A Framework for Discussing Trust in Increasingly Autonomous Systems

Updated June 2017
The MITRE Corporation, McLean Virginia 22102

Background

We are seeing an increase in the complexity, sophistication, and interconnectedness of automation used in a variety of everyday applications from transportation to medical diagnostics. These increasingly autonomous systems depend upon software, data, and networked communications for safe, secure, and efficient operations. They must continue to function appropriately in the face of design defects, unanticipated situations, faulty/missing data, and deliberate attacks because their operational failure could have dire consequences. MITRE recognized that there was a need for a comprehensive framework for discussing trust in these increasingly autonomous systems. This framework shown in the figure on the next page was first documented in a paper presented at the Association for the Advancement of Artificial Intelligence (AAAI) 2014 Spring Symposium.[1]

We have used this framework to discuss the trust humans place in these cyber-intensive systems and trustworthiness of the systems themselves. The goal of this paper is to introduce this framework and explain its elements and relationships.

Framework

Fundamentally we structured our thinking around three major factors: 1) *people*; 2) the increasingly autonomous systems (aka the *machine*) which people are interacting with; and 3) the *environment* in which people and machines operate. We will discuss each factor and their interactions below. Key elements of the framework are identified in the text with *italics*. Humans and machines are both considered to be part of the greater system.

People

We will start with people since they are the most important part. Automation systems are designed and constructed by people; they are managed and operated by people, ultimately for the benefit of people.

People's perspectives of these automation systems will vary based upon their role, which might be that of system operator, developer, acquirer, regulator, or the general society. One thing all roles would have in common is that their perception will be based upon evidence that they can observe; evidence that is presented to them; and their own viewpoint which includes not only their experience but a cultural perspective. How an individual perceives system behavior will be greatly influenced by their *culture*[2]. Culture has many dimensions to include: ethic, religious, national upbringing, professional affiliation, and age. Let's consider an example: A teenager who grew up using tablets, smartphones, and the internet is much more likely to have confidence in the ability of a machine which is capable of automatically parallel-parking a car than a gentleman in his 60s who still changes his own oil.

Environment

The environment a system (i.e., the human-machine team) is intended to operate in will have significant influence. First, environmental *circumstances* will establish the *context* of the operation of both the machine and the people. The machine will determine the circumstances through the *content* of the data received from its sensor inputs. Correctly perceiving the context is an important part of an increasingly autonomous system's ability to respond correctly (i.e., a critical component of its competency).

The Machine

The machine (i.e., the automation system) was designed and constructed to accomplish specific functions to a certain level of performance. The real *competency* of an automation system to accomplish these functions exists whether it can be accurately observed, measured, or assessed. It is a trait of the automation hardware and algorithms. This real competency is going to be dependent upon the system's architecture and the quality of the execution of the development activities. Competency is an engineering trait. There are many ways by which a system's competency can be estimated. For most of today's systems, intended functions can be tested and assessed. People's confidence in a machine's functionality will be only partially dependent upon the results of these tests. *Confidence* is a human trait. If in testing, known inputs produce expected outputs/behaviors, people may be confident that in operation the system will function as intended. Basically, does the automation system do what the human intended, when he intended it to do it?[3] Exhaustive testing is just one method to determine a systems competency. A less formal method might include simply observing how well the system works in operations.

Most systems involve some degree of human interaction or *collaboration*. To be effective, even highly automated and autonomous systems will be interacting with humans. Often people's confidence in a system can be determined not just by the system doing what it should and when it should, but it needs to be doing it for the right reasons[3]. Thus, the nature of the human-machine interaction (i.e., the collaboration) needs to be established so that the human operator understands why the system is functioning the way that it is. This will help determine confidence in that system's ability.

The machine may have bounds or *constraints* on its behavior as part of the system design to limit negative effects or

[1] Lacher, Andrew; Grabowski, Robert; and Cook, Steve, *Autonomy, Trust, and Transportation*, Association for the Advancement of Artificial Intelligence 2014 Spring Symposium, Stanford, CA, March 2014.

[2] Bailey, B. P.; Gurak, L .J.; and Konstan, J. A. 2001. An Examination of Trust Production in Computer-mediated Exchange, *Proceedings of the 7th Conference on Human Factors and the Web*.

[3] Fisher, A. 2013. Inside Google's Quest To Popularize Self-Driving Cars, *Popular Science*, September 2013.

consequences on people of poor system performance. As confidence in the system grows, these constraints could be relaxed. An approach to consider is incrementally integrating an autonomous system by initially constraining system behavior to limit consequences of system failures until confidence grows. Effectively we would be matching authority with the level of demonstrated trustworthiness of the system. This is very similar to how we manage Extended-range Twin Operations (ETOPS). As confidence in the reliability of an aircraft engine grows we give the airline increasing authority to transit further from suitable diversion airports.[4] Using an approach that matches increasingly autonomous system authority with the level of earned trust follows societal norms.

