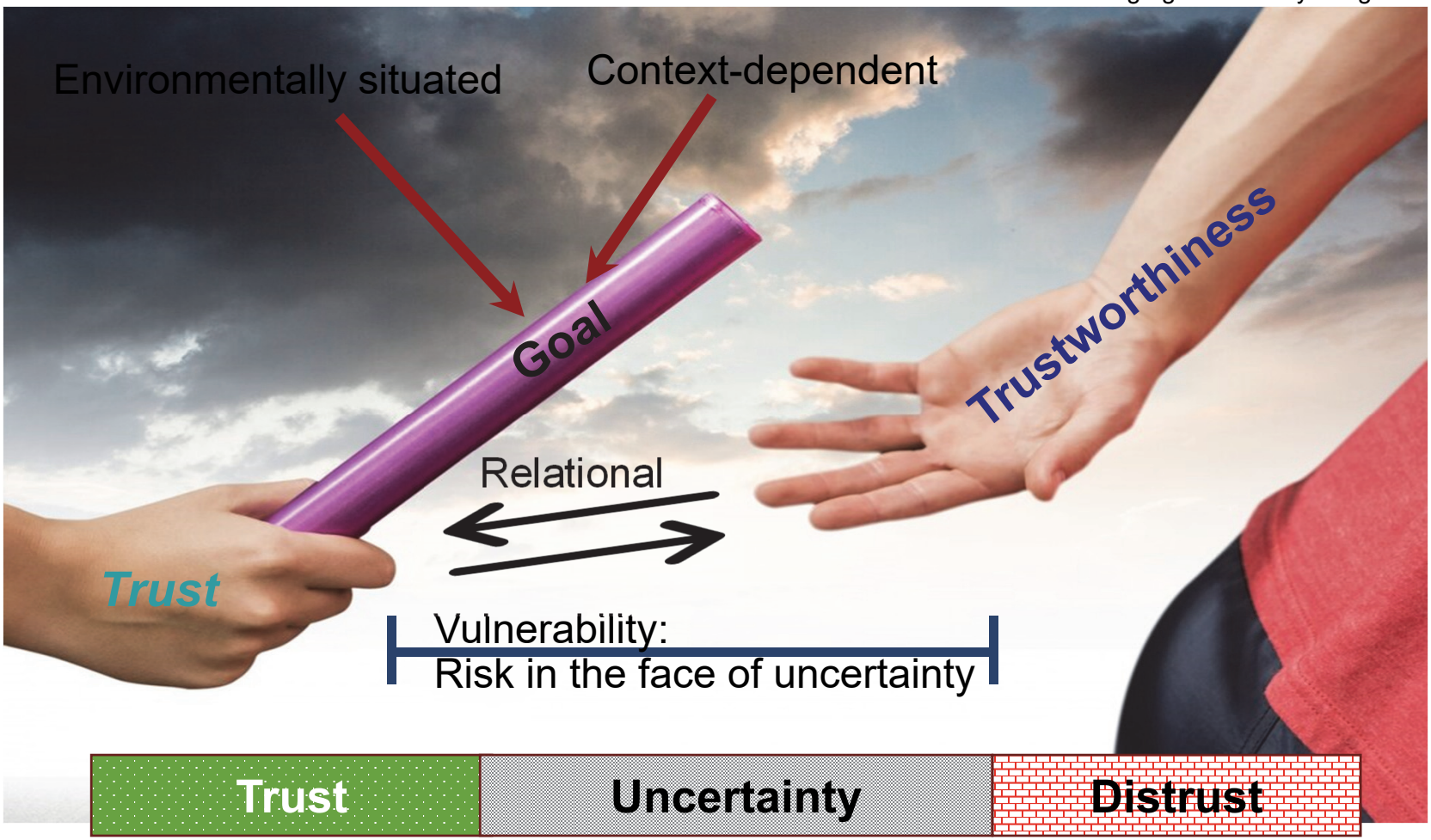
Executive Order 14110

In this ITEA paper on T&E of AI for trust, we…



Trust is a *cognitive state* where a trustor is *willing to delegate a goal* to a trustee

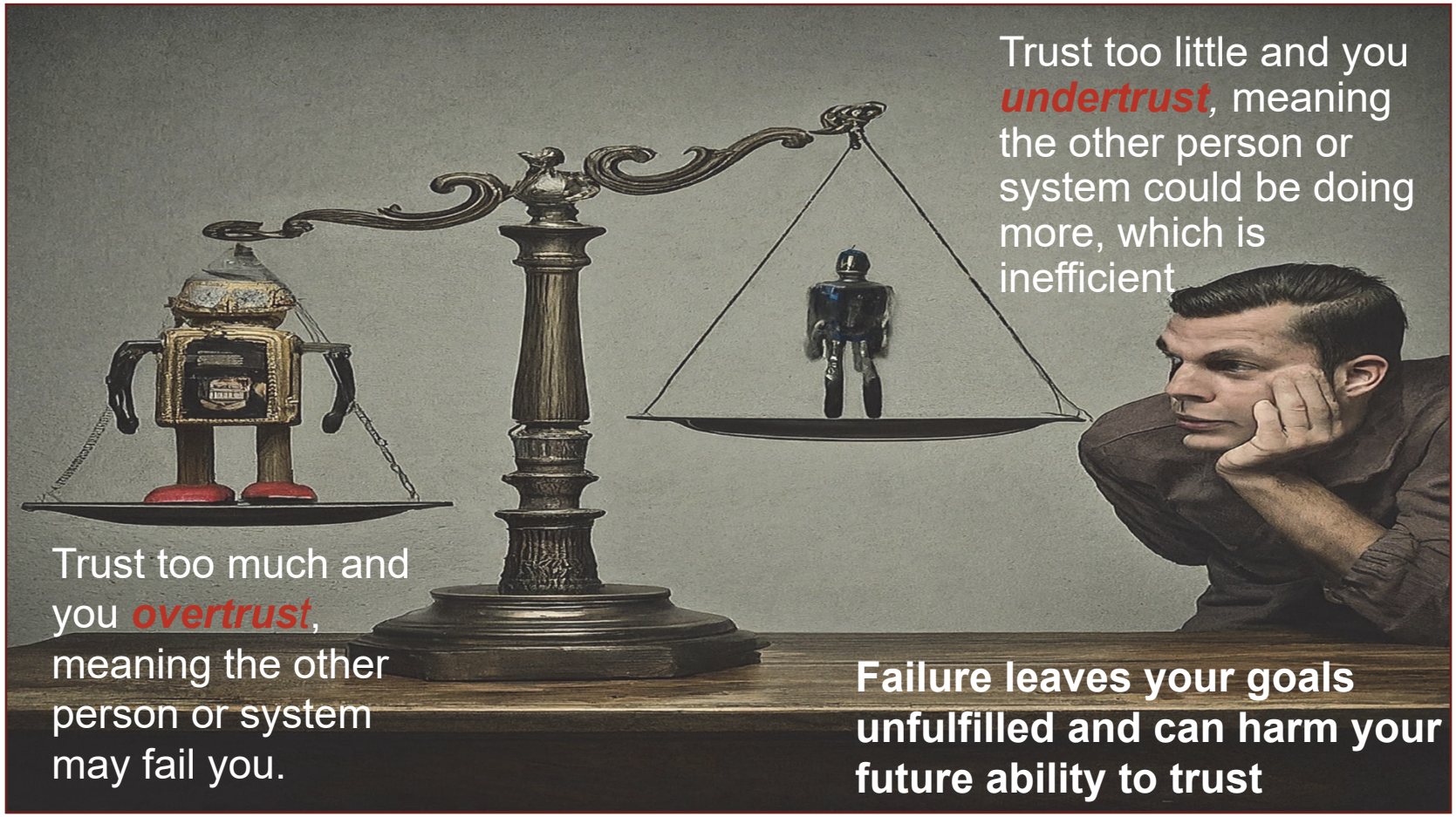
This image generated by ImageFX



Evidence from validated surveys demonstrates that there are three main components of *trust*.

