RFI: Developing a Federal AI Standards Engagement Plan

AGENCY NAME:	National Institute of Standards and Technology (NIST)
SOLICITATION:	RFI: Developing a Federal AI Standards Engagement Plan
PROJECT TITLE:	Algorithmic Assurances as a Basis for AI Standards Development
APPLICANT:	United Technologies Research Center 411 Silver Lane East Hartford, CT 06118
BUSINESS TYPE:	Large Business
APPLICANT NUMBER:	P.E00.0347
TECHNICAL CONTACT:	Brett Israelsen, PhD Phone: (860) 610-7278 Fax: (860) 610-7134 Email: ISRAELBR@utrc.utc.com
BUSINESS CONTACT:	Ms. Alison Gotkin Phone: (860) 610-7728 Fax: (860) 622-0268 Email: GotkinAE@utrc.utc.com
DATE SUBMITTED:	June 10, 2019

This document does not contain any export controlled technical data.

United Technologies
Research Center

CONTENTS

1. Introduction

The task of developing Artificial Intelligence (AI) standards at the level of the federal government is a complex undertaking. Consequently, all of our views cannot be communicated within an easily digestible RFI (much of the technical concepts herein are a summary from [1]). In interest of being clear and concise we limit our response to addressing two key aims of the Executive Order (EO):

1. "Ensure that technical standards minimize vulnerability to attacks from malicious actors and reflect Federal priorities for innovation, public trust, and public confidence in systems that use AI technologies; and develop international standards to promote and protect those priorities."
2. NIST "shall issue a plan for Federal engagement in the development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies."

2. A Tangible Connection Between Trust and AI

It is clear that governments, businesses, and individuals want to have confidence in, and appropriate trust of "AI". From a high level, the goal of appropriate trust in AI will ultimately lead to the other beneficial outcomes frequently referenced in these discussions such as improving safety, transparency, fairness, and minimizing user vulnerability.

But why is this view useful? As a society, we are experiencing a large amount of difficulty translating our notions about these critically important issues into *practical* and *actionable* solutions. More simply stated, without a formal understanding of what we are trying to do, we cannot define how we are going to do it. Until we have that understanding, we will have limited success identifying possible paths forward. Furthermore, given a "promising" approach we can only have limited confidence that our proposed solutions will, in fact, be effective and free of oversight.

A trusting relationship between an AI and user (e.g. government entity, business, community, or individual) is illustrated below. This representation is largely a consensus of many varied studies on trust (see for example [2, 3, 4, 5]).

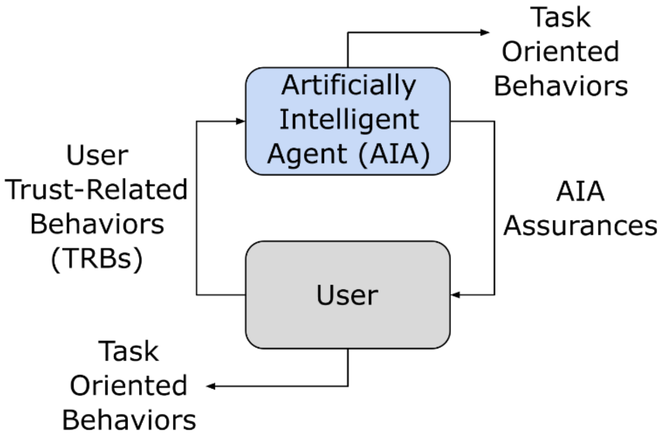


Figure 1: Trust cycle between a user (government, business, individual, et cetera) and an AI system

It is clear in this figure that AI systems affect user trust via "Assurances", and that based on that trust users exhibit certain behaviors (such as turning off a malfunctioning vehicle, or continuing to utilize an algorithm for assisting judges make parole decisions). Using this as a guide we can approach the question of improving trust in AI from a more principled standpoint.

3. Opportunities for Intervention

In developing standards, the government is in the position to intercede between an AI agent and users. The “intervention layer” shown below highlights the level at which government standards can act as a mediator between AI systems and users to ensure trust is not degraded. In words, governments are in a position to drive the development behind standards that guide, and policies that govern, AIA assurances and to safeguard user’s TRBs.

