INSTITUTE FOR DEFENSE ANALYSES

Trust, Trustworthiness, and Assurance of AI and Autonomy

David M. Tate

April 2021

Approved for public release;
distribution is unlimited.

IDA Document D-22631

Log: H 21-000128

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Project AI-2-4837, “Test, Evaluation, Verification, and Validation of Autonomous Systems,” for the Assistant Director (Autonomy), Office of the Director of Defense Research and Engineering for Modernization. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For More Information

David A. Sparrow, Project Leader
dsparrow@ida.org, (703) 578-2992

David M. Tate, Author
dtate@ida.org, (703) 575-1409

Leonard J. Buckley, Director, Science and Technology Division
lbuckley@ida.org, (703) 578-2800

Copyright Notice

INSTITUTE FOR DEFENSE ANALYSES

IDA Document D-22631

Trust, Trustworthiness, and Assurance of AI and Autonomy

David M. Tate

Executive Summary

Introduction

In this paper I argue that it is not useful to speak of trust in artificial intelligence (AI) or AI-enabled autonomous systems (AIAS) as if "trust" were a single thing. I argue further that it is simply wrong to believe that any important kind of trust can be "built in" to AIAS through system design choices and testing. Each of the important kinds of trust associated with AIAS—and there are several—will require additional deliberate action beyond design and testing, and this is why trust cannot be built in to a system. We need to clean up both our language and our thinking about trust in AIAS in order to focus better on the core challenges to successful employment of such systems.

For purposes of argument, I define a system to be *trustworthy* to the extent that:

1. When employed correctly, it will dependably do well what it is intended to do.
2. When employed correctly, it will dependably not do undesirable things.
3. When paired with the humans it is intended to work with, it will dependably be employed correctly.

That last criterion is important, because it is quite possible to design and build AIAS that could function as intended, but that humans cannot interact with in the necessary ways. Considering all three criteria, the authorities who regulate the use of an AIAS must *know* when it is trustworthy, with sufficient justified confidence that they are willing to permit its use. At present, we face a choice between fielding AIAS of unknown trustworthiness or being bounded in what we can do by the limitations of our ability to provide evidence for trustworthiness that is both valid and compelling.

Findings

A system is *assured* when the relevant authorities have sufficient justified confidence in the trustworthiness of the system to authorize its employment in specified contexts. There are three key features here:

- Whose trust is needed (i.e., a regulating authority).
- The level of confidence required (given potential risks and benefits).
- The (context-dependent) level of confidence justified by the available evidence.

An *assurance case* is a collection of explicit arguments that a system is sufficiently trustworthy for its intended use. Assurance cases are routine in the safety and cybersecurity communities and have increasingly been generalized to apply to other aspects of system trustworthiness as well. Leading international standards organizations have adopted assurance cases as the preferred approach to establishing the trustworthiness of engineered systems and software.

To establish trustworthiness, an assurance case must provide explicit evidence-based arguments supporting the three facets of trustworthiness: that the system is sufficiently likely to do the things it is supposed to do and to avoid doing undesired things, given how it is designed to interact with humans. The complexity of the arguments will depend on the complexity of the system, the richness of the human-machine teaming concept, and the range of desired and unacceptable behaviors to be addressed. Furthermore, just as different decisions call for different supporting assurance cases, the nature of the evidence and argument will also be different depending on whose trust matters. Indiscriminate use of the word "trust" can obscure these distinctions. The assurance case model lets us replace fuzzy intuitions about trust with concrete testable requirements. Rather than saying that we need explanations and transparency and trust, we can say "For *these* stakeholders, we need evidence to support *this* argument in support of *that* assurance claim at *this* level of confidence" for a finite list of specific stakeholders and claims.

Viewed from that perspective, the purpose of test, evaluation, verification, and validation (TEV&V) becomes clear: it is the activity that produces the evidence that completes the needed assurance arguments. Some of this evidence will be familiar, such as verification of compliance with design specifications, reliability testing, or human factors evaluations. Other evidence may require novel tools and techniques. This is where "transparency" fits in: ancillary outputs or explanatory models that make the machine behavior more understandable to humans are useful precisely when the explanations they produce improve appropriate reliance by humans or increase stakeholder confidence in assurance claims. Fortunately, software tools already exist to support formal assurance case development and management, including tools designed specifically for AIAS applications.

