INSTITUTE FOR DEFENSE ANALYSES

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

John Haman, Project Leader

Yosef Razin
Kristen Alexander

OED Draft

March 2024

IDA Document 3000512

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, BD-9-229990, "Methods Develop," 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 G. Anderson, Dr. Daniel E. Hellmann, and Dr. Eric R. Schulman from the Operational Evaluation Division, and Dr. David Tate from the Cost Analysis and Research Division. Dr. Arun S. Maiya from the Information Technology and Systems Division.

For more information:
Dr. John T 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 Document 3000512

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

John Haman, Project Leader

Yosef Razin
Kristen Alexander

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

Yosef S. Razin
Institute for Defense Analyses
Alexandria, VA
yrazin@ida.org
www.linkedin.com/in/yosefrazin

Dr. Kristen Alexander
Director, Operational Test and Evaluation
Alexandria, VA
kristen.l.alexander5.civ@mail.mil
www.linkedin.com/in/kristen-alexander-6992753a/

Abstract

This introductory work 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. The paper provides a set of working definitions and works toward dispelling confusion and myths surrounding trust. While also explaining trustworthiness, it moves 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.

Introduction

> ***"At the end of the day the integration of AI technology is about trust and a responsible AI ecosystem is the foundation for that trust. 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 Secretary of Defense Kathleen Hicks (2021)*

Trust in Artificial Intelligence (AI) has become one of the hottest topics in public discourse with the explosion of generative AI, to the point where it has become a buzzword. The 2018 National Department of Defense (DoD) AI Strategy (Blackburn) and the 2020 DoD Data Strategy (OSDPA) started to stress operator trust and data trustworthiness, respectively. By 2020, the National Security Commission on AI final report mentioned trust 132 times (NSCAI 2020), spanning users, managers, the public, Test and Evaluation, Validation and Verification (TEVV), hardware, digital ecosystems, strategic partners, networks, and teaming, among others. Within three years, the DoD's Responsible AI (RAI) strategy states the desired end state for the entire responsible AI project is to engender trust from warfighters, leaders, and the American people in how the Department uses AI (DEPARTMENT OF DEFENSE 2022). By 2023, it had become a primary objective of U.S. national AI development as set forth by President Biden himself (Executive Order 14110, October 30, 2023). Trust has come to be seen as essential at every level from procurement to testing to public acceptance.

The Director, Operational Test and Evaluation (DOT&E) Strategy Implementation Plan identifies the need to develop methods for the adequate assessment of operational and ethical performance of AI-enabled systems. Testing must demonstrate with confidence the DoD AI ethical principles including responsible, ethical, equitable, traceable, reliable, and governable (*DoD Adopts Ethical Principles for Artificial Intelligence*). The DoD RAI policy indicates that when these ethical principles are met, the desired end state of trust is achieved (DEPARTMENT OF DEFENSE 2022). DOT&E, working alongside organizations including the DoD's Chief Data and AI Office (CDAO) and the Office of Developmental Test, Evaluation and Assessment (DTE&A), must ensure that those who work on AI within the DoD – in its design, development, and testing – understand the various usages of 'trust', why the concept of trust remains important, and how to approach it effectively. However, the standard approach of turning toward previous work in academia, government, and industry can confuse more than help (D. H. McKnight and Chervany 2001; Peter A. Hancock et al. 2011; P. A. Hancock et al. 2023; Hoff and Bashir 2015; Probasco, Emelia S., Toney, Autumn S., Curlee, Kathleen T. 2023). We aim to dispel some of that confusion in this paper. Per Probasco, Emelia S., Toney, Autumn S., Curlee, Kathleen T. (2023), while nearly 14 percent of the 322,000 peer-reviewed papers on AI include a trust-related keyword, there are more than a dozen related keywords. And to make matters worse, some use these keywords in nuanced and distinct ways while others use them interchangeably. Beyond keyword confusion, there are more than a hundred definitions of trust in technology in the academic literature (Gefen 2014) and more than 60 proposed trust measures for AI and autonomous systems (Razin and Feigh). Furthermore, 'trust' is pursued as a research topic in a dozen academic fields, from politics to psychology, that each have their own distinct approaches, definitions, and frameworks (D. H. McKnight and Chervany 2001). Finally, this problem is compounded by the basic

reality that 'trust' and related terms are sufficiently common and colloquial (Goldberg 2019), such that attempting a technical definition may be inappropriate, if not impossible.

Despite the lack of convergence around keywords, definitions, and models, research into trust in technology has accumulated a number of scientifically validated findings. These findings are meaningful and shed light on the powerful and concrete ways various kinds of trust affect the adoption, acceptance, use, and misuse of technology. Furthermore, recent research in AI has shed light on some of the unique aspects of trust in AI compared to other technologies.

However, a better scientific grasp of trust has not always come without cost. Trust research has been exploited through persuasive design techniques, such as emotional triggers and gambling-like gamification, which have been implicated in smartphone addiction (Leslie, 2016; X. Chen et al. 2023). Tech companies are leveraging persuasive designs, based on the latest in trust research, to profile and model their users and predict their behaviors in order to sway users to invest evermore time and attention to their products (X. Chen et al. 2023).

Such misappropriate trust in advanced technologies has caused catastrophic failures. The Patriot missile system, a missile defense system developed during the Cold War, has exemplified the importance of accounting for trust within test and evaluation (T&E), and the consequences when it is not. The Patriot system was initially intended as an air defense capability, with both a semi-autonomous and an autonomous mode (Hawley). While there were known issues with the autonomous mode for air defense, the Army decided that the autonomous mode was appropriate for missile defense, though it was still under human oversight.

While the Patriot reported success against the majority of Iraqi ballistic missiles it engaged during Operation Desert Storm, 28 U.S. soldiers in a barracks in Dharan, Saudi Arabia, were killed when the system failed to engage (GAO, 1992). Additionally, there were reports of repeated close calls with friendly aircraft (Hawley 2017). Despite these issues and an unpublished report predicting future trouble (Hawley), the Army and their primary contractor lauded the system's success and placed their trust in its capabilities. Patriot testing leading up to Operation Iraqi Freedom featured a few intercept flight tests against aircraft, tens of passes by live aircraft during tracking exercises, and hundreds of simulated aircraft trajectories; but there were tens of thousands of friendly aircraft sorties during Operation Iraqi Freedom, many of which had features that hadn't been present in the live or simulated aircraft present during Army Patriot testing.

The Patriot was hyped after Operation Desert Storm and then successfully intercepted all nine of the ballistic missiles it engaged during Operation Iraqi Freedom in 2003. However, the confidence built up by the previous successes contributed to two instances of fratricide during Operation Iraqi Freedom when Patriot shot down a Royal Air Force

Tornado and a U.S. Navy F/A-18C, killing two allied airmen and one U.S. Navy personnel (Directorate of Air Staff 2003; Hawley). Patriot operators had the ability and the time to examine the tracks, conclude they were not actually missiles, and either halt the engagement prior to interceptor launch or destroy the interceptors after launch but before they reached the friendly aircraft. In both cases, however, the Patriot operators failed to do this because they had been trained to trust the system instead of questioning the system when it was reporting unusual information. In both cases, there were many contributing factors, the lack of any of which would have prevented these tragedies, but low-probability events and edge cases can and do happen; and system designers and testers, especially autonomous ones, need to take this into account. Understanding how trust can lead to overcompliance (see Further Useful Definitions on compliance) and over trust, as well as studying ways to better calibrate trust and system trustworthiness through training, doctrine, policy, interface design, and incorporation of user feedback, may have mitigated such a failure.