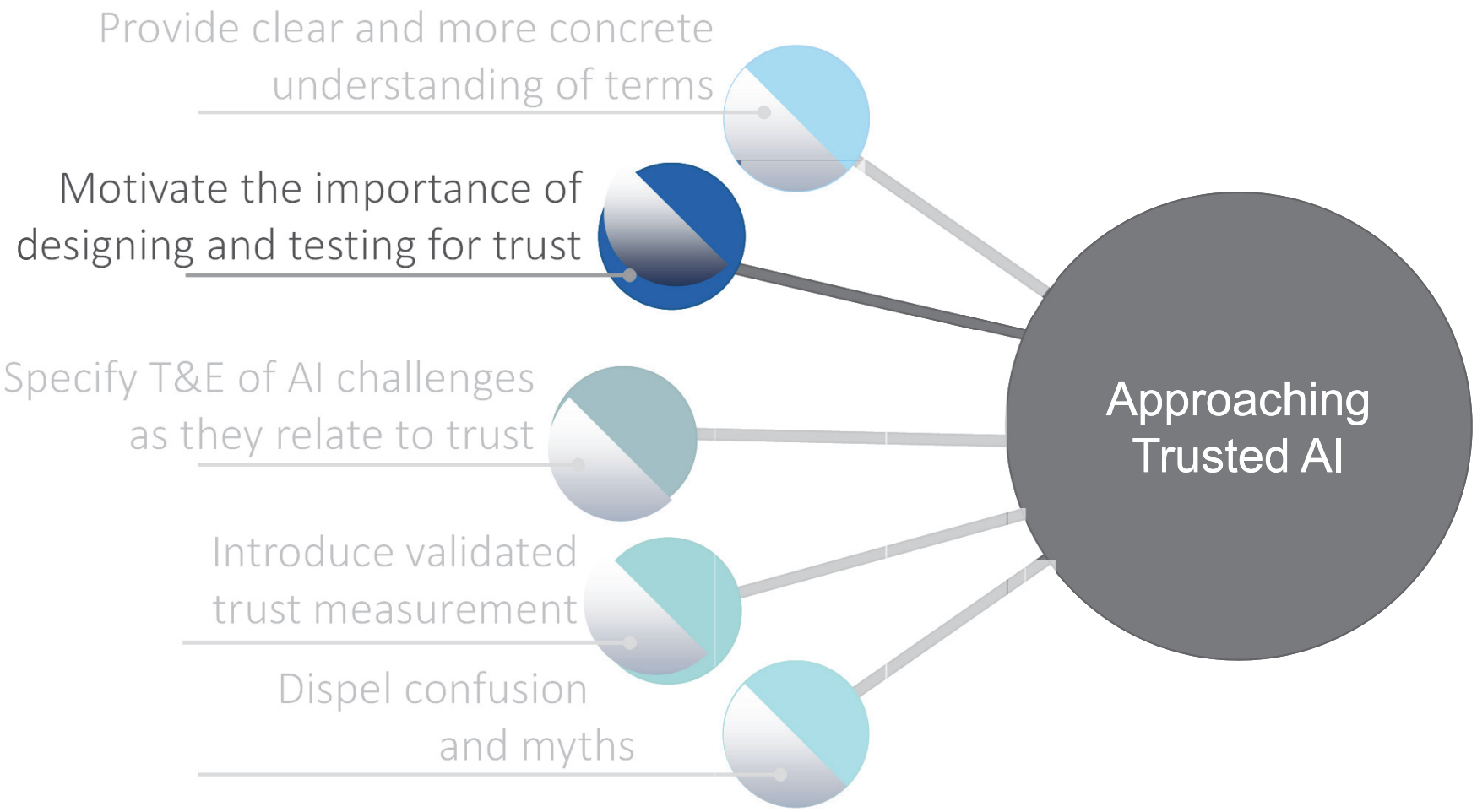
	Capability-based Trust/ Performance	Structural Trust/ Integrity	Affective Trust/ Benevolence
The trustor's expectation that the trustee	possesses the appropriate competence to accomplish the trustor's goal.	will adhere to norms, morals, laws, standards, and ethics that align with the trustor's values.	supports the trustor's goal and *potentially* values the trustor's general well-being.

Our goal is not to *maximize* trust but to achieve *calibrated trust*, to *trust* as much as the other is *trustworthy*

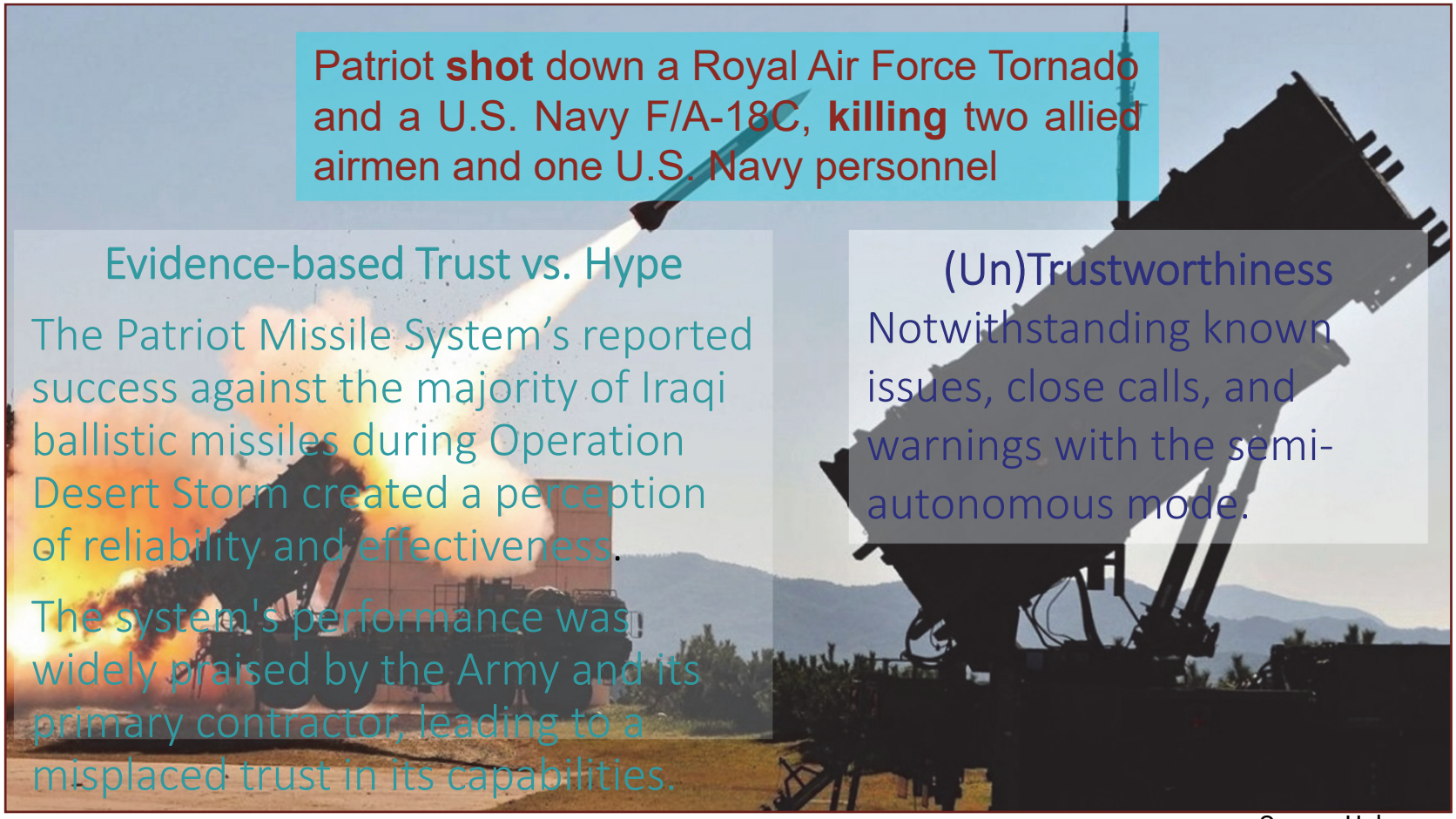


This image generated by ImageFX

In this ITEA paper on T&E of AI for trust, we…



If we do not consider the relationship between *trust* and *trustworthiness*, *mis-calibrated trust* can become a major, even fatal, threat.