Conclusions

Developers of AIAS are now on the verge of having real systems in the development pipeline whose potential employment is limited not by their trustworthiness, but by our ability to understand and characterize that trustworthiness. It is time to stop talking about "building in trust" and to instead start using the analytical methods, tools, standards, and processes that already exist to provide structure and rigor to the TEV&V process. Assured trustworthy AIAS are not beyond our abilities, once we are clear on precisely what that means.

Contents

1. Loose Talk About Trust

Many words have been written about the importance of *trust* in artificial intelligence (AI) and autonomous systems. The "National Artificial Intelligence Research and Development Strategic Plan" (National Science and Technology Council 2016) asserts "the need for explainable and transparent systems that are trusted by their users, perform in a manner that is acceptable to the users, and can be guaranteed to act as the user intended," and adds that "[f]urther progress in research is needed to address this challenge of creating AI systems that are reliable, dependable, and trustworthy." The Office of Management and the Budget (2020) asserts that the "importance of developing and deploying AI requires a regulatory approach that fosters innovation and growth and engenders trust, while protecting core American values" and adds that "the continued adoption and acceptance of AI will depend significantly on public trust and validation." Other writers (Hancock, Billings, and Schaefer 2011) emphasize that society will not realize the benefits of AI-enabled autonomous systems (AIAS) if their operators do not trust them enough to rely on them. "No trust, no use" (Schaefer et al. 2016). Still others note that trust in AIAS is sometimes misplaced, and that our trust in our machines should be calibrated and appropriate, not blind (Metcalfe 2017).

Some commenters go so far as to say that trust must be "built in" when developing and fielding AIAS. Colin Parris, Chief Technology Officer at GE Digital, writes "At GE, we appreciate the tremendous potential and improvements intelligent, autonomous systems can bring to industrial productivity. But none of it matters unless you have built-in trust" (Parris 2019). The President and CEO of Booz Allen Hamilton writes (Rozanski 2019), "Now is the time for experts…to agree on a national strategy for AI that will advance its adoption by building in trust that it is safe, effective, ethical, accountable and transparent." In their recent survey paper on trust in AIAS, authors at the Center for Security and Emerging Technology summarized this need (Konaev, Huang, and Chahal 2021):

> The transparency of the system, the capacity of the system to explain its decisions, the quality of communications between human and machine, and the reliability of the system in the present and future are all critical factors for calibrating trust and enabling effective human-machine teaming. Research and innovation has therefore focused on ways to "build in" trust into autonomous and AI-enabled systems through features and functions that make these systems more transparent, explainable, auditable, reliable, robust, and responsive.

In this paper I argue that it is not useful to speak of trust in AIAS as if it were a single thing and that it is simply wrong to believe that any important kind of trust can be "built in" as a function of system design choices and testing. Each of the important kinds of trust associated with AIAS—and there are several—will require additional deliberate action beyond design and testing, and this is why "trust" cannot be "built in" to a system. We need to clean up both our language and our thinking about trust in AIAS to focus better on the core challenges to successful employment of such systems.

2. What is Trustworthiness?

Everyone agrees that AI-enabled and autonomous systems should be "trustworthy," though not everyone phrases the requirement in that way. Some writers speak of systems being "dependable" or "reliable" or "robust" or "safe and effective"—each of which has a slightly different flavor, with different connotations. Within the Department of Defense, the Director of Operational Test and Evaluation is charged with assessing systems for "effectiveness and suitability." How do these other formulations relate to trustworthiness?

For purposes of this discussion, I propose the following functional definition: a system is *trustworthy* to the extent that

1. When employed correctly, it will dependably do well what it is intended to do.
2. When employed correctly, it will dependably not do undesirable things.
3. When paired with the humans it is intended to work with, it will dependably be employed correctly.