Thus, human factors, as they relate to trust, are potent forces that should not be ignored, but handled responsibly and with care. In the following work, we will endeavor to clarify what trust is and explain how it is used, as well as related human factor concepts and their implications for test and evaluation. Therefore, this work will begin with defining key terms relating to trust and trustworthiness (Section 2), explore what factors underly the various types of trust and what can bias them (Section 3), and consider how to approach testing for trust and trustworthiness for AI-enabled systems (Section 4).

Foundations of Trust

To start, we must first understand that the range of concepts referred to as 'trust' is very broad. At one end there is interpersonal trust, which is trust between two people. At the other end there is public trust, which is broadly a sense of trust the public has in an institution. Between people and institutions, there are multiple levels of trust that can be discussed; between an individual and a group, between groups, between strangers, and in public figures or companies. But trust is not only placed in people. Non-humans are also sometimes trusted, such as pets, tools, and machines, including highly complex systems, such as AI. **How much someone or something should be trusted is called their trustworthiness, and this can be thought of as a property of the trustee. Trust itself, though, is a cognitive state, such as an attitude, belief, or expectation of the trustor.**

This still leaves us with the question, what is trust? While being as broad as possible and acknowledging that it is nearly impossible to find a definition that everyone will accept, for the purposes of this paper, we define trust as

> *a state effectuated by the trustor in which the trustee is given power over some subset of the trustor's goals, which the trustor believes they could not have accomplished better on their own* (Razin and Feigh).

This definition would apply to trust between teammates, but also to a commander's trust that the automatic mode of the Patriot missile system furthers the mission goals without endangering allied or civilian personnel. It further extends to the Army acquisition office's trust that the Integrated Visual Augmentation System (IVAS) as delivered meets the requirements to enhance a soldier's performance in the field, as well as the public's trust that social media applications will try to keep their data reasonably safe. Sometimes the goals the trustor are putting in the hands of the trustee are explicit (e.g., mission objectives) but goals can also be assumed (e.g., data privacy, suitability, legal compliance, or willingness to act in the other's best interest); sometimes incorrectly.

Our definition of trust highlights the importance of understanding that trust is deeply intertwined with power and vulnerability. Trust arises when the trustor recognizes that a trustee is as good as – if not better resourced/situated/skilled than – they are for accomplishing their goal, and therefore they allow themselves to be vulnerable by delegating the accomplishment of their goals to the trustee. Trust's strength is that it allows us to plan in the face of uncertainty, but at the same time, it always comes at the risk of failed goals or even betrayal.

Trust often extends beyond a belief; when we trust enough, that belief crosses a threshold where it becomes an intention on which we act. But how much trust is necessary before it becomes action? This is known as the trust calibration problem (Yue Wang et al. 2024): the trustor should only trust as much as the trustee is trustworthy. Too much trust

(over trust) can lead to unmet expectations, disappointment, and regret. Too little trust (under trust) means the trustor is wasting resources and effort when a more effective way exists to accomplish their goals.

It is worth stressing that if the goal is calibrating trust for optimal effectiveness and suitability, as in the design of many of the DoD's systems, systems should not necessarily be designed to maximize trust. Maximizing trust would lead to over trust, which already is a serious problem when it comes to placing our trust in technology (Merritt et al. 2015). For instance, Tesla invested heavily in public trust, and in early 2022 the company was the single most trusted to develop safe and reliable autonomous vehicles (Cole, March 22, 2022) (see Further Usefull Definitions on reliability); yet Tesla's vehicles account for 91 percent of all self-driving-related crashes. This is not only because of Tesla's popularity. According to National Highway Traffic Safety Administration data, over trust has likely resulted in more than 700 crashes involving autonomous driving capabilities, with a fatality rate 8 times higher than the national average for crashes involving for manually driven vehicles (Saddiqui and Merrill, June 10, 2023). It appears that Tesla had concentrated on maximizing trust alone rather than calibrating trust to trustworthiness. This might be because calibrating trust to actual vehicle trustworthiness would likely decrease sales and investment. Vendors and contractors, in general, have an interest in maximizing trust, whereas test and evaluation practices should be evidence-based and endeavor to accurately assess both trust and trustworthiness in support of achieving calibrated trust.[1]

Although trustworthiness is a property of a system, it is not fixed, but context-relative and dynamic. The trustworthiness of a system depends on the goals of the trustor, the operational environment, the constraints, and the level of risk. The amount of trustworthiness required for a particular system also differs by trustor: the type and level of confidence an operator or commander needs to be willing to employ a system can differ from what a regulator would need to approve the system or what the public needs to accept the system. While they might be similar, the types of evidence and arguments needed by each audience differ. For example, the developmental tester needs a different level of algorithmic transparency than the operational tester or the operator. The operator may need more understandability than transparency. Operators need to understand what the system is designed to do and when it can be expected to do so, but it is less important that they know exactly how it works. On the other hand, the public's trust might be more dependent on perceiving whether the software is fair or biased and understanding the design choices that went into attempting to achieve equity. These might not be as relevant to an operator in a mission context when the bias is not the immediate priority.

[1] Note that public confidence has dropped in Tesla since early 2022, given news of the various accidents. By September 2023, Tesla only ranked #5 among autonomous vehicle companies. Hence, trust develops over time and calibration improves with experience and exposure.

Furthermore, **increased transparency does not necessarily lead to higher or better-calibrated trust**. A system that performs poorly but has high transparency will calibrate trust by lowering it. But there is a widespread myth that if AI were just more transparent, explainable, and understandable, increased trust would automatically follow (Niu, Terken, and Eggen 2018; Weitz et al. 2021; Buçinca, Malaya, and Gajos 2021). Research has found that sometimes transparency, explainability, and understandability can undermine trust, even in trustworthy systems (Zerilli, Bhatt, and Weller 2022; Buçinca, Malaya, and Gajos 2021). They can increase operator workload and lead to confusion and under trust (Helldin 2014). Inaccurate explanatory methods (such as metamodels that oversimplify system decision rationales) can lead to miscalibration of trust in both directions (Zerilli, Bhatt, and Weller 2022) as well as over-reliance (Buçinca, Malaya, and Gajos 2021) and complacency (Razin et al. 2021). It has been suggested that many explanatory frameworks assume that end-users will examine each explanation critically and apply logic to assess each of the AI's explanations (Buçinca, Malaya, and Gajos 2021). However, research has shown that instead, end-users leverage a heuristic mental model of the AI agent to simply judge whether or not to trust (Buçinca, Malaya, and Gajos 2021) and that, unless users are highly motivated to assess content, the value of transparency and its mechanism for increasing trust is more symbolic (Liu 2021); meaning that the very signaling of the existence of transparency, whether it is really being transparent or simply pretending, is enough to assuage users and increase their trust. Since the transparency does not need to be real, this can lead to miscalibrated trust in the AI.