Source Unknown

Patriot demonstrated the danger of *mis-calibrated trust* and accounting for it in training, doctrine, policy, PR, and interface design

Source Unknown

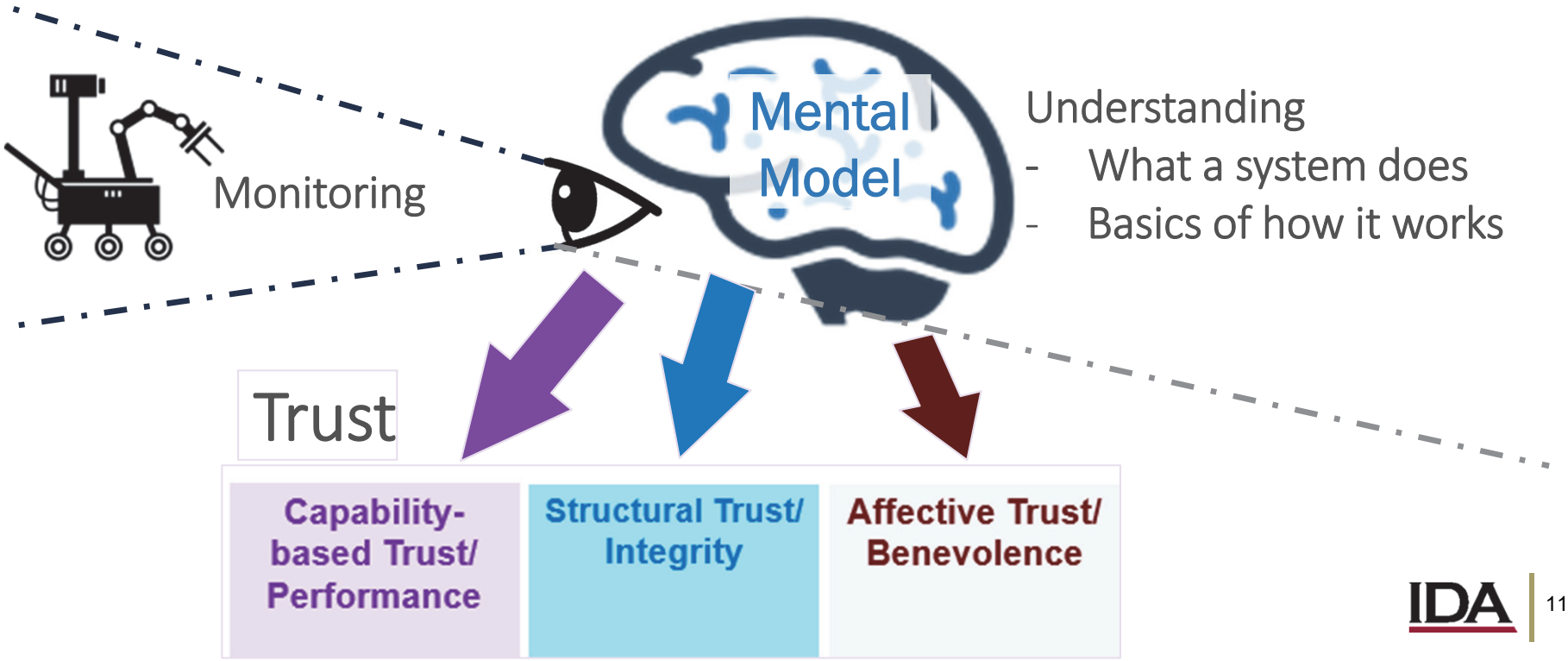
The types of *trust* and desired *trustworthiness* *differ* for different stakeholders and trustees

AP Photo/Andrew Harnik

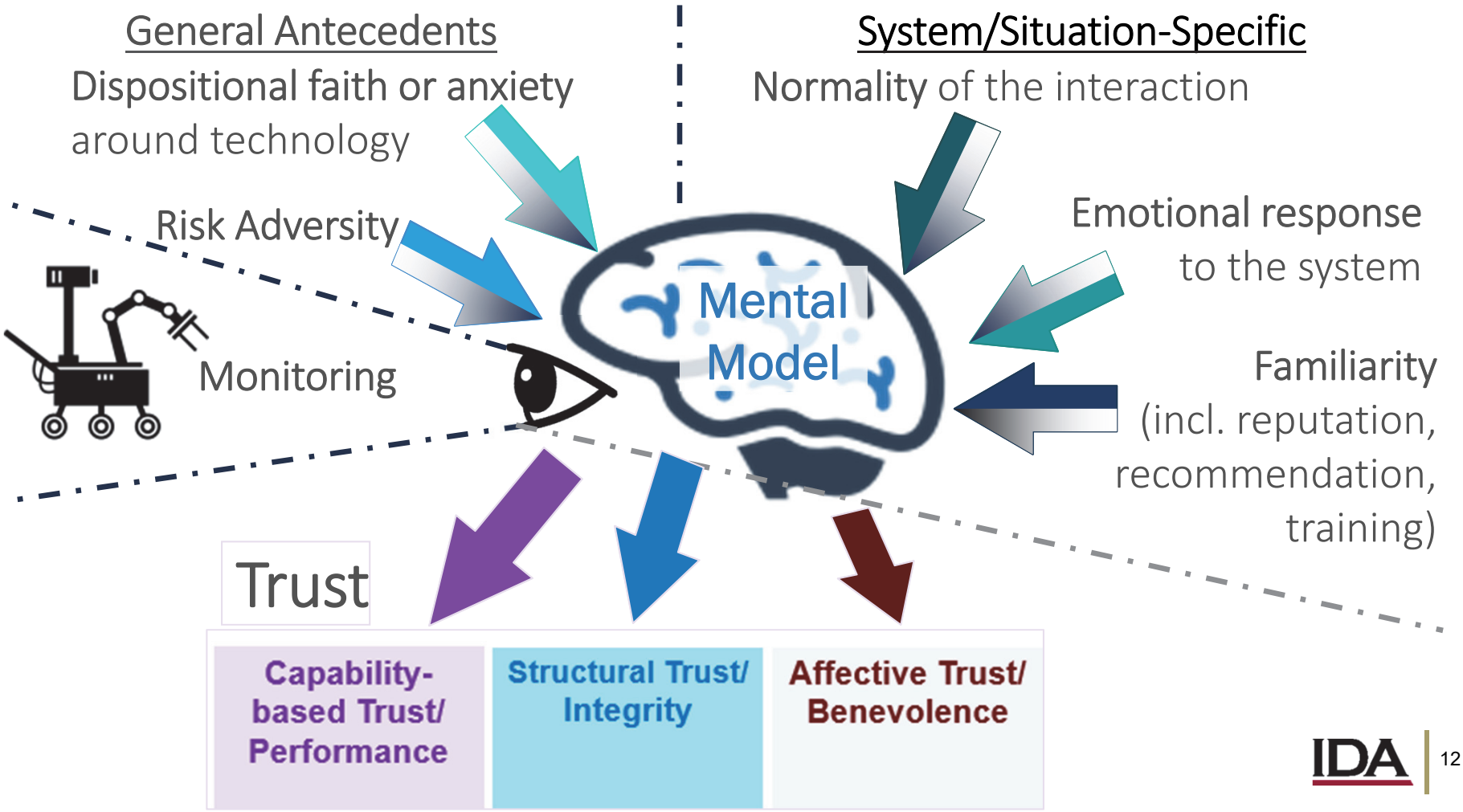
*"Our **operators** must come to trust the outputs of **AI systems**, our **commanders** must come to trust the legal, ethical, and moral foundations of **explainable AI**, and the **American people** must come to trust the values their **Department of Defense** has integrated into every application."*