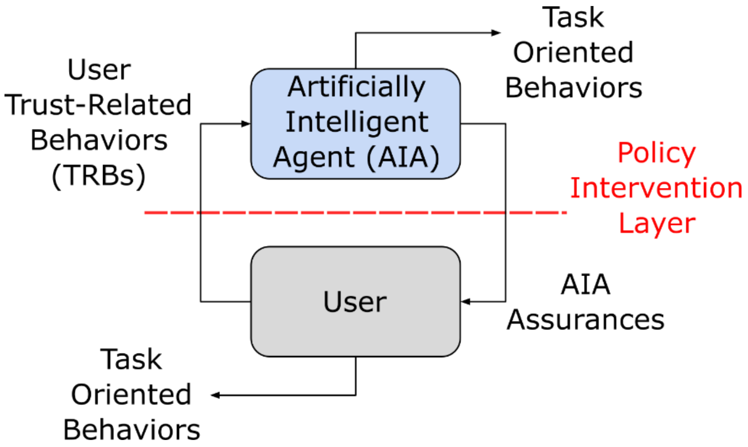


Figure 2: Illustration depicting where standards and policy can intervene in the AI/user trust relationship

AI systems (i.e. autonomous vehicles, personal assistants, decision-making algorithms, et cetera) have different capabilities (see Figure 3 below). Some AI systems possess only one of the capabilities shown, others possess several, and must be treated differently.

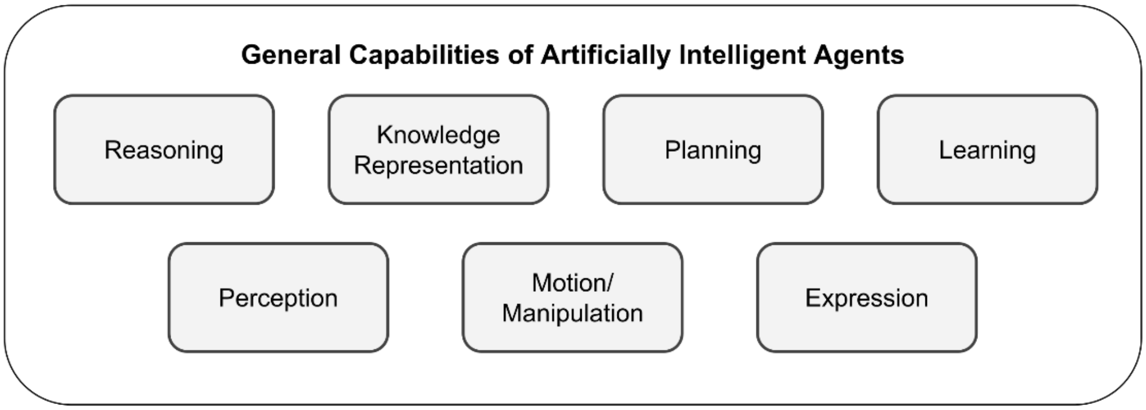
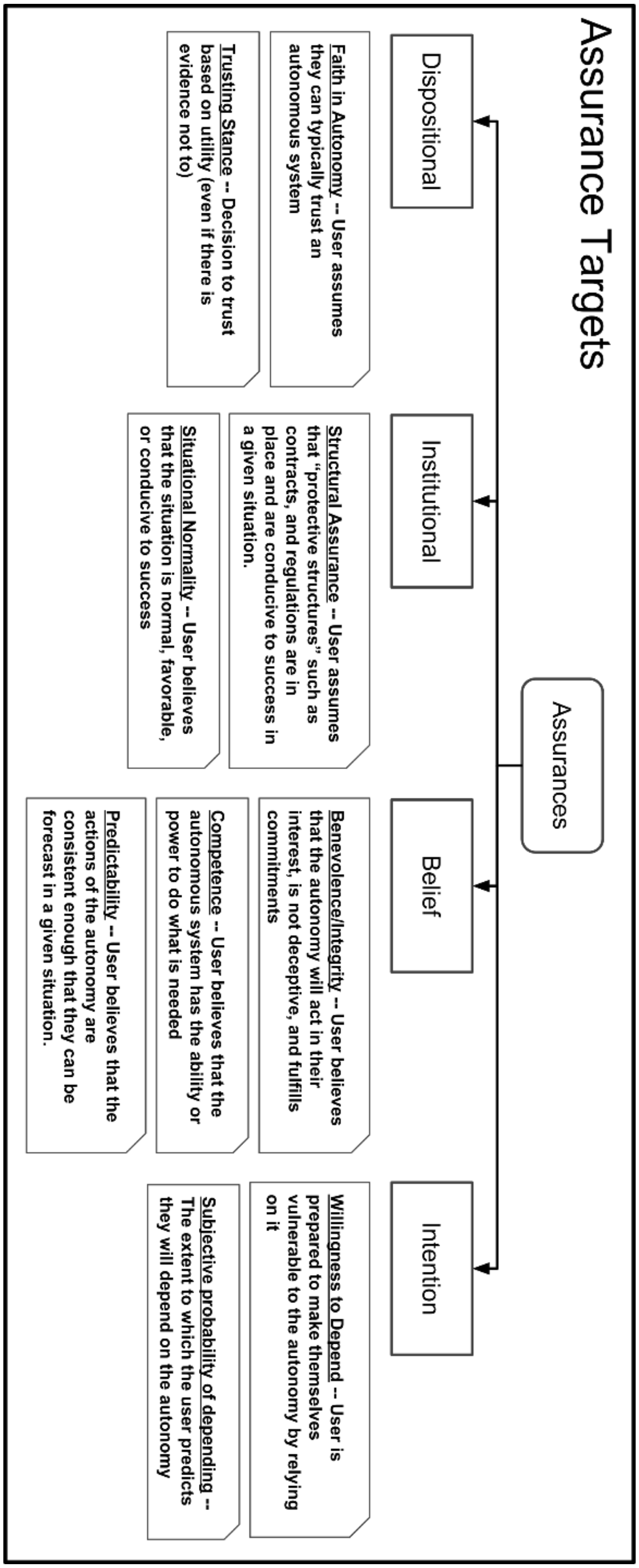