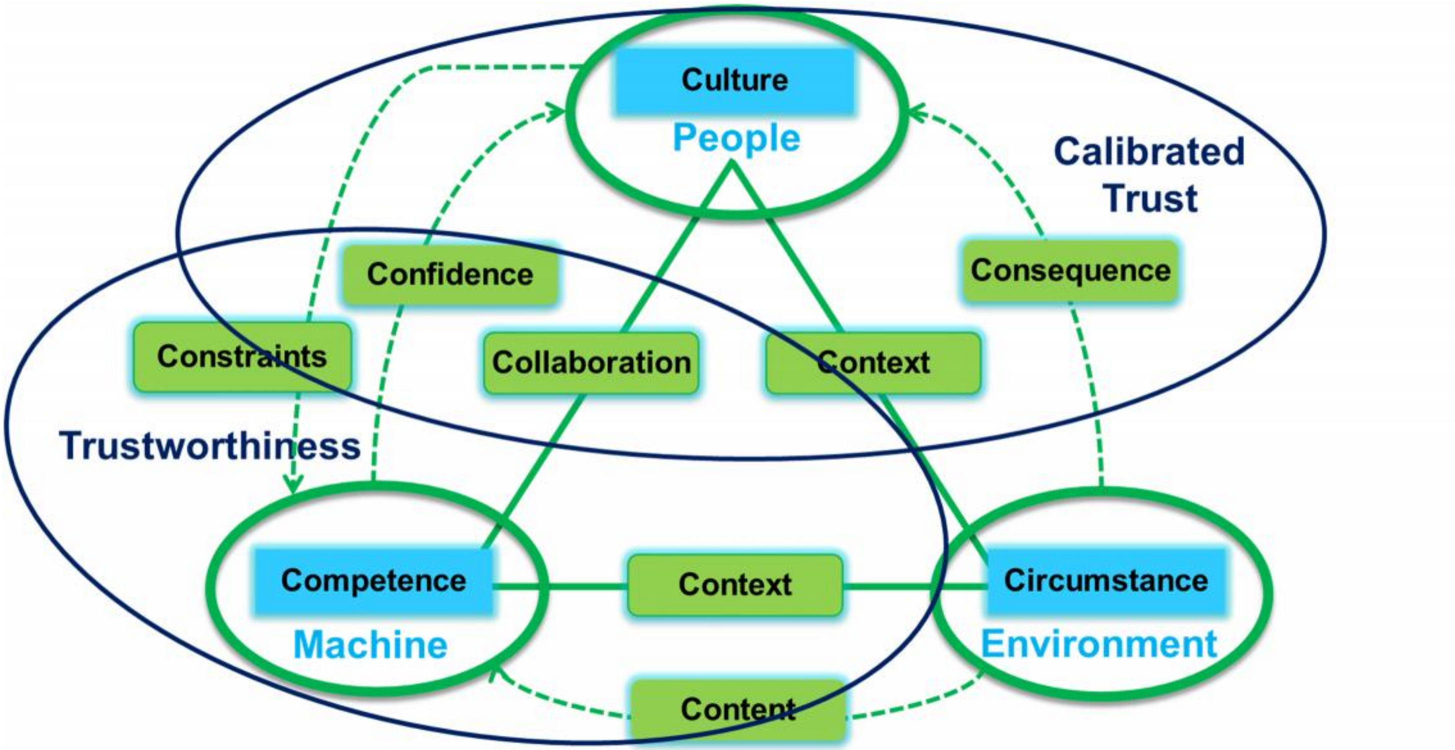
Calibrated Trust and Trustworthiness

If a cyber-intensive system is to operate in situations where the consequences of ineffective performance could cause physical harm to persons or property (e.g., driverless car), humans will need to have a mechanism for both establishing and maintaining appropriate trust in the perception and judgment of these systems. ***Calibrated Trust*** *is not a trait of the system; it is the status the system has in the mind of human beings based upon their perception and expectation of system performance.*[5] Trust is a belief that something is expected to be reliable, good, and effective. Calibrated trust is based upon evidence and perception. Establishing and maintaining trust is not just an engineering challenge, it is a human factors challenge involving cultural, organizational, sociological, interpersonal, psychological, and neurological perspectives[6].

Trustworthiness *is the real competency of a system to perform functions given the extent of the authority it has been granted and the consequences of its potential actions.* Essentially, the level of trust it is worthy of people having in that system. Trustworthiness may not be able to be easily observed or measured. As systems become more complex and more sophisticated is becomes even more difficult to be able to assess the systems trustworthiness.

Conclusion

The trust we place in a system should not exceed the system's trustworthiness. In other words, the authority given to a system should correspond to its demonstrated competency given the circumstances and potential consequences of the systems actions.



Approved for Public Release - Unlimited Distribution - Case: 17-2432
The contents of this material reflect the views of the author and/or the Director of the Center for Advanced Aviation System Development, and do not necessarily reflect the views of the Federal Aviation Administration (FAA) or Department of Transportation (DOT). Neither the FAA nor the DOT makes any warranty or guarantee, or promise, expressed or implied, concerning the content or accuracy of the views expressed herein.
For more information or to comment contact Andrew Lacher, Unmanned and Autonomous Systems Research Strategist (703-983-7812, alacher@mitre.org)

[4] Federal Aviation Administration 2008. Extended Operations (ETOPS and Polar Operations), Advisory Circular No: 120-42B, U. S. Department of Transportation.

[5] The definitions of trust and trustworthiness are based upon National Research Council 2014. *Autonomy Research for Civil Aviation: Toward a New Era of Flight,* ISBN 978-0-309-30614-0 Washington, DC 2014.

[6] Lee, J.D. and See, K. A. 2004. Trust in Automation: Designing for Appropriate Reliance, *Human Factors*, Vol. 46. No. Spring 2004.