Source Unknown

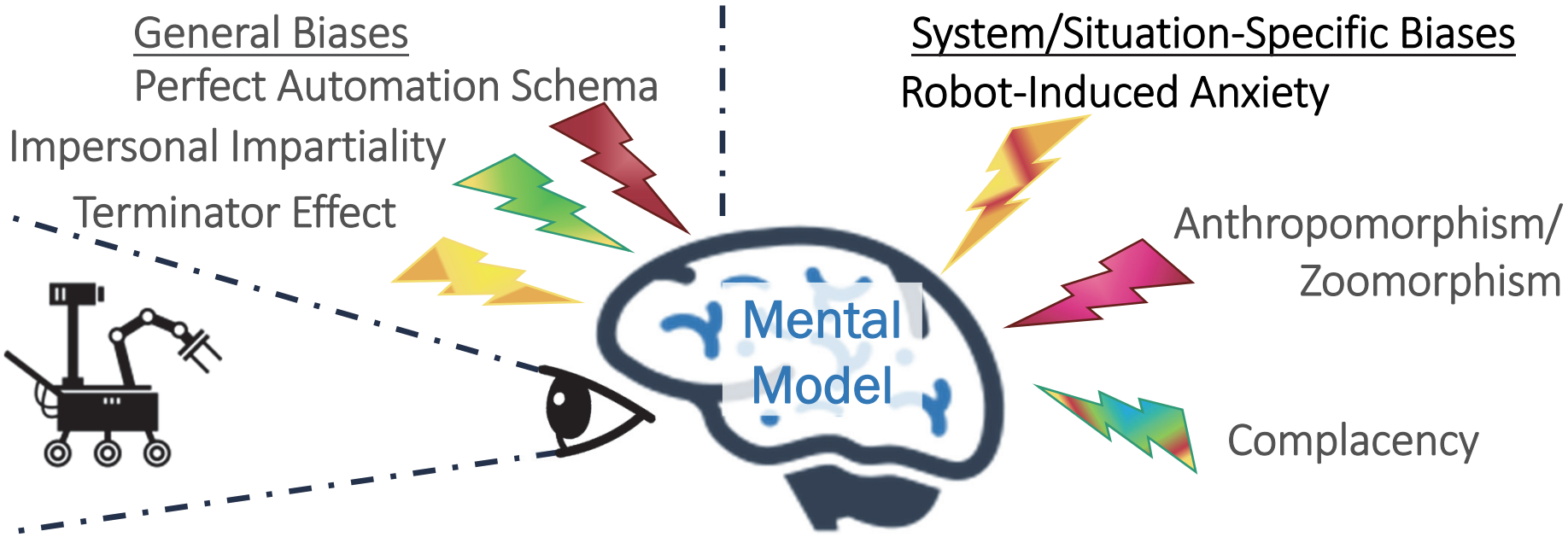
The trustor's *mental model* of the trustee forms their expectations, perceptions, and understanding of a system's ***trustworthiness***.



The trustor's *mental model* of the trustee forms their expectations, perceptions, and understanding of a system's ***trustworthiness***.