That last criterion is important, because it is possible to design and build AIAS that could function as intended, but that humans cannot interact with in the necessary ways. For example, a self-driving car that sometimes requires a human operator to suddenly take decisive action after hours of boredom cannot be trustworthy: "It's this whole issue of human beings [being] very, very bad at monitoring a task if they're only asked to monitor without being engaged" (Bigelow 2018).

Similarly, there have been case studies of AI systems for interpreting medical imagery that improved the performance of human physicians at identifying cancers in controlled experiments—but that resulted in higher costs with no improvement in detection rates. When those AI systems were used in clinical practice, the overall patient outcomes were worse in some cases (Lehman et al. 2015). This unexpected failure has been attributed to the *laboratory effect*, in which people behave differently when they know they are part of a laboratory experiment than they do in their everyday lives (Gur et al. 2008).

Note that, in many cases, how well the human-machine team performs its intended functions depends directly on the nature and extent of the human's propensity to rely on the AIAS in various circumstances—that is, "operator trust." This subjective, personal notion of "trust" comes closest to the ordinary language sense of the word. At the simplest level, the AIAS can't be useful if the humans ignore it or turn it off. At a more nuanced level, team performance suffers when the humans have poorly calibrated notions of the tasks and roles for which the machine is trustworthy. The more complex the teaming interactions become, the more opportunity there is for this kind of miscalibration (Metcalfe et al. 2017). Furthermore, this subjective trust relationship between human and AIAS is dynamic, changing over time in response to observed behavior, increased familiarity, and learning by doing. It can also be influenced by factors external to the teaming relationship (Schaefer et al. 2021). This poses significant challenges for the design of human-machine teaming operational concepts. Since the machine side of these human-machine interaction concepts must be implemented in the AIAS hardware and software, these are design-time challenges that cannot in general be solved by post-deployment tweaks to training or user interfaces. While designers should certainly try to anticipate how their hardware and software choices will affect human-machine teaming, this will not in general be sufficient to ensure appropriate trust by all stakeholders.

Finally, in practice it is not sufficient for a system to be trustworthy and likely to be trusted by end users. The authorities who regulate the use of the system must also *know* when it is trustworthy, with sufficient justified confidence that they are willing to permit its use. In the ongoing conversation about trust in AIAS, this vital step is often overlooked. In the early days of civil engineering, it was often easier to build a bridge than it was to know how much weight that bridge could support. In these early days of machine learning, it is often easier to build an AIAS than to know how trustworthy it is (and under what specific circumstances). For the foreseeable future, we face a choice between fielding AIAS of unknown trustworthiness and being bounded in what we can do by the limitations of our ability to provide evidence for trustworthiness that is both valid and compelling.

3. Trustworthy vs. Trusted vs. Assured

A. Assurance: Justified Objective Trust

It is possible to introduce terminology that resolves some of the confusion alluded to above. Say that a system is *assured* when the relevant authorities have sufficient justified confidence in the trustworthiness of the system to authorize its employment in specified contexts. The level of assurance is determined by three key features:

- Whose trust is needed (i.e., a regulating authority).
- The level of confidence required (given potential risks and benefits).
- The (context-dependent) level of confidence justified by the available evidence.

An *assurance case* is a collection of explicit arguments that a system is sufficiently trustworthy for its intended use. Assurance cases, which are routine in the safety and cybersecurity communities, have increasingly been generalized to apply to other aspects of system trustworthiness as well. Several international standards organizations, including the International Organization for Standardization (ISO) and the Institute of Electrical and Electronics Engineers (IEEE), have adopted assurance cases as the preferred approach to establishing the trustworthiness of engineered systems and software (ISO/IEC 2011).

To establish trustworthiness, an assurance case must therefore at a minimum provide explicit evidence-based arguments supporting our three facets of trustworthiness: that the system is sufficiently likely to do the things it is supposed to do and to avoid doing undesired things, given how it is designed to interact with humans. How much likelihood is "sufficient" will depend on the nature and magnitude of the potential benefits and risks of employing the system, as perceived by the regulating authorities (and perhaps by the public). The complexity of the arguments will depend on the complexity of the system, the richness of the human-machine teaming concept, and the range of desired and unacceptable behaviors to be addressed.