Figure 3: Representative set of capabilities that make up an AI system. Some AI systems may only possess a single capability, others may possess many of them

Similarly, user trust is a multi-dimensional quantity. We most often refer to a system's competence, or predictability as something that influences our trust, but there are many other dimensions that may need to be considered in different situations depending on the specific capabilities of an AI system, and the TRBs involved (for example, predictability may be deemed more important in a high-risk application).



Faith in Autonomy -- User assumes they can typically trust an autonomous system

Trusting Stance -- Decision to trust based on utility (even if there is evidence not to)

Structural Assurance -- User assumes that "protective structures" such as contracts, and regulations are in place and are conducive to success in a given situation.

Situational Normality -- User believes that the situation is normal, favorable, or conducive to success

Benevolence/Integrity -- User believes that the autonomy will act in their interest, is not deceptive, and fulfills commitments

Competence -- User believes that the autonomous system has the ability or power to do what is needed

Predictability -- User believes that the actions of the autonomy are consistent enough that they can be forecast in a given situation.

Willingness to Depend -- User is prepared to make themselves vulnerable to the autonomy by relying on it

Subjective probability of depending -- The extent to which the user predicts they will depend on the autonomy

Figure 4: Diagram illustrating the various dimensions of user trust (i.e. targets for assurances). The three most commonly considered (Competence, Predictability, and Situational Normality) are shown in white.

In a trusting relationship assurances bridge the gap between the AI system capabilities and elements of user trust. They serve as the connection from the different capabilities of an AI system to the relevant trust-dimensions for a given application.

4. The Landscape of Existing Algorithmic Assurances

While *Assurances* include any method by which AI systems affect user trust, *Algorithmic Assurances* are **intentionally designed** properties and behaviors meant to encourage (or certify) appropriate trust. A survey of related technical literature defines and classifies algorithmic assurances (see [1]). Figure 5 illustrates the different technical disciplines involved in creating algorithmic assurances from the perspective of the “level of integration” (i.e. the extent to which the algorithm affects the core functionality of the AI system) of the assurance.

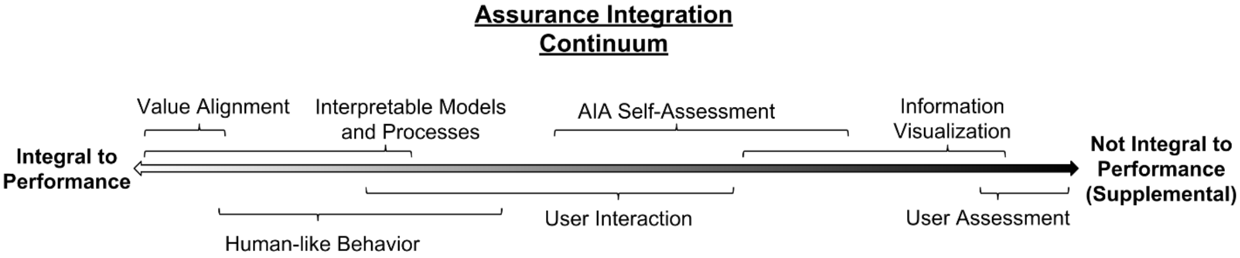

Assurance Integration Continuum
Value Alignment
Interpretable Models and Processes
AIA Self-Assessment
Information Visualization
Integral to Performance
Not Integral to Performance (Supplemental)
User Interaction
User Assessment
Human-like Behavior