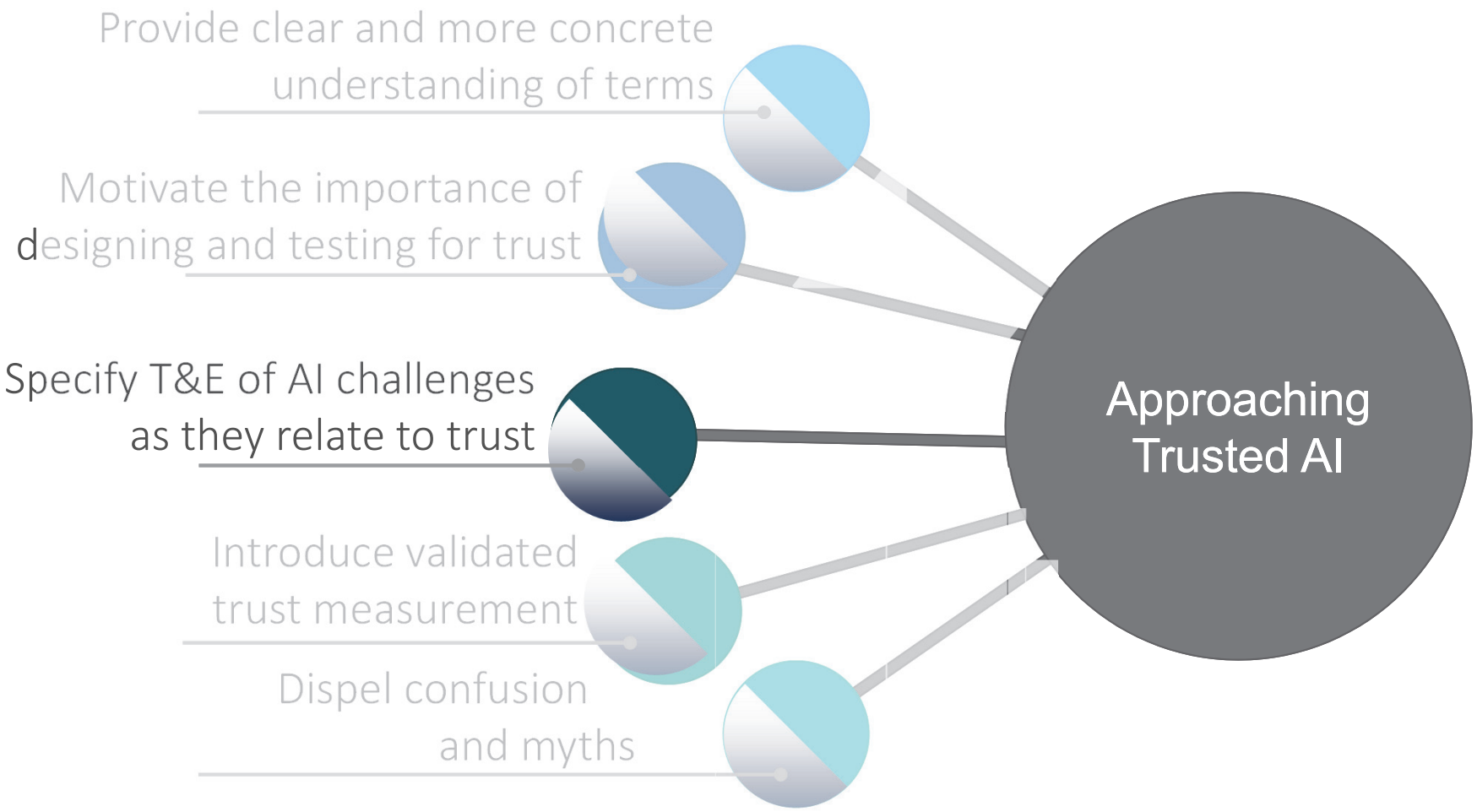


Mental models are powerful but prone to *biases*

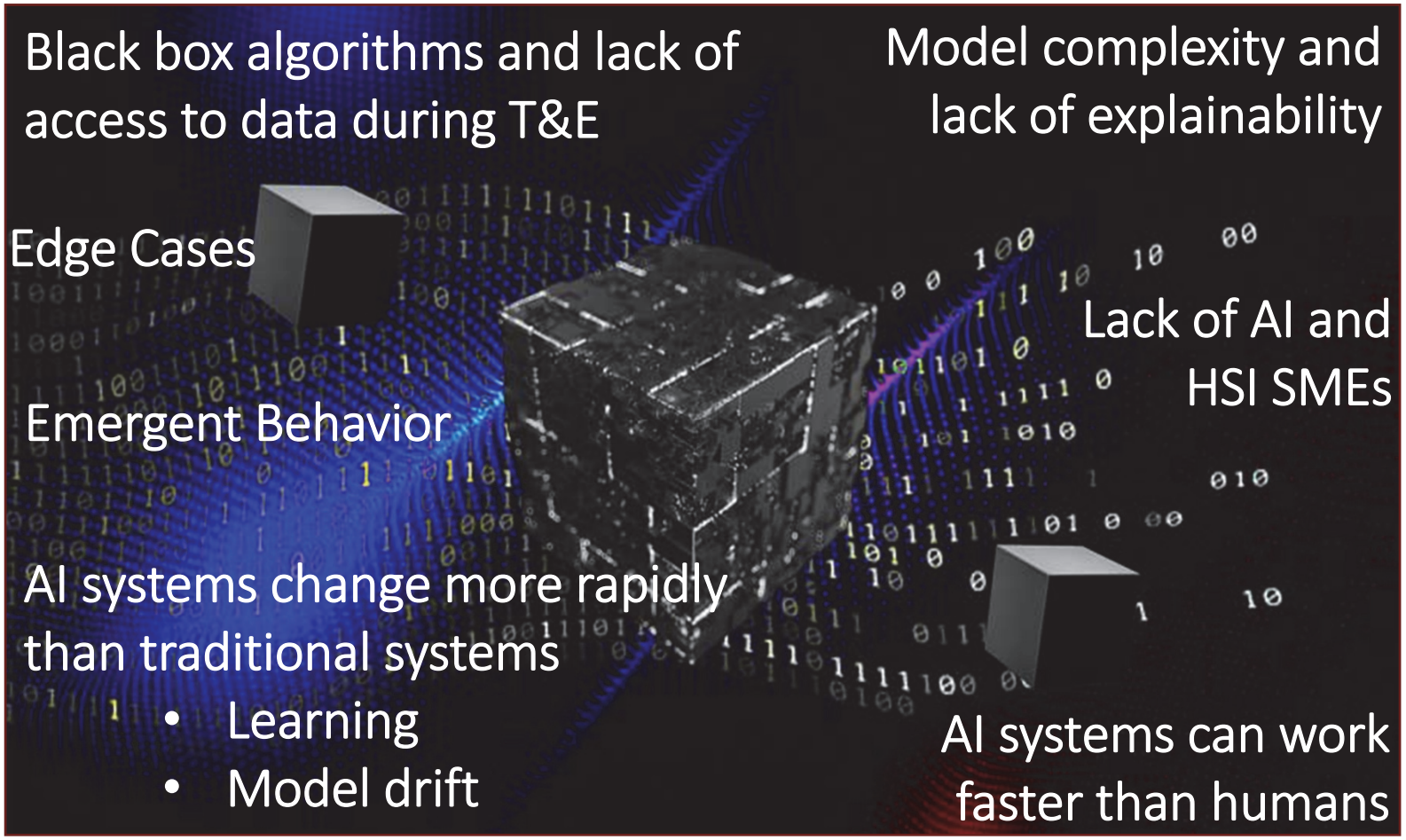


Biases arise from heuristic processes, so mitigating them presents a *tradeoff*.

In this ITEA paper on T&E of AI for trust, we…

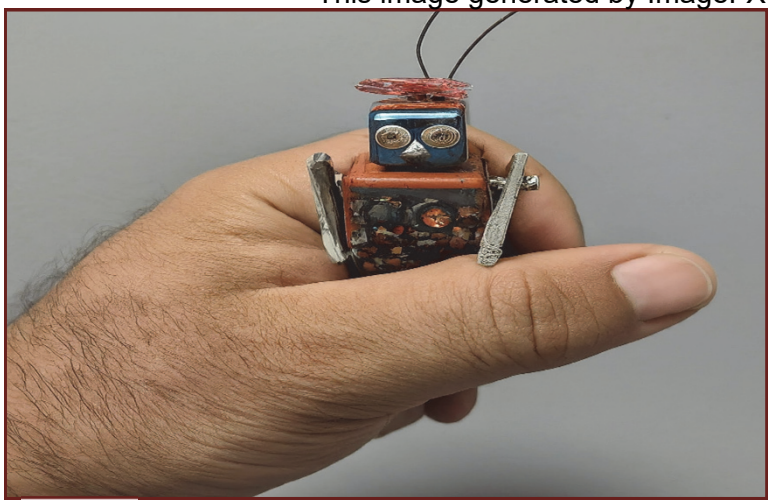


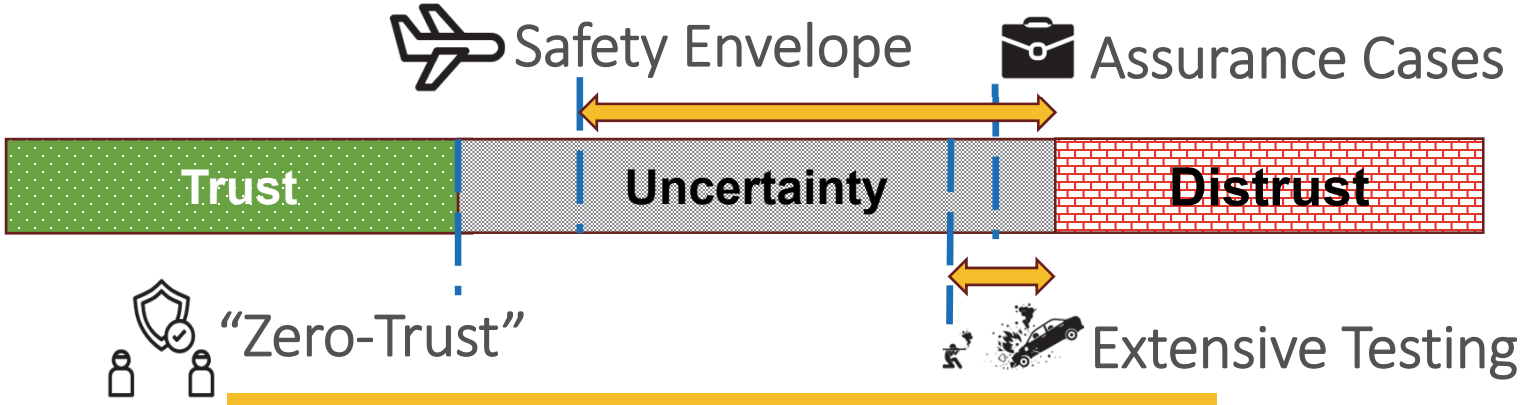
The power, speed, complexity, and size of AI presents unique challenges to evaluating ***trust*** and ***trustworthiness***



The new uncertainties that AI generates requires both higher assurance of ***trustworthiness*** and more ***trust***