Successful fielding of an AIAS involves a succession of decision points. There is an implied assurance case to be made at each decision point, supporting a conclusion of "we're ready to go to the next step." The structure of the required arguments and the nature of the supporting evidence will necessarily be different for (say) awarding safety releases on the test range vs. awarding safety certification for a fielded system.

B. Different Assurance Arguments for Different Audiences

Just as different decisions call for different supporting assurance cases, the nature of the evidence and argument will also be different, depending on whose trust matters. The vendor needs assurance that the system will be profitable and that liability risk is acceptable, given the potential rewards. Buyers want to know that the product will be worth the price, as well as being safe. Regulators need assurance that the system conforms to applicable laws, regulations, and standards. And the public wants to be reassured that the product will not lead to adverse social outcomes for non-buyers.

Use of the word "trust" can obscure these distinctions, as in statements like this one (Flournoy, Haines, and Chefitz 2020):

> The Pentagon cannot let TEVV become a barrier to fielding AI-enabled systems in an operationally relevant time frame, but must do so in a manner that engenders trust in such systems and is consistent with U.S. values and principles. The ultimate goal of any TEVV system should be to build trust—with a commander who is responsible for deploying a system and an operator who will decide whether to delegate a task to such system—by providing relevant, easily understandable data to inform decision-making.

When these authors say "build trust" here, they are actually talking about two very different needs. The commander (and the acquisition executive) require assurance, as defined above. For the operator, the Pentagon's goal is appropriate reliance on the system, which depends not only on TEV&V but also on interface design, training, and familiarization. For the Congress, the Pentagon's goal is to justify the required funding. For the public, the goal is belief that the system furthers national security goals within acceptable ethical and stewardship limits. As we will see below, simply providing data to inform decision-making does not automatically result in appropriate trust.

As an example of the diversity of assurance argument needs, consider a hypothetical self-driving personal car in the United States. The stakeholders whose "trust" matters include the federal government, the state and local governments of any jurisdictions where the car might drive on public roads, potential passengers, the public in general who will be sharing the roads with these cars, the insurance industry, consumer advocacy organizations, and the companies who are potential developers, vendors, maintainers, and operators of such cars. Because these stakeholders have different values and priorities, as well as differing levels of technical savvy, the same arguments and evidence will not be equally persuasive to them all. Some stakeholders may also require higher degrees of confidence than others—for instance, a regulatory body protecting public safety might have a lower tolerance for risk of injury than an insurer or manufacturer. And in the case of potential customers and the general public, the psychology of trust will be as important as the reality of trustworthiness.

A recent example of an assurance argument aimed at the psychology of public trust was seen when American Airlines reintroduced the Boeing 737 Max to its operational fleet of aircraft. American publicized the fact that the first flight after the hiatus carried not only the president of the company but also the captain's wife and the first officer's mother (Hart 2020). While this information was irrelevant to regulators, the public might well incorporate it into its subjective assessments of the trustworthiness of the redesigned aircraft.

As noted above, successful deployment of AIAS can depend on the trust attitudes of the humans who operate, collaborate with, or coexist with the machines. Continuing our example, the overall safety and utility of a self-driving car will greatly depend on how the drivers around that car behave, which in turn will depend on how those drivers expect the self-driving car to behave. As we will see below, this is a key feature of AIAS that will call for a very different design process.

C. Explanation, Transparency, and "Built-In Trust"

A number of commenters have responded to the importance of trust in enabling successful deployment of AIAS by asserting that developers should therefore "build in trust" when designing AIAS. It is useful to drill down into exactly what they might mean by this and whether it is a viable strategy for achieving the goals of AIAS.

Here's a typical example of this way of speaking (Greenwood 2018):

> For AI to benefit both businesses and societies, we need to design in ethical principles like trust and transparency.… Designers have the opportunity to help organizations anticipate negative outcomes, define what good looks like, address bias, and build in trust and transparency.

The link between transparency and trust is made explicit here, but the purported mechanism—how transparency translates into trust—is not. It is assumed that the right kind of transparency will automatically produce the right kind of trust, but no details are provided. As we have seen above, there is no one right kind of trust. Furthermore, the relationship between system transparency and human trust is anything but simple (Palmer and Zwillinger 2016; Schaefer et al. 2021).