Digging Deeper into Trust: Safety and Bias

There is increasing convergence around the concept that trust can be categorized into three main types (Razin and Feigh):

- *Capability-based or Performance Trust*, which refers to the trustor's expectation that the trustee possesses the competence to accomplish the tasks and goals intended by the trustor;
- *Structural Trust or Integrity*, which involves the trustor's expectation that the trustee will adhere to the norms, morals, laws, and ethics that align with the trustor's values. This is also the aspect of trust directly shaped by regulations, laws, and community/industry standards; and
- *Affective Trust or Benevolence*, which pertains to the trustor's expectation that the trustee supports, aligns with, or shares the specific goal being entrusted. This type of trust can also extend to the trustor's general well-being.

These three dimensions of trust are supported by the trustor's mental model of the trustee and how confident the trustor is that they and the trustee share a mental model of the mission, its goals, and relevant information for their successful accomplishment (Razin and Feigh). In the case where the trustee is a technological system, this shared mental model is as much the system's – however much of a model it can form – as it is its creator's. Simultaneously, this shared mental model supports the trustor's situational awareness, allowing them to perceive and understand whether the system is supporting their goals and to predict if the system will continue to do so (Andrews et al. 2023).

The mental model is, in turn, informed by past experience, familiarity, reputation of the product or brand, training, and how similar it is to other systems or technologies the trustor has encountered before (Razin and Feigh).

Trust, as a cognitive state shaped by our expectations, is prone to various cognitive biases, which should be accounted for in system design, testing, and fielding. Cognitive biases are not inherently negative; they serve as mental shortcuts (heuristics) that aid us in making judgments, often more quickly than if we had to consciously reason them out. However, it is essential to recognize that these biases can lead us astray. Several cognitive biases are associated with trust and AI, including:

- *Perfect Automation Schema*, which is holding unfounded high expectations of AI and automation performance but with a very low willingness to forgive errors (Merritt et al. 2015).
- *Complacency*, which arises when individuals become overly reliant on AI systems and overlook potential risks. When this occurs, compliance can slip into complacency, when the trustor stops monitoring or reflecting on the

trustee's behaviors or advice (Merritt et al. 2019), such as with the Patriot system in Operation Iraqi Freedom, as discussed above.

- *Robot-induced anxiety and negative perceptions*, which is related to expectations of robots and AI being dangerous. These expectations may be formed by popular media (Terminator Effect (Cheatham 2022))or may be a result of age, risk aversion, and familiarity with robots, AI, and other complex technological systems (Desai 2012).
- *Impersonal Impartiality*, in which AI is assumed to be unbiased and fair because it is not human and uses logic and not emotions to generate solutions. This impartiality is especially assumed in systems that do not look, sound, or move like a human (Das, Yixiao Wang, and Green 2021; Okoh 2023).
- *Anthropomorphism and zoomorphism*, which arises from designs that evoke human or animal characteristics, both structural (face, body) and functional (naturalness of language, movement, facial expressions, tone of voice). This bias then causes people to attribute human-like and animal-like qualities to AI entities, and to expect more human or animal-like behavior more generally from them (Visser et al. 2016; Verbene, Ham, and Midden 2015; Graesel 2022). It can also induce the "uncanny valley" effect, a secondary cognitive bias in which being too similar to humans or animals actually causes more discomfort, anxiety, and distrust (Verbene, Ham, and Midden 2015).

It is important to be aware of these biases and approach trust in AI with a critical and discerning mindset. Since biases arise from heuristic processes that allow us to make faster decisions, mitigating them always presents a tradeoff. Often the tradeoff is with speed, but sometimes the tradeoff ends up being between two biases. Making an AI seem less human might increase initial trust, but could then lead to over trust and more fragile trust (Visser et al. 2016); whereas for other people, it might make them distrust the machines by inducing more anxiety (Culley and Madhavan 2013). On the other hand, increasing anthropomorphism might lower initial trust and acceptance, but increase forgiveness (Visser et al. 2016); though it could lead to less system use if it enters the uncanny valley (Mathur and Reichling 2009). One of the greatest challenges organizations will face with AI is determining how their values inform and shape these tradeoffs.

One oft-mentioned cognitive effect tied to trust that was not included in the list above is the *sunk cost* fallacy, because with regard to trust this is **not** a fallacy. Sunk cost refers to the tendency to continue trusting despite opposing evidence after one has made significant resource and/or psychological commitments (Olivola 2018; Arkes and Blumer 1985). However, this can also be understood as commitment and a willingness to forgive and trust that ultimately the system will prove worthwhile. Trust requires vulnerability, and that includes making space for flexibility and even failure (Razin and Feigh 2021).

This will be a particularly hard but crucial lesson for traditionally conservative and risk-averse organizations to absorb. A particular challenge that requires much more forethought and development is how to balance the requirements of safety-critical systems against trust in AI and the vulnerabilities it entails.

Further Useful Definitions

This brings us to defining some terms that are closely related to trust and sometimes used interchangeably. Further confusion may arise given these terms are commonly found in other fields, such as statistics.

- *Reliance, reliability,* and *dependence*: Reliance and dependence can result from trust, but can also arise from a lack of viable alternatives or being forced into a situation. Reliance is the extent to which an operator is willing to use a system in mission contexts. Dependence is similar, but stresses the trustor's lack of autonomy and need for the trustee. The NIST definition takes reliability to mean "the ability of an item to perform as required without failure, for a given time interval, under given conditions" (Probasco, Emelia S., Toney, Autumn S., Curlee, Kathleen T. 2023). To understand the difference, Patriot operators have reliance on their radar, but their radar has poor reliability.
- *Robustness, resiliency, and fragility*: Robustness and resiliency are the ability of a system "to maintain its level of performance under a variety of circumstances" and to "withstand unexpected changes", respectively (Probasco, Emelia S., Toney, Autumn S., Curlee, Kathleen T. 2023). Thus, they focus on the system's reliability when under the strain of external conditions and unexpected usage. Some differentiate between the two; with robustness focusing on maintaining performance despite disruptions, and resilience emphasizing the ability to recover after disruptions; but both aspects contribute to the trustworthiness of a system. Fragility is the opposite of robustness.
- *Confidence*: Confidence relates to the strength of one's belief or expectation, and therefore does not necessarily imply trust. For example, someone could express confidence in the failure of a system without actually trusting it. Usually, confidence is taken in the positive sense as directly supporting trust through the strength of one's expectations in the systems reliability, robustness, and resiliency. Furthermore, there is also the confidence in how well one's trust is calibrated. As with many aspects of trust, confidence must be understood at multiple levels and resolutions.
- *Trusting behaviors/Compliance*: Trusting behaviors, such as compliance, are often easier to measure and less subjective than survey responses. Compliance is when the trustor adheres to the trustee's advice, judgement, or decisions. However, it is challenging to determine whether these behaviors stem from trust beliefs or intentions alone. For instance, high time pressure might increase reliance, and therefore perceived compliance, despite a lack of trust.