This image generated by ImageFX





Is the gap an acceptable risk?

Despite the risks, generally the largest AI-specific *bias* we face is initial *overtrust*

Tesla

High Expectations
- Culture
- Brand Reputation
- Recommendation
- Research

Other user experience
- Training
- Experience
- Feedback

@Worst_Username_Evar

Calibrated Trust

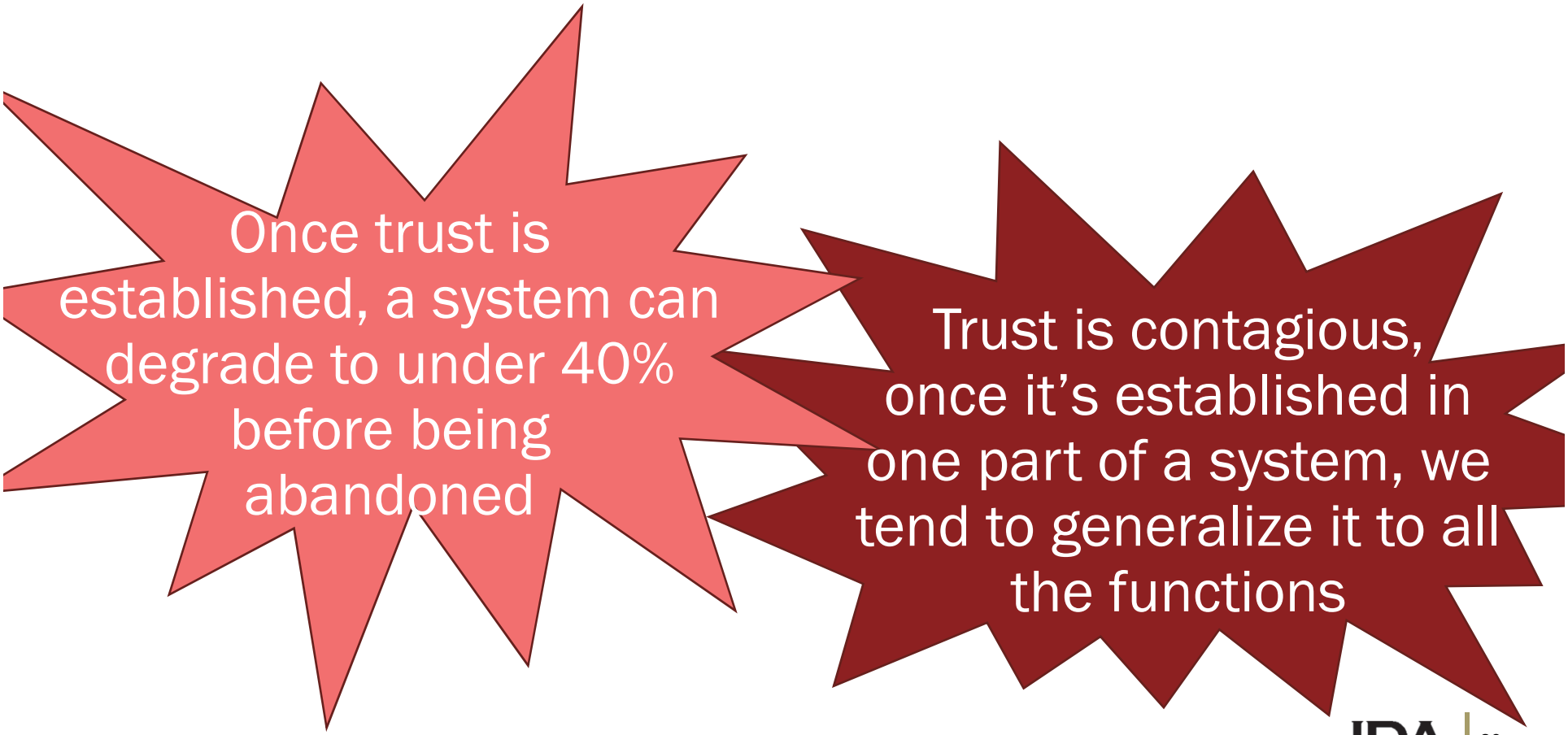
Poorly calibrated high initial *overtrust* is often very *brittle* and *non-resilient.* Once it's lost, it's hard to regain.