Even authors who clearly understand many of these nuances can adopt terminology and language that obscures them. In an otherwise useful discussion of trust-aware systems engineering of autonomous systems, Palmer and Zwillinger (2016) present a section titled "Building Trust In." After making it clear that what they are really talking about is building systems that are "trustable" (i.e., trustworthy), they immediately drop that word and instead talk about "building trust in":

> The underlying concept of our Trust V approach is that trust, like quality, must be built into the system and not "bolted on" after the fact. Starting in

> the concept of operations (ConOps) phase and continuing through the entire Systems Engineering lifecycle, activities to engender trust should be built in. The Trust V approach identifies a selection of trust methods across all phases of the systems lifecycle to instill confidence that the system will perform correctly. Of course, the appropriate trust method(s) used for any specific system must be negotiated between customers, end users, and contractors.

This language gives the strong impression that appropriate trust—by all who require "confidence that the system will perform correctly"—happens automatically, given the proper use of "trust methods" in the design. What this language obscures is that no "trust methods" currently exist that would enable designers to preemptively "engender trust" in the necessary ways. When the authors say that "trust method(s) used…must be negotiated between customers, end users, and contractors," they are sidestepping the question of what that negotiation would look like. In many cases, it will involve significant trial-and-error experimentation, not only to design a trustworthy system but also to discover the concept of operations and forms of transparency that allow customers, users, and the public to achieve appropriate trust in that system. This is a very different notion of "building in" than traditional examples such as building in durability in a garment or building in accuracy in a watch. It's not something the developers can do by themselves in their workshop.

The notion of "transparency" is itself problematic. What exactly constitutes transparency in operations for an AIAS? Many authors frame this in terms of *explanations*—the system must be able to explain why it makes the choices it makes or at least provide ancillary outputs that permit humans to construct such explanations. From the foregoing discussion, we can see that this is really just an incomplete notion of assurance. Explanation will generally not be sufficient to all audiences, either to induce appropriate trust or to make a convincing assurance argument. If we think of explanations as components of assurance cases, we can see immediately that they must be tailored to the recipient. The kind of explanation that best supports appropriate operator reliance will not, in general, be the same kind of explanation that best supports regulatory assurance or public trust. The notion of transparency is an oversimplification of the mechanisms of assurance.

The very metaphor of transparency implies that there is an objective truth that can be seen clearly (and is the same for all observers) if the "window" is clear enough. Trust is not like that; there is no objective trust that will be the same for all observers, even if the proper window into AIAS behavior were provided. Rather, facts about the AIAS must be established and assembled differently to meet the information needs (and account for the emotional needs) of the various observers.

D. Assurance as the Output of TEV&V

Talk of transparency and explanation and trust reflects a partial awareness of the assurance gap for AIAS. Commenters inside and outside the autonomy community have correctly perceived that traditional systems engineering practices cannot guarantee that our AIAS will be sufficiently safe, secure, and dependable. This clearly has something to do with trust, and trust clearly has something to do with transparency and explanation. The assurance case model lets us translate these fuzzy intuitions about trust into concrete testable requirements. Rather than saying that we need explanations and transparency and trust, we can say "For *these* stakeholders, we need evidence to support *this* argument in support of *that* assurance claim at *this* level of confidence" for a finite list of specific stakeholders and claims.

Viewed from that perspective, the purpose of TEV&V becomes clear: it is the activity that produces the evidence that completes the needed assurance arguments. Some of this evidence will be familiar, such as verification of compliance with design specifications, reliability testing, or human factors evaluations. Other evidence may require novel tools and techniques—digital twins, virtual testbeds, formal models, adversarial AI for robustness assessment, etc. This is where "transparency" fits in: any ancillary outputs or explanatory models that make the machine behavior more understandable to humans are useful precisely when the explanations they produce improve appropriate reliance by humans or increase stakeholder confidence in assurance claims.