Figure 5: Illustration of different technical disciplines involved in creating algorithmic assurances, and to what degree they are integrated into the AI system

This figure highlights many of the technical methods and disciplines used to create algorithmic assurances to encourage appropriate trust:

- **Value Alignment:** Includes disciplines such as AI safety, Validation & Verification, and Certification (as an example see [6]).
- **Human-like Behavior:** Includes the design of systems to be similar to a human, i.e. speaking like a human, or moving like a human (see [7, 8]).
- **Interpretable Models and Processes:** Designing AI whose models and processes are able to be inspected and understood, or “transparent” (as an example see [9]).
- **User Interaction:** Includes disciplines that focus on making AI systems more trustable by relying on human interaction. This includes human in-the-loop and human on-the-loop systems. (see [10]).
- **AIA Self-Assessment:** Disciplines that focus on enabling systems to self-assess their competency boundaries and limitations. The capability enables the communication of the level of competency to users (see [11] as an example).
- **Information Visualization:** Disciplines that focus on displaying critical data and information to help users have better awareness (see [12]).

- **User Assessment:** This category represents systems that rely solely on users to develop their own opinions about what appropriate TRBs are. *Unfortunately most current technologies use this ineffective approach*. The ineffectiveness is manifest in myriad reports of misuse and failures of AI systems all over the world. (see [13, 14]).

There are several other important ways by which algorithmic assurances can be classified, which won't be discussed herein. However, each of these areas are discussed in more detail in [1]. In short, the ability to classify assurances will enable creation of principle-based standards, and provide a strategy for further research and development.

5. A Path Forward

Viewing the challenge of developing AI standards from this perspective there are several different opportunities that present themselves in developing a strategy for AI standards. The lists below are incomplete, but highlight some key opportunities.

5.1. Opportunities: Algorithmic Assurances

- AI systems should be classified/recognized by their capabilities (for example, one autonomous robot might be classified as: decision-making, learning, perception, and motion. Another simple, pre-trained, decision-support algorithm may only be classified as: decision-making)
- Users need to have sufficient assurances of each of the varying capabilities based on which properties are relevant in a given application
- Classification of assurances that an AI system possesses will indicate the capability/limitation of an AI to give appropriate assurance (for example, a decision-making, and learning agent that only has "decision-making assurances" should be outfitted with "learning assurances" as well)
- Do algorithmic assurances currently exist for all combinations of "AI capability"/"trust dimension" pairs? The answer is no. Further research and development is required in this area; such research (and associated pipelines to enable it) needs to be encouraged and funded accordingly.

5.2. Opportunities: Trust-Related Behaviors (TRBs)

- Users (i.e. governments, businesses, communities, individuals) must be aware of their specific trust-related behaviors (TRBs). In many applications current methods are not effective enough.
- Users should be able to initiate/terminate TRBs (as an example, this is important to consider with systems that utilize user data. How can a user's data be revoked? After data is revoked should system "forget" that it ever existed?). Predatory AI systems violate a user's ability to consciously initiate/terminate their trust relationship with an AI system. At a minimum, measures must be taken to ensure that fundamental human rights are upheld and protected (see [15]).

- Users should be notified (by algorithmic assurances) when previous appropriate TRBs are no longer appropriate, or vice versa.

6. UTRC Facilities, Experience and Resources

The United Technologies Research Center (UTRC) is the central research organization for United Technologies Corporation (UTC). This facility encompasses a 550,000 sq. ft. complex, and is used to conduct applied research within various technical disciplines, including chemical sciences, embedded electronic systems, materials and structures, product development and manufacturing, information technology, and dynamic systems and controls. UTRC's primary function is to generate engineering knowledge in technical areas having current and future application to the diverse product interest of UTC divisions (Pratt & Whitney, Collins Aerospace, Carrier, and Otis). UTRC has approximately 600 employees, of whom more than 80% are professional scientists and engineers.