Florida Highway Patrol

Again: Our goal is not to *maximize* trust but to achieve *calibrated trust*, to *trust* as much as the other is *trustworthy*

Some fun research findings: Systems need to be at least *70–85% capable* for Capability-based trust to be established

BUT



Systems must be designed to help support proper trust calibration and testing

Instrumenting **systems** to assess both system behavior AND also **humans** and their **interactions**

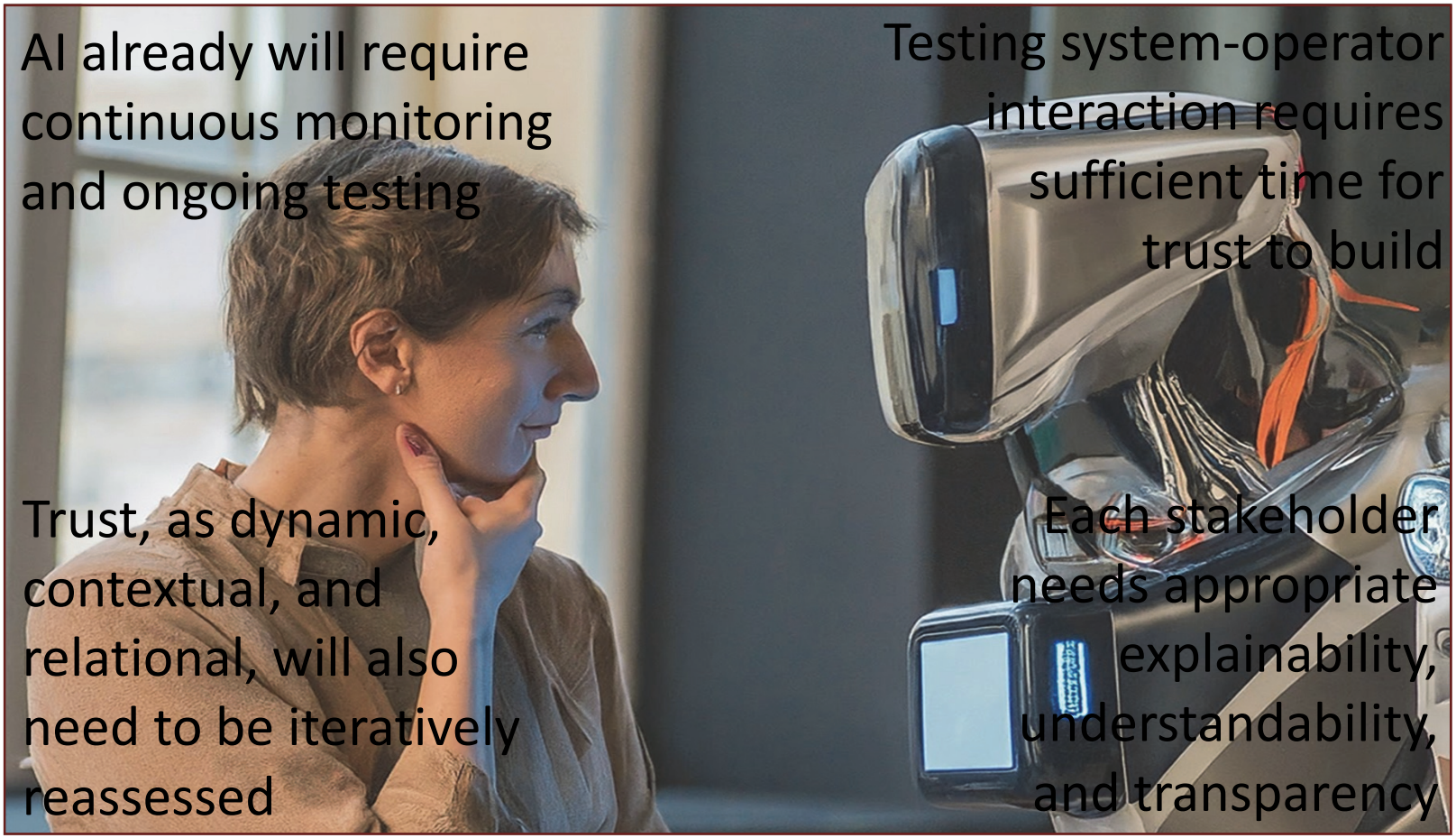
Careful design of training

Thoughtful UX design

Trust calibration is not only about good design and confidence from T&E

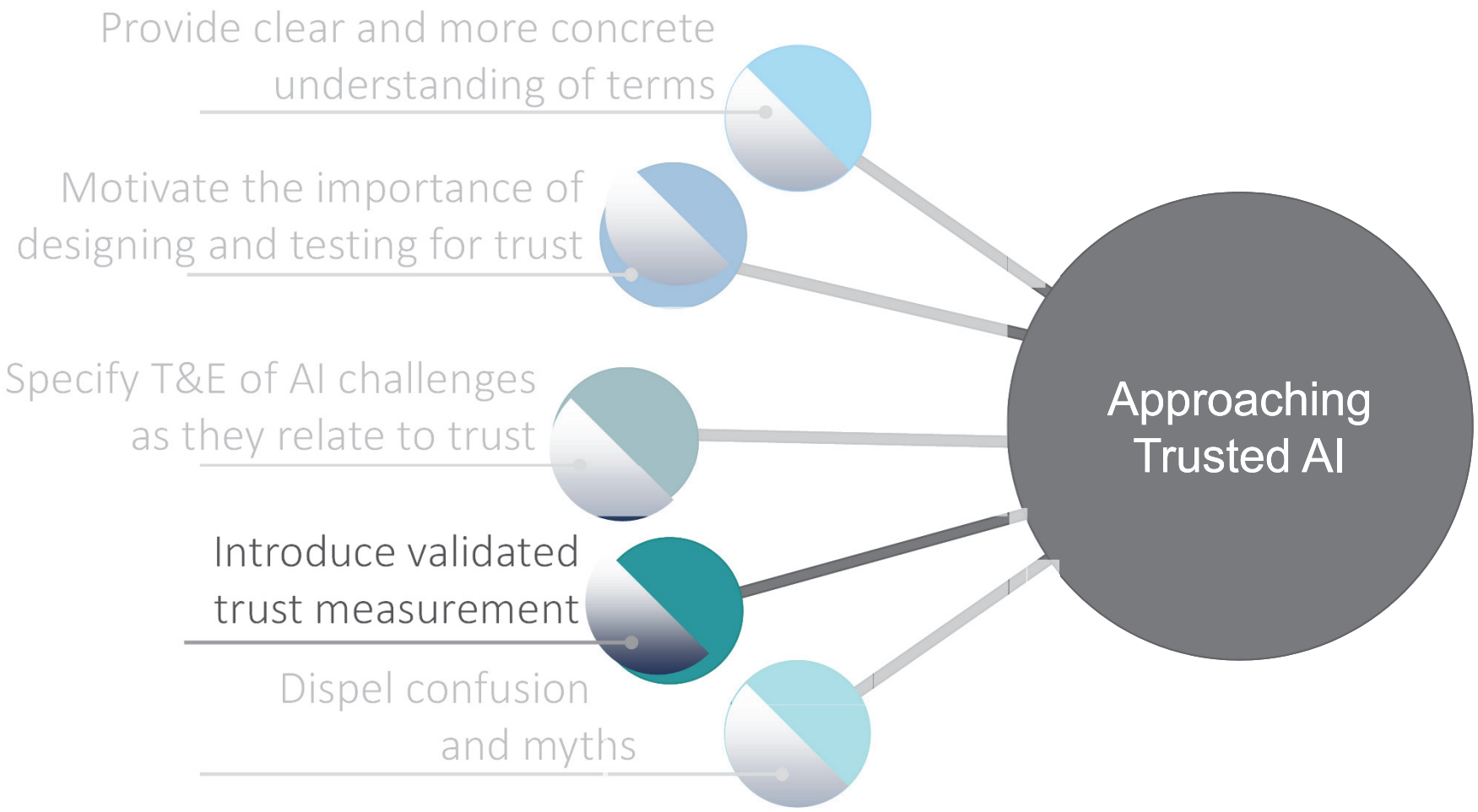
This image generated by Imgflip

Testing trust in AI systems *cannot* be 'one-and-done' or done without their operators



This image generated by ImageFX

In this ITEA paper on T&E of AI for trust, we…



This image generated by ImageFX

This image generated by ImageFX

If we want to measure system *trustworthiness*, we need to instrument systems and tests to evaluate whether the system is

- capable of fulfilling the designer's, operator's, and commander's goal when deployed correctly.
- able to be deployed correctly by those who are meant to operate it (Tate 2021).

Many of the factors for establishing *trustworthiness* in AI are the ***same*** as those for *trustworthy* software…

- Instrumentation
- Data
- Documentation
- User-Centric Design

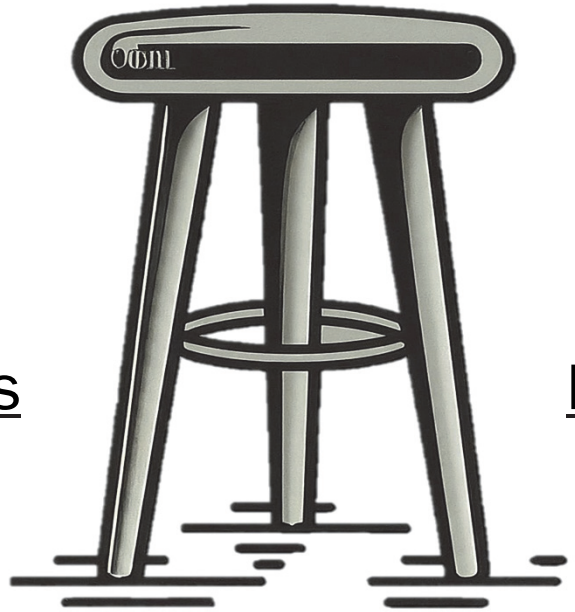
… but some major differences make this more difficult.

- Access to training data
- Black box models
- Programs don't build the algorithms themselves

This image generated by ImageFX

The best practice to capture *trust* is using *mixed methods* in order to triangulate it

<u>Behavioral measures</u>
See if they **act** as if they trust



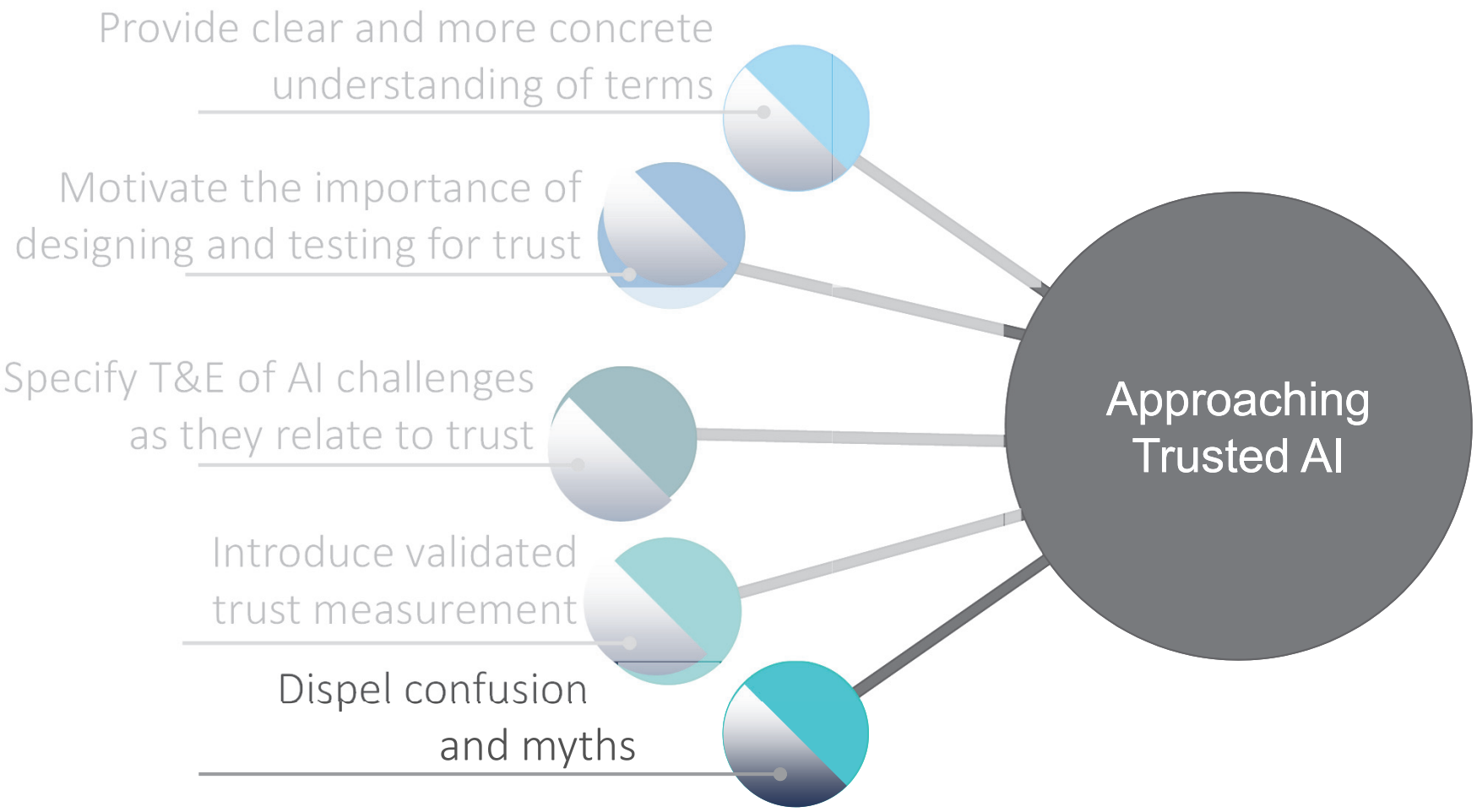
<u>Physiological measures</u>
See if they **react** as if they trust

<u>Survey measures</u>
See if they **believe** that they trust

Navigating the wide array of available trust surveys and models can be overwhelming, but there are some guideposts

This image generated by ImageFX

In this ITEA paper on T&E of AI for trust we



Myth Buster
Many trust measures *exist* and have been *validated*

Measurement Literature Reviews	Number of measures
Measurement of trust in automation: A narrative review and reference guide. Kohn, et al. (2023)	25
Trust Measurement in Human-Autonomy Teams: Development of a Conceptual Toolkit. Krausman, et al. (2022)	24
Converging Measures and an Emergent Model: A Meta-Analysis of Human-Automation Trust Questionnaires Razin and Feigh. (In Process)	62

Myth Buster

Transparency does not Guarantee Trust

It is *necessary*, *if done appropriately*, but not *sufficient*

What level does each stakeholder need?