Historically, the engineering community has treated the different assurance arguments needed—safety, ethics, cybersecurity, robustness, etc.—as separate ad hoc efforts performed by different sets of specialists. For toasters or telephones, that's a reasonable approach—the factors affecting safety are mostly distinct from the factors affecting performance, and so forth. For sophisticated AIAS, however, the autonomous capabilities induce coupling of all dimensions of trustworthiness. In our hypothetical self-driving car, safety, cybersecurity, reliability, and operational effectiveness are inextricably linked. Formulating explicit assurance cases for all stakeholders (including the public) and deriving TEV&V strategies to produce the evidence that supports all of them offers a systematic way to account for all the issues of trust simultaneously, while minimizing redundancy and omissions.

These highly interconnected assurance cases can be complex. Fortunately, software tools already exist to support formal assurance case development and management (Netkachova, Netkachov, and Bloomfield 2014; Denney and Pai 2018), including tools designed specifically for AIAS applications (Bloomfield et al. 2019; Asaadi et al. 2020). There is also much active research in both formal and empirical methods to provide the evidence these assurance-case models are built on (Dreossi et al. 2019).

4. Summary and Conclusions

Too much trust and too little trust are both unhelpful. We want to trust our AI-enabled and autonomous systems in those cases where they are trustworthy, but not trust them in those cases where they are not. We want our regulators to quickly permit the use of trustworthy AIAS, but restrict or forbid the use of untrustworthy systems. We want the public to take advantage of AIAS to improve productivity and innovation, but avoid unintended harm to individuals or social structures. We want the military to be able to exploit the power of AIAS to defend the nation and promote global interests, but without violating the Laws of Warfare, rules of engagement, or fundamental national values. In short, we want people to have justified confidence that AIAS deployed in governmental functions adhere to high standards of safety, fairness, and dependability.

While it is tempting to characterize these desires as being about *trust*, there are good reasons not to. As we have seen above, there are both practical and perceptual differences among these flavors of "trust" that have important consequences for what is needed to achieve these goals. Discourse about the importance of trust can obscure more than it illuminates, misleadingly suggesting similarities (and common solutions) across fundamentally unrelated challenges. Discourse about "building in trust" to AIAS mistakenly implies that trust is an attribute of the AIAS that can be predictably engineered. Instead, we need discourse about assured trustworthiness and appropriate reliance, and how to achieve them.

There is widespread concern that current approaches to TEV&V of engineered systems will not be sufficient to realize the promise of AIAS. This concern is justified—the traditional sequential treatment of systems engineering, developmental testing, operational testing, and field testing will not be able to assure the trustworthiness of human-AIAS collaborations in any but the simplest of cases. However, this does not mean that we lack the technologies to achieve the goal. Assurance cases, as pioneered by the safety and cybersecurity communities, provide a structured mechanism to address these challenges. Formal assurance cases resolve the overloading of the term "trust" by making clear which arguments need to be convincing to whom or need to promote specific patterns of belief and reliance. They also provide a mechanism for translating general notions of the need for explanation and transparency into concrete testable requirements for specific evidence in support of specific claims.

Importantly, assurance cases also clearly establish which outputs of TEV&V are needed to support final acceptance by decision-makers and the public and how this translates into specific measurements and test instrumentation. Assurance cases are thus extremely useful for designing test strategies and test plans, identifying infrastructure needs, incorporating modeling and simulation efficiently into TEV&V, and generally avoiding duplication and wasted effort in the accumulation of evidence toward assurance. Software tools already exist to support assurance case development and management for AIAS (Bloomfield et al. 2019; Asaadi et al. 2020). Assured trustworthiness of AIAS can be achieved, but it will require not only these new tools but also a change of management and of culture in the integration of assurance into the development process.

For more than a decade now, the autonomy community has been aware of the challenges AIAS would pose to TEV&V (Macias 2008). Developers of AIAS are now on the verge of having real systems in the development pipeline whose potential employment is limited not by their trustworthiness, but by our ability to understand and characterize that trustworthiness. It is time to stop talking about "building in trust" and to instead start using the assurance methods, tools, standards, and processes that already exist to provide structure and rigor to the TEV&V process. Assured trustworthy AIAS are not beyond our abilities, once we are clear on precisely what that means.