Challenges for Testing Trust in AI

AI and advanced autonomy present a new range of challenges to trust in systems and their trustworthiness. Traditional systems are assumed to not change, and therefore testing can have a hard stop after which the trustworthiness of the system is taken to change slowly, with decreasing reliability often likened to a bathtub curve. Changes in reliability may be tracked and communicated by the industry manufacturer or service program office or after major incidents. However, many AI-enabled systems change much more rapidly over time, not just degrading, with algorithmic drift or hardware failure, but also improving as learning occurs. Furthermore, these changes may not be as generalizable as traditional systems, as learning and drift are highly contextual. These differences from traditional systems require iterative testing of AI-enabled systems, which is radically different and far more difficult than current testing strategies. Furthermore, these systems can be "black boxes" compared to other software. Even if testers have access to the code and model parameters, which they rarely do, some AI-enabled systems are famously non-explainable (Saeed and Omlin 2023; Maclure 2021). AI-enabled systems assign relevance to and generalize patterns that do not match human salience and pattern finding. They do not generalize in the same way as humans do. Given their size and processing speeds, modern AI models can detect patterns that elude humans. These systems are often very large and complex, even for AI subject matter experts, who are typically in high demand and short supply.

The speed, size, and complexity of AI-enabled systems makes trust calibration difficult because we have a hard time forming the necessary mental models. Edge cases and emergent properties will continue to surprise us. That same emergence in AI also inhibits formally proving safety and implementing zero-trust policies. If we want to use AI, we must be aware that we are asking for trust in inherently uncertain systems and must be open to and accept the fact that we are vulnerable. It is a tradeoff we must acknowledge. We can manage it through tools such as risk frameworks, extensive testing, safety envelopes, and assurance cases, but ultimately there will always be a gap in what we can know and control and what we cannot.

One might expect that gap to lead us to avoid and under trust AI. Ironically, the bigger problem among the public according to the research is over trust, at least at first (Capgemini Research Institute 2023). To understand this better, we need to think about trust as a dynamic state that continuously changes over time and over interactions with the system. Early on, our expectations of technology in general are formed by our culture, public opinion, and education. We might develop more specific beliefs about trust in a given system based on brand, reputation, recommendations, and research. Finally, we calibrate our trust with actual experience: the more we train, interact, or get feedback on a system's actual performance in the tasks and missions, the more we develop a sense of the system's

level of trustworthiness. With technology, most of us are not experts and just assume that the computer or application or car or drone will work as advertised until it does not. And when it fails, our trust can be brittle. For serious failures, users might never want to deploy that system again. For lower-risk failures, we might balance that risk against the system's cost, the possibility the system will get updated or fixed or will learn, and whether there are viable alternatives.

Designing for better AI trust calibration

Sometimes we might wish to mitigate fragility, whereas other times fragility could be better for the system design. There can be benefits to people assuming that a system is fairer and less prone to mistakes when they must first accept the technology and form a trusting relationship with it. However, this will only work until the system fails. Depending on the magnitude of the failure, the level of risk, and the consequences, the technology might be abandoned because of broken trust. From humans, we would call this betrayal.

Unlike humans, technological systems are generally seen as much more expendable, especially by operators, who until recently were using alternative technologies or were grappling with the same mission goals and tasks manually. This calculus changes for officers who might be aware that the system could give their operators an edge, and it changes drastically the more advanced the technology, the larger the organization, and the higher up the chain of command one looks (Manez et al. 2009; Renshon 2015). Leaders are especially subject to the sunk cost fallacy, often feeling pressure for specific programs to succeed, and instead of abandoning them after repeated failures, committing further resources to their improvement (Renshon 2015). Therefore, it is important that people train on systems, that the limits of systems are well-characterized through robust testing, that their operating envelope is communicated during training to properly set a prior expectation, and that tasks start small or low risk.

Training can be bolstered by simulation, where the algorithms – either standalone or in some combination of hardware-in-the-loop or software-in-the-loop simulation – can be experienced with controlled levels of pressure and lower risk. Training with simulations can inform both plans for gradual adoption as well as iterative designs, where the operators slowly adapt their mental model and expand trust in AI over time, handing off systems one by one. This phased handoff could also mitigate the well-known cognitive bias called *contagion*, in which a failure in one function or set of functions lowers trust in other parts of the system, regardless of whether they are actually correlated.

Therefore, establishing trust in AI-enabled systems faces the following challenges:

- The system itself must perform up to par. Research has found that depending on the task, if the system performance is below a certain threshold (70 percent–85 percent), users will not be able to initially establish trust (J. D. Lee and See 2004).
  - The system does not need to be perfect, but it must be good enough that confidence in its trustworthiness is justified.
  - This can be established in contractor testing (CT) and development testing (DT), to some extent, but subject matter expert and operator feedback is important to ensure that the system and the operators have goal alignment. If the system is doing something well at 93 percent accuracy, but it is not what the operator thinks it's doing or what they need it to do, then the system is not suitable for that user for this task, and trust will be degraded.
  - However, once trust has initially been established, it might persist despite obviously untrustworthy behavior and even spread to previously unseen behaviors. This has been shown by (Holbrook et al. 2023) where interactive robots were trusted to guide people outside of buildings during simulated but believable emergencies despite making observed mistakes that would have resulted in serious injury and fatalities.
- Systems must be designed to help support proper trust calibration. Trust cannot be forced; it must be earned.
  - This includes instrumenting systems to understand performance and how they are being used.
  - This also includes the design of the user interface, the information given to the operator to support their understanding, and the design of the training and familiarization with the system.
    - How the AI system is introduced and brought online matters!
- In operational testing, systems cannot be tested without their operators. Whether we are testing training or fielded operations, testing AI in general requires sufficient time and interaction for the trust to build. It is not 'one and done'. This is just as true for the testers, for them to properly asses how much trust operators and commanders might place in the systems under test compared to how trustworthy they are.
  - The dynamic learning and adapting that occurs also implies the need for iterative re-assessment during sustainment.
  - The need for iterative-reassessment amplified by AI-enabled systems, both because their models can drift over time by becoming increasingly

decoupled from the world on whose data they were initially trained, and conversely because of changes to the system as they learn.

- Furthermore, as real-world field tactics evolve quickly during conflict and new applications for the AI-enabled systems arise, continuous testing strategies assure continued effectiveness in a rapidly changing operational landscape, especially as such systems can exhibit unexpected emergent behaviors.

In AI systems that continue to be updated or to learn, the operator and testers should ensure that their level of previously established trust is still warranted (Yaxley et al. 2021). A further point is that the operator will need a different level of understandability than the tester or regulator, who might need more explainability, or to assess potential threats to the system, such as adversarial AI. At the highest level of transparency, there might be required transparency down to the hardware and code itself. This supports the idea from assurance cases that each party who requires assurance will need arguments and evidence specific to their goals and needs for the system.