6.1. UTC Code of Ethics

As a corporation UTC is committed to operating by five key values: Trust, Integrity, Respect, Innovation, and Excellence. In pursuing development of evermore advanced and autonomous systems to serve our customers, and benefit the public, it is critical that this be done in a way that is ethical and responsible—in a way such that the technology we develop can be trusted appropriately.

6.2. UTRC Autonomous and Intelligent Systems Department

The Autonomous and Intelligent Systems Department within UTRC has expertise in the areas of Machine Intelligence, Controls, Robotics, and System Modeling, Design and Optimization. Members of the group have experience in human-AI trust, and areas related to it, such as human-robot interaction, interpretable machine learning methods, and data visualization. Put simply, we are experienced in implementing (semi-)autonomous technologies that are trustworthy.

7. Conclusions

Viewing the overall goal of federal engagement in AI standards from the perspective of ensuring appropriate trust between users (governments, businesses, communities, and individuals) and AI systems enables us to have a principled understanding of the real challenges that we face. With a principled understanding of the problem we can move forward with confidence that the strategies implemented will be comprehensive and that the possibility of oversight will be minimized.

We have shown that the perspective presented herein encompasses many of the ideas of “transparency”, "safety", and "explainability" that are currently receiving a lot of attention in this sphere. At the same time this framework also highlights some less-considered concepts. Using the model of the human-AI trust cycle we have identified a “Policy Intervention Layer” that shows that government/business-driven standards and policy can intervene in the trust relationship via “assurances” and “user trust-related behaviors”. Further investigation into specific interventions is required for successful navigation of this complex undertaking.

Please feel free to contact us for clarification of these ideas, or further information.

References

[1] B. W. Israelsen and N. R. Ahmed, "``Dave...I Can Assure You ...That It's Going to Be All Right ..." A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships," *ACM Comput. Surv.,* vol. 51, pp. 113:1--113:37, 1 2019.

[2] A. Baier, "Trust and Antitrust," *Ethics,* vol. 96, pp. 231-260, 1986.

[3] D. H. McKnight and N. L. Chervany, "What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology," *International Journal of Electronic Commerce,* vol. 6, pp. 35-59, 2001.

[4] J. D. Lewis and A. Weigert, "Trust as a Social Reality," *Soc. Forces,* vol. 63, pp. 967-985, 6 1985.

[5] N. Luhmann, "Trust and Power," *Stud. Sov. Thought,* vol. 23, pp. 266-270, 1982.

[6] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, "Concrete Problems in AI Safety," 6 2016.

[7] M. Kwon, S. H. Huang and A. D. Dragan, "Expressing Robot Incapability," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction,* New York, NY, USA, 2018.

[8] J. Tripp, D. H. McKnight and N. K. Lankton, "Degrees of Humanness in Technology: What Type of Trust Matters?," *AMCIS,* 2011.

[9] Z. C. Lipton, "The Mythos of Model Interpretability," 6 2016.

[10] N. Sweet and N. Ahmed, "Structured synthesis and compression of semantic human sensor models for Bayesian estimation," in *2016 American Control Conference (ACC),* 2016.

[11] B. W. Israelsen, N. Ahmed, E. Frew, B. Argrow and D. Lawrence, "Machine Self-Confidence in Autonomous Systems via Meta-Analysis of Decision Processes," in *AHFE 2019,* Washington D.C., 2019.

[12] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North and D. A. Keim, "Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis," *IEEE Trans. Vis. Comput. Graph.,* vol. 23, pp. 241-250, 1 2017.

[13] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics,* vol. 39, pp. 429-460, 3 1996.

[14] M. Desai, M. Medvedev, M. Vázquez, S. McSheehy, S. Gadea-Omelchenko, C. Bruggeman, A. Steinfeld and H. Yanco, "Effects of changing reliability on trust of robot systems," in *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI),* 2012.

[15] High Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, 2019.

[16] A. Lillard, E. W. Frew, B. Argrow, D. Lawrence and N. Ahmed, "Assurances for Enhanced Trust in Autonomous Systems," in *Proceedings: 2015 AAAI Fall Symposium,* 2015.

[17] B. Abdollahi and O. Nasraoui, "Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems," in *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent,* J. Zhou and F. Chen, Eds., Cham, Springer

International Publishing, 2018, pp. 21-35.

[18] C. Castelfranchi and R. Falcone, "Trust and control: A dialectic link," *Appl. Artif. Intell.,* vol. 14, pp. 799-823, 2000.

[19] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man. Mach. Stud.,* vol. 27, pp. 527-539, 11 1987.