Appendix A. References

Asaadi, Erfan, Ewen Denney, Jonathan Menzies, Ganesh Pai, and Dimo Petroff. 2020. “Dynamic Assurance Cases: A Pathway to Trusted Autonomy.” *Computer* 53.12: 35–46.

Bigelow, Pete. 2018. “Video of Fatal Uber Crash Reveals Shortcomings of Human and Automated Driving.” *Car and Driver*, March 22, 2018. Accessed April 23, 2021, at https://www.caranddriver.com/news/a19563095/video-of-fatal-uber-crash-reveals-shortcomings-of-human-and-automated-driving/.

Bloomfield, R., H. Khlaaf, P. Ryan Conmy, and G. Fletcher. 2019. “Disruptive Innovations and Disruptive Assurance: Assuring Machine Learning and Autonomy.” *Computer* 52, no. 9: 82–89. doi: 10.1109/MC.2019.2914775.

Denney, E., and G. Pai 2018. “Tool Support for Assurance Case Development.” *Autom Softw Eng* 25:435–99. https://doi.org/10.1007/s10515-017-0230-5.

Dreossi, Tommaso, Daniel J. Fremont, Shromona Ghosh, Edward Kim, Hadi Ravanbakhsh, Marcell Vazquez-Chanlatte, Sanjit A. Seshia. 2019. “VERIFAI: A Toolkit for the Design and Analysis of Artificial Intelligence-Based Systems.” February 2019. arXiv:1902.04245 [cs.AI].

Flournoy, Michèle A., Avril Haines, and Gabrielle Chefitz. 2020. “Building Trust through Testing: Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, Including Deep Learning Systems.” October 2020. Accessed March 25, 2021, at https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.

Greenwood, Tom. 2018. “AI’s Next Technological Breakthrough? Ethics.” *Designit* (blog), December 2018. Accessed at https://medium.com/@Designit/ais-next-technological-breakthrough-ethics-144e36815ec9 on March 30, 2021.

Gur, D., A. I. Bandos, C. S. Cohen, C. M. Hakim, L. A. Hardesty, M. A. Ganott, R. L. Perrin, W. R. Poller, R. Shah, J. H. Sumkin, and L P. Wallace. 2008. “The ‘Laboratory’ Effect: Comparing Radiologists’ Performance and Variability during Prospective Clinical and Laboratory Mammography Interpretations.” *Radiology* 249 (1): 47–53.

Hancock, P. A., D. R. Billings, and K. E. Schaefer. 2011. “Can You Trust Your Robot?” *Ergonomics Des.: Q. Hum. Factors Appl.* 19 (3): 24–29.

Hart, Robert. 2020. “Boeing 737 Max Completes First New Flight With Captain’s Wife, American Airlines President Aboard.” Accessed March 25, 2021, at https://www.forbes.com/sites/roberthart/2021/12/29/boeing-737-max-completes-first-new-flight-with-captains-wife-american-airlines-president-aboard/.

ISO/IEC. 2011. "Systems and Software Engineering - Systems and Software Assurance - Part 2: Assurance Case." ISO/IEC standard 15026-2:2011(E), February 2011.

Konaev, Margarita, Tina Huang, and Husanjot Chahal. 2021. "Trusted Partners: Human-Machine Teaming and the Future of Military AI." CSET issue brief, February 2021.

Lehman, C. D., R. D. Wellman, D. S. Buist, K. Kerlikowske, A. N. Tosteson, and D. L. Miglioretti. 2015. "Diagnostic Accuracy of Digital Screening Mammography with and without Computer-Aided Detection." *JAMA Internal Medicine* 175 (11): 1828–37.

Macias, Fil. 2008. "The Test and Evaluation of Unmanned and Autonomous Systems." White Sands Missile Range, NM.

Metcalfe, J. S., A. R. Marathe, B. Haynes, V. J. Paul, G. M. Gremillion, K. Drnec, C. Atwater, J. R. Estepp, J. R. Lukos, E. C. Carter, W. D. Nothwang. 2017. "Building a Framework to Manage Trust in Automation." *Proc. SPIE 10194, Micro- and Nanotechnology Sensors, Systems, and Applications IX*, 101941U (May 18, 2017). doi: 10.1117/12.2264245.