Measuring Trust and Trustworthiness

One particular challenge to the test and evaluation of trust and trustworthiness in any system, with AI or not, is how to measure them. Trustworthiness is the more straightforward of the two conceptually but is often more difficult to measure in practice. Recall that trustworthiness is essentially how well a system allows the trustor to accomplish their goals. In order to measure this as a property of the system, one needs to instrument the system such that one can answer 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).

Measuring trustworthiness might require internal instrumentation, access to code, outputs, and external sensors that monitor system movement; but it also can include human factors evaluation of the user interface, task analysis, and other measures of suitability. These components are the baseline for all software testing, and if we cannot get it right for traditional systems, we will never be prepared for such testing for AI-enabled ones.

Trust is commonly measured through surveys that capture an individual's explicit perceptions, knowledge, and expectations. Other ways of measuring trust include physiological and behavioral metrics, but surveys only probe what the user is consciously aware of. Best practice is to use a combination of these methods in order to triangulate whether a person consciously trusts the system and whether they act or react as if they do.

Over 60 trust surveys have been proposed and used for human-automation, human-robot, and human-computer trust; but there is wide variance in their proven reliability, validity, and usability (Razin and Feigh). When choosing a scale, it is important to determine first why you want to measure trust and what resolution is required. Trust scales can be divided into two main categories: single-factor and multi-factor. Single-factor scales are good for answering "Does this person trust the system?", whereas multi-factor scales decompose trust further and can help answer "Why does the person trust the system, and in which ways?"

The other major decision when choosing a trust survey is how long the survey should be. Short surveys capture less data but are suitable when one wants to ask about trust during a task, especially if they will be asked multiple times throughout the task or when the task cannot be interrupted for long. Long surveys capture more data and can yield more statistically meaningful results. These are better suited for before and after tasks and between missions. However, whether using a short survey many times or using a very long survey once, testers should be mindful of survey fatigue, as too many questions can be tiring and yield lower-quality results.

A series of well-validated, single-factor trust surveys can be found in (Merritt et al. 2015; Merritt et al. 2019); a shorter multi-factor one in (Wojton et al. 2020); a longer, but not overly long, multi-factor one in (McKnight, D, Harrison et al. 2011), and both a single- and multi-factor scale in (Schaefer 2016). The U.S. Army Combat Capabilities Development Command Army Research Laboratory has a good review of both validated surveys and non-survey methods for measuring trust (Krausman et al. 2022), and a large review of the 60 validated and published scales including guidance on how to assesses surveys for reliability and validity can be found in (Razin and Feigh).

Although some have expressed doubts that trust is measurable or useful to measure, we hope this review demonstrates that not only can trust be measured, but that serious consequences can arise if it is not.

The Challenge of Safety and Assurance

Can measuring trust and trustworthiness be used to further manage the risks and vulnerabilities trust entails? This is the question that safety and assurance address.

Safety processes manage trust by putting bounds on how much risk is acceptable. The National Institute of Standards and Technology (NIST) and the DoD define system safety as the governability of these systems, meaning the "ability to disengage or deactivate deployed systems that demonstrate unintended behavior" (Depertment of Defense 2020). A different conception of safety that is often conflated with trust is *psychological safety*, which is how much members of an organization or team perceive their groups are open to risk-taking and mistakes (Gallo, 2023). Thus, system safety puts hard limits on the amount

of risk allowed, supporting robustness. On the other hand, (perceived) robustness and resilience lead to psychological safety (see Further Useful Definitions).

One proposed approach to help manage risk and implement governability is *assurance cases*, which combine risk analysis with evidence and tailored arguments to assure specific audiences (Tate 2021). It is not always possible to gather sufficient evidence to cover every possible scenario given the complexity of AI and how it interacts with systems, environments, and operators. Thus, additional arguments are employed to extend the existing evidence that remaining risks are at least accounted for and manageable. This approach aims to provide the trustor with the assurance that the system will perform as expected, or at least not exceed a safe operational space or envelope (Neto et al. 2022).

Another critical tool for supporting assured safety and building trust is modeling and simulation (M&S). M&S can be used to identify the system's operational envelopes within which operational testing can occur safely. Simulations can also be used to test how human users interact with AI-enabled systems and how systems properly communicate with their users or adapt to them (Neto et al. 2022). One particular challenge is understanding how users behave and use the system near or beyond its operational envelope (Freeman, Rahman, and Batarseh 2021). Identifying the edges of operational envelopes for AI-enabled systems can be difficult given the size, speed, and complexity of their calculations and the further possibility of emergent behavior that might not be discovered until end-to-end operational testing occurs, if then. M&S provides one way to identify some problems early on, if it is sufficiently realistic in the relevant dimensions. This does not mean every simulation needs to have the highest fidelity possible. Instead, the realism of the simulation must be suited to the questions or behavior being analyzed. A simulation with very rudimentary representation of the physical environment might be adequate to exercise many AI decision functions and evaluate human-AI teaming CONOPS in useful ways.

Just like the systems themselves, trust must be established in the M&S. Trustworthiness can be assessed through validation, verification, and accreditation (VV&A) of M&S (Elele and Hall 2016), but we must not forget to measure how the simulated systems are perceived and used by operators; in short, how much they trust them.

Increased Transparency and Understandability Do Not Guarantee Trust

Trust is rather closely related to understanding and explaining AI, situational awareness, mental models, fairness/bias, and workload. There is a widespread myth that all of these factors are correlated positively with trust. The reality is much more complex. The wrong level of transparency or wrong kind of explanation can undermine trust calibration and spread trust contagion. Too much transparency or explanation can also burden the operator with information, increasing workload and forcing the operator to rely on the system, not out of trust as much as out of lack of ability to handle more workload (Helldin 2014).

Furthermore, some experiments suggest that robots can be over-trusted in particularly stressful situations to the extent that even when the human does not trust the robot or believe it to be proficient, they will follow it into fatal situations (Holbrook et al. 2023). Shared mental models are positively correlated not with trust as much as calibrated trust. But as mentioned above, too much confidence in one's mental model can be detrimental (Razin et al. 2021; Gigerenzer, Hoffrage, and Kleinbölting 1991). Situational awareness is also supported by the shared mental model, but its correlation with trust/calibrated trust is rather messy (Razin and Feigh; Endsley). There is strong evidence supporting that the ability to predict how a system or environment will change over time correlates with better calibrated capability-based trust (McKnight, D, Harrison et al. 2011; Tussyadiah and Park 2018; Söllner, Pavlou, and Leimeister 2013).

Fairer AI Does Not Guarantee Trust

Algorithmic fairness is the practice of reducing unwanted bias in the data, model, and system behavior. Of particular interest is the practice of reducing bias against historically disadvantaged groups and preventing discrimination based on categories as defined by law (AI Fairness 360 2020; NIST AIRC 2024; ALTAI portal 2024)). There is a general assumption that if fairness is achieved, trust will follow. This assumption, however, may be truer of trust by the public than of operator trust. The public might not trust a company or government that is caught using unfair algorithms, but the impact of fairness on operator trust calibration remains less clear (Angerschmid et al. 2022).