The wrong level of transparency or wrong kind of explanation can undermine trust calibration and spread trust contagion

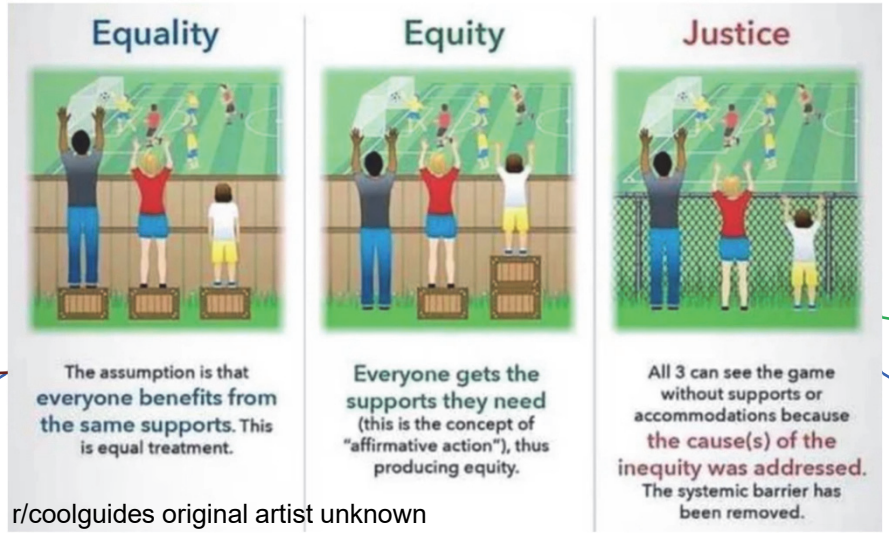
Too much transparency can overload the operator to rely on the system even if they do not trust it

Good transparency supports trust calibration, NOT trust

Myth Buster
Fairer AI Does Not Guarantee Trust

There are multiple types of fairness and they are mutually exclusive

We disagree on which form of fairness should be prioritized



r/coolguides original artist unknown

The type of fairness necessary to gain public trust
IS NOT
the type of fairness necessary to gain organizational trust
IS NOT
the type of fairness necessary to gain operator trust

Some concrete recommendations

Publish a shared HSI glossary

Encourage the use of quality trust surveys

Validate new measures for triangulation

Research measuring contextual trust calibration

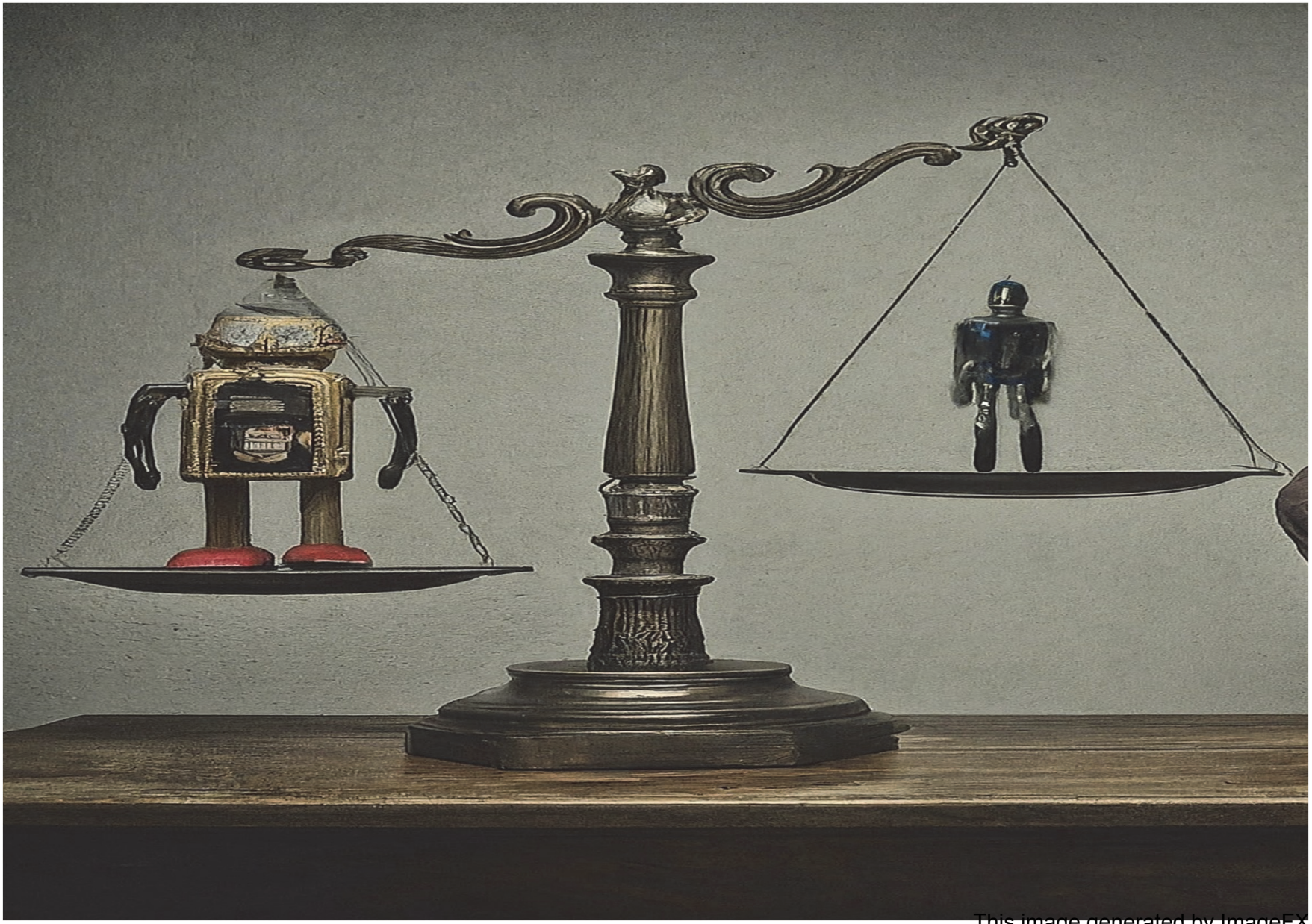
Some concrete recommendations

Shift Left: Bring operators and SMEs in earlier

Shift Right: Testing iteratively into sustainment

Budgeting for trust-related data collection

Train operators, testers, and decision makers



This image generated by ImageFX

Trust Me

Yosef S. Razin, IDA

yrazin@ida.org

Back up slides

Developing AI Trust:
From Theory to Testing and the Myths in Between

Yosef S. Razin, IDA

yrazin@ida.org

Dr. Kristen Alexander, DOT&E

kristen.l.alexander5.civ@mail.mil

Why Trust for AI is Particularly Hard

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
01-03-2024	Final	

4. TITLE AND SUBTITLE

DATAWorks 2024: Developing AI Trust: From Theory to Testing and the Myths in Between

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Yosef S Razin

5d. PROJECT NUMBER

BD-9-2299(90)

5e. TASK NUMBER

229990

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

3001946

3001946

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

Director, Operational Test and Evaluation
1777 N. Kent Street, Room 12063
Arlington, VA, 2220
(703) 697-3655

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release: distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: John Haman

14. ABSTRACT

This introductory presentation aims to provide members of the Test and Evaluation community with a clear understanding of trust and trustworthiness to support responsible and effective evaluation of AI systems. I provide a set of working definitions and works toward dispelling confusion and myths surrounding trust. While also explaining trustworthiness, we move beyond trustworthiness' techno-centricity to focus on how people develop trust in AI systems. In particular, this work highlights trust's relational and context dependence and how this gives rise to varying testing requirements for different stakeholders, including users, regulators, testers, and the general public. Therefore, trustworthiness and trust cannot be tested separately from their users and other stakeholders; nor can they be assessed just once, but require continuous assessment. By understanding trust and trustworthiness, the Test and Evaluation community can more confidently assess whether systems are reliable and meet the expectations and needs of users, regulators, and the general public.

15. SUBJECT TERMS

Human-System Interaction, Human-Machine Teaming, Trust, Trustworthiness, AI, Behavioral Measures

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John T Haman
Unclassified	Unclassified	Unclassified	SAR	30	19b. TELEPHONE NUMBER *(Include area code)* 703-845-2132