National Science and Technology Council. 2016. "The National Artificial Intelligence Research and Development Strategic Plan," October 2016, 15.

Netkachova, K., O. Netkachov, and R. Bloomfield. 2014. "Tool Support for Assurance Case Building Blocks." In *Computer Safety, Reliability, and Security*, edited by F. Koornneef and C. van Gulijk. SAFECOMP 2014. *Lecture Notes in Computer Science* 9338. Springer, Cham. https://doi.org/10.1007/978-3-319-24249-1_6.

Office of Management and the Budget. 2020. "Guidance for Regulation of Artificial Intelligence Applications." Memorandum M-21-06, November 17, 2020.

Palmer, G., A. Selwyn, and D. Zwillinger. 2016. "The 'Trust V': Building and Measuring Trust in Autonomous Systems." In *Robust Intelligence and Trust in Autonomous Systems,* edited by R. Mittu R., D. Sofge, A. Wagner, and W. Lawless. Boston: Springer. https://doi.org/10.1007/978-1-4899-7668-0_4.

Parris, Collin. 2019. "Digital Twin 2.0 and the Emergence of 'Humble AI.'" Weblog article, January 16, 2019. https://www.linkedin.com/pulse/digital-twin-20-emergence-humble-ai-colin-parris/.

Rozanski, Horacio. 2019. "The AI Boom: Why Trust Will Play a Critical Role." Weblog article, May 9, 2019. https://knowledge.wharton.upenn.edu/article/coming-ai-breakout-need-rules-road-now/.

Schaefer, K. E., J. Y. C. Chen, J. L. Szalma, and P. A. Hancock. 2016. "A Meta-analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems." *Hum. Factors* 58 (3): 377–400.

Schaefer, Kristin E., Brandon S. Perelman, Gregory M. Gremillion, Amar R. Marathe, and Jason S. Metcalfe. 2021. "Chapter 12 - A Roadmap for Developing Team Trust

Metrics for Human-Autonomy Teams.” In *Trust in Human-Robot Interaction*, edited by Chang S. Nam and Joseph B. Lyons. Academic Press, 261–300.

Appendix B.
Abbreviations

AI	Artificial Intelligence
AIAS	Artificial-Intelligence-Enabled Autonomous Systems
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
TEV&V	Test, Evaluation, Verification, and Validation

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE April 2021	2. REPORT TYPE FINAL	3. DATES COVERED *(From–To)*

4. TITLE AND SUBTITLE Trust, Trustworthiness, and Assurance of AI and Autonomy	5a. CONTRACT NUMBER HQ0034-19-D-0001 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Tate, David M.	5d. PROJECT NUMBER AI-2-4837 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Institute for Defense Analyses Systems and Analyses Center 4850 Mark Center Drive Alexandria, VA 22311-1882	8. PERFORMING ORGANIZATION REPORT NUMBER IDA Document D-22631
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Assistant Director for Autonomy Under Secretary of Defense for Research & Engineering Modernization	10. SPONSOR/MONITOR'S ACRONYM(S) OUSD(R&E) 11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (22 June 2021).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The importance of "trust" in artificial intelligence and autonomous systems (AIAS) is widely touted, but not necessarily in productive ways. This paper argues that discourse about the importance of trust can obscure more than it illuminates, misleadingly suggesting similarities (and common solutions) across fundamentally unrelated challenges. In particular, it is important to distinguish whose trust is required, and whether that trust is subjective or objective. The paper also pushes back against the idea that trust can be "built in" as a property of the AIAS, and argues that assurance cases, as originally developed by the safety and security communities, provide a framework for understanding trustworthiness, disambiguating different concepts of trust, and enabling test, evaluation, verification, and validation (TEV&V) approaches to AIAS development and employment.

15. SUBJECT TERMS

Artificial Intelligence; assurance case; Autonomous Systems; explanation; TEV&V; transparency; trust; trustworthiness

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Nickols, Wayne
a. REPORT Uncl.	b. ABSTRACT Uncl.	c. THIS PAGE Uncl.	SAR	22	19b. TELEPHONE NUMBER *(include area code)* (571) 372-6431