Three major problems have arisen with algorithmic fairness and its assumed relation with trust in AI:

- Achieving algorithmic fairness generally requires tradeoffs in the performance of the model or the quality of the data. For safety-critical or zero-trust operations, this tradeoff is not always seen as acceptable or might present its own ethical dilemmas.
- There are multiple types of fairness, such equity and equality. Mathematically, it has been proven that many of these types are mutually exclusive (Barocas, Hardt, and Narayanan 2023), causing debate on which types of bias are acceptable in a given system or application (Simons, Adams Bhatti, and Weller 2021).
- Finally, although the law or corporate policy may dictate how fairness is designed into a system and what tradeoffs are acceptable, these same choices might not be acceptable to the public, commanders, or operators. Thus, implementing a particular concept of fairness does not guarantee either public or operator trust, and might indeed put them in conflict with one another.

Recommendations

To achieve trusted and trustworthy AI, we recommend a number of tangible steps. Previous work on assuring system quality, while critical, alone is insufficient to guarantee or assess trust or its calibration. First, a shared consistent glossary for human-system integration (HSI) should be disseminated that establishes a common terminology for trust and related concepts. Second, organizations need to encourage the use of reliable and validated trust surveys and behavioral metrics. This must include resourcing for instrumentation and survey collection. If appropriate surveys for their applications are not available or have not been validated in the necessary domain, organizations should conduct or invest in research to create and validate such surveys. Furthermore, measurement of trust from physiology and trust calibration are still active areas of research that should be supported.

Once a common terminology and strong metrics are established, they need to be incorporated into development and T&E. Others have already argued that AI test and evaluation will need to be a spectrum, shifting to both the left and the right. This is just as true for trust and trustworthiness measurement. Trust and trustworthiness design and testing needs must be considered from the very beginning, with both HSI subject matter experts and, more importantly, operators brought into the process early. At the other end, trust's relational nature and AI's ability to both learn and drift will require continuous monitoring and ongoing testing into fielding and sustainment. Finally, all this will require education and training – testers should be aware of the range of metrics available and how to assess their appropriateness and quality; and commanders and managers need to be made aware of the new risks and vulnerabilities AI entails and how to understand trustworthiness, fairness, and transparency when it comes to responsible AI, as well as training for both end users and developers which accounts for calibrated trust design. In the case of the government, this may include incorporating trust into AI, HSI, or human-machine teaming modules in the Defense Acquisition University system; NIST refining its AI glossary, including trust and assurance among its AI measurement and evaluation projects and AI workshops; and providing training on AI.gov and AI.mil.

Conclusion

Despite its ubiquity and importance, trust research is still developing as a field. However, as it develops it is critical to establish a common lexicon and disseminate the current understanding of what trust and trustworthiness are and why they matter. In many ways, AI makes us confront our own understandings of ourselves; to better come to grips with the implicit assumptions we make about the world and the words we use. Trust is not the same as trustworthiness and testers should work to achieve both trustworthy systems and well-calibrated trust. Focusing on trustworthiness or maximizing trust alone can have

disastrous consequences. Trust is hard to measure, and it is very personal. Different stakeholders need different evidence and assurances with regard to the transparency, explainability, mental model, and types of trust the system supports. Contrary to some management fallacies, trustworthiness, transparency, and fairness alone do not guarantee trust, much less calibrated trust. Calibrating trust also takes time: it emerges relationally through personal interaction, personal history, and evidence from fielding. Thus, training and building familiarity are critical. AI and supported simulations might allow for a smoother transition between manual and AI-supported operations, over which operators, commanders, testers, and decision makers can learn how much to trust systems. This process can and should be designed to account for tradeoffs among cognitive biases, including trust contagion, the perfection automation schema, complacency, the Terminator effect, anthropomorphism, and sunk cost. Finally, it is clear that for AI-enabled systems to be trusted, they cannot be fully tested without their operators and without sufficient time for familiarization, training, and operational testing; and such assessments need to be continuous if the AI will continue to learn or to be updated after initial fielding.

Author Bios

Yosef S. Razin is a Research Associate at the Institute for Defense Analyses and a Robotics Ph.D. candidate at the Georgia Institute of Technology, specializing in human-machine trust and the particular challenges to trust that AI poses. His research has spanned the psychology of trust, ethical and legal implications, game theory, and trust measure development and validation. He has applied his research to telerobotics, autonomous cars, AI assistants, and decision support, with a focus on improving our understanding of human-machine teaming.

Kristen Alexander, Ph.D., is the Chief Learning and Artificial Intelligence Officer at DOT&E and focuses on adequate testing of AI-enabled systems and developing curriculum to support the T&E workforce. Prior to that, she served as the Technical Advisor for Deputy Director, Land and Expeditionary Warfare at DOT&E. Dr. Alexander received her B.S. from University of Rochester and her Ph.D. from Carnegie Mellon University in chemical engineering and is the recipient of the Secretary of Defense Medal for Exceptional Civilian Service.

References

AI Fairness 360. 2020. “AI Fairness 360.” Accessed January 04, 2024. https://ai-fairness-360.org/.

ALTAI portal. 2024. “ALTAI Portal.” Accessed January 04, 2024. https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal.

Andrews, Robert W., J. M. Lilly, Divya Srivastava, and Karen M. Feigh. 2023. “The Role of Shared Mental Models in Human-AI Teams: A Theoretical Review.” *Theoretical Issues in Ergonomics Science* 24 (2): 129–75.

Angerschmid, Alessa, Kevin Theuermann, Andreas Holzinger, Fang Chen, and Jianlong Zhou. 2022. “Effects of Fairness and Explanation on Trust in Ethical AI.” In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 51–67.

Arkes, Hal R., and Catherine Blumer. 1985. “The Psychology of Sunk Cost.” *Organizational Behavior and Human Decision Processes* 35 (1): 124–40. doi:10.1016/0749-5978(85)90049-4.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness in Machine Learning: Limitations and Opportunities* 1. Massachusettes Institute of Technology: MIT Press. Accessed February 08, 2024. https://fairmlbook.org/pdf/fairmlbook.pdf.

Blackburn, R. A. “Summary of the 2018 Department of Defense Artificial Intelligence Strategy.” https://media.defense.gov/2019/Feb/12/2002088963/-1/-1/1/SUMMARY-OF-DOD-AI-STRATEGY.PDF. Accessed January 04, 2024.

Buçinca, Zana, Maja B. Malaya, and Krzysztof Z. Gajos. 2021. “To Trust or to Think.” *Proc. ACM Hum.-Comput. Interact.* 5 (CSCW1): 1–21. doi:10.1145/3449287.

Cheatham, Kyler. 2022. “The "Terminator Effect": The Human Side of Artificial Intelligence.” Accessed November 26, 2023.

Chen, Xiaowei, Anders Hedman, Verena Distler, and Vincent Koenig. 2023. “Do Persuasive Designs Make Smartphones More Addictive? A Mixed-Methods Study on Chinese University Students.” *Computers in Human Behavior Reports* 10: 100299.

Cole, Craig. 2022. “Study Says Tesla the Most-Trusted Brand to Develop Autonomous Vehicles.” *CNET*, March 22. Accessed October 20, 2023. https://www.cnet.com/roadshow/news/tesla-most-trusted-to-develop-autonomous-vehicles-self-driving-cars/.

Culley, Kimberly E., and Poornima Madhavan. 2013. “A Note of Caution Regarding Anthropomorphism in HCI Agents.” *Computers in Human Behavior* 29 (3): 577–79.

Das, Kaustav, Yixiao Wang, and Keith E. Green. 2021. “Are Robots Perceived as Good Decision Makers? A Study Investigating Trust and Preference of Robotic and Human Linesman-Referees in Football.” *Paladyn Journal of Behavioral Robotics* 12 (1): 287–96. Accessed November 26, 2023.

DEPARTMENT OF DEFENSE. 2022. “DoD Responsible AI Strategy.” Unpublished manuscript, last modified January 04, 2024. https://www.ai.mil/docs/RAI_Strategy_and_Implementation_Pathway_6-21-22.pdf.

Desai, Munjal. 2012. “Modeling Trust to Improve Human-Robot Interaction.” University of Massachusetts Lowell.

Directorate of Air Staff. 2003. “Aircraft Accident to Royal Air Force Tornado GR MK4A ZG710.”.

Department of Defense. “DoD Adopts Ethical Principles for Artificial Intelligence.” Accessed February 09, 2024. https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

DOD Adopts Ethical Principles for Artificial Intelligence. Depertment of Defense. 2020. Accessed October 20, 2023. https://www.defense.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

Elele, James N., and David H. Hall. 2016. “M&S Requirements and VV & a Requirements: What's the Relationship?” *ITEA*, 333–41. https://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=10540229&an=120515580&h=ggaa3oi%2bc90iwuqmvs1pvllqmxihyeood6xntmmkqjj0j7lrz%2ftjcjjlrdnk0cxc9uocucnhqosvu4yfvrobya%3d%3d&crl=c.

Endsley, Mica R. “Automation and Situation Awareness.” In *Automation and Human Performance 2018*, 163–81. https://aerohabitat.eu/uploads/media/Automation_and_Situation_Awareness_-_Endsley.pdf.

Executive Order 14110. 2023. “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” *The White House*, October 30. Accessed January 04, 2024. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Freeman, Laura, Abdul Rahman, and Feras A. Batarseh. 2021. “Enabling Artificial Intelligence Adoption Through Assurance.” *Social Sciences* 10 (9): 322.

Gallo, Amy. 2023. “What Is Psychological Safety?” *Harvard Business Review*, 2023. Accessed October 18, 2023. https://hbr.org/2023/02/what-is-psychological-safety.

Gefen, David. 2014. “Trust and TAM in Online Shopping: An Integrated Model.” *MIS Quarterly* 24 (4): 665–94. doi:10.2307/3250951.

Gigerenzer, Gerd, Ulrich Hoffrage, and Heinz Kleinbölting. 1991. “Probabilistic Mental Models: A Brunswikian Theory of Confidence.” *Psychological review* 98 (4): 506.

Goldberg, Ken. 2019. “Robots and the Return to Collaborative Intelligence.” *Nature Machine Intelligence* 1 (1): 2–4. https://www.nature.com/articles/s42256-018-0008-x.

Graesel, Hannah S. 2022. “Trust in Artificial Intelligence? The Role of Mental Models, Openness, and Anthropomorphism in Human-Agent Teams.” BS Thesis, University of Twente.

Hancock, P. A., Theresa T. Kessler, Alexandra D. Kaplan, Kimberly Stowers, J. C. Brill, Deborah R. Billings, Kristin E. Schaefer, and James L. Szalma. 2023. “How and Why Humans Trust: A Meta-Analysis and Elaborated Model.” *Frontiers in Psychology* 14.

Hancock, Peter A., Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. “A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction.” *Human factors* 53 (5): 517–27. doi:10.1177/0018720811417254.

Hawley, John K. “Patriot Wars: Automation and the Patriot Air and Missile Defense System.” https://www.cnas.org/publications/reports/patriot-wars.

Helldin, Tove. 2014. “Transparency for Future Semi-Automated Systems: Effects of Transparency on Operator Performance, Workload and Trust.” PhD Thesis, Örebro Universitet.

Hicks, Kathleen. *DoD Artificial Intelligence Symposium and Tech Exchange*, 2021. Accessed January 31, 2024. https://www.youtube.com/watch?v=f_HW7sA6AaE.

Hoff, Kevin A., and Masooda Bashir. 2015. “Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust.” *Human factors* 57 (3): 407–34. doi:10.1177/0018720814547570.

Holbrook, Colin, Daniel Holman, Joshua Clingo, and Alan Wagner. 2023. *Overtrust in AI Recommendations to Kill:* Center for Open Science.

Krausman, Andrea, Catherine Neubauer, Daniel Forster, Shan Lakhmani, Anthony L. Baker, Sean M. Fitzhugh, Gregory Gremillion, Julia L. Wright, Jason S. Metcalfe, and Kristin E. Schaefer. 2022. “Trust Measurement in Human-Autonomy Teams: Development of a Conceptual Toolkit.” *ACM Transactions on Human-Robot Interaction (THRI)* 11 (3): 1–58.

Lee, John D., and Katrina a. See. 2004. “Trust in Automation: Designing for Appropriate Reliance.” *Human factors* 46 (1): 50–80.

Leslie, Ian. 2016. “The Scientists Who Make Apps Addictive.” *The Economist: 1843 Magazine*, 2016. Accessed October 20, 2023. https://www.economist.com/1843/2016/10/20/the-scientists-who-make-apps-addictive.

Liu, Bingjie. 2021. “In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction.” *J Comput Mediat Commun* 26 (6): 384–402. doi:10.1093/jcmc/zmab013.

Maclure, Jocelyn. 2021. “AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind.” *Minds and Machines* 31 (3): 421–38.

Manez, Juan A., Maria E. Rochina-Barrachina, Amparo Sanchis, and Juan A. Sanchis. 2009. “The Role of Sunk Costs in the Decision to Invest in R&D.” *The Journal of Industrial Economics* 57 (4): 712–35.

Mathur, Maya B., and David B. Reichling. 2009. “An Uncanny Game of Trust: Social Trustworthiness of Robots Inferred from Subtle Anthropomorphic Facial Cues.” In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 313–14.

McKnight, D. H., and Norman L. Chervany. 2001. “What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology.” *International Journal of Electronic Commerce* 6 (2): 35–59.

McKnight, D, Harrison, Michelle Carter, Jason B. Thatcher, and Paul F. Clay. 2011. “Trust in a Specific Technology: An Investigation of Its Components.” 2 (2).

Merritt, Stephanie M., Alicia Ako-Brew, William J. Bryant, Amy Staley, Michael McKenna, Austin Leone, and Lei Shirase. 2019. “Automation-Induced Complacency Potential: Development and Validation of a New Scale.” *Frontiers in Psychology* 10 (FEB): 1–13.

Merritt, Stephanie M., Jennifer L. Unnerstall, Deborah Lee, and Kelli Huber. 2015. “Measuring Individual Differences in the Perfect Automation Schema.” *Human factors* 57 (5): 740–53.

Neto, Antonio V. S., João B. Camargo, Jorge R. Almeida, and Paulo S. Cugnasca. 2022. “Safety Assurance of Artificial Intelligence-Based Systems: A Systematic Literature Review on the State of the Art and Guidelines for Future Work.” *IEEE Access*.

NIST AIRC. 2024. “NIST AIRC.” Accessed January 04, 2024. https://airc.nist.gov/Home.

Niu, Dongfang, Jacques Terken, and Berry Eggen. 2018. “Anthropomorphizing Information to Enhance Trust in Autonomous Vehicles.” *Human Factors and Ergonomics in Manufacturing & Service Industries* 28 (6): 352–59.

NSCAI. 2020. “National Security Commission on Artificial Intelligence Final Report.” https://www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf. Accessed January 04, 2024.

Okoh, Chukwunoyenim. 2023. “Robotic Judges: A Future to Desire or Not?” Accessed November 26, 2023.

Olivola, Christopher Y. 2018. “The Interpersonal Sunk-Cost Effect.” *Psychol Sci* 29 (7): 1072–83.

OSDPA. “DOD Data Strategy.” https://media.defense.gov/2020/Oct/08/2002514180/-1/-1/0/DOD-DATA-STRATEGY.PDF. Accessed January 04, 2024.

Probasco, Emelia S., Toney, Autumn S., Curlee, Kathleen T. 2023. “The Inigo Montoya Problem for Trustworthy AI: The Use of Keywords in Policy and Research.”.

Razin, Yosef S., and Karen M. Feigh. *Converging Measures and an Emergent Model: A Meta-Analysis of Human-Automation Trust Questionnaires*. https://arxiv.org/abs/2303.13799.

———. 2021. “Committing to Interdependence: Implications from Game Theory for Human-Robot Trust.” *Journal of Behavioral Robotics* 12 (1): 481–502.

Razin, Yosef S., Jack Gale, Jiaojiao Fan, Jaznae’ Smith, and Karen M. Feigh. 2021. “Watch for Failing Objects: What Inappropriate Compliance Reveals About Shared Mental Models in Autonomous Cars.” In *Proceedings of the Human Factors*. Vol. 65, 643–47. Los Angeles, CA: SAGE Publications.

Renshon, Jonathan. 2015. “Losing Face and Sinking Costs: Experimental Evidence on the Judgment of Political and Military Leaders.” *International Organization* 69 (3): 659–95.

Saddiqui, Faiz, and Jeremy B. Merrill. 2023. “17 Fatalities, 736 Crashes: The Shocking Toll of Tesla's Autpilot.” *The Washington Post*, June 10.

Saeed, Waddah, and Christian Omlin. 2023. “Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities.” *Knowledge-Based Systems* 263: 110273.

Schaefer, Kristin E. 2016. “Measuring Trust in Human Robot Interactions: Development of the “Trust Perception Scale-HRI”.” In *Robust Intelligence and Trust in Autonomous Systems*, 191–218: Springer.

Simons, Joshua, Sophia Adams Bhatti, and Adrian Weller. 2021. “Machine Learning and the Meaning of Equal Treatment.” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 956–66.

Söllner, Matthias, Paul A. Pavlou, and Jan M. Leimeister. 2013. “Understanding Trust in IT Artifacts-a New Conceptual Approach.” *Available at SSRN 2475382*.

Tate, David M. 2021. *Trust, Trustworthiness, and Assurance of AI and Autonomy*. https://www.researchgate.net/profile/david-tate-8/publication/355479055_trust_trustworthiness_and_assurance_of_ai_and_autonomy.

Tussyadiah, Iis P., and Sangwon Park. 2018. “When Guests Trust Hosts for Their Words: Host Description and Trust in Sharing Economy.” *Tourism Management* 67:261–72. doi:10.1016/j.tourman.2018.02.002.

Verbene, Frank M. F., Jaap Ham, and Cees J. H. Midden. 2015. “Trusting a Virtual Driver That Looks, Acts, and Thinks Like You.” *Human factors* 57 (5): 895–909. Accessed November 26, 2023.

Visser, Ewart J. de, Samuel S. Monfort, Ryan McKendrick, Melissa A.B. Smith, Patrick E. McKnight, Frank Kreuger, and Raja Parasuraman. 2016. “Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents.” *Journal of Experimental Psychology* 22 (3): 331.

Wang, Yue, Fangjian Li, Huanfei Zheng, Longsheng Jiang, Maziar F. Mahani, and Zhanrui Liao. 2024. “Human Trust in Robots: A Survey on Trust Models and Their Controls/Robotics Applications.” *IEEE Open J. Control. Syst.* 3: 58–86.

Weitz, Katharina, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2021. ““Let Me Explain!”: Exploring the Potential of Virtual Agents in Explainable AI Interaction Design.” *Journal on Multimodal User Interfaces* 15 (2): 87–98.

“Why Consumers Love Generative AI.” 2023. Unpublished manuscript, last modified February 09, 2024. https://prod.ucwe.capgemini.com/wp-content/uploads/2023/05/Final-Web-Version-Report-Creative-Gen-AI.pdf.

Wojton, Heather M., Daniel Porter, Stephanie T. Lane, Chad Bieber, and Poornima Madhavan. 2020. “Initial Validation of the Trust of Automated Systems Test (TOAST).” *The Journal of social psychology* 160 (6): 735–50.

Yaxley, Kate J., Keith F. Joiner, Jean Bogais, and Hussein A. Abbass. 2021. “Life Learning of Smart Autonomous Systems for Meaningful Human-Autonomy Teaming.” In *a Framework of Human System Engineering: Applications and Case Studies*, edited by Holly A. Handley and Andreas Tolk, 43–61. Hoboken New Jersey: Wiley-IEEE Press.

Zerilli, John, Umang Bhatt, and Adrian Weller. 2022. “How Transparency Modulates Trust in Artificial Intelligence.” *Patterns* 3 (11). Accessed November 26, 2023.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

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

4. TITLE AND SUBTITLE

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 and Kristen Alexander

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

3000512

3000512

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

Director, Operational Test and Evaluation
1700 Defense Pentagon
Room 3E1088 Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

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: John T Haman

14. ABSTRACT

With the explosion of generative AI, trust in AI has quickly gained significant attention in public discourse, to the point where it is becoming a buzzword. Consequently, it is crucial to ensure that everyone involved in the design, development, and testing of AI is aware of what trust is, why it matters, and how to operationalize and measure it. This introductory paper aims to provide members of the T&E community with a clear understanding of trust and trustworthiness to support responsible and effective evaluation of AI systems. It also seeks to dispel common myths and misconceptions around trust. The paper focuses on the particular ways AI challenges trust and specific considerations that this raises for the T&E of AI-enabled systems. It provides a set of working definitions and outlines the major debates that have previously caused confusion. While defining trustworthiness and explaining its importance, the paper also moves beyond the techno-centric concept of trustworthiness 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

15. SUBJECT TERMS

trust; Trustworthiness; artificial intelligence (AI)

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